



Thèse

2016

Open Access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

Méta-évaluation de la traduction automatique de la parole (TAP) dans le domaine médical

Starlander, Marianne

How to cite

STARLANDER, Marianne. Méta-évaluation de la traduction automatique de la parole (TAP) dans le domaine médical. Doctoral Thesis, 2016. doi: 10.13097/archive-ouverte/unige:85032

This publication URL: <https://archive-ouverte.unige.ch/unige:85032>

Publication DOI: [10.13097/archive-ouverte/unige:85032](https://doi.org/10.13097/archive-ouverte/unige:85032)



**UNIVERSITÉ
DE GENÈVE**

**FACULTÉ DE TRADUCTION
ET D'INTERPRÉTATION**

Méta-évaluation de la traduction automatique de la parole (TAP) dans le domaine médical

Thèse

Présentée à la Faculté de traduction et d'interprétation
de l'Université de Genève
pour obtenir le grade de Docteur en Traitement Informatique Multilingue

par

Marianne Starlander

Jury :

- Prof. Pierrette Bouillon, Faculté de traduction et d'interprétation, Université de Genève (Directrice de thèse)
- Prof. Sonia Halimi, Faculté de traduction et d'interprétation, Université de Genève (Présidente du Jury)
- Prof. Paula Estrella, FaMAF, Université de Cordoba (Jurée externe)
- Prof. Pascaline Merten, LTC, Université libre de Bruxelles (Jurée externe)
- Dr. Lucia Morado, Faculté de traduction et d'interprétation, Université de Genève (Jurée)

Soutenue le 8 avril 2016 à l'Université de Genève

Thèse No 27

Résumé

La traduction automatique de la parole (TAP) est en plein essor, comme en témoigne l'apparition de nombreuses applications au cours de la dernière décennie. Cependant, dans le domaine médical, l'utilisation de ce type de technologie, qui serait pourtant d'un grand secours dans la pratique hospitalière d'urgence, pose encore problème. En particulier, l'évaluation de ces systèmes, notamment de leur qualité à l'usage, revêt une grande importance.

Ce travail vise par conséquent deux objectifs. Premièrement, il s'agit de proposer une méthode qui permet de comparer des systèmes de TAP dans le domaine médical de manière fiable. Pour atteindre cet objectif, nous avons évalué deux systèmes de TAP pour le domaine médical comparables du point de vue de leur fonctionnement linguistique (hybride fondé sur une grammaire interlangue) dans des conditions d'utilisation proche de la réalité à l'Hôpital de Dallas, en employant la méthode EAGLES-ISO.

Deuxièmement, l'objectif est de proposer une nouvelle mesure pour la TAP qui serait équivalente à l'utilisation de la méthode EAGLES-ISO, mais plus rapide et plus économique à mettre en place. Nous avons par conséquent effectué une deuxième série d'expériences qui débouche sur une méthode d'évaluation alternative. Celle-ci se concentre sur la qualité à l'usage de la TAP produite. Elle est fondée sur l'évaluation humaine mais fait intervenir le *crowdsourcing* afin d'augmenter l'efficacité, la rapidité et la fiabilité de l'évaluation de la TAP dans le domaine médical tout en comprimant les coûts engendrés. La méthode d'évaluation proposée, qui constitue la principale contribution de ce travail, porte le nom *TURKoise*.

Abstract

Spoken Language Translation (SLT) is a booming area of research and development, as the increasing offer of SLT systems in the later ten years shows. In the medical domain its use unfortunately still remains problematic, although they would be of much help in the emergency reception of many hospitals, helping to break the language barrier in health care. This is why the evaluation of SLT system, specifically in terms of quality in use requires all our attention.

The present work has thus two main objectives: firstly, we would like to propose a methodology that would enable evaluators to compare SLT systems in the medical domain in a reliable way. In order to achieve this initial goal we have evaluated two comparable SLT systems in the medical domain (both hybrid, rule-based interlanguage systems) in a close to real usage situation at the Dallas Children's Hospital, using the EAGLES-ISO method.

Secondly, our objective is to offer a new measure for SLT in the medical domain, which would be equivalent to the EAGLES-ISO method, with a faster and more cost-effective implementation. We have therefore run a second set of experiments which has lead us to an alternative evaluation method focused on the quality in use of the SLT output. This human evaluation method – baptized as *TURKoise* – relies on crowdsourcing in order to improve the efficiency, rapidity and reliability of current SLT evaluation in the medical domain while reducing drastically the costs of classic evaluation methods.

Remerciements

L'aboutissement de cette thèse a été possible grâce au concours de nombreuses personnes à qui je voudrais témoigner toute ma reconnaissance.

Je voudrais tout d'abord adresser toute ma gratitude à la directrice de cette thèse, la Prof. Pierrette Bouillon, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion et à améliorer le produit final. Je souhaite aussi remercier les membres du jury d'avoir accepté, pour raison exceptionnelle, les délais courts pour la lecture de la thèse.

Je tiens à remercier spécialement mes collègues de TIM et de la FTI, d'aujourd'hui et d'hier : Lucia, Silvia, Lucie, Victoria, Claudia, Tobias, Violeta, Nikos, Manny, Aurélie, Donatella, Paula, Samantha, Klea, Philippe, Selja Seppälä, Agnès Lisowska, Marianne Santaholma, Bruno Cartoni, Andrei Popescu-Belis, Maria Georgescu, Gisela Anspach, Susan Armstrong et Maghi King. Je voudrais exprimer aussi toute ma reconnaissance envers Véronique, Mélanie et Julie, sur qui j'ai pu compter pour la relecture en flux tendu ! Un grand merci à Johanna Gerlach pour son aide dans la mise en place des évaluations AMT et pour la relecture de certains chapitres.

J'aimerais aussi remercier le Prof. Glenn Flores et son équipe du Dallas Children's Hospital, qui m'ont permis d'effectuer mon expérience dans leur service.

Je tiens également à témoigner toute ma gratitude envers tous les évaluateurs bénévoles : Alba, Clara, Lucia, Silvia, Sara, Paola, Magda, Gladys, Miriam, Maria, Adrian, Lucas, Daniel, Ashley, Diego, Adrian, Tahona, Laura, Ana, James, Emma, Rachel, Kathryn, Melissa et aussi tous ceux que j'aurais malencontreusement oubliés ici.

Je réserve également dans ces remerciements une place de choix à mes amies de toujours : Arancha Pinedo, qui a toujours répondu présente que ce soit pour la relecture, le WOZ ou les évaluations, et Sonia Kalinowski pour la patience témoignée lors de la relecture, mais aussi Vanessa, Caroline, Hélène, Clara, Laura, Lorraine et Miki.

Je veux aussi remercier mes colocataires successives (notamment Heba, Heidrun, Lucie et Rosanna), qui m'ont toujours appuyée. Toute ma gratitude va aussi à mes amis gaillards (Régis, Nathalie, Juan, Agnès, Susanne, Yves, les Freland, Aldine et Vincent) mais aussi à mes chers voisins (Nadine, Vicky, Laure, Claude, Jean-Claude, Aivett et Justine), toujours prêts à me donner un coup de main.

Au fil des années, tant de personnes m'ont témoigné leur soutien que je n'arrive pas à les nommer toutes ici. J'aimerais toutefois les remercier chaleureusement de m'avoir entourée dans ces moments souvent douloureux de la réalisation de cette thèse.

Je remercie ma famille Starlander, Boucher-Kregine mais aussi tous les Ybargüen pour leur soutien. Un grand merci à mes parents, Rose et Sven, pour leur soutien et leur patience.

Je tiens à dédier cette thèse à Imyllia, qui a toujours connu sa maman comme thésarde et qui a elle-même réalisé dès sa tendre enfance au moins trois thèses (composées de dessins).

Pour finir, j'aimerais remercier mon mari Daniel pour son aide précieuse, ainsi que le nouveau membre de la famille qui est en route, grâce à qui le processus de finalisation de la thèse a été accéléré !

1	Introduction	13
1.1	Objectifs	15
1.1.1	Premier objectif : méthode d'évaluation EAGLES - ISO	15
1.1.2	Deuxième objectif : TURKoise.....	16
1.2	Contributions.....	17
1.3	Plan.....	18
2	Traduction automatique de la parole dans le domaine médical	23
2.1	TAP médicale linguistique ou statistique.....	24
2.1.1	Traduction automatique statistique	27
2.1.2	Traduction automatique linguistique.....	28
2.2	Différentes architectures des systèmes de TAP du médical.....	30
2.3	Systèmes de TAP dans le domaine médical.....	32
2.3.1	Phraselator	33
2.3.2	De Janus à Jibbigio en passant par Speechalator et Nespole !	34
2.3.3	Spoken Translation : <i>Converser</i>	35
2.3.4	Fluential : S-Minds.....	36
2.3.5	De <i>IBM Mastor</i> à <i>IBM TRANSTAC</i>	37
2.3.6	BBN Transtalk.....	38
2.4	MedSLT: Traduction automatique de la parole dans le domaine médical	40
2.4.1	Fonctionnement général du système MedSLT	44
2.4.2	Résolution de l'ellipse dans le système bidirectionnel.....	50
2.4.3	Module d'aide dans le bidirectionnel	53
2.4.4	Différentes versions du bidirectionnel : V1 et V2.....	55
2.5	Conclusion.....	56
3	Evaluation de la traduction automatique de la parole (TAP)	59
3.1	Normes ISO sur l'évaluation et EAGLES : principes de base	61
3.2	Qualité à l'usage.....	64
3.2.1	Efficacité : accomplissement de la tâche.....	68

3.2.2	Productivité	68
3.2.3	Sécurité : absence de danger	69
3.2.4	Satisfaction : questionnaire	69
3.3	Qualité interne et externe	70
3.3.1	Mesures de la précision issues de l'évaluation de la RV	74
3.3.2	Mesures de la précision issues de l'évaluation de la TA.....	77
3.3.2.1	Evaluation humaine	78
3.3.2.2	Évaluation automatique.....	85
3.4	Vue d'ensemble de l'évaluation des autres systèmes de TAP	94
3.4.1	Campagnes d'évaluation IWSLT	95
3.4.2	NIST, MITRE et DARPA : de Babylon à TRANSTAC.....	96
3.4.3	Systèmes de TAP dans le médical : Nespole, S-minds, Transonics, Converser 99	
3.5	Conclusion.....	99
4	Méthodologie de la recherche	101
4.1	Question de recherche 1 : Méthode d'évaluation EAGLES-ISO pour la TAP médicale 107	
4.1.1	Modèle de qualité	107
4.1.1.1	Qualité à l'usage.....	108
4.1.1.2	Qualité externe	109
4.1.2	Mise en place de l'expérience EXPQ1	113
4.1.2.1	Participants	114
4.1.2.2	Protocole.....	114
4.1.2.3	Etude pilote	117
4.1.2.4	Données	117
4.1.2.5	Statistique : mesures employées.....	118
4.1.3	Récapitulatif des variables pour Q1_principale : grille d'interprétation.....	121

4.2	Question de recherche 2 : TURKoise.....	124
4.2.1	Étape 1 : Evaluation humaine pour TAP médicale : échelle d'utilité à 4-points 124	
4.2.1.1	Objectif.....	124
4.2.1.2	Mise en place de l'expérience EXPQ2_S1	124
4.2.2	Étape 2 : Corrélation entre mesures automatiques et échelle HMS	127
4.2.2.1	Objectif.....	127
4.2.2.2	Mise en place de l'expérience EXPQ2_S2	128
4.2.2.3	Mise en place de l'expérience EXPQ2_S3	129
4.2.3	Étape 3 : TURKoise	130
4.2.3.1	Objectif.....	130
4.2.3.2	Mise en place de l'expérience pilote : EXPQ2_S4-S6.....	132
4.2.3.3	Mise en place de l'expérience EXPQ2_S7	134
4.2.4	Récapitulatif des variables pour Q2_principale	135
4.3	Conclusion.....	137
5	Résultats de Q1_principale : évaluation EAGLES-ISO :.....	141
5.1	Qualité à l'usage.....	142
5.1.1	Efficacité (Q1_S1).....	143
5.1.1.1	Accomplissement de la tâche	144
5.1.1.2	Efficacité de la tâche	146
5.1.2	Productivité : Temps nécessaire pour effectuer la tâche (Q1_S2)	150
5.1.3	Satisfaction (Q1_S3)	152
5.1.4	Sécurité : étude de l'échec des interactions (Q1_S4).....	153
5.1.5	Bilan et tableau récapitulatif des résultats de la qualité à l'usage	157
5.2	Qualité externe : capacité fonctionnelle de la RV, la rétro-translation et la TA	158
5.2.1	Reconnaissance vocale (Q1_S5)	160
5.2.1.1	WER et SER.....	161

5.2.1.2	Taux d'erreur sémantique (SemER).....	164
5.2.2	Rétro-traduction (Q1_S6).....	165
5.2.3	Traduction automatique (Q1_S7).....	171
5.2.3.1	Evaluation humaine.....	171
5.2.3.2	Evaluation automatique.....	173
5.2.4	Bilan et tableau récapitulatif : Précision de la RV, RetroTra, TA.....	174
5.3	Qualité externe : Facilité d'utilisation.....	176
5.3.1	Facilité de compréhension (Q1_S8).....	176
5.3.2	Facilité d'apprentissage (Q1_S9).....	179
5.3.2.1	Médecin : Facilité d'apprentissage.....	182
5.3.2.2	Patient : Facilité d'apprentissage.....	183
5.3.3	Bilan et tableau récapitulatif de la qualité externe : facilité d'utilisation....	183
5.4	Conclusion : méta-évaluation de la méthode EAGLES-ISO	184
6	Résultats de Q2_principale : méthodologie alternative d'évaluation de la TAP (TURKoise).....	189
6.1	Etape 1 : Evaluation humaine pour TAP médicale (HMS) par évaluateurs experts (EXPQ2_S1).....	191
6.1.1	Échelle <i>Human Medical Scale</i> - HMS	191
6.1.2	Résultats de l'expérience EXPQ2_S1	193
6.1.2.1	Application de HMS.....	193
6.1.2.2	Cohérence des évaluateurs experts.....	195
6.1.3	Conclusion de l'étape 1 : Q2_S1.....	196
6.2	Etape 2 : Corrélation entre mesures automatiques et échelle HMS	197
6.2.1	Mesures automatiques : critiques	197
6.2.2	Résultats de l'expérience EXPQ2_S2 : Corrélation entre mesures automatiques et échelle HMS	199
6.2.3	Résultats de l'expérience EXPQ2_S3 : <i>MedSLT</i> comparé à <i>GT</i>	200
6.2.3.1	Evaluation humaine (HMS)	201

6.2.3.2	Evaluation automatique : BLEU, TER, WER.....	204
6.2.3.3	Corrélation entre mesures automatiques et humaines – MedSLT et GT 206	
6.2.4	Conclusion de l'étape 2 : Q2_S2 et Q2_S3.....	207
6.3	Étape 3 : crowdsourcing pour évaluation TAP- TURKoise.....	208
6.3.1	Crowdsourcing	209
6.3.2	Résultats de l'expérience EXPQ2_S4-S6	212
6.3.2.1	Q2_S4 : Faisabilité du recrutement sur AMT	212
6.3.2.2	Q2_S5 : Fiabilité des travailleurs AMT	213
6.3.2.3	Q2_S6 : Compétitivité au niveau du temps et des coûts	217
6.3.3	Résultats de l'expérience EXPQ2_S7	219
6.3.4	Conclusion de l'étape 3 : TURKoise.....	220
6.4	Conclusion pour Q2_principale	221
7	Conclusion générale	223
7.1	Principaux résultats	223
7.2	Contributions.....	226
7.3	Limites.....	227
7.4	Perspectives	228
8	Bibliographie	229
9	Liste des tableaux	249
10	Liste des figures	253
11	Liste des annexes.....	254

1 Introduction

La traduction automatique de la parole (TAP) est en plein essor. À l'orée des années 90, de grands projets de recherche tels que Verbmobil (Wahlster 2000) ont été lancés. Du fait du récent avènement de la téléphonie mobile intelligente, l'offre « d'interprètes automatiques » proposée comme application pour téléphone intelligent (smart phone) a explosé à l'instar de *Google Translation App* ou *Jibbigo*¹. Nous ne sommes plus très loin des applications qui relevaient jusqu'à présent des films d'anticipation, comme le traducteur de StarTrek². Toutefois, une communication parfaite et sans faille avec ce type de technologie reste pour l'heure du domaine de la science-fiction. Malgré les avancées fulgurantes de ces dix dernières années, il subsiste de nombreux problèmes et difficultés inhérents à la complexité technologique.

La TAP combine trois technologies : la reconnaissance vocale (RV), la traduction automatique (TA) et la synthèse vocale (SV) (Waibel et Fügen 2008). En interagissant, ces procédés permettent de traduire de manière instantanée des phrases prononcées oralement. Elles rendent ainsi possible le dialogue entre personnes de langues différentes, même si la traduction n'est pas totalement exempte d'erreurs. En effet, la TAP cumule les difficultés des différentes sous-tâches ; ainsi, elle doit traiter des entrées imparfaites du fait de l'oralité (hésitations, répétitions...), mais aussi dues aux erreurs de RV, et ce, de manière quasi-simultanée. Le recours à la post-édition pour garantir une meilleure qualité de la sortie, comme pour la TA écrite, est de fait impossible. La qualité de la TA produite doit impérativement être suffisante pour permettre un dialogue.

Malgré ces difficultés, la TAP représente indéniablement une avancée primordiale dans la communication entre êtres humains, qui peuvent se parler par ce truchement dans des contextes divers : commerce international, échanges interculturels ou, comme dans le cadre du Projet Fond National Suisse *MedSLT* (Rayner, Bouillon et al. 2008) que nous décrirons ici, pour la communication médicale. L'objectif de *MedSLT* est de permettre à un médecin de communiquer avec son patient au moment de poser un diagnostic. Ce logiciel, appelé *MedSLT*, fonctionne de la manière suivante : le médecin pose une question, par exemple « Depuis quand avez-vous mal ? ». Cette phrase sera reconnue puis traduite dans la langue cible, afin de la faire entendre au patient. À son tour, le patient pourra répondre par le biais du logiciel. Il s'agit d'un

¹ Le guide en ligne sur les nouvelles technologies *Digital Trends* donne une liste d'applications de ce type, <http://www.digitaltrends.com/mobile/best-language-translation-apps>, consulté le 05.08.2014.

² Blog Yahoo, <http://ca.news.yahoo.com/blogs/right-click/google-instant-translation-devices-could-turn-science-fiction-172350808.html>, consulté le 30 juillet 2013.

système bidirectionnel fondé sur une approche linguistique. La principale raison de ce choix est la volonté de garantir une certaine fiabilité du système malgré l'existence d'un corpus d'entraînement très restreint. Jusqu'à présent, il semble avéré que des systèmes contraints par des grammaires sont plus prévisibles que des systèmes statistiques entraînés sur un ensemble de données aussi restreint que le nôtre (Rayner, Estrella et al. 2011). Il faudrait entraîner un système statistique avec des données bien plus importantes et par conséquent moins précises qui s'étendraient au-delà de notre domaine d'application (Costa-Jussà, Farrús et al. 2012). C'est la solution employée par de nombreux systèmes (évoqués plus haut), notamment les applications proposées par *Google Translate* ou *Jibbig* de *Silicon Valley Campus*³.

Dans la pratique hospitalière, il est à noter que dans l'urgence, le personnel hospitalier a recours à des systèmes généralistes comme *Google Translate* (GT)⁴, soit à des outils ciblés sur le diagnostic médical proposant des phrases préenregistrées. L'avantage de GT ou d'autres applications installées sur une tablette ou un téléphone intelligent, c'est qu'elles sont disponibles facilement à un large public. GT propose de nombreuses combinaisons linguistiques (dont certaines langues minoritaires). Ainsi, tout en étant un système de TA statistique généraliste, il est souvent utilisé dans la pratique médicale pour pallier le manque d'interprètes, notamment en cas d'urgence.

Plusieurs articles décrivent ces pratiques et abordent le problème du manque d'interprètes pour les cas d'urgence (Kaliyadan et Gopinathan Pillai 2010, Khanna, Karliner et al. 2011, Patil et Davies 2014). Celui-ci avait déjà été mentionné en 2003, dans l'étude menée par Glenn Flores (Flores, Laws et al. 2003), en raison du manque d'interprètes professionnels dans les hôpitaux, la traduction est souvent effectuée par des proches ou des personnes prises au hasard qui parlent les deux langues. Quoique relativement bilingues, ces personnes n'ont aucune formation médicale, elles ne sont donc pas à l'abri de commettre des erreurs. Les chiffres avancés par les études effectuées jusqu'à présent dans ce domaine sont très parlants : le taux moyen d'erreurs par visite médicale est de 31% (Flores, Laws et al. 2003). La plus fréquente est l'omission d'informations qui peut mener à de graves erreurs de diagnostic.

Dans le cas de l'hôpital cantonal de Genève, il existe un accord avec la Croix-Rouge pour assurer un service d'interprétation, mais le coût croissant encourage à trouver une alternative. (Cuvilliez et Tiger 2014) fait état d'une étude comparative des solutions numériques

³ Site de présentation de Jibbig, <http://www.cmu.edu/silicon-valley/news-events/news/2009/jibbig.html>, consulté le 05.08.2014.

⁴ Site de support de Google Translate, <https://support.google.com/translate/answer/6142474>, consulté le 24.04.2015.

actuellement utilisées. L'étude met en évidence le manque de fiabilité des systèmes évalués en comparant la précision des traductions obtenues sur un corpus.

Notre approche consiste à garantir des traductions fiables au dépend de l'exhaustivité. En effet, *MedSLT* ne traitant que les phrases couvertes par la grammaire, il est moins robuste qu'un système statistique. Néanmoins, dans un domaine comme le diagnostic médical, il est primordial de privilégier la sécurité. Dès lors que toute une palette de systèmes existent souvent fondés sur des approches différentes (linguistique, statistique ou hybride) et dont la couverture diffère beaucoup (systèmes généralistes vs systèmes spécialisés, limités à des domaines restreints), comment évaluer la TAP ? Peut-on arriver à une méthodologie capable de donner une véritable information sur la qualité de chaque système par rapport à son utilité pour les potentiels utilisateurs ?

1.1 Objectifs

Ce travail poursuit deux objectifs. Notre question de recherche générale est : « Comment comparer des systèmes de TAP dans le domaine médical de manière fiable ? ». Cette question générale est décomposée en deux questions complémentaires : *Peut-on évaluer les systèmes de TAP du médical en utilisant la méthode EAGLES-ISO ?* Pour répondre à cette question, nous allons effectuer une série d'expériences avec deux versions comparables du système *MedSLT* (V1 et V2, décrites dans la section 2.4).

Notre deuxième question de recherche est la suivante : *Peut-on proposer une nouvelle mesure pour la TAP qui serait équivalente à l'utilisation de la méthode EAGLES, mais plus rapide et plus économique à mettre en place ?* Nous effectuerons par conséquent une deuxième série d'expériences afin de proposer une méthode d'évaluation efficace, rapide et à moindre coût pour la TAP dans le domaine médical.

1.1.1 Premier objectif : méthode d'évaluation EAGLES - ISO

Nous allons dans un premier temps tester une méthode reposant sur la définition d'un modèle de qualité. Celui-ci est constitué de caractéristiques de qualité externe et de qualité à l'usage (*quality in use*). Il repose sur les normes ISO (la série ISO/IEC 9126 et ISO/IEC 14598, mais également la nouvelle série ISO 25000 *Software Product Quality Requirements and Evaluation* (SQuaRE)). Les deux séries de normes offrent un cadre suffisamment général pour être appliquées à toutes sortes de logiciels. Nous devons donc nous approprier ce cadre afin de proposer notre propre méthode adaptée à notre besoin : savoir laquelle des deux versions de

notre système de TAP dans le domaine médical (V1 ou V2) est mieux adaptée à notre contexte d'utilisation. Nous prenons ainsi résolument la direction d'une évaluation orientée vers l'utilisateur, même s'il s'agit dans notre cas de déterminer pour le développeur dans quelle direction poursuivre le développement de *MedSLT*.

Pour la partie qualité à l'usage, nous avons organisé en février 2008 une simulation d'utilisation réelle de nos deux systèmes à l'Hôpital des enfants de Dallas, dans le service du Prof. Dr Flores. Ce type d'évaluation orientée vers les utilisateurs finals a le mérite de répondre à la question de l'utilité finale de notre logiciel dans la perspective d'une véritable utilisation future dans un cadre hospitalier. Elle nécessite toutefois un important investissement en temps et argent. Nous verrons dans le chapitre sur l'état de l'art de l'évaluation de la TAP (Chapitre 3) que ce type d'évaluation n'est que rarement effectuée pour les raisons invoquées⁵ et est remplacée par des méthodes plus rapides et moins coûteuses qui se limitent le plus souvent à la qualité externe de type évaluation orientée vers le développeur (*developer-oriented evaluation*)⁶. Pour ce type de méthode, l'évaluation se cantonne le plus souvent à une seule caractéristique de qualité : la « capacité fonctionnelle », ce qui se traduit dans le cadre de la TAP par l'évaluation des sous-tâches qui la composent, c'est-à-dire la qualité de la production de la reconnaissance vocale (RV) et de la traduction automatique (TA).

1.1.2 Deuxième objectif : TURKoise

La méthode appliquée pour atteindre le premier objectif est certes complète, mais elle est longue et coûteuse à mettre en œuvre. Nous aimerions donc, dans un second temps, proposer une méthode d'évaluation qui, comme le suggère Maghi King dans (King 2007 : p.141), serait un procédé externe permettant de refléter la qualité à l'usage. Notre méthode a recours à des évaluateurs humains, ce qui peut désormais être envisagé à moindre coût et dans des délais acceptables grâce au *crowdsourcing* (anglicisme communément employé au lieu du terme « externalisation ouverte⁷ » proposé par le Grand Dictionnaire québécois et le bureau de

⁵⁵ "The explanation routinely given for not carrying out extrinsic evaluations is that they are too time-consuming and expensive.", p.115, Belz, A. (2009). "That's Nice... What Can You Do With It?" *Computational Linguistics* 35(1): 111-118.

⁶ "Developer-oriented evaluations focus on functionality (just one component in the ISO standards) and seek to assess the quality of a system's (or component's) outputs. The user-oriented vs. developer-oriented distinction concerns evaluation purpose.", p. 111, *ibid*.

⁷ Définition du terme « externalisation ouverte » issue de Termdat <https://www.termadat.bk.admin.ch/Search/Search>, (site terminologique de l'administration suisse), et du Grand dictionnaire <http://www.granddictionnaire.com/> : « Pratique qui consiste pour une organisation à externaliser une activité, par l'entremise d'un site Web, en faisant appel à la créativité, à l'intelligence et au savoir-faire de la communauté des internautes, pour créer du contenu, développer une idée, résoudre un problème ou réaliser un projet innovant, et ce, à moindre coût. », sites consultés le 28.07.2014.

terminologie de l'administration fédérale suisse et au lieu du terme « production participative » préconisé par les autorités terminologiques françaises et publié dans le journal officiel⁸).

Nous proposons la méthode *TURKoise* qui repose sur deux piliers principaux : une méthode d'évaluation humaine hybride (parce qu'elle contient des éléments d'évaluation externe et de qualité à l'usage) de la production de notre système de TAP pour le médical *MedSLT*. Nous évaluerons la TA issue de la RV avec en ligne de mire la question de l'utilité de cette traduction produite dans notre contexte d'utilisation. Le deuxième pilier est l'exploitation des avantages du *crowdsourcing* pour arriver à une mise en œuvre plus fiable et moins coûteuse de notre évaluation humaine hybride. C'est cet élément qui donne son nom à la méthode. Celui-ci est dérivé de l'*Amazon Mechanical Turk* (une des plateformes de *crowdsourcing* les plus connues) et de la couleur turquoise pour faire allusion à la métrique automatique BLEU⁹. Elle est le fruit de l'expérience acquise dans la première partie de la thèse (Chapitre 5) et consiste à apprécier uniquement des caractéristiques externes du logiciel tout en donnant des informations sur la qualité à l'usage de celui-ci sans avoir à procéder à une évaluation aussi complète et coûteuse que celle proposée au chapitre 5.

1.2 Contributions

Les principales contributions de cette thèse sont les suivantes : tout d'abord, c'est la première fois qu'une méthode d'évaluation de cette envergure - comprenant des tests en milieu hospitalier - est mise sur pied pour évaluer des systèmes linguistiques. En effet, les expériences antérieures menées dans le cadre du département américain de la défense (DARPA) portaient uniquement sur des systèmes statistiques et n'appliquaient pas directement les normes ISO (section 2.3). Nous apportons donc au domaine un modèle de qualité réutilisable pour des systèmes similaires à *MedSLT*, comme *BabelDr*¹⁰. Par ailleurs, s'il s'agit d'évaluer un système de TAP dans un autre contexte d'utilisation, il suffira d'adapter le modèle proposé à celui-ci. Par exemple, pour étudier si le système est utilisable dans le cadre du triage hospitalier dans un service d'urgence, il faudra ajouter les caractéristiques adéquates (comme celle de l'adéquation et de la conformité) et accentuer encore les caractéristiques de facilité d'utilisation (en ajoutant la facilité d'exploitation et l'attractivité, notamment de l'interface) pour coller au contexte

⁸ Définition du terme « production participative » issue du site FranceTerme : <http://www.culture.fr/franceterme/>, consulté le 28 juillet 2014 : Mode de réalisation d'un projet ou d'un produit faisant appel aux contributions d'un grand nombre de personnes, généralement des internautes.

⁹ Papineni, K., S. Roukos, T. Ward et Z. Wei-Jing (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, USA.

¹⁰ Site du projet BabelDr, <http://babeldr.unige.ch/project-description/>, consulté le 01.03.2016.

spécifique d'utilisation (interface adapté au public cible, langues nécessaires, domaines, application fixe ou mobile...).

Ensuite, les expériences que nous avons effectuées ont également permis de confirmer certains faits sur l'évaluation de la TAP, qui peuvent également s'appliquer à la TA. Nous avons notamment mis en relief que les mesures automatiques classiques ne sont pas adaptées à l'évaluation de la TAP linguistique. En outre, nous avons pu vérifier que les évaluateurs experts (traducteurs ou interprètes) sont plus sévères dans l'évaluation de la TAP (et de la TA) que les évaluateurs non-experts, c'est-à-dire des personnes bilingues qui ne sont ni traducteur ni linguiste.

Enfin, nous avons développé la méthode d'évaluation TURKoise qui est adaptée à la TAP tant linguistique que statistique dans le domaine médical, et axée sur la qualité à l'usage du résultat produit. TURKoise repose sur l'évaluation humaine, mais intègre le *crowdsourcing* (CS). Avec nos expériences, nous montrons que le CS peut convenir dans le cadre de tâches cognitives assez complexes. Si le CS a déjà été appliqué dans divers domaines, y compris pour l'évaluation de la TAP, dans le cadre des campagnes d'évaluation IWSLT (section 3.4.1), il n'a pas été utilisé pour une évaluation aussi complexe que la nôtre. En effet, dans le cas de l'IWSLT, il ne s'agissait que d'une simple tâche binaire. Les résultats des expériences menées avec des évaluateurs d'une plateforme de CS et de celles réalisées avec des petits groupes d'évaluateurs experts sont comparables.

1.3 Plan

Cette thèse sera organisée de la manière suivante. Au chapitre 2, nous expliquerons en détail en quoi consiste la TAP en insistant sur les différences entre la TAP reposant sur des techniques linguistiques et statistiques. Le but est d'expliquer pourquoi nous avons développé un système de TAP reposant sur des techniques linguistiques, quand la plupart des équipes de recherche choisissent des techniques statistiques. Nous décrirons, tout d'abord, en quoi consiste la TAP en décrivant les deux technologies qui la composent. Nous ferons brièvement la distinction entre la TA linguistique et la TA statistique. Ensuite, nous évoquerons les fondements du fonctionnement de la reconnaissance vocale (RV). Nous proposerons, enfin, une classification des différents systèmes de TAP selon leur architecture et leur fonctionnement (section 2.3). Dans la section 2.4, nous décrirons le système de TAP pour le médical *MedSLT* et les deux versions que nous voulons départager grâce à notre évaluation complète. Dans la section 2.4.1, nous verrons comment fonctionne *MedSLT* en général. Dans la section 2.4.2, nous

aborderons une particularité de notre système qui est la résolution de l'ellipse. En effet, *MedSLT* ne traduit pas littéralement les phrases elliptiques mais procède à la résolution de l'ellipse grâce à une substitution des éléments manquants. Dans la section 2.4.3, nous détaillerons comment le module d'aide guide les utilisateurs vers la couverture du système en employant des méthodes statistiques. Dans la section 2.4.4, nous expliquerons en détails les différences entre les deux versions évaluées. Version 1 (V1) est une version plus restreinte dans les phrases couvertes : lorsque le patient répond, il ne peut utiliser que des phrases elliptiques en réponse directe à la question. Le deuxième système (V2) permet l'utilisation de phrases entières en plus des ellipses. Selon (Waibel et Fügen 2008), il est plus facile pour des utilisateurs de s'adapter à un vocabulaire donné (domaine couvert par le système : maux de gorge, de tête ou autre domaine) qu'à un type de structure de phrase. Nous terminerons ce chapitre par un tableau récapitulatif des principaux systèmes de TAP dans le monde académique ou présents sur le marché.

Au chapitre 3, nous commencerons par les fondements de l'évaluation de logiciels en décrivant les principes de base des normes ISO (distinction entre qualité à l'usage, qualité interne et externe) et comment définir un modèle de qualité (section 3.1). Ensuite, nous passerons en revue les différentes façons d'évaluer la TAP, en distinguant nettement les deux méthodes appliquées dans le cadre de cette thèse. Nous expliquerons en quoi consiste l'évaluation de qualité à l'usage (section 3.2). Dans la section 3.3, nous nous attarderons sur les caractéristiques de qualité externe notamment sur la capacité fonctionnelle de précision en différenciant les différentes méthodes d'évaluation humaines et automatiques. Parmi les mesures humaines, nous décrirons celles qui sont le plus souvent utilisées en TA depuis le rapport ALPAC (Pierce et Carroll 1966) : la fidélité et la lisibilité que l'on retrouve sous la forme de l'évaluation de l'adéquation et de fluidité dans de nombreuses campagnes d'évaluation de la TAP (Dorr, Olive et al. 2011). Au niveau des métriques automatiques, nous avons choisi de décrire plus en détail les métriques automatiques suivantes : WER, SER ; BLEU ; METEOR et TER (cf. section 3.3.2.2). Nous clorons ce chapitre par un récapitulatif des modes d'évaluation des systèmes de TAP commerciaux et de recherche.

Au chapitre 4, nous allons décrire la méthodologie de notre travail de recherche. Dans la section 4.1, nous définirons notre première question principale de recherche (**Q1_principale**) : *peut-on évaluer des systèmes de TAP avec la méthode EAGLES-ISO que nous proposons ?* Nous décrirons celle-ci en expliquant le cadre de notre évaluation. Cela consistera à définir notre modèle de qualité en suivant les sept étapes d'EAGLES (section 4.1.1). Nous préciserons aussi les détails de la mise en place de l'expérience (**EXPQ1**) à Dallas (participants, protocole,

mesures employées) dans la section 4.1.2. Dans la section 4.2, nous définirons notre deuxième question principale de recherche (**Q2_principale**) : peut-on proposer une méthode d'évaluation moins coûteuse en temps et argent qui permettrait d'évaluer uniquement des caractéristiques externes du logiciel tout en donnant des informations sur sa qualité à l'usage et sans avoir à procéder à une évaluation aussi complète et coûteuse que celle appliquée pour notre première question de recherche ? Nous détaillerons par quelles expériences nous répondrons à cette deuxième question en proposant pas à pas la méthode TURKoise (section 4.2.3).

Au chapitre 5, nous présenterons les résultats concernant notre première question principale de recherche après application de notre méthodologie d'évaluation complète fondée sur les normes ISO et EAGLES en prenant en compte des caractéristiques externes et de qualité à l'usage. Cette méthode EAGLES-ISO sera appliquée aux deux versions de notre système de TAP dans le domaine médical (*MedSLT*) afin de déterminer laquelle de ces deux versions est plus adaptée à notre contexte d'utilisation. Les résultats seront présentés caractéristique de qualité par caractéristique de qualité en distinguant celles du ressort de la qualité à l'usage (satisfaction, efficacité, productivité et sécurité, section 5.1), et celles de qualité externe du produit obtenu par les sous-tâches du système (capacité fonctionnelle et précision, section 5.2 et facilité d'utilisation, section 5.3). Pour la capacité fonctionnelle, nous appliquerons les méthodes d'évaluation classiques de la précision de la RV et de la TA. Les mesures choisies pour la RV sont le taux d'erreur des mots et au niveau de la phrase ainsi que du sens (section 5.2.1). Dans la section 5.2.2, nous évaluerons la sous-tâche de la rétrotraduction (voir 2.4). Pour la TA, nous proposerons une évaluation humaine de l'adéquation et de la fluidité de la traduction produite à partir des résultats de RV (section 5.2.3.1) et une évaluation automatique classique intégrant WER, SER, BLEU et TER (section 5.2.3.2). Pour la facilité d'utilisation, nous étudierons la facilité d'apprentissage (section 5.3.1) et la facilité de compréhension (section 5.3.1) des deux versions en termes de nombre de phrases hors couverture et de progression en temps entre la première utilisation et la dernière utilisation des systèmes étudiés. Toutes ces mesures appliquées correspondent à des attributs de qualité qui font partie de notre modèle de qualité défini dans la section 4.1.1. Elles nous permettront d'arriver à un score final pour chacune des deux versions de *MedSLT* comparées. Nous pourrons, par conséquent, conclure ce chapitre par un résultat clair mais qui sera sujet à discussion au vu de l'investissement en temps et argent lié à cette méthode d'évaluation EAGLES-ISO complète pour la TAP.

Au chapitre 6, nous présenterons les résultats de notre deuxième question principale de recherche. Nous commencerons ce chapitre en proposant notre propre échelle d'évaluation hybride (section 6.1), dont le but est d'inclure une perspective d'évaluation de qualité à l'usage dans une mesure de type qualité externe, en nous concentrant sur l'utilité de la traduction fournie par notre système de TAP. Nous expliquerons le choix de l'échelle d'évaluation (section 6.1.1) avant de l'appliquer en effectuant une évaluation classique par des évaluateurs experts (section 6.1.2). Puis, nous chercherons à automatiser notre évaluation humaine en recherchant une métrique classique automatique (telle que BLEU) qui serait en corrélation avec notre échelle hybride (section 6.2). Nous mettrons ainsi en relation notre échelle hybride avec une série plus étendue de métriques automatiques issues du développement de systèmes statistiques (METEOR, TER) afin de déterminer si ces méthodes peuvent convenir à la TAP dans le domaine médical et remplacer ou compléter notre échelle hybride humaine tout en évitant le problème de la variation entre les juges humains. Cette expérience a été effectuée sur *MedSLT* (sans distinction entre V1-V2) et sur un système statistique (*Google Translate*¹¹) afin de vérifier si nos résultats s'appliquent aussi à des systèmes reposant sur une approche distincte, en utilisant un sous-ensemble de nos données. Dans la section 6.3, nous proposerons de combiner l'échelle HMS proposée au *crowdsourcing* en utilisant la méthode des *Amazon Mechanical Turk* (AMT) afin de fournir une alternative aux évaluations automatiques qui n'ont pas atteint les résultats escomptés s'agissant de la corrélation avec notre mesure. Nous arriverons ainsi à notre méthode *TURKoise*. Dans cette section, nous expliquerons tout d'abord les raisons de ce choix et sa mise en œuvre (section 6.3.1) avant de présenter les résultats obtenus par cette méthode (section 6.3.2 et 6.3.3).

Enfin, au chapitre 7, nous conclurons cette thèse par un examen critique de notre travail, en nous arrêtant sur les avantages et inconvénients de chaque méthode. Nos critères d'évaluation seront, d'une part, la corrélation entre les résultats obtenus avec chaque méthode et, d'autre part, la cohérence entre les juges (inter-rater agreement) pour les mesures humaines. Cependant, nous prendrons également en compte le coût (en temps et argent) comme le préconise (Koehn 2010) ainsi que les critères classiques d'évaluation des mesures elles-mêmes (Hovy, King et al. 2002, King 2007, Popescu-Belis 2007), à savoir si « les métriques d'évaluation [utilisées] reflètent le plus précisément possible la qualité recherchée ou, à défaut, des combinaisons de métriques complémentaires » (Popescu-Belis 2007 : p.85). Cela permettra plus globalement de conclure si la méthode d'évaluation aura atteint son objectif. Nous

¹¹ Site de Google Translate, <https://translate.google.com/>, consulté le 19.12.2014.

commencerons par rappeler les différents résultats décrits lors de ce travail (section 7.1) en indiquant quelles sont les limites (section 7.2) et les perspectives (section 7.3) de notre travail. Nous finirons ce chapitre par notre contribution au domaine : l'élaboration et l'application d'une méthode d'évaluation adaptée aux systèmes de traduction automatique de la parole *TURKoise*.

2 Traduction automatique de la parole dans le domaine médical

La traduction automatique de la parole (TAP) combine deux technologies : la reconnaissance vocale (RV) automatique, qui consiste à convertir la parole en texte, suivie de la traduction automatique (TA). Ces deux technologies ont fait l'objet d'énormes avancées ces dernières années, mais elles demeurent complexes. Ainsi cela reste un défi de proposer un système de TAP suffisamment fiable et précis pour être utilisé dans le domaine médical.

Dans la pratique médicale, le besoin d'un outil performant est pressant. En attendant, comme brièvement mentionné dans l'introduction du chapitre précédent, les utilisateurs ont recours soit à des outils généralistes de TAP (proposants une véritable RV et TA) comme *Google Translate* (GT)¹², *Jibbig*¹³, *Voice4Tra*¹⁴ (Matsuda, Xinhui et al. 2013) ou *SayHi Translate*, soit à des outils ciblés sur le diagnostic médical proposant des phrases préenregistrées. Ils sont une sorte de collections électroniques de phrases utiles au personnel médical dans des situations d'urgence, comme l'accueil des migrants. Nous allons en citer quelques-uns ici, comme *Tradassan*¹⁶ recommandé par l'organisme MigrantHealth, *Traducmed*¹⁷ (Cuvilliez et Tiger 2014) ou encore *Universal Doctor Speaker*. Ils peuvent être considérés comme des alternatives électroniques aux brochures ou fiches de traductions utilisées en milieu hospitalier lorsqu'il y a besoin de communiquer avec des patients étrangers.

Les principaux avantages de ces applications sont d'une part leur grande disponibilité et d'autre part, la grande variété de combinaisons linguistiques proposées et notamment les langues minoritaires.

Ainsi, tout en étant un système de TA statistique généraliste, GT est souvent utilisé dans la pratique médicale pour pallier au manque d'interprètes, notamment en cas d'urgence. Comme cette pratique existe, GT fait l'objet de certaines évaluations dans le milieu hospitalier pour étudier si cette solution est valable ou si une traduction numérique (pas forcément avec RV) peut remplacer les solutions actuelles mises en place par l'hôpital (Randhawa, Ferreyra et al. 2013, Cuvilliez et Tiger 2014). Les différences inhérentes au niveau des fonctionnalités et de

¹² Site de support de Google Translate, <https://support.google.com/translate/answer/6142474>, consulté le 24.04.2015.

¹³ Site de Jibbig, <http://jibbig.com/>, consulté le 19.01.2015.

¹⁴ Site du Consortium U-star à l'origine de VoiceTra4U, <http://www.ustar-consortium.com/app/app.html>, consulté le 17.04.2015.

¹⁵ Site de SayHi Translate, <http://www.sayhitranslate.com/>, consulté le 17.04.2015.

¹⁶ Site de Tradassan hébergé par MigrantHealth.eu, <http://migranthealth.eu/index.php/en/news/40-tradassan-an-app-to-facilitate-communication-with-foreign-patients>, consulté le 16.02.2016.

¹⁷ Site Traducmed (créé pour permettre au personnel médical d'estimer une situation en l'absence de traducteur), <http://traducmed.fr/>, consulté le 24.04.2015.

¹⁸ Site de Universaldoctor, <http://www.universaldoctor.com/>, consulté le 16.02.2015.

la technologie impliquées ne sont souvent pas prises en compte dans l'évaluation de ces outils qui est principalement axée sur l'adéquation et la précision des traductions fournies (Patil and Davies 2014). De ce point de vue, les outils de phrases fixes préenregistrées peuvent servir de traductions de référence (Cuvilliez and Tiger 2014), comme les traductions ont été effectuées par des traducteurs humains spécialisés. Néanmoins, le principal inconvénient des systèmes à phrases fixes, à part le manque de flexibilité, est qu'ils sont plus longs et fastidieux à manier. L'utilisateur doit naviguer dans des hiérarchies de catégorie avant d'arriver à la phrase qu'il recherche. En outre, ces outils sont souvent assez limités au niveau de la traduction prévue des réponses des patients (Cuvilliez et Tiger 2014 : p.14)

Dans ce chapitre nous allons poursuivre par une distinction des différents types de logiciels de TAP dans le domaine médical. La première distinction porte sur la technologie sur laquelle repose le système qui peut être statistique (section 2.1.1) ou linguistique (section 2.1.2). La seconde distinction concerne l'architecture du système, c'est-à-dire s'il s'agit d'un système à phrases fixes, unidirectionnel ou bidirectionnel (section 2.2). Dans la section 2.3, nous passerons en revue les systèmes de TAP du domaine médical. Dans la section 2.4, nous décrirons *MedSLT* et en quoi consistent les particularités de ce logiciel de TAP dans le domaine médical qui fait l'objet de notre thèse. Nous commencerons par expliquer le fonctionnement général de notre système linguistique fondé sur l'interlangue (section 2.4.1), puis nous expliquerons comment fonctionne la résolution des ellipses (section 2.4.2) et le module d'aide (section 2.4.3) pour finir avec les différences entre les deux versions étudiées (section 2.4.4). Dans la conclusion de ce chapitre (section 2.5) nous présenterons un tableau récapitulatif des principaux systèmes de TAP mentionnés.

2.1 TAP médicale linguistique ou statistique

La TAP dans le domaine médical comme la TAP générale dépend de plusieurs composantes complexes la RV et la TA.

En ce qui concerne la **reconnaissance vocale** (RV), on retrouve la distinction entre RV statistique ou linguistique. La principale différence réside dans la nature du modèle de langage utilisé. Nous allons faire une très brève description du fonctionnement de base de la RV grâce au schéma suivant (Figure 2.a) et en nous basant sur les ouvrages (Jurafsky et Martin 2000, Lamel et Gauvain 2003, Haton, Cerisara et al. 2006, Rayner, Bouillon et al. 2006) :

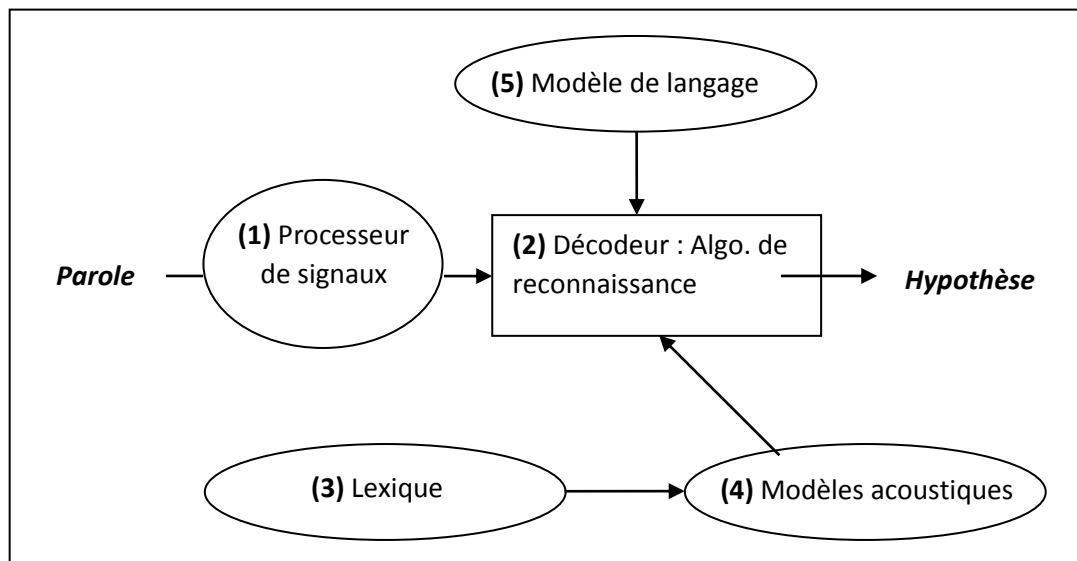


Figure 2.a : Architecture d'un système de reconnaissance vocale

Pour traduire la parole en texte (hypothèse), les principaux éléments sont, premièrement, les outils du traitement du signal ((1) et (2)) et, deuxièmement, le lexique (3) et le modèle acoustique (4). Grâce à ces éléments, le signal sonore est tout d'abord traduit en son, ensuite en mots puis en phrases. C'est le modèle du langage qui déterminera comment se composent les mots et les phrases. Le modèle acoustique (4) définit la forme orale de chaque mot en modélisant le signal acoustique. Un modèle acoustique (MA) est créé le plus souvent en utilisant des modèles de Markov cachés (MMC) grâce à un apprentissage basé sur des exemples de prononciations pour chaque son. Les MMC sont des modèles statistiques complexes que nous n'expliquerons pas en détail ici (cf. chapitre 7 de Jurafsky et Martin (2000)). Au niveau supérieur, le modèle du langage (ML) (5) décrit comment les mots s'assemblent pour former une phrase correcte, en estimant la vraisemblance de leur séquence. Il permet de restreindre le champ de la recherche de correspondance entre le signal acoustique et le mot en excluant les combinaisons peu probables. En effet, seuls les mots autorisés dans une position spécifique par le modèle de langage ont besoin d'être pris en compte. Le modèle du langage aide également à choisir entre deux mots à la prononciation similaire ou même identique (homophones). Les modèles du langage peuvent être statistiques ou bien linguistiques.

Le but du **modèle linguistique du langage** (MLL) est de déterminer quelles structures sont permises dans l'application, grâce à une grammaire. Le modèle statistique du langage (MSL) est fondé sur la probabilité qu'un mot $P(W)$ apparaisse à côté d'un autre mot. Un **modèle statistique du langage** (MSL) est basé sur un corpus duquel on extrait la fréquence avec

laquelle un mot en suit un autre (bigramme) ou plusieurs autres (n-grammes). Il calcule ainsi la probabilité d'une chaîne de mots en fonction de son contexte. Par exemple, un modèle à trigrammes calcule la probabilité de la présence d'un mot en fonction des deux mots le précédant. Cependant, déjà un modèle trigramme requiert une quantité importante de données (sous forme de texte) pour estimer le modèle. Ainsi, pour entraîner un trigramme pour un vocabulaire de 1 000 mots, on a besoin d'estimer environ 10^9 probabilités. Cet exemple montre bien que cette technologie est gourmande en ressources. Ce qui est le principal inconvénient de cette approche pour notre type de système pour lequel il faudrait tout d'abord constituer un corpus de taille suffisamment importante. Ces dernières décennies, en revanche, les ressources vocales générales ont explosé permettant le développement de systèmes de RV généraux de plus en plus performants, notamment pour les systèmes de dictée vocale comme *Dragon Naturally Speaking*¹⁹.

Toute approche - linguistique ou statistique - présente des avantages et des inconvénients. Certes les MSL sont généralement plus adaptés au traitement des discontinuités de la parole spontanée et surtout pour le traitement de la parole hors couverture parce que ces modèles sont reconnus comme étant plus robustes (Rayner, Hockey et al. 2006 : p.12). Cependant, les MLL possèdent le grand avantage de pouvoir capturer les dépendances à longue distance mais aussi de ne pas dépendre de l'entraînement sur un large corpus. Toutefois, les MLL sont difficiles à écrire dans le format exigé (*Context Free Grammar, CFG*), ce qui explique pourquoi on préfère les écrire avec des formalismes plus complexes comme les grammaires d'unification (GU) et de revenir vers des CFG après une phase de compilation (Rayner, Hockey et al. 2006 : p.12) et souvent de spécialisation intégrant des probabilités, ce qui s'apparente au final à une approche hybride. C'est pourquoi il n'est pas si aisé de classer les systèmes de TAP actuels dans des catégories bien définies : statistique, linguistique ou hybride.

Pour plus de détails sur les développements récents en RV, l'article suivant peut être consulté (Xiaodong et Li 2011).

Au niveau de la **traduction automatique**, jusqu'à la fin des années 80, c'est la TA linguistique qui était l'approche prédominante, avant que n'apparaissent les bases de la TA statistique par le Groupe IBM (Brown, Cocke et al. 1988). Nous allons dans la suite faire une brève introduction.

¹⁹ Site de Nuance, <http://www.nuance.com/naturallyspeaking/>, consulté le 16.01.2015.

2.1.1 Traduction automatique statistique

La traduction statistique implique deux modules séparés : un modèle du langage statistique (SLM) et un modèle de traduction (Arnold 2003 : pp.138-141). Le premier fournit les probabilités pour une suite de mots dans une langue, calcul qui se fait en estimant la probabilité d'un premier mot multiplié par les probabilités conditionnelles des mots suivants. Ce modèle fournit la probabilité qu'une certaine suite apparaisse dans une langue donnée. Le second modèle fournit les probabilités conditionnelles d'une phrase cible C dans un texte cible qui traduit un texte contenant la phrase source S. Le produit de cette probabilité et de celle de S (fournie par le modèle de langue) donne la probabilité de la paire de phrase source et cible. Le modèle de traduction exige un large corpus bilingue aligné. La dernière décennie a vu les ressources de ce type se multiplier, notamment grâce à quantité d'informations disponibles sur le Web et qui peuvent être extraites pour former d'énormes corpus. C'est ainsi que, dès 2007, *Google* propose le premier système de TA automatique statistique accessible à un large public²⁰. Depuis lors, le nombre de combinaisons linguistiques a explosé pour compter 90 langues traitées²¹.

Cette grande variété de langues est l'un des principaux avantages du système. De manière générale, la TA statistique a particulièrement été testée pour l'arabe et le chinois, notamment dans les grandes campagnes de compétition entre systèmes de TA : GALE (Dorr, Olive et al. 2011), IWSLT (2006, 2007, 2008, 2011) ou (Koehn 2010).

Un autre avantage communément reconnu de la TA statistique est que, comme les modèles de traductions sont calculés sur des traductions humaines, ou des textes bilingues rédigés par des humains, les traductions obtenues sont plus naturelles et plus fluides que le résultat obtenu par la plupart des systèmes de TA linguistique (Costa-Jussà, Farrús et al. 2012). Un autre avantage de la TA statistique est sa robustesse. Là où la TA linguistique est limitée par la couverture de la grammaire, la TA statistique va toujours proposer une traduction par rapport aux probabilités calculées par le modèle de langage.

Cependant, l'inconvénient de ce type de systèmes est le caractère imprévisible de la traduction statistique (Avramidis et Koehn 2008). C'est l'une des raisons pour lesquelles la récente évolution dans le domaine de la TA statistique est l'intégration de certains éléments « linguistiques », comme c'est le cas pour l'approche de TA statistique fondée sur des

²⁰ Blog de Google, <http://googlesystem.blogspot.ch/2007/10/google-translate-switches-to-googles.html>, consulté le 16.02.2016.

²¹ Site de Google, <http://translate.google.com/about/>, consulté le 19.01.2015.

syntagmes (Tillmann, Vogel et al. 2000, Koehn 2010) et l'approche utilisant des modèles factorisés (*Factored based models*) décrite dans (Koehn, Hoang et al. 2007, Koehn 2010). Ce type de traduction statistique est maintenant proposé à un plus large public notamment par l'Université d'Edinburgh sous la forme du système de TA statistique interactif *Caitra* (Koehn 2009), et avec le projet européen EU-Bridge (Freitag, Wuebker et al. 2014). L'UE a rénové ses systèmes de traduction automatique en proposant de la SMT intégrée dans son système de mémoire de traduction²².

L'hybridation des systèmes statistiques s'est accentuée ces dernières années en TA dans le cadre du projet européen Euromatrix²³ (Eisele, Federmann et al. 2008) et avec une application dans le domaine médical dans (Labaka, España-Bonet et al. 2014, Turner, Brownstein et al. 2015). (Bowen 2013) donne plus de détails sur la SMT au service de la TAP en particulier. Nous allons maintenant brièvement passer en revue les technologies de base de la TA linguistique.

2.1.2 Traduction automatique linguistique

La traduction automatique linguistique repose sur une représentation linguistique d'un texte par le biais de grammaires. Le rôle d'une grammaire pour la TA linguistique est double. D'une part, elle est descriptive puisqu'elle définit les règles de combinaison des mots en phrases correctes. D'autre part, elle permet d'associer à une phrase sa représentation syntaxique et/ou sémantique (selon le type de grammaire). Ainsi, l'analyse d'un texte d'entrée est effectuée grâce à la grammaire qui décrit les phrases et les structures permises²⁴.

De nombreuses grammaires ont été développées pour les besoins de la formalisation de la langue afin d'arriver à une description utilisable par des machines. Cependant, aucune d'entre elles n'arrive à décrire entièrement une langue naturelle²⁵, en rendant explicites ses ambiguïtés. C'est sans doute pourquoi la plupart des systèmes de TA commerciaux qui se veulent généraux se contentent d'une base théorique minimale et utilisent une description linguistique sous forme de dictionnaires bilingues et de règles de grammaires peu élaborées, avec un degré plus ou moins profond d'analyse (Abeillé et Blache 2000). Le développement de grammaires à grande

²² *Language and Translation*, journal en ligne sur les langues et la traduction de la Commission européenne, consulté le 23.06.2015 : http://ec.europa.eu/dgs/translation/publications/magazines/languagetranslation/documents/issue_06_en.pdf

²³ Site du projet Euromatrix, <http://www.euromatrix.net/publications/>, consulté le 23.06.2015.

²⁴ "A grammar of a language is a set of rules which says how these part of speech can be put together to make grammatical or 'well-formed' sentences", p. 39, Arnold, D., L. Balkan, R. L. Humphreys, S. Meijer et L. Sadler, Eds. (1994). *Machine Translation : An Introductory Guide*. London, Blackwell-NCC.

²⁵ "Aucun système ne pourra jamais traiter une langue naturelle dans sa totalité", p. 60, Abeillé, A. et P. Blache (2000). *Grammaires et analyseurs syntaxiques. Ingénierie des langues*. J.-M. Pierrel. Paris, Hermès Science Publications: 51-76.

échelle pose en effet des problèmes de partage des tâches, de maintenance et de mise à jour, ce qui implique un coût important en termes d'heures de travail (Abeillé et Blache 2000 : p.60).

Toutefois, au niveau de la recherche, où l'on se limite souvent à des domaines plus restreints, il est possible de représenter dans une grammaire les structures nécessaires et un lexique pour un domaine donné. Dans le cadre de notre projet *MedSLT* (section 2.4), nous avons utilisé la plateforme *Regulus* (Rayner, Bouillon et al. 2003, Rayner, Hockey et al. 2006), qui permet d'écrire des grammaires qui pourront être adaptées aux différents domaines spécialisés notamment pour éviter les principaux écueils de la TA linguistique : l'ambiguïté de la langue et le décalage et les divergences existants entre les langues naturelles.

Le traitement automatique de l'ambiguïté demande un grand effort parce qu'il faut intégrer de nombreuses informations syntaxiques et sémantiques dans les grammaires et lexiques afin de donner les connaissances nécessaires pour choisir la bonne interprétation d'une phrase ambiguë. Cette formalisation implique de nombreuses heures d'encodage du lexique et d'écriture de règles de grammaire afin de décrire tous les mots ambigus et donner les moyens de choisir le sens approprié.

Une solution à ce problème déjà mentionnée est de restreindre le système de TA à un domaine très précis, où la petite taille du vocabulaire contourne le problème de l'ambiguïté. C'est l'approche adoptée par *MedSLT*, où les questions de diagnostic sont divisées en petits domaines restreints (maux de tête, de gorge, thoraciques) d'environ 400 mots. En plus de découper le corpus en sous-domaines, nous verrons plus loin que, dans le cadre de notre projet, le problème de l'ambiguïté est également évité grâce à des méthodes de spécialisation par domaine (cf. section 2.4.1, méthodes EBL). L'approche par des sous-domaines est communément suivie en TA. Le premier exemple le plus connu est celui de TAUM – METEO. Or, depuis, d'autres sous-domaines (Somers 2003) prouvent l'efficacité de cette approche, notamment la TA d'annonces de travail, de rapport techniques (CRITTER, (Isabelle, Dymetman et al. 1988)), médicaux (ANTHEM, (Deville 2001).), de bulletins d'avalanches (Bouillon et Clas 1993), etc.

Des systèmes de TA linguistique se différencient en fonction de la profondeur d'analyse et de représentation. Nous distinguons ainsi les systèmes directs et les systèmes indirects de transfert ou par interlangue. Les systèmes par interlangue qui poussent plus loin l'analyse afin d'obtenir une représentation intermédiaire la plus abstraite possible et commune à toutes les langues. Les systèmes par interlangue produisent des traductions moins littérales comme ils extraient le sens de la langue source (LS). L'analyse produit une représentation interlangue du

texte source, et le module de génération produit une phrase en langue cible à partir de cette dernière. On parle de traduction par paraphrase (Nirenburg 1993). Cette approche suit l'hypothèse consistant à dire que plus l'analyse du texte source est profonde et abstraite, plus le transfert sera léger, ce qui permet de supprimer le module de transfert. Idéalement, le passage par l'interlangue doit rendre directement possible la traduction de la langue source à la langue cible grâce à une représentation interlangue neutre et universelle. Formellement, on peut définir une interlangue comme étant une sorte de métalangue qui se compose d'unités (vocabulaire), de relations (rôles et fonctions), de syntaxe (qui permet d'exprimer les combinaisons acceptables) et de sémantique (afin de permettre l'interprétation) (Seligman et Dillinger 2006).

Cependant, le principal problème de cette approche interlangue est l'importance du temps de développement puisqu'il faut procéder à une analyse et à une formalisation sémantique des différents domaines (Jurafsky et Martin 2009). Cet inconvénient peut être résolu à condition d'utiliser une interlangue restreinte, soit par rapport à un certain groupe de langues proches ou par rapport à un certain domaine bien défini. Ainsi, une approche plus restreinte est née dans le cadre de projets de traduction automatique de la parole, tel que l'« *Interchange Format* » de *C-Star* (Levin, Gates et al. 1998) et l'interlangue de *MedSLT*. Dans *MedSLT*, que nous allons décrire dans la section 2.4, la représentation interlangue est constituée d'un ensemble de structures de traits plates avec un seul niveau d'enchâssement (Bouillon, Rayner et al. 2005), pour évoluer vers une structure plus complexe (Bouillon, Rayner et al. 2011). Les attributs sont définis de manière intuitive et leur valeur dérivée de la forme canonique de l'anglais. L'avantage de cette représentation est qu'elle est très simple. Elle facilite l'écriture des règles de traduction tout en étant assez expressive pour refléter les nuances des concepts clés pour le domaine traité. En outre, elle permet aussi de contraindre la reconnaissance vocale. L'inconvénient est le coût que représente le développement pour parvenir à décrire tout un domaine. L'avantage de ce type d'architecture est la précision et la fiabilité qu'elle apporte.

2.2 Différentes architectures des systèmes de TAP du médical

Au-delà de la technologie inhérente que nous avons décrite brièvement ci-dessus, plusieurs autres caractéristiques permettent de classer les systèmes de TA de la parole en différentes catégories : la couverture du système, le degré de liberté d'expression, l'architecture, l'interface et le support choisi. D'abord, le traitement sera différent s'il s'agit d'un système de traduction de la parole limité à un domaine très spécifique, comme la réservation de transports aériens ou le diagnostic médical, ou à un domaine plus étendu ou encore un système totalement

général. Par exemple, le système *Voice Response Translator* (VRT)²⁶ ou *Phraselator*²⁷ obligent l'utilisateur à utiliser des **phrases fixes** tout en lui permettant d'adapter la couverture aux besoins. Les nouveaux systèmes, tels que la TAP proposée par *Skype Translator* (Lewis 2015) ou par les applications de TAP proposées sur les smartphones (application *Google Translate* ou *Jibbig*²⁸), sont des systèmes généraux libres. Il existe également des systèmes intermédiaires comme *MedSLT*, qui est un système **contrôlé**, couvrant un sous-ensemble d'une langue naturelle, c'est-à-dire un domaine spécifique et restreint au niveau du sens, comme celui des maux de tête ou de maux de gorge mais aussi au niveau de la couverture de certaines formes syntaxiques.

La « liberté » accordée à l'utilisateur dépendra beaucoup de l'architecture de base du système : si celui-ci initie le dialogue (**system initiative**), le champ de réponse de l'utilisateur est très restreint, ce qui est un gage de réussite. Si c'est l'utilisateur qui a l'initiative (**user initiative**), celui-ci peut formuler ses phrases librement. La première option permet de guider plus facilement l'utilisateur vers la couverture du système. Cependant, elle implique souvent un dialogue peu naturel. La deuxième option laisse plus de liberté, mais l'inconvénient dans ce cas est qu'il est difficile de garantir une reconnaissance vocale efficace et précise.

Si le système est **unidirectionnel**, c'est-à-dire que le système ne traite que l'entrée d'un utilisateur unique puisque les réponses se font de façon non verbale ou par « oui » et par « non », un seul utilisateur doit être formé au maniement du logiciel. Si le système est **bidirectionnel**, c'est-à-dire qu'il permet également de reconnaître et de traduire la réponse du deuxième utilisateur, arriver à un emploi efficace du système de TAP est plus difficile parce que les réponses sont plus difficilement contrôlables, en particulier dans le domaine qui nous intéresse : le diagnostic médical. Le médecin posera toujours le même type de questions et aura l'habitude d'utiliser un système de TAP. À l'inverse, le patient qui répond au médecin sera confronté pour la première fois à ce type de système. De plus, il sera plus difficile de le restreindre à des réponses couvertes par le système, à moins qu'il puisse suivre des règles d'utilisation strictes (par exemple : utiliser des phrases sans verbe en rapport direct avec la question posée) ?

On peut également distinguer les systèmes de TAP par leur interface – multimodale ou non, ou par le support : ordinateur personnel (PC) ou portable, assistant digital (PDA),

²⁶ Site de la nouvelle version de Voice Response Translator, système employé à des fins militaires, <http://www.miltrans.com/>, consulté le 16.02.2016.

²⁷ Site de vente de Phraselator, <http://www.voxtec.com/phraselator/features/>, consulté le 27.06.2014.

²⁸ Site de Jibbig, <http://jibbig.com/>, consulté le 19.01.2015.

téléphone portable (*smarthphone*) ou support spécifiquement manufacturé. *Phraselator*²⁹, par exemple, tourne sur un support spécifiquement conçu pour le terrain militaire. La version de *MedSLT*, développée en 2009, fonctionne aussi sur un assistant personnel digital (PDA) ou téléphone portable, grâce aux travaux de (Tsourakis, Bouillon et al. 2009). De nombreux systèmes sont maintenant développés comme application pour les smartphones. Nous n'en citerons que trois : *Jibbig* (produit issu de la recherche au CMU) (Eck, Lane et al. 2010) ou *Voice4Tra*³⁰ (Matsuda, Xinhui et al. 2013) ou encore le système commercial *SayHi Translate*³¹. Google propose aussi un système de TAP basé sur *Google Translate* classique mais incluant la RV³². Ce type de traducteur pour le grand public se multiplie avec des niveaux de qualité très variables. Ils témoignent de l'essor de la TAP mobile.

L'objectif de ces systèmes est de permettre à un large public de communiquer dans une grande variété de langues, principalement dans le domaine du tourisme mais aussi pour le secteur médical (par exemple *Jibbig*).

Nous allons maintenant voir plus en détail des systèmes plus particulièrement destinés au domaine médical, dans la mesure où c'est ce domaine qui nous intéresse dans le cadre de cette thèse.

2.3 Systèmes de TAP dans le domaine médical

Comme nous venons de le voir, il existe actuellement de nombreux systèmes de TAP sur le marché. Rien que dans le domaine médical, plusieurs projets sont développés en parallèle. Nous ne parlerons pas de tous les systèmes, mais nous évoquerons les systèmes développés par les groupes de recherche les plus importants tels que l'Université de Carnegie Mellon (CMU), ou IBM, ou encore ceux qui émanent de consortiums comme celui pour la recherche avancée en traduction de la parole (Consortium for Speech Translation Advanced Research, *C-star*), ou son successeur TC-star (Technology and Corpora for Speech to Speech Translation, *TC-star*) ou encore le projet *Nespole*!³³ cofinancé par l'Union européenne et les USA (NSF).

²⁹ Site de vente de Phraselator, <http://www.voxtec.com/phraselator/features/>, consulté le 27.06.2014.

³⁰ Site du Consortium U-star à l'origine de VoiceTra4U, <http://www.ustar-consortium.com/app/app.html>, consulté le 17.04.2015.

³¹ Site de SayHi Translate, <http://www.sayhitranslate.com/>, consulté le 17.04.2015.

³² Site de téléchargement de Google Translate App pour Android,

<https://play.google.com/store/apps/details?id=com.google.android.apps.translate&hl=fr>, consulté le 23.06.2015

³³ Page décrivant le système Nespole, sur le site de l'Institut Fondazione Bruno Kessler, <http://hlt.fbk.eu/projects/nespole-speech-speech-translation-e-commerce-service-scenarios>, consultée le 27.06.2014.

D'autres systèmes tels que *Transonics* (Narayanan, Ananthakrishnan et al. 2004, Ettelaie, Gandhe et al. 2005) ou le prototype du SRI³⁴ pour l'anglais-pachto (Precoda, Franco et al. 2004) ou *CCLINC*³⁵ pour le coréen-japonais (Lee, Sinder et al. 2002) sont des prototypes limités à une seule paire de langue. Le développement de *Transonics* se poursuit désormais dans le cadre de *Speechlinks*³⁶ de l'Université South California, financé par le département américain de la défense DARPA (Shin et al, 2010). Il comporte une application au domaine médical pour la pose de diagnostic pour la combinaison farsi-anglais. Nous poursuivrons ce chapitre avec des descriptions plus détaillées des systèmes *Phraselator*, *Nespole*, *Speechalator*, *Jibbig*, *Converser*, *S-Minds* et *MASTOR*, qui comportent tous une application dans le médical.

Nous traiterons de *MedSLT* dans la section 2.4.

2.3.1 Phraselator

*Phraselator*³⁷ est un système à phrases fixes développé selon une idée de Ace Sarich (Sarich 2004) grâce à des fonds issus de la défense américaine (DARPA³⁸) et commercialisé par *Voxtec*³⁹. Ce système mobile et compact (*hand-held*) à commande vocale est unidirectionnel et a été utilisé sur le terrain en Afghanistan et en Irak. Ce système est disponible de l'anglais vers plus de soixante langues (Ehsani, Frederking et al. 2008) dont le farsi, le pachto, l'ourdou, les langues des Balkans ainsi que différents dialectes arabes.

Étant donné qu'il s'agit d'un système à phrases fixes, les domaines sont divisés en modules très spécifiques d'une taille d'environ 500 à 1000 mots traitant de sous-domaines comme l'évaluation de la douleur, la vérification des fonctions vitales, l'obtention d'un frottis de la gorge, etc...⁴⁰ Une version spécialisée pour le triage en urgence dans une clinique pour enfants existe pour la combinaison linguistique anglais-espagnol. Elle comprend les sous-domaines de l'accueil, vérification des liens familiaux, et des questions de triage (*Close Triage Interaction*). La particularité de ce type de systèmes à phrases fixes est que la reconnaissance vocale peut se faire sur la phrase entière, qui est considérée comme une unité indivisible puisque les phrases à reconnaître sont acoustiquement distinctes et très peu nombreuses. L'outil utilisé

³⁴ Page du SRI sur la recherche en TAP, <http://www.speech.sri.com/>, consultée le 27.06.2014.

³⁵ Page concernant CCLINC sur le site de recherche militaire, <http://www.dtic.mil/docs/citations/ADA460574>, consultée le 27.06.2014.

³⁶ Site de Speechlinks, de l'Université South California, USC, <http://sail.usc.edu/transonics/s2s.php>, consulté le 17.04.2015.

³⁷ Site de vente de Phraselator, <http://www.voxtec.com/phraselator/features/>, consulté le 27.06.2014.

³⁸ Site de l'agence du développement de la défense américaine, <http://www.darpa.mil/default.aspx>, consulté le 26.06.2014.

³⁹ Site de vente de Phraselator, <http://www.voxtec.com/phraselator/features/>, consulté le 27.06.2014.

⁴⁰ Page décrivant les différents modules disponibles pour le Phraselator 2, <http://www.voxtec.com/VTP/Modules.aspx>, consultée le 27.06.2014.

est *DynaSpeak*⁴¹ de SRI. Cela présente l'avantage de rendre la RV très simple, rapide et efficace. Quant à la TA, il s'agit simplement d'une base de données sous forme de table contenant les phrases du sous-domaine et leur traduction correspondante. Même la synthèse vocale automatique est évitée car ce sont des enregistrements d'humains qui sont utilisés pour toutes les langues cibles. Comme *Phraselator* est une application militaire, les langues ciblées sont liées aux conflits mondiaux actuels et surtout la couverture du système est axée sur des domaines militaires. Ainsi, des sous-domaines médicaux sont inclus, tout particulièrement pour les urgences, mais ce n'est pas l'orientation principale de *Phraselator*.

2.3.2 De Janus à Jibbiga en passant par Speechalator et Nespole !

Les systèmes que nous décrivons dans cette section sont issus de la collaboration de plusieurs universités et instituts de recherche regroupés en consortiums. Au cœur des consortiums successif *C-star*, *TC-star* et *Nespole !*, souvent dans le cadre de grands projets des USA (DARPA) ou de l'UE, se trouvent les Universités de Carnegie Mellon (CMU) et l'Institute de Technologie de Karlsruhe (KIT) qui forment l'entité *Interactiv systems laboratory* (ISL).

Le système JANUS est un précurseur en matière de recherche en TAP issue du CMU et UKA (maintenant KIT). *Speechalator* (Waibel , Badran et al. 2003) pour l'arabe et le prototype de l'ISL pour le thai⁴² (Schultz, Alexander et al. 2004), tous les deux appliqués au médical, se basent directement sur les méthodes issues de JANUS I-III (Levin, Lavie et al. 2000), tout en intégrant les technologies les plus récentes. Ainsi, *Speechalator* (Waibel , Badran et al. 2003) est un système bidirectionnel de TAP dans le domaine du diagnostic médical pour la paire anglais-arabe. Il tourne sur un assistant personnel digital (PDA). Il se base sur une RV statistique. Pour la traduction, les techniques ont évolué d'une approche purement basée sur les connaissances (*knowledge based*) comme celles utilisées dans le cadre de *Nespole !* (Lavie, Metze et al. 2002) à une traduction toujours basée sur l'interlangue en format *interchange-format* (IF) mais couplée avec des méthodes d'entraînement statistiques afin de transformer la représentation IF en texte cible (Waibel , Badran et al. 2003). Les systèmes de TAP issus du consortium *Nespole !*⁴³ qui regroupe les instituts de recherche du KIT, du CMU, de l'institut de recherche du Trento en Italie (ITC/IRST) et la société italienne Aethra ainsi que le laboratoire CLIPS de Grenoble et l'agence de tourisme du Trentin (APT) fonctionnent de façon similaire.

⁴¹ Page du site du centre de recherche SRI qui décrit le module de reconnaissance vocale Dynaspeak, <http://www.speechatsri.com/products/dynaspeak.shtml> consulté le 27.06.2014

⁴² Site du journal universitaire de Carnegie Mellon, The Tartan, <http://www.thetartan.org/2005/9/26/scitech/interact>, consulté le 27.06.2014.

⁴³ Page décrivant le système Nespole, sur le site de l'Institut Fondazione Bruno Kessler, <http://hlt.fbk.eu/projects/nespole-speech-speech-translation-e-commerce-service-scenarios>, consulté le 27.06.2014.

Le but premier de l'application de *Nespole!* est de permettre à des touristes navigant sur les pages Internet de l'APT d'entrer directement en contact avec un agent de tourisme italien via la TAP. Cette communication est proposée grâce à une vidéoconférence intégrée. Il existe cependant une version pour le diagnostic médical qui a été testée dans le cadre d'une collecte de données pour ce domaine (Mana, Burger et al. 2003). L'approche de cette application est également très proche de JANUS, puisque la RV utilise la boîte à outils de reconnaissance vocale de JANUS (Janus Speech Recognition Toolkit, JRTk). Elle est donc statistique.

Dans *Nespole!*, le module principal d'analyse est un analyseur syntaxique qui utilise une grammaire indépendante de contexte. S'il échoue, une analyse par patrons sémantiques toujours basée sur la représentation interlangue en *Interchange-format* (IF) (Blanchon 2004 : p.161) intervient. Ces outils ainsi que d'autres développés par la CMU ont ensuite été regroupés dans une boîte à outils appelée SPICE (*Speech Processing – Interactive Creation and Evaluation Toolkit for new languages*) afin de permettre de déployer rapidement de nouveaux systèmes de TAP comme cela a été le cas pour le bulgare ou le vietnamien ou encore l'afrikaans (Schultz et Black 2006). SPICE est également utilisé pour le système anglais-farsi développé par le CMU dans le cadre de TRANSTAC (Bach, Eck et al. 2007). La recherche dans le domaine a continué au CMU et a débouché sur la commercialisation de *Jibbig* (Eck, Lane et al. 2010) pour le grand public (Hyman 2014) mais également sur une version robuste du système anglais-thaï dédiée aux militaires et médecins sur le terrain (Schultz, Black et al. 2006).

2.3.3 Spoken Translation : *Converser*

Converser, développé par Spoken Translation⁴⁴ est un système bidirectionnel et interactif. Le dialogue est initié par l'utilisateur qui peut à tout moment vérifier la qualité de la reconnaissance. La couverture du système est large ; le but de l'équipe qui développe *Converser* est d'aller au-delà du domaine médical restreint en permettant de parler librement, raconter des anecdotes, tisser un lien entre patient et professionnel de la santé, ce qui implique de traiter également la langue générale. Dans la version décrite dans (Seligman et Dillinger 2006), le système couvre environ 300 000 mots. L'interface de *Converser* prévoit plusieurs moyens d'interaction outre la parole et la commande vocale, puisque l'utilisateur peut également taper des phrases à l'aide d'un clavier ou écrire avec un stylo. C'est donc un système multimodal.

⁴⁴ Site de *Converser*, <http://spokentranslation.com/technology/>, consulté le 16.02.2016.

Converser se base sur la technologie de *Dragon*⁴⁵ pour la reconnaissance vocale (statistique) et sur des systèmes commerciaux de TA écrite (Word Magic⁴⁶ et Lingenio⁴⁷) pour la traduction (Dillinger et Seligman 2006). Ces systèmes de traduction sont des systèmes généraux, non-restreints au domaine médical (Zong et Seligman 2005).

Comme il s'agit d'un système à large couverture, les risques d'ambiguïtés sont accrus, de sorte que les utilisateurs doivent contrôler et corriger le système, ce qui peut ralentir le dialogue. *Converser* est un système interactif qui inclut cependant une fonction de raccourci (*translation shortcuts*) (Dillinger et Seligman 2006 : p.3, Seligman et Dillinger 2008) qui permet de sauvegarder des phrases préalablement vérifiées dans une sorte de base de données. Ainsi, la phase de vérification est court-circuitée pour les phrases récurrentes.

2.3.4 Fluential : S-Minds

*S-MINDS*TM (*Speaking Multilingual Interactive Natural Dialog System*), développé par *Fluential*⁴⁸ (anciennement Sehda), est un système bidirectionnel, limité à des domaines restreints de gestion de catastrophes et de triage des patients. L'initiative du dialogue revient au médecin, qui mène un dialogue de type questions-réponses. La reconnaissance vocale se fait à l'aide de différents outils commerciaux : *Nuance*⁴⁹, *SRI EduSpeak*⁵⁰ et *HTK d'Entropics*⁵¹, selon le domaine d'application et les langues, en fonction des ressources disponibles. Dans le contexte du système coréen-japonais décrit dans (Ehsani, Kinzey et al. 2006), les modèles de langage utilisés pour la RV sont linguistiques à cause de la pénurie en corpus adéquats pour créer un modèle statistique. En effet, l'une des préoccupations de *Fluential* est d'arriver à construire un système de TAP rapidement (un à trois mois) et surtout avec peu de ressources en corpus, puisque les langues visées sont des langues minoritaires.

La traduction repose sur l'étiquetage sémantique. Chaque phrase prononcée est découpée en concepts clés, traduits ensuite par une simple mise en correspondance dans la langue cible. Il s'agit d'une approche plutôt statistique, basée sur les syntagmes (*phrase-based approach*). Celle-ci permet de réduire le corpus d'entraînement à quelques milliers de phrases au lieu de milliards de mots séparés. La synthèse est assurée par un module de génération développé par

⁴⁵ Site de Nuance, <http://www.nuance.com/naturallyspeaking/>, consulté le 16.02.2016.

⁴⁶ Site de Wordmagicsoft, <http://www.wordmagicsoft.com/>, consulté le 16.02.2016.

⁴⁷ Site de Lingenio, <http://www.lingenio.de/>, consulté le 16.02.2016.

⁴⁸ Site de Fluential, <http://www.fluentialinc.com/>, consulté le 16.02.2016.

⁴⁹ Site de Nuance, <http://www.nuance.com/nuancerecognition/>, consulté le 16.02.2016.

⁵⁰ Site de SRI EduSpeak, <http://www.speechsri.com/products/eduspeak.shtml>, consulté le 16.02.2016.

⁵¹ Site de HTK d'Entropics, <http://htk.eng.cam.ac.uk>, consulté le 16.02.2016.

Fluential qui permet d'assembler des enregistrements humains. Si un mot n'a pas été préenregistré, le système retombe sur un système de synthèse texte à parole classique.

Selon une récente étude comparative (Soller, Chan et al. 2012), le système basé sur des concepts plutôt que sur des statistiques obtient des résultats supérieurs aux autres concurrents sur le marché pour les besoins de traduction dans un domaine restreint (diabète)⁵².

2.3.5 De *IBM Mastor* à *IBM TRANSTAC*

Le système *MASTOR*⁵³ développé par IBM pour traduire du mandarin, ainsi que plus récemment de l'arabe et de l'hébreu à l'anglais, utilise des techniques statistiques pour la reconnaissance vocale et pour la traduction automatique. Pour la RV, *MASTOR* est basé sur le module d'IBM ViaVoice⁵⁴. Pour l'arabe et l'hébreu, la RV utilise des modèles acoustiques basés sur l'intégration des contextes des trente graphèmes différents nécessaires pour décrire ces langues. Le modèle du langage est basé sur la méthode n-grammes, les trigrammes sont calculés automatiquement sur des corpus étendus, hors domaine issus d'Internet, pour compenser la rareté des corpus adéquats pour les langues telles que le chinois ou l'arabe (Zhou, Besacier et al. 2007). Pour permettre l'adaptation rapide à de nouveaux utilisateurs et à l'environnement acoustique, *MASTOR* utilise des algorithmes et des modèles acoustiques qui donnent une bonne précision de RV malgré le bruit environnant comme c'est le cas dans une salle d'hôpital en particulier avec une configuration prévoyant un microphone ouvert. *MASTOR* emploie des techniques de modélisation du langage très avancées qui permettent de limiter la taille du corpus d'entraînement.

Pour la TA, *MASTOR* intègre des méthodes automatiques d'extraction du sens principal contenu dans la phrase traitée (NLU, *Natural language understanding*). Le sens extrait est sauvegardé sous forme de représentation en arbre de concepts de type "actions" et "besoins". Les concepts inclus dans les arbres, produits par un analyseur sémantique robuste, sont ensuite transformés en concepts équivalents dans la langue cible. Puis, un générateur statistique (modèle statistique de n-grammes) de langue naturelle génère une phrase cible à partir de ces concepts. Une autre technique, employée pour la TA, utilise des transducteurs à états finis qui

⁵² "A concept-based English-Spanish speech translation system has been successfully developed in prototype form that can accept long utterances (up to 20 words) with limited to no degradation in accuracy. The functionality of the system is superior to leading commercial speech translation systems", p. 927, in R. William Soller, Ph.D., Philip Chan, Pharm.D., and Amy Higa, Performance of a New Speech Translation Device in Translating Verbal Recommendations of Medication Action Plans for Patients with Diabetes, *Journal of Diabetes Science and Technology*, Volume 6, Issue 4, July 2012

⁵³ Mastor, d'IBM, <http://domino.research.ibm.com/comm/research.nsf/pages/r.uit.innovation.html?Open&printable>, consulté le 16.02.2016.

⁵⁴ ViaVoice, d'IBM, http://www-306.ibm.com/software/pervasive/embedded_viavoice/, consulté le 16.02.2016.

traitent des syntagmes. Cette technique présente l'avantage d'être rapide et efficace. Elle est décrite plus en détails dans (Zhou, Besacier et al. 2007).

La participation d'IBM au programme TRANSTAC (Zhou, Cui et al. 2013) a permis de développer ces techniques pour déboucher sur le système *IBM TRANSTAC* pour de nouvelles langues, d'abord la combinaison arabe d'Irak et anglais afin de parfaire les technologies statistiques de développement rapide pour des langues à faibles ressources comme le dari et le pachto (Zhou, Cui et al. 2013).

2.3.6 BBN Transtalk

BBN Transtalk est un système de TAP bidirectionnel anglais, iraquien, pashto, dari, farsi, malaisien, indonésien et arabe, développé dans le cadre du projet TRANSTAC. Comme l'indique la combinaison de langues, ce système est destiné aux zones de conflits. (Prasad, Kumar et al. 2012). Il s'agit d'un système statistique limité au domaine du déploiement militaire, mais qui intègre de ce fait une composante médicale. Pour la reconnaissance vocale, le système utilise Byblos (Chow, Dunham et al. 1987) et la TA est de type statistique.

BBN transtalk fait partie des systèmes de TAP qui ont pu être développés dans le cadre du programme de recherche du DARPA (section 3.4.3).

Le programme *Translation for Tactical use* (TRANSTAC), financé par le département de la défense américaine (DARPA), regroupe ainsi plusieurs groupes de recherche (notamment l'entreprise BBN (Stallard et al., 2007) avec le système de TAP Transtalk (Prasad, Natarajan et al. 2013), ou encore le CMU (Bach, Eck et al. 2007), IBM (Gao et al., 2006), SRI (Akbacak et al., 2009), et USC (Belvin et al., 2005)) afin de développer des logiciels de TAP notamment dans le domaine médical à l'usage des militaires sur le terrain.

Les grands programmes de recherche américain comme TRANSTAC mais aussi les programmes de recherche européens comme *TC-star*⁵⁵, projet financé par le sixième programme cadre de l'union européenne, ont fait progresser le domaine de la TAP de manière phénoménale comme le montre le Tableau 2.1 et la Figure 2.b : Évolution de la technologie de la TAP, tiré de l'article (Nakamura 2009).

⁵⁵ Site du consortium TCStar, <http://tcstar.org/pages/info.htm>, consulté le 19.01.2015

Projet / campagne	Période d'activité	Objectifs, type de système de TAP visé et plateforme
C-STAR	1992-2004	Uni/bidirectionnel limité à un domaine pour la parole spontanée
VERBMOBIL	1993-2000	Unidirectionnel limité à un domaine (conversation)
DARPA BABYLON	2001-2003	Bidirectionnel limité à un domaine pour la parole spontanée Portable et PDA
DARPA TRANSTAC	2005-2010	Bidirectionnel pour domaine restreint (dialogue spontané) Mobile/Smartphone
IWSLT	2004-	Unidirectionnel, pour domaine restreint et pour parole libre (<i>unlimited free-style talk</i>)
TC-STAR	2004-2007	Unidirectionnel, domaines discours politiques et télévision. Sur serveur
DARPA GALE	2006-2011	Unidirectionnel, Télévision et conversations Sur serveur
DARPA BOLT	2011-	Uni/bidirectionnel Conversations Sur serveur
U-STAR	2010-2016	Uni/bidirectionnel, domaines non-limités, toutes les langues Sur Smartphone, sur serveur

Tableau 2.1 : Récapitulatif des campagnes et consortiums en TAP (Agency 2008 : p.98)

Depuis le consortium C-STAR (début années 90) au consortium U-STAR, la TAP est passée d'un système unidirectionnel ou bidirectionnel limité à un domaine pour la parole spontanée à un système couvrant « toutes les langues » pour des domaines non-limités accessible par smartphone via un serveur. L'image reprise de (Nakamura 2009) ci-dessous illustre bien l'évolution considérable de la TAP dans le cadre des consortiums européens et asiatiques, (C-STAR, (Akiba, Federico et al. 2004)), et maintenant « universels » comme le consortium U-star à l'origine de VoiceTra4U⁵⁶.

⁵⁶ Site du Consortium U-star <http://www.ustar-consortium.com/app/app.html>, consulté le 17.04.2015.

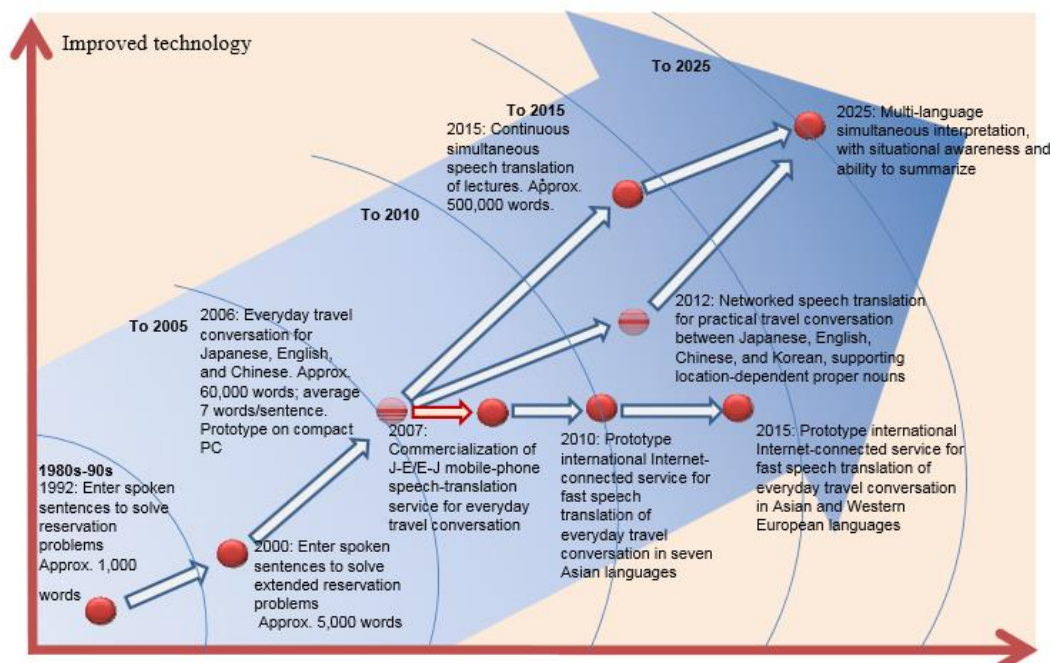


Figure 2.b : Évolution de la technologie de la TAP, tiré de l'article (Nakamura 2009).

Nous pouvons conclure que ces vingt-cinq dernières années, l'évolution de systèmes linguistiques fondés sur l'interlangue vers des systèmes statistiques s'est accentuée (ex. CMU, (Lavie, Levin et al. 2001) (Waibel et Fügen 2008), grâce à la collecte d'une quantité énorme de données audio et grâce aux progrès réalisés dans les technologies inhérentes à la TAP grâce aux efforts de recherche dans le cadre institutionnel.

Nous allons maintenant voir où se situe notre système *MedSLT* par rapport aux différents systèmes de TAP que nous avons décrits dans cette section 2.3. Nous commencerons par une présentation du fonctionnement général de *MedSLT* avant d'expliquer quelles sont les différences entre les deux versions du système qui seront étudiées dans la partie empirique de ce travail.

2.4 MedSLT: Traduction automatique de la parole dans le domaine médical

Medical Spoken Language Translator (*MedSLT*) est un système de traduction automatique de la parole (TAP) dans le domaine médical. Le but de *MedSLT* est de traduire des questions de diagnostic dans des domaines spécialisés afin de permettre un **dialogue fiable** entre un médecin et son patient qui ne parlent pas la même langue. Notre préoccupation principale réside par conséquent dans la précision du système, ce qui explique notre choix d'un traitement linguistique pour la reconnaissance vocale (RV) et pour la traduction automatique

(TA). Concernant la RV, nous utilisons la boîte à outils commerciale *Nuance*⁵⁷. Elle est contrainte par des modèles de langage linguistique (MLL) dérivés de grammaires d'unification grâce à la plateforme *Regulus* (Rayner, Hockey et al. 2006). S'agissant de la TA, puisque le système est résolument multilingue, nous avons adopté une approche linguistique interlangue (Bouillon, Rayner et al. 2006). Il couvre les sept langues suivantes : anglais, français, catalan, espagnol, japonais, arabe et finnois.

Afin que les utilisateurs puissent vérifier que les tours de paroles reconnues par le système correspondent bien à ce qu'ils ont dit, nous offrons à l'utilisateur une rétro-traduction (RétroTra) de la phrase dite. Au lieu de soumettre le résultat brut de la reconnaissance vocale, nous présentons une paraphrase de la phrase dite. La RétroTra permet de vérifier la qualité de la RV, mais elle va au-delà de cette étape puisqu'elle est le produit de la RV combinée à la traduction vers l'interlangue puis de retour vers la langue source. Elle est donc le moyen de fournir une sécurité accrue aux utilisateurs.

En nous fondant sur les critères définis à la section 2.2, concernant la couverture d'un système, nous pouvons classer le système *MedSLT* dans la catégorie des systèmes de TAP contrôlé. Il est le résultat d'un compromis entre un système de traduction de phrases fixes et un traitement linguistique profond. En effet, le système est limité à la couverture de domaines restreints en rapport direct avec les questions de diagnostic découpés par sous-domaines médicaux afin d'éviter les problèmes d'ambiguïté tout en étant plus général qu'un système à phrases fixes puisqu'il permet de varier la formulation des questions. Pour guider les utilisateurs novices vers la couverture du système, *MedSLT* est complété par un module d'aide, basé sur un modèle statistique du langage (MSL). Celui-ci propose à l'utilisateur des exemples de phrases possibles dérivées à partir de la reconnaissance statistique. Ainsi, le système s'inscrit dans la tendance actuelle de l'hybridation. *MedSLT* innove cependant puisque toutes les grammaires du système (pour la reconnaissance, l'analyse et la génération) sont compilées par la plate-forme *Regulus* à partir de grammaires générales d'unification, motivées linguistiquement. Celles-ci sont d'abord spécialisées automatiquement pour les différents domaines du diagnostic par des méthodes d'apprentissage basées sur des corpus, puis compilées pour les différentes tâches du système : reconnaissance et analyse avec *Nuance* après conversion dans le format de grammaires CFG (*Context-Free Grammar*) requis par *Nuance*, et génération. Les différentes grammaires requises par le système sont ainsi plus faciles à maintenir et restent cohérentes entre elles, c'est ainsi notamment que toute phrase reconnue sera nécessairement bien analysée.

⁵⁷ Site de Nuance, <http://www.nuance.com/>, consulté le 16.01.2015.

Construire ces grammaires manuellement serait évidemment très fastidieux et source d'erreurs et d'incohérences. Avec la technologie de spécialisation employée par *Regulus*, il devient beaucoup plus facile de développer plusieurs versions de *MedSLT* : unidirectionnelle, bidirectionnelle restreinte (V1) et bidirectionnelle moins restreinte (V2).

La version **unidirectionnelle** est orientée vers le médecin. Elle lui permet de poser un ensemble de questions *fermées*, ce qui laisse le patient répondre uniquement par *oui*, par *non* ou par des gestes, comme dans la Tableau 2.2.

1. Médecin : « Do you have a fever? »
MedSLT : « ¿Tiene fiebre? »
Patient : « sí »
2. Médecin : « More than 100 degrees Farenheit? »
MedSLT : « ¿Tiene más de 38 grados de fiebre? »
Patient : « sí »
3. Médecin : « More than 102 degrees Farenheit? »
MedSLT : « ¿Tiene más de 39 grados de fiebre? »
Patient : « no »
4. Médecin : « Do you have a sore throat? »
MedSLT : « ¿Le duele la garganta? »
Patient : « sí »
5. Médecin : « Cough? »
MedSLT : « ¿Tiene tos? »
Patient : « no »

Tableau 2.2 : Exemple de dialogue avec la version unidirectionnelle

Se cantonner à des questions fermées rallonge le dialogue puisque le patient est limité dans sa marge de manœuvre. Dans l'exemple ci-dessus, il ne peut pas répondre simplement par « *Tengo 38.5 grados de fiebre* » (J'ai 38.5 degrés de fièvre). Le médecin est obligé de poser deux questions avant de parvenir à la même conclusion. Cet inconvénient est partiellement compensé par la possibilité de poser des questions elliptiques au lieu de répéter la question entière, comme l'illustrent les exemples 2, 3 et 5 du Tableau 2.2 ci-dessus. L'avantage de cette version de *MedSLT* est que le système peut être opéré uniquement par le médecin, sans que le patient n'ait besoin de voir l'interface du système. En effet, les questions traduites sont synthétisées, ce qui lui permet de répondre par « oui » ou « non » ou par des gestes sans l'intermédiaire du système. Cela présente un certain avantage lorsque les utilisateurs ne sont pas à l'aise avec les ordinateurs.

L'ajout de la version **bidirectionnelle** s'est principalement fait à la demande de notre partenaire, le Prof. Glenn Flores de l'Hôpital pour enfants de Dallas⁵⁸, qui soulignait l'avantage d'une architecture permettant au patient d'interagir directement. L'avantage de la version bidirectionnelle est évidemment qu'elle donne la parole au patient, qui peut répondre sans détour à des questions ouvertes par des ellipses ou des phrases entières (selon la version employée) en rapport direct avec la question du médecin (Tableau 2.3, ci-dessous).

1. Médecin : « Do you have a fever? »
MedSLT : « ¿Tiene fiebre? »
Patient : « sí, tengo mucha fiebre »
2. Médecin : « How high is your fever? »
MedSLT : « ¿Cúantos grados de fiebre tiene? »
Patient : « 38.5 »
2bis. Médecin : « How high is your fever? »
MedSLT : « ¿Cúantos grados de fiebre tiene? »
Patient : « Tengo 38.5 grados de fiebre. »
3. Médecin : « Do you have a sore throat? »
MedSLT : « ¿Le duele la garganta? »
Patient : « Sí, mucho. »
4. Médecin : « Congestion? »
MedSLT : « ¿Tiene la nariz tapada? »
Patient : « No, tengo mucho moco. »
5. Médecin : « Cough? »
MedSLT : « ¿Tiene tos? »
Patient : « Un poco. »

Tableau 2.3 : Exemple de dialogue avec la version bidirectionnelle

L'inconvénient de cette version est que le patient oublie qu'il s'agit d'un système contrôlé et donc restreint et qu'il s'écarte de sa couverture. En outre, le développement de la version actuelle du bidirectionnel est assez complexe.

Dans la suite de ce chapitre, nous décrirons tout d'abord le fonctionnement général de *MedSLT* (section 2.4.1) en détaillant les technologies de RV et TA impliqués. Nous expliquerons le fonctionnement de la traduction, en décrivant la représentation intermédiaire sous forme d'interlangue et les règles de traduction bidirectionnelles espagnol-interlangue et interlangue-espagnol. Nous aborderons principalement les ressources concernant la paire de langues anglais-espagnol puisque, d'une part, les versions de *MedSLT* étudiées dans le cadre de

⁵⁸ Dallas Children's Hospital, <https://www.childrens.com/> et University Southwest, <http://www.utsouthwestern.edu/>, sites consultés le 18.02.2016.

ce travail sont de cette combinaison linguistique. D'autre part, nous avons contribué à leur élaboration, à savoir la grammaire de l'espagnol et les règles de traduction espagnol-interlangue et interlangue-espagnol. Dans la section 2.4.2, nous expliquerons comment fonctionne la résolution de l'ellipse dans *MedSLT*. Ensuite, dans la section 2.4.3, nous donnerons des détails sur le fonctionnement du module d'aide. Enfin, pour conclure cette description de *MedSLT*, nous terminerons sur les différences entre les deux versions du système qui feront l'objet de l'évaluation comparative dans la suite de cette thèse : V1 bidirectionnelle restreinte et V2 bidirectionnelle moins restreinte (section 2.4.4).

2.4.1 Fonctionnement général du système MedSLT

La reconnaissance vocale dans *MedSLT* est linguistique. L'utilisation d'un modèle linguistique du langage (MLL) présente un premier grand avantage, celui d'éviter de dépendre de l'existence de vastes corpus d'entraînement. Son deuxième grand avantage réside dans la précision qu'il garantit. En effet, si une structure n'est pas décrite dans la grammaire, elle ne pourra pas être reconnue. Par exemple, dans le contexte du système bidirectionnel, le reconnaiseur se limite, pour le patient, à des structures de phrases affirmatives et à la première personne, ce qui permet d'éviter la confusion entre « *me duele la garganta* » (j'ai mal à la gorge) et « *le duele la garganta* » (« avez-vous mal à la gorge » ou « il a mal à la gorge »), très proches phonétiquement. Ce type de reconnaiseur est par conséquent plus fiable dans le contexte d'une application contrôlée circonscrite à des domaines d'applications bien précis, comme c'est le cas de *MedSLT*. En outre, en utilisant un MLL, la RV est contrainte au niveau d'une phrase entière, c'est-à-dire à un niveau supérieur que celui de bigrammes ou de trigrammes. Ce niveau de contrainte donne à l'utilisateur une sécurité supplémentaire par rapport à un reconnaiseur statistique qui pourrait reconnaître des portions de phrases et les traduire incorrectement. Toutefois, le reconnaiseur est moins robuste qu'un reconnaiseur à modèle statistique du langage (MSL) lorsqu'il est face à des mots et structures inconnus. Un utilisateur novice aura alors plus de difficultés à manier le système. C'est pourquoi nous avons complété notre système linguistique par un module d'aide plus robuste fondé sur un MSL au cas où la reconnaissance linguistique échouerait (section 2.4.3 pour plus de détails).

La présence de ce module d'aide est indispensable pour le patient qui fera sans doute une utilisation unique du système. Relevons encore que, si une phrase a été reconnue correctement avec le MLL, elle sera très probablement traduite correctement grâce à l'attribution d'une représentation correcte par le reconnaiseur. Enfin, le développement des ressources nécessaires à un système de TAP basé sur des MLL ne dépend pas de l'existence d'un vaste

corpus, comme c'est le cas pour l'entraînement d'un MSL. L'obtention d'un tel corpus est particulièrement difficile dans ce domaine. Elle peut être atteinte par la mise en place d'une expérience de magicien d'Oz, lors de laquelle les utilisateurs pensent qu'ils interagissent avec un véritable système alors que le système est remplacé par des humains. L'objectif est de récolter des données proches de l'utilisation finale afin d'obtenir un corpus exploitable pour construire un système de TA statistique par exemple, ou pour étudier comment s'adaptent les utilisateurs à un système sans même l'avoir construit (Starlander 2007). Ce procédé est utile mais coûteux et peut être évité par l'écriture de grammaires (Rayner, Bouillon et al. 2004).

Il faut souligner, cependant, que nous n'aurions pas pu tirer profit de ces avantages liés à l'utilisation de MLL sans l'existence de *Regulus* (Rayner, Hockey et al. 2006), qui permet d'écrire des grammaires à un niveau d'abstraction supérieur à celui des grammaires formelles de type CFG requises pour la RV. Ce formalisme de grammaire est fastidieux à écrire à la main à cause de sa redondance. Cela prendrait trop de temps d'écrire des grammaires CFG suffisamment étendues pour couvrir les domaines de *MedSLT*. La plateforme *Regulus* (Rayner, Hockey et al. 2006), implémentée en *SICStus Prolog*⁵⁹, a été développée dans le cadre du projet *MedSLT*, mais elle découle des travaux effectués par l'équipe de *Netdecision Ltd* et *Fluency Tech Ltd*, ainsi que dans le cadre du projet *Spoken Language Translation* (SLT) (Rayner, Carter et al. 2000), puis au sein de la NASA. Cette plateforme contient notamment un compilateur qui permet de passer d'une grammaire d'unification à une grammaire CFG et de spécialiser automatiquement une grammaire générale d'unification (GGU) à l'aide de corpus. Comme on peut le voir dans la Figure 2.c, la même grammaire sert pour la reconnaissance (1), l'analyse (2) et la génération (4). C'est ce qui rend un traitement linguistique possible et attrayant pour le type de qualité que nous visons dans le domaine médical dans lequel nous travaillons.

⁵⁹ Site de SICS Swedish ICT, développeur de Sicstus prolog, <http://www.sics.se/isl/sicstuswww/site/index.html>, consulté le 18.02.2016.

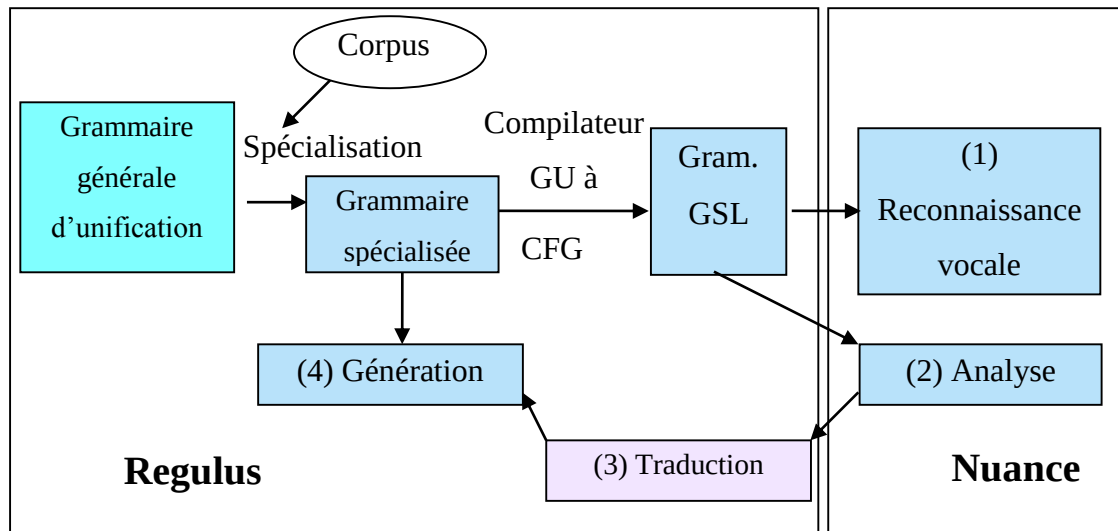


Figure 2.c : Fonctionnement de MedSLT

Pour *MedSLT*, l'avantage de faire appel à des grammaires est non négligeable : comme illustré par la Figure 2.c, *Regulus* requiert une grammaire par sous-tâche (reconnaissance, analyse et génération) et de préférence une par sous-domaine de diagnostic comme celui du diagnostic de l'hypoglycémie ou des maux de gorge afin d'éviter le problème de l'ambiguïté de la langue. Sans la possibilité que nous offre *Regulus* d'écrire une seule grande grammaire d'unification par langue, et de la spécialiser ensuite semi-automatiquement sur la base de corpus, nous aurions déjà dû écrire au moins vingt-huit grammaires différentes pour couvrir les deux domaines cités pour les sept langues de *MedSLT*.

En utilisant des grammaires écrites avec *Regulus*, nous pouvons combiner différentes spécificités propres aux tâches qu'elles doivent accomplir sans multiplier la création de règles. La reconnaissance doit notamment être suffisamment contrainte pour éviter les erreurs mais rester assez générale pour permettre au locuteur des variations. Alors que pour la génération la grammaire doit être très contrainte de manière à ne produire qu'un seul résultat pour chaque représentation. Comme la reconnaissance et la génération ont des exigences différentes, nous dérivons des grammaires spécialisées pour ces tâches à partir de grammaires générales. Afin d'obtenir le meilleur résultat possible pour la reconnaissance et la génération du système bidirectionnel, il résulte, après l'étape de spécialisation, six grammaires spécialisées différentes, à savoir :

- une grammaire de reconnaissance anglaise pour les questions du médecin
- une grammaire de génération anglaise pour la traduction de la réponse espagnole du patient
- une grammaire anglaise pour la rétro-traduction de la réponse du médecin
- une grammaire de reconnaissance espagnole pour les réponses du patient
- une grammaire de génération espagnole pour la traduction de la question anglaise du médecin
- une grammaire de génération espagnole pour la rétro-traduction de la réponse du patient.

Elles sont toutes soit issues de la grammaire générale de l'anglais, soit de la grammaire générale partagée pour les langues romanes (GGROM).

Pour spécialiser ces grammaires générales, *Regulus* emploie la méthode de l'apprentissage par explication (*explanation based learning*, EBL). Celle-ci se base sur un corpus pour former une sous-grammaire contenant uniquement les règles nécessaires au traitement des phrases du corpus (Rayner, Hockey et al. 2002), (Hockey, Lemon et al. 2003).

Comme illustré dans la Figure 2.d, ci-dessous, la grammaire *Regulus* intervient tout d'abord lors de la reconnaissance vocale (RV) pour transformer l'entrée orale en un texte source. Celui-ci est analysé, grâce à *Regulus*, et transformé une fois l'étape de résolution des ellipses achevée en une représentation sémantique source. Puis, grâce à l'application des règles de traduction, cette représentation sémantique est transformée dans notre formalisme interlangue, et après vers l'équivalent en langue cible, qui pourra être généré en un texte cible avant d'être synthétisé grâce à la boîte à outils *Nuance*. La première étape, effectuée par le reconnaiseur, consiste à vérifier que la phrase en entrée est conforme à la grammaire de reconnaissance en analysant la phrase avec celle-ci. Cette analyse donnera lieu à une représentation de la phrase source. Pour la phrase source « *do you have a fever ?* » précédant la phrase elliptique « *congestion ?* », on obtient la représentation source suivante :

```
[ [utterance_type,ynq], [voice,active], [pronoun,you], [tense,present],
  [state,have], [secondary_symptom,fever] ]
```

Comme le montre l'exemple ci-dessus, l'analyse est une simple concaténation de la sémantique des mots de la phrase, où sont retenus principalement la sémantique des mots pleins ainsi que certains éléments comme le type de phrase qui est ici de type question oui/non ([utterance_type,ynq]), la voix ([voice,active]) et le temps du verbe ([tense,present]). Les informations morphologiques telles que le nombre ne sont pas retenues parce qu'elles sont la source de trop d'erreurs au niveau de la RV, elles sont donc reconstituées lors de la génération. Nous obtenons ainsi une représentation plate sous forme d'attributs-valeurs.

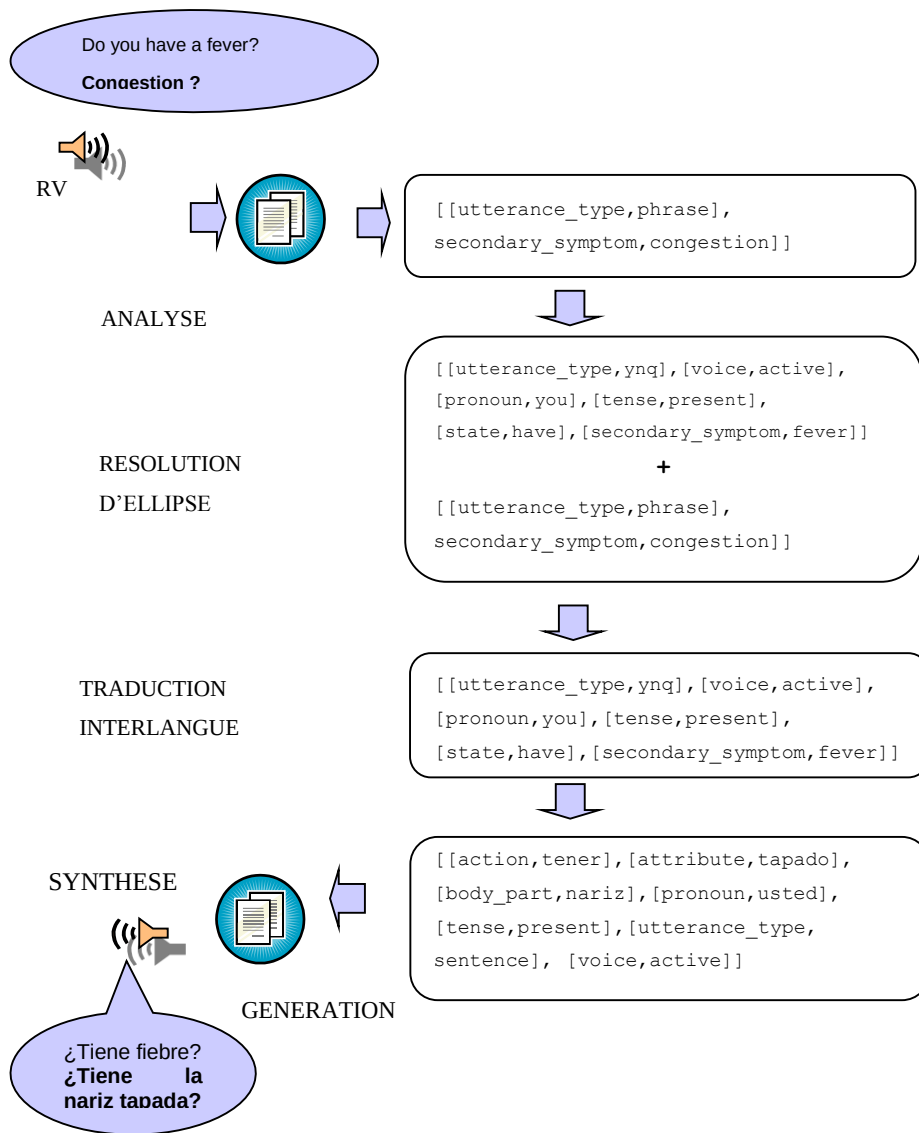


Figure 2.d : Flux de traitement avec MedSLT

La représentation sémantique obtenue sert de base à la traduction. Le système par interlangue permet d'utiliser une seule représentation, commune à plusieurs langues. Cela facilite notamment l'ajout de langues supplémentaires. Pour une nouvelle langue, il suffira alors de créer les règles de correspondance vers l'interlangue, plutôt que de créer des règles de transfert pour chaque combinaison de langue (Levin, Gates et al. 1998).

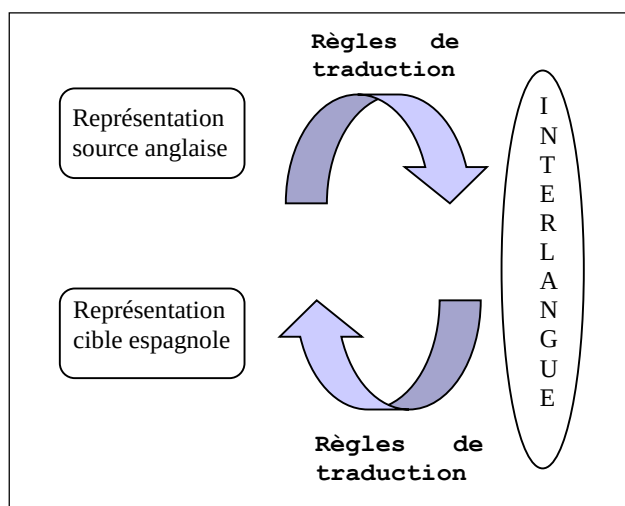


Figure 2.e : Règles de traduction pour le bidirectionnel anglais-espagnol⁶⁰

Grâce aux règles de traduction (Figure 2.e) de l'anglais à l'interlangue, nous obtenons la représentation interlangue suivante, qui est très proche de la représentation anglaise donnée à la page antérieure :

```
[[pronoun,you],[secondary_symptom,fever],[state,have_symptom],
[tense,present],[utterance_type,ynq],[voice,active]]
```

Les règles de transfert de l'interlangue à l'espagnol transforment cette interlangue commune à toutes les langues en la représentation cible équivalente suivante :

```
[[action,tener],[pronoun,usted],[symptom,fiebre],[tense,present],
[utterance_type,sentence],[voice,active]]
```

À partir de cette représentation, la génération s'effectue grâce à une variation spécifique de la même grammaire générale Regulus que celle qui a servi de base à l'analyse. La surgénération est évitée grâce aux contraintes syntaxiques et sémantiques décrites dans la grammaire.

La dernière étape du traitement consiste en la synthèse à l'oral de la phrase en texte issue de la génération grâce au logiciel texte-parole de *Nuance* (Nuance Vocalizer 4.0⁶¹).

Nous avons vu dans cette section qu'afin d'obtenir les représentations nécessaires aux étapes successives du traitement décrit ici nous avons recours à des grammaires d'unification Regulus, spécialisées pour les différentes tâches requises : la reconnaissance et la génération.

⁶⁰ Adapté de Bouillon, P., M. Rayner, N. Chatzichrisafis, B. A. Hockey, M. Santaholma, M. Starlander, H. Isahara, K. Kanzaki et Y. Nakao (2005). A generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation., 30-31, May, 2005, Budapest, Hungary. Tenth Conference of the European Association of Machine Translation, Budapest, Hungary.

⁶¹ <http://www.nuance.com/realspeak/vocalizer/>

Nous allons maintenant décrire une autre particularité de *MedSLT*, qui est la résolution de l'ellipse.

2.4.2 Résolution de l'ellipse dans le système bidirectionnel

Dans le type de dialogue que couvre *MedSLT*, les ellipses sont fréquentes et naturelles. Elles interviennent à deux niveaux : pour les questions du médecin et pour les réponses du patient. Du côté des questions posées par le médecin, les ellipses accélèrent la prise d'information en évitant de répéter les phrases en entier, ce qui est très souhaitable surtout dans le contexte de l'unidirectionnel où le médecin est limité à des questions fermées. Du côté des réponses du patient, les ellipses étaient la seule forme de réponse possible dans la première version bidirectionnelle du système, puisque les patients pouvaient uniquement répondre par des phrases elliptiques en rapport direct avec la question du médecin. C'était un moyen de simplifier la tâche puisqu'en forçant le patient à répondre de cette façon, il est plus aisé de prévoir les réponses du patient et de le guider vers les phrases couvertes par le système.

L'emploi d'ellipses entraîne la nécessité soit de les résoudre avant de traduire, soit de traduire les ellipses telles quelles. Mais une traduction mot à mot des ellipses peut prêter à confusion, même dans un domaine limité comme le nôtre, comme l'atteste l'exemple ci-dessous tiré de (Bouillon, Rayner et al. 2007).

Question 1: Do you have a pain in your stomach?
Traduction: ¿Le duele el estomago?
Question 2: in your head?
Traduction 2:*¿en la cabeza?

Tableau 2.4 : Exemple d'ellipse

Dans cet exemple, « *in your stomach* » devient le sujet de la phrase en espagnol, ce qui implique que, lors de la traduction, l'ellipse devrait changer de catégorie syntaxique (syntagme prépositionnel → syntagme nominal). Comme des exemples similaires sont fréquents, il semble souhaitable de traduire les ellipses dans leur contexte. Il existe principalement deux types d'ellipses dans notre corpus : les omissions et les substitutions. Dans le Tableau 2.5 : Résolution d'ellipse par substitution, il s'agit d'une substitution.

La méthode employée dans *MedSLT* est inspirée du dialogue homme-machine. Elle repose sur l'apprentissage, à partir d'un corpus d'exemples, des éléments intrinsèques d'une

ellipse et de la façon de la résoudre. La résolution de l'ellipse s'effectue juste avant le passage à l'interlangue, au moment de la représentation source. Elle consiste en une manipulation au niveau des listes d'attributs de la représentation sémantique plate. L'omission est considérée comme une simple concaténation de listes et la substitution, telle une substitution de listes d'éléments similaires, comme l'illustre l'exemple dans le Tableau 2.5 ci-dessous. Lorsque l'analyseur détecte qu'il s'agit d'une ellipse, il procède tout d'abord à la substitution. Si celle-ci n'est pas possible, il tente une omission.

Do you have fever ?		[[utterance_type,ynq],[voice,active],[pronoun,you],[tense,present],[state,have],[secondary_symptom,fever]]
		
A congestion ?		[[utterance_type,phrase] [secondary_symptom,congestion]]
		
Do you have a congestion ?		[[utterance_type,ynq],[voice,active],[pronoun,you],[tense,present],[state,have],[secondary_symptom,congestion]]

Tableau 2.5 : Résolution d'ellipse par substitution

Afin de permettre à l'analyseur de reconnaître une ellipse, le corpus d'exemples décrit chaque type d'ellipse possible. La Tableau 2.6 montre au moins un exemple de chaque type d'ellipse autorisé pour exprimer le lieu de la douleur.

<pre> ellipsis_class(where, ['en la garganta', 'en un solo lado', 'en ambos lados', 'en el lado izquierdo', 'en el lado derecho', 'en el lado', 'en la frente', wh-'dónde']). </pre>

Tableau 2.6 : Classes d'ellipses pour le lieu

Ces exemples sont analysés par la grammaire pour être généralisés afin d'extraire un certain nombre de patrons qui définissent la bonne formation d'une classe d'ellipses.

```

compiled_ellipsis_class(
  where,
  [[[number, _],[part, _],[prenom, _],[prep, _]],
   [[part, _],[prenom, _],[prep, _]],
   [[body_part, _],[part, _],[prep, _]],
   [[part, _],[prep, _]],
   [[body_part, _],[prep, _]],
   [[number, _],[part, _],[prenom, _]],
   [[part, _],[prenom, _]],
   [[body_part, _],[part, _]],
   [[part, _]],
   [[body_part, _]]]).

```

Tableau 2.7 : Généralisation des patrons à partir des exemples du Tableau 2.6

La Tableau 2.7 montre les patrons dérivés des classes d'ellipses correspondant au Tableau 2.6. Ces patrons permettent d'identifier les éléments substituables les uns aux autres dans la phrase elliptique et la phrase complète. Lorsque la représentation d'une phrase elliptique concorde avec celle d'un des patrons, elle pourra se substituer à la représentation correspondante de la phrase complète qui correspond à un patron de la même classe. Cette méthode est assez robuste pour apprendre des patrons suffisamment généraux pour les domaines contrôlés de *MedSLT*.

La résolution d'ellipses se complique un peu dans le contexte du système bidirectionnel, puisque la réponse elliptique du patient est résolue par rapport à la question du médecin. Ainsi, dans l'exemple du Tableau 2.8, la réponse « *ayer* » est résolue par rapport à la question du médecin « *When did you visit the doctor* », par concaténation de la représentation de la phrase elliptique avec celle de la question du médecin.

Médecin : When did you visit the doctor? Traduction : cuándo ha consultado un médico? Patient : ayer Traduction : I visited the doctor yesterday

Tableau 2.8 : Exemple d'interaction avec le bidirectionnel

Le Tableau 2.9, ci-dessous, illustre la résolution d'ellipse par concaténation de l'exemple vu plus haut.

Q : When did you visit a doctor ?	[[action,visit],[agent,doctor], [pronoun,you],[tense,past],[time,when], [utterance_type,whq], [voice,active]]resolution_[[utterance_
R : Ayer	 [[utterance_type,phrase] [temporal,ayer]
Résolution de l'ellipse	 [[action,consultar],[person,médico], [pronoun,usted],[temporal,ayer], [temporal,cuándo], [tense,past],[utterance_type,sentence], [voice,active]]
T : I visited a doctor yesterday	[[action,visit],[agent,doctor], [pronoun,you],[tense,past],[time,when], [time,yesterday],[utterance_type,ynq], [voice,active]]

Tableau 2.9 : Résolution de l'ellipse de l'interaction de Fig. 3.10 par concaténation

L'étude (Bouillon, Rayner et al. 2007) a montré que la méthode de résolution d'ellipse adoptée convient à un système de TAP comme *MedSLT*. Cependant, pour assurer l'efficacité d'un tel système, bâti sur l'utilisation d'un langage contrôlé, plus particulièrement dans le contexte du bidirectionnel limité aux ellipses, il est primordial de pouvoir guider l'utilisateur vers la couverture du système. Nous allons voir comment l'intégration du modèle d'aide *MedSLT* nous aide à accomplir cette tâche.

2.4.3 Module d'aide dans le bidirectionnel

Ce module d'aide est intégré dans les différentes versions de *MedSLT*, mais nous décrivons en particulier celui du bidirectionnel. Sa fonction est d'aider l'utilisateur (médecin et patient) à retrouver la formulation adéquate pour obtenir une traduction. Il existe deux types d'aides différentes, l'aide pour le médecin exclusivement issue du modèle de langage statistique (MSL), et l'aide qui indique des pistes de réponse au patient dans le cadre du bidirectionnel. À chaque fois qu'un utilisateur de l'interface médecin du système clique sur le bouton « *Push to talk* » (Figure 2.f), le reconnaiseur produit deux résultats : la reconnaissance principale issue du modèle de langage linguistique (MLL), qui sera utilisée pour la traduction, et la reconnaissance secondaire produite par un MSL, qui servira de point de départ au module d'aide.

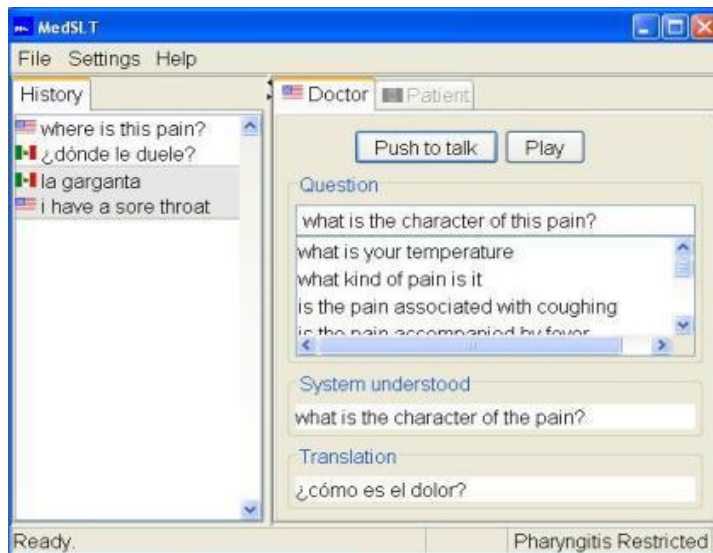


Figure 2.f : Capture d'écran du bidirectionnel anglais-espagnol, onglet médecin

Le MSL n'a pas dû être créé spécifiquement pour le module d'aide parce qu'une version statistique du système existait déjà. Elle avait été développée pour comparer de manière équitable un traitement statistique et linguistique pour la TAP (Rayner, Bouillon et al. 2004). Actuellement, le MSL nous sert uniquement de ressource pour notre module d'aide. Pour l'entraîner, nous avons utilisé le module *SayAnything* de Nuance (Nuance 2003).

À partir de cette reconnaissance MSL, même partielle, une liste de questions plausibles est extraite d'un corpus précompilé de questions et de réponses jugées exactes par les développeurs, au moyen d'une simple comparaison n-grammes. Ce corpus est dynamique, il est recompilé à chaque reconstruction du système et correspond à l'état de développement de la grammaire. Pour les réponses du patient, le système d'aide cherche dans le corpus d'aide la question la plus proche de celle qui a été envoyée à la traduction et propose les réponses correspondantes. Ainsi une liste de phrases d'aide est proposée au patient, avant même qu'il ne lance la reconnaissance, comme l'illustre la Figure 2.g ci-dessous.



Figure 2.g : Capture d'écran du bidirectionnel anglais-espagnol, onglet patient

Pour créer le corpus d'aide auquel est comparée la sortie du MSL, il faut fournir au module deux ressources principales : un fichier de traductions produites par le système et un fichier créé manuellement contenant les classes d'aide. La fonction de ces classes d'aide est d'améliorer la comparaison n-grammes en apportant des informations plus ciblées sur le sens.

Au final, ces classes sont en lien direct avec les classes du lexique de la grammaire *Regulus*. Il reste à définir dans ce fichier quels sont les mots qui, au contraire, ne doivent pas être retenus, parce que ce sont des mots vides, comme les articles dans notre cas, puisque nous ne les retenons pas dans la représentation. Le module d'aide décrit ci-dessus est primordial pour guider les utilisateurs vers la couverture du système. Notamment dans le cas des patients, qui n'auront sans doute qu'une occasion unique d'utiliser le système *MedSLT*, c'est le seul moyen pour qu'ils puissent utiliser de manière efficace notre outil, surtout si les versions diffèrent d'une utilisation à l'autre comme ce sera le cas dans notre expérience décrite à la section 4.1.2 et au chapitre 5. Nous allons maintenant voir en quoi se distinguent les deux versions utilisées dans notre partie expérimentale.

2.4.4 Différentes versions du bidirectionnel : V1 et V2

Ainsi que nous l'avons annoncé plus haut, nous comparerons deux versions différentes de notre système *MedSLT*. Ces deux versions bidirectionnelles (restreinte et non-restreinte) que nous évaluerons au chapitre 5 ont été développées pour la paire anglais-espagnol et le sous-domaine des maux de gorge. Comme nous l'avons vu, il est assez aisé de dériver différentes versions du système *MedSLT*. Les deux versions bidirectionnelles que nous utilisons ne diffèrent que par la grammaire de reconnaissance du patient. Il suffit donc de construire deux

corpus différents pour produire deux grammaires de reconnaissance pour chaque version. Celle pour la version restreinte est spécialisée pour ne reconnaître que des phrases elliptiques ; celle pour la version non-restreinte peut reconnaître, en outre, des phrases déclaratives complètes. Pour l'espagnol, la spécialisation de la version non-restreinte est très importante parce qu'elle limite l'ambiguïté de la grammaire générale. Du fait que la grammaire spécialisée ne contient pas de règles pour les phrases interrogatives, une phrase comme « *tiene fiebre* » (« il a de la fièvre ») ne recevra que l'interprétation correcte : « Il a de la fièvre », et non l'autre interprétation possible : « Avez-vous de la fièvre ? »).

La version 1 (V1) est plus restreinte en ce sens qu'elle ne couvre que les phrases elliptiques en rapport direct avec les questions posées par le médecin ou les réponses courtes comme « *sí* », « *no* », « *no sé* », « *no entiendo* » (« oui », « non », « je ne sais pas », « je ne comprends pas »). La version 2 (V2), moins restreinte, permet aux patients de répondre par oui ou par non, des ellipses et des phrases complètes comme « *Tengo fiebre por la mañana* » (« J'ai de la fièvre le matin ») ou « *Me duele la garganta desde hace más de una semana* » (« J'ai mal à la gorge depuis plus d'une semaine »). Cependant, les phrases entières couvertes par V2 sont seulement affirmatives et, dans une mesure restreinte, relatives et non interrogatives. Ces dernières sont réservées aux médecins, c'est-à-dire uniquement, en espagnol, pour la traduction des questions posées par les médecins en anglais.

Pour illustrer la différence entre ces deux versions avec des chiffres, précisons que la version V1 est entraînée sur un corpus ne comportant que 141 phrases tandis que le corpus d'exemples de V2 comporte 453 phrases, c'est-à-dire presque quatre fois plus d'exemples.

C'est sur ces deux versions (V1 et V2) que sera testée notre méthode d'évaluation EAGLES-ISO dont le but est de comparer de manière fiable des systèmes de TAP dans le domaine médical.

2.5 Conclusion

Dans ce chapitre, nous avons tout d'abord expliqué en quoi consiste la traduction automatique de la parole dans le domaine médical. Elle peut-être décomposée en trois sous-tâches principales : la RV, la TA et la synthèse vocale. Nous avons mis en avant les différences entre les technologies linguistiques et statistiques, en insistant sur les avantages et inconvénients de chaque type de système. Pour la TA statistique, les principaux avantages sont la fluidité de la TA obtenue en se fondant sur de vastes corpus de traductions humaines et la robustesse du système. L'inconvénient de la TA statistique est son caractère imprévisible

(section 2.1.1). Pour la TA linguistique, l'inconvénient principal est le coût que représente le développement pour parvenir à décrire tout un domaine. L'avantage de ce type de TA est la précision et la fiabilité qu'elle apporte (section 2.1.2).

Nous avons également vu qu'il n'est pas toujours facile de classer les systèmes de TAP selon les techniques mises en œuvre : statistiques ou linguistiques. D'autant plus que l'évolution actuelle tend vers une hybridation des systèmes en combinant les deux technologies.

Ensuite, nous avons exposé d'autres manières de classer les systèmes de TAP : par leur architecture (unidirectionnelle, bidirectionnelle ou à phrases fixes), par la couverture du système (générale ou limitée à un domaine), le degré de liberté d'expression, l'interface et le support choisi (section 2.2).

Nous avons poursuivi ce chapitre par un tour d'horizon des systèmes de TAP spécialisés dans le domaine médical. Nous avons vu comment ont évolué les principaux systèmes dans le temps grâce aux moyens mis à disposition par les grands consortiums internationaux successifs (C-STAR, TC-STAR, U-STAR) et les grands projets américains du DARPA (BABYLON à BOLT, en passant par TRANSTAC). Cette évolution a impliqué le développement de systèmes résolument statistiques durant la dernière décennie par le fait de la disponibilité accrue de données nécessaires au développement de ce type de technologie.

Nous avons terminé le présent chapitre par une description du système *MedSLT* en mettant en lumière les particularités de ce système de TAP dans le médical (résolution d'ellipse et module d'aide) par rapport aux autres systèmes décrits dans la section 2.3. Nous avons tout d'abord expliqué le fonctionnement général de ce système linguistique fondé sur une grammaire de type interlangue. Puis, dans la section 2.4.4, nous avons mis en avant quelles sont les différences entre les deux versions du système que nous étudierons dans la partie empirique de ce travail. En conclusion, notre système est un système linguistique de type interlangue alors que la majorité des systèmes actuels du domaine médical sont statistiques (*Jibbig*, *IBM Transtac*, *BBN Transtalk*, *Transonics*). C'est également le cas de *Google Translate* qui, comme nous l'avons vu dans l'introduction, est souvent utilisé dans la pratique hospitalière sans être un système spécialisé du domaine.

Le Tableau 2.2, à la page suivante, reprend les informations les plus importantes concernant la plupart des systèmes de TAP du domaine médical décrits dans ce chapitre.

Au chapitre suivant, nous effectuerons un état de l'art de l'évaluation de la TAP, puisque c'est l'élément essentiel de notre thèse.

Système	Couverture	Architecture	Reconnaissance	Traduction	Langues	Matériel
Converser (Spoken Translation)	300 000 mots, général	Bidirectionnel et uni Multimodal	Statistique <i>Dragon</i> , dépendant de l'utilisateur	Linguistique	Spa – Eng (prototype pour Jap et Ger)	portable
IBM MASTOR IBM Transtac	30 000 mots, limité au triage médical	Bidirectionnel	Statistique	Statistique	Ara - Eng Chi - Eng	portable ou PDA
S-MINDS (Fluential)	limité au triage médical	Bidirectionnel	Linguistique Nuance, SRI, HTK	Hybride : statistique basée sur les syntagmes	Kor - Eng Esp - Eng	portable ou PDA
Transonics Speechlinks	Médical militaire	Bidirectionnel	Statistique	Statistique	Ang-Pachtou Ang-Coréen	portable ou PDA
Speechalator (CMU)	Limité	Bidirectionnel	Statistique (JANUS)	Linguistique (Interlangue) + entraînement statistique	Egyptien	PDA
Jibbigio (CMU)	Général	Bidirectionnel	Statistique	Statistique	Grande variété de langues	smartphone
TransTac Program prototype (CMU)	Militaire, médical et formation	Bidirectionnel	Statistique (trigrammes) (JANUS)	Statistique pour iraquien Phrases fixes et Statistique pour farsi	Farsi et Iraquien – Eng	portable pas de GUI (eyes-free, hands free)
BBN Transtalk	Militaire et médical	Bidirectionnel	Statistique	Statistique	Ang, iraquien, pashto, dari, farsi, malaisien, indonésien	Portable, conception spéciale
Nespole medical SLT	Tourisme et médical	Bidirectionnel	Statistique (JANUS)	Modules multiples (Interlangue)	Eng, Ger, Fre, Ita, Spa	portable
Phraselator	500 -1000 mots par domaine	Phrases fixes Unidirectionnel	Dynaspeak, SRI	Phrases pré-traduites	60 langues	PDA, conception spéciale
MedSLT	1000 - 2000 mots (selon langue) langage contrôlé	Uni et bidirectionnel	Linguistique	Linguistique, langage contrôlé	Module d'aide	Portable, PDA ou smartphone

Tableau 2.10 : Récapitulatif des principaux systèmes de TAP du domaine médical

3 Evaluation de la traduction automatique de la parole (TAP)

La tâche qui consiste à évaluer des systèmes de TAP est compliquée du fait qu'il n'existe pas de méthodologie simple et claire. Dans la pratique, de nombreuses méthodes peuvent être employées. Pourtant, l'évaluation est une étape indissociable du développement de logiciels (de manière générale et en TAP aussi) afin de déterminer la bonne évolution du système. Généralement, à ce stade, l'évaluation se restreint aux résultats produits par le logiciel, dans son ensemble ou, le plus souvent, par chacune des sous-tâches qui le composent. La perspective de l'utilisateur final n'est souvent prise en compte qu'en bout de chaîne de production, à l'obtention d'un prototype fonctionnel du logiciel en question.

Notre objectif est d'évaluer des systèmes en cours de création mais dans une optique centrée sur l'utilisateur final, afin de guider le développeur pour qu'il arrive à un produit qui correspond aux attentes de ceux qui vont véritablement utiliser le logiciel. Nous recherchons par conséquent une méthode d'évaluation adaptée à cet objectif. Cette quête nous mènera à explorer, dans un premier temps, un cadre théorique pour observer ensuite la pratique de l'évaluation dans le domaine qui nous intéresse, soit la TAP et, plus particulièrement, la TAP dans le domaine médical. Notre état de l'art de l'évaluation ne sera pas exhaustif, mais ciblé sur nos intérêts et les systèmes de TAP. Nous constaterons dans la section 3.4, que la TAP est très souvent évaluée par rapport à la performance de ses composantes, c'est-à-dire la RV et la TA (cf. campagnes IWSLT, section 3.4.1). Cependant, d'autres caractéristiques comme la facilité d'utilisation ou la fiabilité d'un système sont particulièrement intéressantes à évaluer dans notre contexte d'utilisation en milieu hospitalier ou médical. C'est ainsi que nous nous sommes tournée vers un cadre tel que celui proposé par les normes ISO (ISO/IEC 2001, ISO/IEC 2003, ISO/IEC 2004, ISO/IEC 2014) sur l'évaluation des logiciels, dont le rôle est de guider l'évaluateur en proposant une série de caractéristiques de qualité normées. Le but est de construire une évaluation complète et conforme à un contexte d'utilisation précis.

Cette approche coïncide avec l'objectif de l'évaluation dans le contexte des logiciels de traitement automatique des langues proposée par (Jones et Galliers 1995 : p.179) : « *Evaluation in NLP is not only to identify and describe variables at some generic level, but to capture variables and values in a form that can be directly tied to a system working* ». Cette citation témoigne de l'importance des paramètres et de l'environnement d'exploitation du logiciel. Elle fait allusion à un concept important de l'évaluation de logiciels : le **contexte d'utilisation**, qui

est déterminé par les utilisateurs, les tâches et l'équipement, ainsi que par l'environnement dans lequel le logiciel sera utilisé (ISO/IEC 2011)⁶². C'est donc le contexte d'utilisation qui déterminera le modèle de qualité du logiciel évalué. Les normes ISO offrent un cadre général à l'évaluation des logiciels qui doit être adapté et interprété pour chaque type de logiciel, mais dont le but reste de standardiser les pratiques d'évaluation. Cet effort de standardisation a été accompagné de travaux d'application à des domaines propres au traitement automatique de la langue (TAL), à l'instar des travaux de *l'Expert Advisory Group on Language Engineering Standards* (EAGLES - projet européen des années 90) et de l'ISLES (projet européen postérieur à EAGLES spécialisé sur l'évaluation de la TA), suivi de FEMTI (King, Popescu-Belis et al. 2003).

Les recherches effectuées par le groupe européen EAGLES⁶³, dont l'objectif était de définir des standards dans le secteur des industries de la langue, ont abouti à une méthodologie d'évaluation pour les outils de TALN, qui a été mise en œuvre dans les domaines de la correction orthographique (EAGLES 1996), annexe D) et pour la traduction assistée par ordinateur (EAGLES 1996), annexe E). Ces travaux ont débouché sur un modèle de qualité décrit dans le Standard ISO/IEC 9126 (ISO/IEC 2001) et (King 2003), et résumé dans une version simplifiée en sept étapes : *the seven steps recipe*⁶⁴. Cette recette met en avant l'utilité d'un système par rapport à un besoin précis. Les travaux de ce groupe ont été poursuivis dans le cadre du Projet International Standards for Language Engineering, (ISLE – 1999-2002), puis se sont concentrés sur la TA dans le cadre du projet *Framework for the Evaluation of MT in ISLE* (FEMTI) (Hovy, King et al. 2002).

Il existe certes des centaines de mesures de qualité proposées ces dernières décennies pour évaluer des logiciels ; toutes sont issues de la pratique et fondées principalement sur des intuitions (Desharnais and Abran 2012). Toutefois, il reste de nombreux domaines d'application pour lesquels une méthode d'évaluation fondée sur ces principes n'a pas encore été définie ou adoptée, notamment celui de la traduction automatique de la parole (TAP). Afin d'adapter le cadre ISO à nos besoins (chapitre 5), nous nous proposons d'employer la méthodologie d'EAGLES (EAGLES 1996) en suivant les sept étapes que le groupe de travail a proposées.

⁶² Source : section définition d'ISO/IEC 25010:2011 : « context of use : users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used ». Cette définition est identique à celle de ISO/IEC 9126-1 : 2001.

⁶³ Site du groupe EAGLES, , consulté le 30.06.2015.

⁶⁴ Site du rapport final EAGLES, présentation des sept étapes d'évaluation de logiciel TAL, , consulté le 02.08.2015.

Nous décrirons, dans la section suivante, les principes fondamentaux de cette méthode et des normes ISO sur lesquelles elle repose. Tout d'abord, nous reprendrons brièvement le cadre général d'ISO, en nous référant aux désormais classiques ISO/IEC 9126, ISO/IEC 14598 et à la série de normes *Software Product Quality Requirements and Evaluation* (SQuaRE) – ISO 25000 (qui remplacera graduellement les normes précitées) pour les définitions des concepts de base de l'évaluation. Notre optique sera de mettre en évidence les nouveautés et différences entre la nouvelle et l'ancienne série sur l'évaluation des logiciels. Ensuite, nous détaillerons la structure hiérarchique du modèle d'évaluation ISO, qui distingue l'évaluation de la qualité à l'usage (section 3.2) de celle de la qualité interne et externe (section 3.3) des logiciels. Enfin, nous clorons ce chapitre par une revue des pratiques d'évaluation des systèmes de TAP en général et plus particulièrement dans le domaine médical.

3.1 Normes ISO sur l'évaluation et EAGLES : principes de base

Les normes ISO (les séries ISO/IEC 9126 et ISO/IEC 14598 mais également la nouvelle série ISO 25000 *Software Product Quality Requirements and Evaluation* (SQuaRE) (Azuma, 2001)) proposent un cadre général en matière d'évaluation des logiciels. Ces normes donnent la définition suivante de la **qualité d'un logiciel** : « *features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs* » (ISO/IEC 9126, 2001), qui met en avant les besoins (explicites ou implicites) de l'utilisateur. SQuaRE (ISO/IEC 25000:2014) fournit la définition suivante de l'évaluation de la qualité d'un logiciel : « *systematic examination of the extent to which a software product is capable of satisfying stated and implied needs* »⁶⁵. Cette définition met en exergue la nécessité d'étudier de manière systématique si le logiciel étudié satisfait aux besoins (explicites et implicites) des utilisateurs.

Selon (ISO/IEC 9126-1:2001), l'évaluation d'un logiciel passe par un **modèle de qualité**⁶⁶ qui va décomposer la qualité d'un logiciel de manière hiérarchique en caractéristiques et sous-caractéristiques. Ainsi, le modèle de qualité doit nous permettre de trouver des moyens efficaces de mesurer, c'est-à-dire de quantifier, la qualité de nos logiciels. Cette approche rejoint celle de Sparck Jones et Galliers (1994 : p. 21), selon qui « (...) les besoins doivent être définis de manière suffisamment concrète afin d'être traduisibles en des mesures quantitatives. »

Au-delà de la définition du modèle de qualité, le groupe de travail européen EAGLES, qui a appliqué les normes ISO 9126 notamment aux logiciels d'aide à la traduction (correction

⁶⁵ Source : ISO/IEC 25000:2014, p.6.

⁶⁶ Source : ISO/IEC 25021:2012, définition de «quality model» : defined set of characteristics, and of relationships between them, which provides a framework for specifying quality requirements and evaluating quality.

d'orthographe et de grammaire ou mémoire de traduction), propose de guider l'évaluateur grâce à une méthode en sept étapes⁶⁷, décrites dans le Tableau 3.1 ci-dessous.

n°	Description des sept étapes EAGLES
1	Définition du but de l'évaluation : Quel est l'objectif de l'évaluation ? Est-ce le système en entier ou une composante du système qui est évalué ?
2	Définition du modèle des tâches : Identifier qui va utiliser le système et pour quelle utilisation. Quel est le profil de l'utilisateur ?
3	Définition des caractéristiques de haute qualité : Identifier les caractéristiques à évaluer dans le contexte d'utilisation donné.
4	Définition du modèle de qualité hiérarchique : Choisir les caractéristiques et sous-caractéristiques. Le modèle de qualité doit aboutir à une subdivision finale appelée attribut, lequel pourra être quantifié par une mesure.
5	Spécification des mesures à appliquer : Comment mesurer les attributs identifiés à l'étape précédente ? Définition des mesures et de l'interprétation des échelles.
6	Préparation de l'évaluation : Mise en place de l'évaluation. Définition du protocole d'évaluation
7	Exécution de l'évaluation et rédaction du rapport

Tableau 3.1 : Principales étapes de l'évaluation selon EAGLES⁶⁸

Comme le montre le Tableau 3.1, cette méthode propose une véritable marche à suivre illustrée par des exemples d'application. Son étape centrale est de définir le modèle de qualité selon les besoins identifiés, comme nous le ferons dans le chapitre 4 (section 4.1.1). Ce modèle de qualité correspond à un ensemble de caractéristiques, reliées entre-elles, qui déterminent le cadre de définition des exigences de qualité et de l'évaluation de la qualité d'un logiciel⁶⁹. Le but n'est pas d'évaluer chaque caractéristique systématiquement, mais de déterminer quelles sont les caractéristiques de qualité importantes dans un contexte d'utilisation particulier et comment les mesurer. Par conséquent, chaque caractéristique finale de qualité retenue doit pouvoir faire l'objet d'un certain nombre de mesures calculables. Ainsi, le rôle de ces normes est de guider l'évaluateur dans la création d'un modèle de qualité adapté au contexte d'utilisation. Selon (Popescu-Belis 2007), nous pourrions définir l'évaluation d'un système de TAL en suivant les principes des normes ISO de la manière suivante : « Évaluer un système, c'est donc mesurer sa qualité en utilisant une décomposition fondée sur un modèle de qualité accompagné de mesures pour chaque attribut pertinent dans le contexte d'utilisation prévu. » (Popescu-Belis 2007 : p.71).

67 Site du projet EAGLES, <http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html>, accédé le 02.07.2015

68 Site du projet EAGLES, <http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html>, accédé le 02.07.2015.

69 Source : section définition de la norme (ISO 25010 :2011) : « quality model: defined set of characteristics, and of relationships between them, which provides a framework for specifying quality requirements and evaluating quality »

Selon la norme ISO/IEC 9126, le modèle de qualité est constitué de deux parties : d'une part, la qualité interne et externe et, d'autre part, la qualité à l'usage (*quality in use*, aussi traduit par qualité d'utilisation). Chaque type de qualité se décompose en caractéristiques et sous-caractéristiques pour former un modèle de qualité hiérarchique (Figure 2.a).

La **qualité interne** peut être mesurée sans exécution du logiciel et correspond à un ensemble d'attributs statiques (architecture du logiciel, nombre de règles, structure et composants d'un logiciel)⁷⁰. Il s'agit généralement de mesures statiques de composantes du système, le plus souvent cette évaluation intervient avant même la commercialisation du produit. Les attributs de **qualité externe**⁷¹ doivent être mesurés en exécutant le logiciel évalué puisqu'il s'intéresse aux résultats produits (Popescu-Belis, 2007) durant son fonctionnement. Enfin, la **qualité à l'usage**⁷² « doit être mesurée en plaçant le système dans un contexte d'utilisation, expérimental ou final, et en observant dans quelle mesure le système aide ses utilisateurs à accomplir leurs tâches » (Popescu-Belis 2007 : p.70).

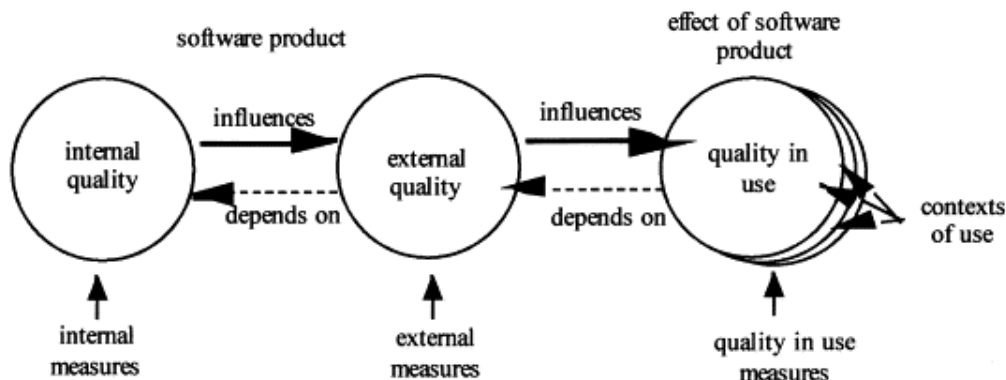


Figure 3.a : Relations entre qualité interne, externe et qualité à l'usage⁷³

Comme nous pouvons le constater dans la Figure 3.a, reprise des normes ISO pour illustrer la relation entre les différents types de qualité, l'objectif final est d'obtenir un produit qui corresponde à un contexte d'utilisation final. La qualité à l'usage adopte le point de vue de l'utilisateur final. Les caractéristiques externes (telles que l'adéquation, la précision, la

⁷⁰ Source: section definition de la norme ISO/IEC 2500:2014: « internal measure of software quality: measure of the degree to which a set of static properties of a software product satisfies stated and implied needs for the software product to be used under specified conditions ».

⁷¹ Source: *ibid.* : « External measure of system or software quality : measure of the degree to which a system or software product enables its behaviour to satisfy stated and implied needs for the system including the software to be used under specified conditions ».

⁷² Source: *ibid.* : «Quality in use: degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, freedom from risk and satisfaction in specific contexts of use ».

⁷³ Schéma classique repris de : Bevan, N. (1999). «Quality in use: Meeting user needs for quality. » *Journal of Systems and Software* 49(1) : 89-96.

résistance à l'erreur ou l'efficacité-temps) auront un impact sur la qualité à l'usage. De la même façon, une mauvaise qualité à l'usage (non-accomplissement d'une tâche) peut provenir de la qualité externe (comme l'adéquation et l'opérabilité) et de la qualité interne correspondante (Bevan 1999). Nous détaillerons à présent les caractéristiques qui relèvent de la qualité à l'usage, puis discuterons la qualité interne et externe dans la section 3.3.

3.2 Qualité à l'usage

Selon la norme ISO 9126-1 (ISO/IEC 2001), la notion de qualité à l'usage d'un système correspond à l'aptitude du produit à permettre à des utilisateurs identifiés d'atteindre des buts définis avec efficacité, productivité, sécurité et satisfaction, dans un contexte d'utilisation précis. La qualité à l'usage se mesure en utilisant le logiciel ou le système en question. Elle s'intéresse aux résultats fournis par le logiciel plutôt qu'aux propriétés du logiciel. Il s'agit de vérifier à quel point le système convient aux besoins de l'utilisateur dans son environnement de travail donné.

La notion de qualité à l'usage est décomposée dans la norme (ISO/IEC 9126 : 4, 2004) en quatre caractéristiques : efficacité, productivité, sécurité et satisfaction. Pour la nouvelle série de normes SQuaRE (série ISO 25000), le modèle de qualité à l'usage est composé désormais des cinq caractéristiques suivantes : **efficacité** (*effectiveness*- précision et degré d'achèvement avec lesquels l'utilisateur atteint des objectifs spécifiés⁷⁵), **efficience** (*efficiency* - rapport entre les ressources dépensées et la précision et le degré d'achèvement avec lesquels l'utilisateur atteint des objectifs spécifiés⁷⁶), **satisfaction** (*satisfaction* - absence d'inconfort, et attitudes positives dans l'utilisation du produit⁷⁷), **absence de risque** (*freedom from risk*) et **conformité** au contexte d'utilisation⁷⁸ (*context coverage*). Cette série de normes n'a pas encore été traduite en français. Pour plus de détails sur ces caractéristiques et sous-caractéristiques de la qualité à l'usage (ainsi que les définitions) selon SQuaRE, il convient de se référer au Tableau 3.2. Il en ressort également que les deux séries de normes relatives à la qualité à l'usage sont proches l'une de l'autre. Les principales différences sont l'ajout de la caractéristique de conformité au contexte d'utilisation et des sous-caractéristiques pour la satisfaction et pour le nouvel équivalent de la caractéristique sécurité (absence de risque) ainsi que le changement de

⁷⁴ ISO/IEC 9126-1: 2001, p.12: «The capability of the software product to enable specified users to achieve specified goals with effectiveness, productivity, safety and satisfaction in specified contexts of use ».

⁷⁵ Source de la définition : ISO 9241-210:2010 (fr) - Ergonomie de l'interaction homme-système — partie 210 : Conception centrée sur l'opérateur humain pour les systèmes interactifs

⁷⁶ Source de la définition : *ibid.*

⁷⁷ Source de la définition : *ibid.*

⁷⁸ Il s'agit plutôt ici d'une glose sur ce concept qu'une véritable traduction.

nom de productivité au profit d'efficience dans un souci de cohérence avec les normes ISO/IEC 25062 et ISO 9241-11 (ISO/IEC 25010 : 2011).

Caractéristiques 9126_1 Définitions (en anglais)	Caractéristiques Square Sous-caractéristiques	Définitions (en anglais)
Efficacité (Effectiveness) : The capability of the software product to enable users to achieve specific goals with accuracy and completeness in a specified context of use.	Efficacité (<i>Effectiveness</i>)	Accuracy and completeness with which users achieve specified goals
Productivité (Productivity) : The capability of the software product to enable users to expend appropriate amounts of resources in relation to the effectiveness achieved in a specified context of use.	Efficience (<i>Efficiency</i>)	Resources expended in relation to the accuracy and completeness with which users achieve (time to complete the task, materials, or the financial cost of usage.)
Satisfaction : The capability of the software product to satisfy users in a specified context of use.	Satisfaction	Degree to which user needs are satisfied when a product or system is used in a specified context of use
	Utilité (<i>Usefulness</i>)	Degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the results of use and the consequences of use
	Confiance (<i>Trust</i>)	Degree to which a user or other stakeholder has confidence that a product or system will behave as intended
	Plaisir (<i>Pleasure</i>)	Degree to which a user obtains pleasure from fulfilling their personal needs
	Confort (<i>Comfort</i>)	Degree to which the user is satisfied with physical comfort

Sécurité (Safety) : The capability of the software product to achieve acceptable levels of risk of harm to people, business, software, property or the environment in a specified context of use.	Freedom from risk	Degree to which a product or system mitigates the potential risk to economic status, human life, health, or the environment
	<i>Economic risk mitigation</i>	Degree to which a product or system mitigates the potential risk to financial status, efficient operation, commercial property, reputation or other resources in the intended contexts of use
	<i>Health and safety risk mitigation</i>	Degree to which a product or system mitigates the potential risk to people in the intended contexts of use
	<i>Environmental risk mitigation</i>	Degree to which a product or system mitigates the potential risk to property or the environment in the intended contexts of use
	Context coverage	Degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in both specified contexts of use and in contexts beyond those initially explicitly identified
	<i>Context completeness</i>	Degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in all the specified contexts of use
	<i>Flexibility</i>	Degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in contexts beyond those initially specified in the requirements (Flexibility can be achieved by adapting a product for additional user groups, tasks and cultures).

Tableau 3.2 : Comparatif des caractéristiques et sous-caractéristiques de la qualité à l'usage selon ISO/IEC 9126 :1 et SQuaRE ISO/IEC 25010 ⁷⁹

Puisque la norme ISO/IEC 25024 – l'équivalent de la norme (ISO/IEC 9126 :4 2004) – n'a pas encore été approuvée et est encore en cours d'élaboration et que le travail sur la norme

⁷⁹ Source : ISO/IEC 25010 (2011).

ISO 9126 :4 est plus répandu⁸⁰ au moment de la rédaction de ce travail, nous fondons notre travail sur les quatre caractéristiques initiales : efficacité, productivité, sécurité et satisfaction.

Il n'est pas toujours facile de quantifier une notion aussi abstraite que la qualité à l'usage (Aberdeen, Oshika et al. 2005). La difficulté de ce type d'évaluation est qu'elle nécessite de mettre en place une simulation d'utilisation du logiciel à évaluer dans le but de s'approcher des conditions réelles.

Dans le domaine médical, la mise en place d'une telle expérience de simulation n'est pas aisée à organiser. Tout d'abord, le problème des sujets-évaluateurs se pose : le mieux serait d'effectuer des tests en conditions réelles, mais ce n'est pas toujours possible, en particulier dans un domaine à risque, comme celui du médical. Il est, en effet, difficile dans un premier temps de parvenir à une couverture suffisante dans tous les domaines pour pouvoir utiliser un système avec des patients réels. Il n'est pas vraiment envisageable d'attendre dans la salle d'urgence jusqu'à ce qu'un cas correspondant à la couverture du système se présente. De plus, évaluer si un système est efficace avec de véritables patients suppose des problèmes de sécurité médicale, tout au moins dans les premières phases d'évaluation. Cette solution est donc à priori difficile à appliquer. Il reste la possibilité de mettre en place un scénario proche de la réalité, en procédant à une simulation. Cela empêche de se heurter aux problèmes liés aux particularités du diagnostic médical. Une solution plus sûre en attendant la maturité d'un système est d'utiliser des patients standardisés, soit des personnes formées pour simuler des pathologies prédéterminées, jouer le rôle des patients et des étudiants en fin de médecine dans le rôle des médecins. Sans être idéale, cette solution peut constituer une première étape. Dans le cas de l'évaluation conduite par (Aberdeen, Oshika et al. 2005), on trouve dans le rôle des patients des personnes dont la langue maternelle est la langue cible et qui ont été formées par des professionnels de la médecine (PM) mais qui n'ont aucune connaissance préalable du système. C'est le PM qui montre à l'utilisateur le fonctionnement du système. D'autres études que nous verrons dans la dernière partie de ce chapitre ont mis en place des simulations notamment dans le domaine militaire (Jones, Gatewood et al. 2014) ou dans le cadre de TRANSTAC (Translation for Tactical use) financé par le département de la défense américaine (DARPA) (Sanders, Weiss et al. 2013) (cf. section 4.4.2). Une fois les participants identifiés, il reste à déterminer des attributs quantifiables. ISO donnent des pistes qui coïncident le plus souvent avec la pratique : le taux d'accomplissement de la tâche et le temps consacré. Nous allons

⁸⁰ Plus de 10 000 résultats pour la série de normes 9126 sur Google Scholars contre 2860 pour la nouvelle série. Recherche effectuée le 26.03.2015.

décrire les mesures proposées par la Norme ISO/IEC 9126-4 : 2004 pour chaque caractéristique de qualité à l'usage.

3.2.1 Efficacité : accomplissement de la tâche

L'efficacité concerne la capacité du logiciel à aider l'utilisateur à accomplir certaines tâches avec précision dans un contexte d'utilisation particulier.

Selon la norme ISO/IEC 9126-4 :2004, les mesures de l'efficacité sont typiquement l'accomplissement de la tâche, l'efficacité de la tâche s'agissant de la proportion des objectifs de la tâche atteints correctement par l'utilisateur, mais aussi la fréquence de l'erreur (ISO/IEC 9126-4 :2004 : p.6).

Dans la pratique, cette mesure se traduit le plus souvent par le taux d'accomplissement de la tâche, ainsi que le montrent les campagnes d'évaluation du MITRE (Aberdeen, Oshika et al. 2005) et de TRANSTAC (Sanders, Weiss et al. 2013) (section 3.4.2). Dans le domaine de la TAP, il s'agira d'effectuer une tâche précise grâce à l'utilisation du système testé.

3.2.2 Productivité

La productivité porte sur la capacité du logiciel à attribuer les ressources en accord avec l'efficience atteinte par l'utilisateur. Elle se mesure en temps consacré à l'exécution de la tâche, mais également au niveau de l'investissement pécuniaire et matériel. Cette caractéristique s'appelle d'ailleurs efficience dans la nouvelle norme SQuaRE (ISO/IEC 25010 :2011).

La mesure classique pour la productivité est le temps consacré à l'exécution de la tâche. En TAP, cette mesure n'est pas toujours facile à interpréter ; un dialogue trop court pouvant éventuellement provenir d'un abandon. Cette mesure doit par conséquent être reliée à l'accomplissement de la tâche.

La norme 9126-4 préconise également de mesurer la productivité par l'efficience de la tâche. Dans l'étude (Aberdeen, Oshika et al. 2005), celle-ci a été mesurée par l'efficacité du dialogue calculée en fonction du nombre d'interactions réussies par rapport au nombre de tours de parole total. L'efficience a aussi été évaluée par le nombre de répétitions engendrées par un mauvais fonctionnement ou une mauvaise utilisation du système.

Les trois dernières mesures proposées par la norme 9126-4 sont plus difficiles à mettre en œuvre ; il s'agit de la productivité économique, du taux de productivité, c'est-à-dire le temps véritable d'actions productives menées par l'utilisateur (ce qui nécessite d'enregistrer les tests effectués afin de les analyser) et de l'efficacité relative de l'utilisateur. Cette dernière consiste

à comparer la performance entre l'utilisateur de l'étude et un utilisateur expert (p.8 (ISO/IEC 2004)).

3.2.3 Sécurité : absence de danger

La sécurité s'applique à la capacité du logiciel d'atteindre des niveaux acceptables de risque lors de l'utilisation du logiciel dans un contexte d'utilisation précis⁸¹. Cette caractéristique est assez vaste et a notamment été divisée pour plus de clarté en sous-caractéristiques dans la nouvelle norme SQuaRE (ISO/IEC 25010 :2011).

Les mesures relatives à la sécurité selon la norme 9126-4 s'attachent à trois dimensions : la mise en danger de la santé de l'utilisateur (*user health and safety*), la sécurité des personnes touchées par l'utilisation du système (*safety of people affected by use of the system*) et le dommage économique (*economic damage*). Comme l'explique Bevan (2009), il n'est pas facile de mesurer la sécurité d'un logiciel. Cela revient la plupart du temps à évaluer les conséquences d'erreurs du système (*product failure*) ou d'erreur humaine lors de l'utilisation du système (*human error*) sur les trois dimensions mentionnées supra. Souvent, le renforcement de la sécurité du système repose sur l'amélioration de l'interface ou du système lui-même⁸². Dans le cadre de l'évaluation de la TAP, il s'agira de mesurer les erreurs d'interaction (dus à l'humain ou au logiciel) avec le système qui peuvent mettre en danger la communication établie grâce au système de TAP étudié.

3.2.4 Satisfaction : questionnaire

La satisfaction est une des caractéristiques principales de la qualité à l'usage. Elle est divisée en quatre sous-caractéristiques dans la nouvelle série de norme ISO : utilité (*usefulness*), confiance (*trust*), plaisir (*pleasure*) et confort (*comfort*). L'utilité d'un logiciel peut être décrite selon (Sanders et al. 2013) comme étant la « valeur qu'apporte un système à son utilisateur final » (Scholtz, 2006).

La satisfaction correspond à la capacité du logiciel de satisfaire les utilisateurs dans un contexte d'utilisation particulier. Conformément à la définition de la norme 9126-4, les mesures de la satisfaction sont le reflet de l'attitude de l'utilisateur par rapport à l'utilisation d'un

⁸¹ Définition de « safety » : The capability of the software to achieve acceptable levels of risk of harm to people, business, software, property or the environment in a specific context of use (ISO 9126-1: 2001).

⁸² P. 18, Bevan, N. (2009). Extending Quality in Use to Provide a Framework for Usability Measurement. Human Computer Interaction International Conference (HCI). San Diego, California, USA, Springer: 13-22.:«It is not easy to measure product safety, but it is usually possible to list the potential adverse consequences of product failure or human error, based on previous experience with similar products or systems. Multiplying the estimated frequency by the estimated impact of the potential failures provides a basis for prioritization of the need to find design solutions to minimize the probability of failures occurring. The solutions can include improving the user interface to reduce the probability of human error».

produit. Alors que les mesures évoquées dans les sections 3.2.1, 3.2.2 et 3.2.3 relèvent plutôt d'une évaluation objective, la détermination de la satisfaction de l'utilisateur est réalisée au travers d'un questionnaire de satisfaction qui permet de collecter des informations plus subjectives. Les utilisateurs peuvent ainsi s'exprimer sur la qualité à l'usage du système dans son ensemble ou sur certains points particuliers (caractéristiques de qualité) en répondant à des questions précises.

Il existe des questionnaires clés en main afin d'assurer la comparabilité des résultats, en particulier dans le cas de questionnaires très courts à remplir après une étude reposant sur un scénario d'utilisation (Lewis 1991). La pratique la plus répandue est le questionnaire de satisfaction rempli juste après l'accomplissement de la tâche sous la forme de questions à répondre sur une échelle de Likert (Brooke 1996). Cette échelle est composée de cinq degrés exprimant l'accord ou le désaccord du répondant, comme suit :

1.	Strongly disagree	Pas du tout d'accord
2.	Disagree	Pas d'accord
3.	Neither agree nor disagree	Sans avis
4.	Agree	D'accord
5.	Strongly agree	Entièrement d'accord

Tableau 3.3 : Échelle de Likert et sa traduction en français

L'avantage de cette échelle est qu'elle est devenue un standard et qu'elle est par conséquent connue par la plupart des répondants, ce qui facilite son utilisation. Elle est, par ailleurs, communément admise dans la recherche, en psychologie notamment. Son inconvénient majeur est lui aussi connu. L'échelle de Likert peut influencer les réponses des participants par l'effet de biais de la tendance centrale. Dans ce cas, les répondants montrent leur réticence à donner des réponses « extrêmes » et restent par conséquent au milieu de l'échelle proposée. En construisant avec soin le questionnaire et en répartissant de manière égale le nombre de questions positives et négatives, un autre biais peut être évité, celui de toujours vouloir acquiescer à une question (Lewis 1991, Brooke 1996)

3.3 Qualité interne et externe

Pour la notion de qualité de produit (interne ou externe), la norme ISO/IEC 9126 :1 (2001) propose un modèle de qualité de produit qui se décompose en six caractéristiques, illustrées dans la Figure 3.b, ci-dessous.

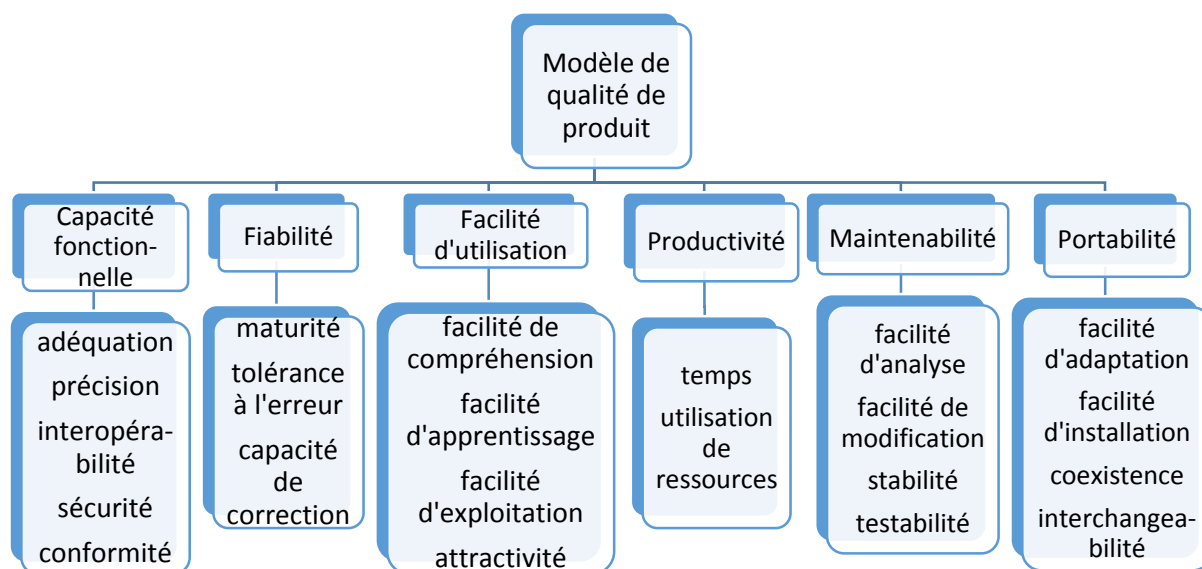


Figure 3.b : Modèle de qualité externe de produit selon ISO/IEC 9126 :1 (2001)

À partir des caractéristiques et sous-caractéristiques choisies en fonction du contexte d'utilisation, les attributs à évaluer seront sélectionnés pour finalement être quantifiés.

Nous reprendrons brièvement chaque caractéristique de la norme ISO/IEC 9126 (1991), en précisant comment chacune est à son tour décomposée en sous-caractéristiques, afin d'établir le plus précisément quelles sont les fonctions indispensables du système évalué. L'importance de chaque caractéristique varie selon le système à évaluer.

La **capacité fonctionnelle** – ou fonctionnalité – concerne l'activité du logiciel. Pour l'évaluer, il faut vérifier l'existence d'un ensemble de fonctions et leurs propriétés données. La capacité fonctionnelle se décompose en plusieurs sous-caractéristiques : l'**adéquation** concerne la capacité d'un logiciel de proposer un ensemble de fonctionnalités pour effectuer une tâche précise et atteindre les objectifs des utilisateurs (King 2007). La **précision**, quant à elle, se réfère à la capacité d'un logiciel de produire les résultats ou l'effet voulus. La caractéristique la plus communément évaluée est celle de la précision du logiciel sous l'angle de la sortie du logiciel. Dans le cas de la TAP, elle est mesurée par la précision de ses sous-composantes : la précision de la RV et la précision de la TA. En résumé, il s'agit d'évaluer si le résultat obtenu par le système est conforme aux attentes. L'**interopérabilité** se rapporte à la capacité d'un logiciel d'interagir avec un autre système. La **sécurité** correspond à la capacité du logiciel de protéger l'accès aux données et d'empêcher les personnes non autorisés d'y accéder. Enfin, la dernière sous-caractéristique de la capacité fonctionnelle est la **conformité**.

La **fiabilité** regroupe un ensemble d'attributs portant sur l'aptitude du logiciel à maintenir son niveau de service dans des conditions précises et pendant une période déterminée. Ses sous-caractéristiques sont la maturité d'un logiciel (les principales erreurs d'une version beta sont-elles corrigées ?), la tolérance à l'erreur, qui informe sur la robustesse d'un logiciel, et la capacité de correction, qui correspond à la capacité d'un logiciel de fonctionner à nouveau sans perte d'information suite à une panne. La fiabilité va pouvoir être mesurée par rapport au nombre d'erreurs commises par le système : si trop d'erreurs persistent et, surtout, si le logiciel ne permet pas de continuer de travailler ou de revenir en arrière après une erreur, le logiciel ne sera pas considéré comme fiable.

La caractéristique de **facilité d'utilisation** ou d'**utilisabilité** (*usability*) concerne l'ensemble des attributs portant sur l'effort à fournir lors de l'utilisation et sur l'évaluation individuelle de cette utilisation par un ensemble défini ou implicite d'utilisateurs.

À propos de la notion d'utilisabilité, cette caractéristique est souvent confondue avec la définition de « *usability*⁸³ » de la norme (ISO 9241-11, 1998), qui reprend de multiples concepts⁸⁴ (Abran et al. 2003) que nous classons dans la qualité à l'usage, comme le préconise la norme ISO 9126.

En effet, Abran, Khelifi et al. (2003) reprennent les différents concepts de facilité d'utilisation (cf. section 3.3) issus des normes ISO 9126 (ISO/IEC 2001 : p.9) sur lesquelles nous nous fondons) : « *a set of attributes that bear on the effort needed for use and on the individual assessment of such use, by a stated or implied set of user* » mais également sur la série des normes 9241 qui donne la définition suivante pour la facilité d'utilisation « *usability* » : « *software is usable when it allows the user to execute his task effectively, efficiently and with satisfaction in the specified context of use* » ce qui correspond plutôt à la définition de qualité à l'usage dans la série 9126. Cela s'explique que la norme 9241-11 se situe dans une approche orientée vers le processus. Il faut noter que, dans les campagnes DARPA décrites dans la section 3.4.2, il est fait référence à « *system usability testing* » (Sanders, Weiss et al. 2013) dans le sens de la norme 9241-11 et que par conséquent nous l'interprétons comme faisant partie du modèle de qualité à l'usage, dans la mesure où ce concept reprend les mêmes

⁸³ Définition de la facilité d'utilisation (*usability*) selon (ISO 9241-11, 1998) « extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use » cited in (Abran et al.2003).

⁸⁴ Source (Abran et al. 2003 : p.326) : « multiple concepts taken together : execution time, performance, user satisfaction and ease of learning ».

caractéristiques que celles décrites dans la section 3.2. Nous ajoutons cette précision parce que ces deux concepts sont souvent à juste titre (selon la norme suivie) confondus.

Concrètement, selon la définition de la norme 9126⁸⁵, l'utilisabilité se traduit par l'effort investi par l'utilisateur pour comprendre la « logique » et l'application (Popescu-Belis, 2007 : p.95). Les sous-caractéristiques de la facilité d'utilisation sont la **capacité de compréhension**, la **capacité d'apprentissage** de l'utilisateur et la **facilité d'exploitation** (*operability*) ainsi que l'**attractivité** du logiciel. Ces dernières concernent plus particulièrement l'interface et le fonctionnement d'un logiciel. S'agissant de la **capacité de compréhension**, ce principe repose sur la capacité de l'utilisateur de comprendre comment le logiciel fonctionne et ce qu'il est supposé faire. Généralement, cette capacité est renforcée grâce à la mise à disposition d'un manuel d'utilisation ou de brèves instructions ou encore grâce à une interface claire. La **capacité d'apprentissage** correspond à la progression d'apprentissage de l'utilisateur, soit les moyens mis à disposition de l'utilisateur pour lui permettre d'apprendre à utiliser un nouveau logiciel. Cette caractéristique peut être mesurée grâce à l'observation d'une progression dans l'efficacité d'utilisation ou de la diminution de l'accès à l'aide en ligne. Enfin, la **facilité d'exploitation** et l'**attractivité** du logiciel dépendent beaucoup de l'efficacité de l'interface. Il est particulièrement important d'optimiser ces caractéristiques lorsque les utilisateurs sont amenés à utiliser un logiciel une seule fois (Tsourakis 2013 : p.211).

La **productivité** ou le **rendement** (*productivity*) du système (également traduit par efficacité du système dans Popescu-Bellis 2007) concerne l'ensemble des attributs portant sur le rapport existant entre le niveau de service d'un logiciel et la quantité de ressources utilisées dans des conditions déterminées. Pour mesurer ce point, il faudrait pouvoir comparer la situation en présence et en l'absence du logiciel.

La **maintenabilité** se rapporte à l'effort nécessaire pour faire des modifications données. Ses sous-caractéristiques sont la capacité d'analyse (des erreurs commises), la capacité de changement, la stabilité et la capacité de tester. Ce sont des caractéristiques plutôt orientées vers le développeur, comme l'est la caractéristique de portabilité, qui concerne l'aptitude du logiciel à être transféré d'un environnement à l'autre.

⁸⁵ Définition de “usability” selon ISO 9126-1, 2000 : « The capability of a software product to be understood, learned, used and attractive to the users, when used under specified conditions. »

Nous avons brièvement vu toutes les caractéristiques pour la qualité externe. Dans les sections suivantes, nous nous concentrerons sur l'évaluation de la **capacité fonctionnelle** du logiciel, approche qui est très courante dans la pratique pour mesurer la qualité de la TAP (section 3.4). Nous allons plus précisément nous focaliser sur la sous-caractéristique de la capacité fonctionnelle la plus étudiée : la précision. Celle-ci est mesurée par le biais de l'évaluation de la précision des sous-tâches qui composent la TAP (reconnaissance vocale et traduction automatique), pour lesquelles il existe de nombreuses mesures établies. Dans la suite de ce chapitre (section 3.3.1), nous décrirons les mesures⁸⁶ issues de la reconnaissance vocale appliquées à la TAP, en nous arrêtant sur les problèmes précis rencontrés dans notre cadre qu'est la TAP médicale (section 3.3.1). Dans la section 3.3.2, nous dresserons un tour d'horizon des mesures de la TA, notamment celles employées lors de l'évaluation de la TAP, faisant la distinction entre mesures externes humaines (section 3.3.2.1) et mesures externes automatiques (section 3.3.2.2). Nous ne ferons donc pas un état de l'art exhaustif mais adapté à nos besoins. Dans la section 3.4, nous proposerons une vue d'ensemble de l'évaluation de la TAP en nous penchant sur les pratiques des grandes campagnes d'évaluations conduites régulièrement (section 3.4.1), celles des grands projets de financement de la TAP (section 3.4.2) et pour finir celles des systèmes de TAP dans le médical (section 3.4.3).

3.3.1 Mesures de la précision issues de l'évaluation de la RV

Les mesures les plus couramment utilisées pour l'évaluation de la reconnaissance vocale appliquée à la TAP sont les taux d'erreur détectés au niveau des mots (**WER**, *Word Error Rate*) et de la phrase (**SER**, *Sentence Error Rate*) (Vogel, Och et al. 2000), (Rossato, Blanchon et al. 2002).

Le WER s'apparente à la distance minimum d'édition, c'est-à-dire au nombre minimal d'opérations (insertion, substitution ou omission) nécessaires pour arriver à une correspondance exacte entre la phrase reconnue et la phrase dite. Il se calcule de la manière suivante (Jurafsky and Martin 2009) :

$$\text{WER} = 100 * (\text{insertions} + \text{substitutions} + \text{omissions}) / \text{nombre total de mots dans la transcription de la phrase dite}$$

Il est important de noter qu'étant donné que le modèle inclut les insertions, le taux d'erreurs au niveau du mot peut excéder les 100%.

⁸⁶ Les normes ISO SQuaRE (ISO/IEC (2014) préconisent l'utilisation de *mesure* au lieu de *métrique*. Nous suivons donc cette recommandation.

Le SER est l'équivalent du WER, mais il est étendu à la phrase.

$$\text{SER} = 100 * \text{nombre de phrases comportant au moins un mot incorrect} / \text{nombre total de phrases}$$

Ainsi, lorsqu'on mesure le taux d'erreur au niveau de la phrase (**SER**, *Sentence Error Rate*), une phrase est considérée comme correcte uniquement si elle est **exactement** identique à la phrase originale (dite par l'utilisateur). À titre d'exemple, un système comme *MedSLT* obtient un taux de SER entre 75% et 100% pour des phrases non couvertes par la grammaire et autour de 20% pour celles qui sont couvertes (Rayner, Bouillon et al. 2004). Cette mesure est en général considérée comme trop stricte pour véritablement apporter un éclairage sur les performances réelles de la qualité du système. Selon Somers et Sugita (2003), le SER est trop sévère et n'a pas beaucoup de sens dans le contexte de la TAP puisqu'une seule erreur de reconnaissance dans une phrase suffit pour rejeter une phrase entière alors que pour la plupart des tâches comprenant de la reconnaissance vocale quelques erreurs de reconnaissance n'ont pas d'effet sur la compréhension globale (Somers et Sugita 2003 : p. 371)⁸⁷. Nous l'illustrons ci-dessous par des exemples issus de la reconnaissance avec *MedSLT*. Dans l'exemple du Tableau 3.4, l'article « un » n'a pas été reconnu par le système, mais, pour les deux phrases, la représentation sémantique aurait été identique parce que l'article n'y est pas représenté (section 2.4).

Transcription : el dolor de cabeza está causado por un movimiento brusco de la columna vertebral (<i>Les maux de tête sont causés par un mouvement brusque de la colonne vertébrale.</i>) Reconnaissance : el dolor de cabeza está causado por [omission] movimiento brusco de la columna vertebral Représentation sémantique : [[utterance_type,ynq],[symptom,dolor],[body_part,cabeza],[tense,present], [action,causar],[cause,movimiento_brusco],[body_part,columna_vertebral]]

Tableau 3.4 : Omission d'un article lors de la reconnaissance par MedSLT

De la même manière, comme notre représentation sémantique ne fait souvent pas la différence entre le pluriel et le singulier, une erreur de reconnaissance comme dans l'exemple du Tableau 3.5 (*dolor – dolores*) n'empêche pas les deux phrases d'obtenir la même

⁸⁷ Somers, H. and Y. Sugita (2003). Evaluating Commercial Spoken Language Translation Software. MT Summit IX, New Orleans.

représentation. De ce fait, ces erreurs n'ont pas d'impact sur la traduction avec un système tel que *MedSLT*.

Transcription : cuándo aparecen los dolores (*Quand apparaissent les douleurs*)

Reconnaissance : cuándo aparece un dolor

Représentation sémantique :

[[utterance_type,whq],[symptom,dolor],[tense,present],[action,aparecer]]

Tableau 3.5 : Substitution d'un article lors de la reconnaissance par *MedSLT*.

Ainsi, pour tempérer les résultats du WER ou du SER, la mesure du **SemER** est utilisée. Le taux d'erreurs sémantiques (*SemER*, *Semantic Error Rate*) est plus conforme aux réalités de la reconnaissance vocale (Ehsani, Frederking et al. 2008 : p. 232). Il mesure la qualité sémantique des phrases en partant de la qualité du contenu sémantique. Il existe différentes manières de mesurer le SemER. Dans notre cas, il correspond au nombre de phrases dont la rétro-traduction a été rejetée par l'utilisateur, c'est-à-dire au nombre de phrases qu'il n'a pas jugé qu'elles correspondaient à la phrase originale au niveau du sens (Tsourakis, Lisowska et al. 2010). Des phrases, qui sont en général seulement altérées par des erreurs minimales comme l'omission d'un article, sont reformées correctement au moment de la rétro-traduction vers la langue originale du fait que la représentation sémantique n'a pas été touchée par cette omission. Ce type d'erreur ne devrait donc pas avoir d'impact négatif sur le score obtenu, ce qui explique pourquoi on préfère le SemER au WER comme mesure d'évaluation pour la RV de la TAP. Le SemER pour *MedSLT* est d'environ 10% pour les phrases couvertes par la grammaire dans le domaine des maux de tête (Rayner, Bouillon et al. 2008), mais varie selon la langue, le domaine et l'évaluation effectuée. Le SemER se situe autour de 30% pour tout le corpus dans le domaine des maux de gorge, comme nous le verrons plus loin (Starlander, Bouillon et al. 2008).

Hors ligne, le SemER peut aussi être jugé manuellement, par un humain, en comparant le sens de la transcription à celui de la phrase reconnue. Il est considéré comme similaire ou non. Cependant, l'appréciation de ce jugement binaire dépend beaucoup du juge. En effet, un évaluateur familier avec la façon dont fonctionne le système reconnaîtra mieux les erreurs qui n'ont pas d'impact sur la représentation sémantique comme celle de l'exemple 5.1 de la page précédente. Il saura quelles phrases sont acceptables, car elles obtiendront un bon résultat au final, c'est-à-dire une traduction fidèle, malgré les quelques erreurs de RV incluses. Par contre, un évaluateur « novice », même s'il a eu comme consigne de seulement s'attacher au sens de la phrase, rejettera plus de phrases en pensant qu'elles étaient incorrectes. Pour éviter ces

différences de jugement, il est préférable de calculer le SemER lors de l'utilisation véritable du système, c'est-à-dire au moment où l'utilisateur du système accepte ou non une phrase sur la base de la rétro-translation qui lui est fournie en temps réel. Cette mesure peut fonctionner seulement dans les cas où la TAP comprend une étape intermédiaire de présentation de la rétro-translation (section 2.4), qui donne un certain contrôle de qualité à l'utilisateur. Ce n'est par exemple pas le cas de *Google Translate*, mais bien le cas d'autres systèmes de TAP dans le militaire/médical comme le système décrit dans (Jones, Gatewood et al. 2014). Cette mesure reflète mieux la qualité sémantique d'une phrase puisque c'est l'utilisateur lui-même qui décide si une phrase est conforme à ce qu'il voulait dire. Ce chiffre reflète donc mieux ce qui se passait en cours d'utilisation. Une autre possibilité serait d'automatiser cette mesure, en envisageant de comparer la représentation sémantique obtenue avec une représentation de référence. Le nombre de différences devrait renseigner sur la détérioration de la reconnaissance vocale qui a véritablement un impact sur la suite du traitement, à savoir la traduction automatique, ce qui permettrait de ne plus comparer les mots, mais le contenu sémantique. Cela ressemble un peu à ce qui a été fait par (Aberdeen, Oshika et al. 2005), comme nous le verrons plus en détail dans la section 3.4.2, qui proposent une méthode où les interactions ne font pas seulement l'objet d'une transcription littérale, mais également d'une transcription au niveau des actes de parole afin de vérifier le contenu sémantique.

Nous verrons les résultats obtenus par *MedSLT* avec ces mesures issues de la reconnaissance vocale au chapitre 5. La suite décrit les principales mesures issues de la TA qui sont communément appliquées à la TAP.

3.3.2 Mesures de la précision issues de l'évaluation de la TA

Parmi les mesures externes de la traduction automatique (TA) appliquées couramment à la TAP, on peut distinguer, d'une part, les mesures humaines et, d'autre part, les mesures automatiques apparues plus tard. Nous n'allons pas effectuer un historique complet de l'évaluation de la TA ici mais rappeler certaines dates clés. La naissance de l'évaluation de la TA est étroitement liée à celle de l'apparition des systèmes de TA. La première démarche date de la publication du rapport ALPAC (Pierce and Carroll 1966), qui fournit en annexe une procédure standard pour l'évaluation humaine de la qualité de la TA basée sur l'intelligibilité et la fidélité de la traduction produite sur une échelle de 9 points. Une autre source d'avancement dans le domaine de l'évaluation de la TA provient des grands projets comme EUROTRA (Hutchins and Somers 1992) et notamment le rapport de Van Slype (Van Slype 1979) qui introduit des notions importantes comme celle de la prise en compte du but de la TA (diffusion

ou compréhension minimale) et de son destinataire pour déterminer la qualité nécessaire, et aussi le type de texte produit.

Dans les années 90, deux groupes de travail, l'un japonais (JEIDA- Japanese Electronic Industry Development Association, (Isahara 1995)) et l'autre européen (EAGLES - Expert Advisory Group on Language Engineering Standards,), ont contribué à faire avancer la méthodologie de l'évaluation. Ceux-ci définissent la qualité de la traduction par rapport au contexte d'utilisation. Dans cette approche, la qualité de la TA est mesurée en fonction de la capacité d'exécuter une tâche en utilisant la TA.

Deux types d'événements ont ensuite précipité les recherches dans le domaine. D'une part, les différentes campagnes d'évaluation organisées aux USA par le département de la défense (ARPA puis DARPA) et par le National Institute of Standards and Technology (NIST) afin de comparer des systèmes de TA commerciaux ou de recherche, pour la plupart linguistiques. D'autre part, l'avènement de la TA statistique, qui a nécessité la création de méthodes d'évaluation plus rapides et moins coûteuses afin de pouvoir tester les résultats au gré des changements effectués dans les algorithmes de traduction. Cela explique l'apparition des premières mesures automatiques dès les années 90. Elles ont trouvé un large public seulement à partir de la création de la mesure phare BLEU - Bilingual Evaluation Understudy (Papineni, Roukos et al. 2002), sur laquelle nous reviendrons dans la section 3.3.2.2.1.

De nombreuses campagnes d'évaluation ont aussi régulièrement eu lieu dans le cadre d'ateliers dans le domaine de la TA (MetricsMatr lancé par le NIST (Callison-Burch, Koehn et al. 2010), le projet GALE⁸⁸ du DARPA et le projet français EVALDA-CESTA⁸⁹), mais aussi directement en TAP (campagnes d'évaluation compétitives annuelles IWSLT, depuis 2004 dans le cadre du consortium *C-Star*, et du projet européen *TC-Star*⁹⁰). Leur objectif est le plus souvent de comparer des systèmes entre eux, mais aussi les différentes méthodes d'évaluation, notamment en mesurant le degré de corrélation entre les diverses mesures automatiques et les mesures humaines, que nous décrirons dans les sections suivantes.

3.3.2.1 Évaluation humaine

L'évaluation humaine est réalisée par des juges humains qui doivent donner des scores à des traductions. On peut distinguer quatre différents types d'évaluation humaine. Elle peut être basée sur une évaluation subjective (1), sur une comparaison entre deux traductions produites

⁸⁸ Site du projet Gale , consulté le 02.08.2011.

⁸⁹ Site du projet Evalda-Cesta, , consulté le 02.08.2011.

⁹⁰ Site du projet TC-STAR, , consulté le 02.08.2011.

(2), sur l'accomplissement d'une tâche (3) ou enfin par rapport à l'effort de post-édition requis pour arriver à une traduction de qualité suffisante pour son utilisation (4). Nous examinerons ces quatre catégories d'évaluation humaine dans cette section.

Les critères de qualité de la TA retenus dans le rapport ALPAC sont l'**intelligibilité** et la **fidélité**. Ils relèvent de la première catégorie (évaluation subjective). L'intelligibilité se mesure uniquement sur le texte cible. Elle concerne la clarté d'un texte et son degré de compréhensibilité. Mais elle ne peut être un critère suffisant pour déterminer la qualité de la traduction dans son ensemble. En effet, un élément indispensable reste la fidélité au texte source. L'évaluation de l'intelligibilité se fait en premier sans avoir recours au texte source, celle de la fidélité en second. Les scores des deux mesures sont liés, mais la différence est que, pour la fidélité, les évaluateurs se réfèrent au texte source.

Dans le Rapport ALPAC (Pierce and Carroll 1966), une échelle allant de 1 (désespérément inintelligible) à 9 (parfaitement clair et intelligible) a été adoptée pour la mesure d'intelligibilité. Pour la fidélité, l'échelle allait de 0 à 9 par rapport au contenu de l'original préservé dans la traduction. Cette échelle n'était pas très clairement définie et difficilement applicable (Estrella 2008). Plus une échelle comporte de points, plus elle peut être difficile à interpréter. Cependant, une échelle à deux niveaux uniquement (ex : fidèle / infidèle) semble trop restrictive, même si elle est plus facile à appliquer. L'étude de (White, O'Connell et al. 1993), menée dans le cadre de la campagne DARPA, montre qu'une échelle de 5 points est plus précise, par rapport à une échelle binaire (du type oui/non) ou à 3 points.

3.3.2.1.1 *Mesures classiques : Fluidité / adéquation et classement binaire*

En 2003, NIST propose un protocole d'évaluation appliqué dans le cadre du projet TIDES fondé sur les critères de **fluidité** (fluency) et d'**adéquation** (adequacy), qui sont devenus les grands classiques de l'évaluation humaine en remplaçant les critères d'intelligibilité et de fidélité proposés dans le rapport ALPAC. L'adéquation d'un énoncé dépend du degré auquel les informations présentes dans l'original sont bien transmises par la traduction. Afin d'éviter de faire appel à des juges humains bilingues, on peut utiliser une traduction de référence au lieu de l'original en langue source. Cela revient à demander d'évaluer la proportion du sens exprimé dans la référence qui est aussi présente dans la traduction (Blanchon et Boitet 2007 : p.44) , ce qui se rapproche beaucoup de la mesure de fidélité du rapport ALPAC et de (Van Slype 1979).

La fluidité est mesurée selon le degré de bonne formation de l'énoncé selon différents critères : respect de la grammaire, de l'orthographe, usage commun du vocabulaire, degré

d'interprétation par un locuteur natif (Blanchon et Boitet 2007). La fluidité n'est pas la même chose que l'intelligibilité (caractère de ce qui est intelligible⁹¹) ou la compréhensibilité (caractère de ce qui est compréhensible⁹²) puisqu'une phrase peut-être tout à fait fluide mais obscure au niveau du sens. La fluidité et l'adéquation s'évaluent souvent sur une échelle de cinq points, comme suggéré par (White, O'Connell et al. 1993), mais elle peut aussi s'articuler sur une échelle de sept points (Sanders, Przybocki et al. 2011) et (Dorr, Olive et al. 2011).

Le Tableau 3.6 reprend les 5 points de l'échelle tels qu'appliqués pour la fluidité et l'adéquation décrite dans le Projet TIDES⁹³ et inspirée de (Blanchon et Boitet 2007 : p.44) pour la traduction en français. Cette échelle est très proche de celle appliquée à l'intelligibilité.

Points	Fluidité	Adéquation
5	Langue cible sans défaut	Toute l'information présente
4	Bonne langue cible	Presque toute l'information
3	Langue cible non-native	Beaucoup d'information
2	Langue cible non-standard	Peu d'information
1	Incompréhensible	Aucune

Tableau 3.6 : Échelle NIST pour la fluidité et l'adéquation

Par la suite, d'autres variantes de ces mesures humaines sont apparues, notamment la notion de **clarté** d'une traduction qui combine l'intelligibilité, la lisibilité et le style en une mesure évaluée sur une échelle de 4 points (Miller et Vanni 2005). Cette mesure, qui s'approche de l'intelligibilité, est issue des travaux ISLE (Hovy, King et al. 2002). Il faut noter que cette mesure ne prend pas du tout en compte la fidélité. Une traduction peut ainsi être très claire tout en étant porteuse d'un sens différent. La clarté est donc accompagnée d'une mesure de cohérence avec le texte source, comme dans NIST lorsque la fluidité est complétée par l'adéquation.

Ce résumé non-exhaustif nous montre que la variation de ce type de mesures est potentiellement infinie. Au niveau des mesures humaines basées sur une évaluation subjective, l'adéquation et la fluidité se sont cependant imposées (Dorr, Olive et al. 2011 : p.748). Dans la plupart des recherches, ce sont ces mesures qui servent de standard pour la comparaison avec les mesures automatiques. Dans le chapitre 6, nous proposerons une échelle plus adaptée aux besoins de la TAP, inspirée de (Somers et Sugita 2003), qui propose d'utiliser une échelle de 7 points basée sur l'utilité de la TA produite.

⁹¹ Définition du Petit Robert, version 2012.

⁹² Définition du Petit Robert, version 2012.

⁹³ Site LDC décrivant le projet TIDES, , consulté le 18.08.2011.

Dans la deuxième catégorie de méthodes d'évaluation humaine, le classement par comparaison de qualité est devenu une mesure populaire lors des grandes campagnes d'évaluation de TA statistique. Il s'agit simplement de classer des traductions par rapport à d'autres. L'évaluateur doit émettre un jugement binaire : la traduction X est-elle meilleure que la traduction Y ? L'avantage de ce type d'évaluation est qu'il est plus intuitif à appliquer que les échelles de fluidité et d'adéquation. De plus, ce type d'évaluation manuelle obtient un meilleur taux d'accord entre évaluateurs (Callison-Burch, Fordyce et al. 2007).

Enfin, il existe les mesures humaines basées sur l'**utilité** (task-based), c'est-à-dire en évaluant à quel point une traduction permet d'effectuer une tâche, que ce soit de compréhension en utilisant des tests à choix multiples ou une tâche plus spécifique comme la recherche d'information sur la base de textes traduits automatiquement comme dans (Tate 2008).

Ces mesures humaines comportent cependant un inconvénient non négligeable qui est celui de devoir fournir un travail supplémentaire artificiel important pour évaluer des traductions automatiques dont le but est de faire gagner du temps aux utilisateurs. Ainsi l'idée d'évaluer la traduction automatique en demandant à des post-éditeurs d'apporter les corrections minimales afin d'arriver à un texte juste d'un point de vue syntaxique, orthographique (Dabbadie, Hartley et al. 2002, Blanchon et Boitet 2007) et sémantique a fait son apparition. Cette mesure peut se baser sur une comparaison de temps mis pour effectuer la post-édition ou sur le nombre d'édicions nécessaires pour arriver à une traduction correcte. De nos jours, avec la généralisation de la post-édition après la TA, cette approche a pris de l'importance (Specia 2011).

3.3.2.1.2 *Cohérence entre évaluateurs*

En outre, les mesures humaines présentent un autre inconvénient mais à des degrés divers. Comme le démontrent de nombreuses études (Turian, Shen et al. 2003, Akiba, Federico et al. 2004, Kulesza et Shieber 2004), il est très rare d'avoir des juges cohérents. L'accord entre évaluateurs demeure un problème majeur. Afin d'éviter des écarts trop importants, les instructions données aux évaluateurs doivent être les plus claires possibles. Idéalement, il faudrait répéter la procédure de test jusqu'à ce que les évaluateurs parviennent véritablement à interpréter l'échelle des scores de manière uniforme. Le nombre minimum d'évaluateurs (E) est de $P+1$, P correspondant au nombre de points sur une échelle, à condition d'arriver à un nombre impair afin d'être en mesure de calculer le mode (c'est-à-dire, la classe d'une distribution

statistique qui présente la fréquence la plus élevée⁹⁴, la tendance centrale) (Hamon, Fügen et al. 2009). Plus le nombre d'évaluateurs est élevé, plus l'évaluation est fiable. En outre, les évaluateurs devraient connaître le domaine sur lequel portent les textes concernés (Arnold, Balkan et al. 1994). Il est préférable de faire des séances d'entraînement de sorte que les évaluateurs s'accordent sur l'échelle, afin d'éviter des erreurs d'interprétation qui donnent lieu à des disparités au niveau de la sévérité. Enfin, il ne faut pas oublier lorsque l'on entreprend une évaluation humaine que le profil de l'évaluateur va influencer les résultats. Nous avons fait l'expérience avec *MedSLT* : si une traduction automatique est évaluée par un traducteur, son jugement sera beaucoup plus sévère que s'il s'agit d'un utilisateur non traducteur (Chatzichrisafis, Bouillon et al. 2006). Lors d'une évaluation de la version anglais-français de *MedSLT*, certains évaluateurs qui n'étaient pas du milieu médical étaient beaucoup plus sévères que les médecins. En effet, ceux-ci ont accordé de meilleurs scores à *MedSLT* que les traducteurs, lorsqu'il s'est agi d'évaluer les mêmes phrases traduites dans le domaine des maux de tête.

De toutes les manières, en présence de mesures humaines, il est indispensable de déterminer si les évaluateurs sont d'accord les uns avec les autres, en calculant le degré de cohérence entre les évaluateurs (*inter-rater agreement*) et, dans le meilleur des cas, de l'évaluateur avec lui-même, à l'intérieur d'une même évaluation (*intra-rater agreement*). Une option simple pour évaluer l'accord entre évaluateurs serait de calculer un simple pourcentage d'accord. L'avantage d'une telle mesure est qu'elle est facilement compréhensible et interprétable. Toutefois, ce pourcentage ne tient pas compte des accords fortuits, ce qui aurait tendance à surestimer le véritable niveau d'accord. La mesure la plus couramment utilisée pour vérifier cette cohérence est le **kappa**.

Kappa

Cette mesure initialement proposée dans le domaine médical par (Cohen, 1960) pour introduire une mesure capable de quantifier la cohérence entre des observations faites par deux médecins a été transposée à d'autres domaines pour arriver dans la pratique de la recherche en TAL, souvent sans être bien adaptée ou comprise.

⁹⁴ Définition du Petit Robert, version 2012.

Hallgren (2012) donne un aperçu des bonnes pratiques et des erreurs communément commises dans le cadre de ces méthodes d'évaluation de la cohérence entre évaluateurs. En effet, il n'est pas facile de reconnaître, parmi toutes les variantes décrites et existantes, laquelle appliquer dans quel contexte, en particulier lorsqu'il s'agit d'analyser la cohérence non plus entre deux évaluateurs mais entre plusieurs évaluateurs (notamment en TAL). Le kappa a été adapté au domaine de la linguistique computationnelle par Jean Carletta (Carletta 1996), mais il n'est pas toujours utilisé à bon escient. Nous résumerons les principales erreurs relevées par (Hallgren, 2012) lorsqu'il s'agit d'évaluer la fiabilité des évaluateurs.

Dans la littérature, il est souvent fait référence à un chiffre pour kappa sans préciser la variante utilisée. Il est probablement sous-entendu qu'il s'agit du kappa de Cohen (Cohen, 1960), mais il est souhaitable de préciser la variante adoptée lors du calcul. Par ignorance des différences, la variante de kappa ou autre mesure de fiabilité appliquée n'est souvent pas la plus adaptée à l'étude selon que l'échelle choisie est nominale (l'évaluateur doit choisir une catégorie nommée, ex. homme-femme), ordinale (les catégories qui sont naturellement ordonnées, ex. type Likert) ou à intervalle (échelle numérique). Un autre aspect à prendre en compte concerne le corpus de données. Si les évaluateurs ont évalué l'ensemble des données, il s'agit d'un design complètement croisé (*fully crossed design*). Si la tâche d'évaluation a été partagée par différents évaluateurs nous sommes en présence d'un *multiple-rater design*.

De la même façon, les données auront une influence sur le choix de la mesure. Ainsi, si le résultat de l'évaluation est très homogène (ce sont toujours les mêmes catégories qui sont choisies par les évaluateurs), les données présenteront une prévalence qui a tendance à baisser artificiellement les estimations de fiabilité des évaluateurs. Il faudra choisir la mesure la plus adaptée parmi celles décrites dans le Tableau 3.7 ci-dessous.

Le dernier point soulevé par (Hallgren 2012) s'intéresse à l'entraînement des évaluateurs, qui est souvent négligé, alors qu'il faudrait prévoir plusieurs séances d'entraînement pour améliorer la cohérence entre évaluateurs.

Famille statistique	Nom de la variante	Quand l'utiliser	Références
Kappa (2 évaluateurs)	Kappa de Cohen	Pas de correction de biais ou de la prévalence Echelle nominale	Cohen (1960)
	Kappa de Siegel et Castellan	Correction du biais	Siegel et Castellan's (1988, pp 284-291)
	Kappa de Byrt et al	Correction de la prévalence	Byrt et al. (1993)
	Kappa de Cohen pondéré	Possibilité de pénaliser les grands désaccords	Cohen (1968)
Statistiques proches de kappa	Kappa de Fleiss	Évaluateurs choisis au hasard pour chaque élément évalué	Fleiss (1971); Gross (1986)
	Kappa de Light	Moyenne de kappa calculée par paire d'évaluateurs	Light (1971)
	Kappa de Davies et Fleiss	Coefficient de type kappa pour toutes les paires d'évaluateurs avec une moyenne P(e)	Davis et Fleiss (1982)
ICC	Model	Accord pour variables ordinales, à intervalle ou de ratios (avec un zéro naturel)	McGraw et Wong (1996); Shrout et Fleiss (1979)
		<i>One-way</i> : Les évaluateurs sont choisis au hasard pour chaque objet évalué	
		<i>Two-way</i> : mêmes évaluateurs pour tous les objets évalués	

Tableau 3.7 : Tableau récapitulatif traduit et extrait de Hallgren, 2012

Les mesures de kappa varient de 0 à 1, 1 correspondant à un accord parfait. La difficulté de cette mesure réside dans la façon d'interpréter les résultats. La lecture des données par (Landis et Koch 1977) est la plus communément admise en TAL. Elle est reprise dans le Tableau 3.8, ci-dessous.

Échelle	Interprétation
< 0	Désaccord
0.0 - 0.20	Accord très faible
0.21 - 0.40	Accord faible
0.41 - 0.60	Accord modéré
0.61 - 0.80	Accord fort
0.81 - 1.00	Accord presque parfait

Tableau 3.8 : Échelle d'interprétation de (Landis et Koch, 1976) traduite en français.

Il faut cependant savoir que le nombre de catégories influe fortement sur le résultat du kappa, initialement conçu pour comparer deux catégories par deux évaluateurs différents. Dans notre cas, nous avons au minimum trois évaluateurs et quatre catégories. Nous verrons dans les chapitres suivants l'impact de ce genre d'évaluation humaine par rapport à des évaluations de type binaire de plus en plus pratiquées en complétant le kappa de Fleiss avec une mesure plus

simple à interpréter : le pourcentage d'accord entre humains ou encore le **coefficient de corrélation intraclasse**, qui est généralement utilisé pour estimer en même temps l'accord et la fidélité (ICC – Intraclass Correlation Coefficient; Shrout et Fleiss, 1979). Ce dernier est inclus dans la plupart des logiciels de statistique (ex. SPSS ou Agreestat) (Gwet, 2012). Nous préciserons à chaque application de ces mesures quelle mesure de cohérence nous avons utilisée (variante de kappa, pourcentage brut de cohérence) et pourquoi, en suivant les conseils de Hallgren (2012).

Pour conclure sur l'évaluation humaine de la TA, nous dirons que, si elle est conduite par un grand nombre d'évaluateurs, cela permet de pondérer statistiquement les résultats pour supprimer les incohérences inévitables entre évaluateurs. En outre, si les évaluateurs sont bien entraînés ou effectuent une tâche simple (classement binaire), l'évaluation humaine peut atteindre une cohérence raisonnable tout en présentant l'avantage d'être effectuée par de potentiels utilisateurs du système évalué. Elle reflète ainsi une appréciation humaine. Malheureusement, ce type d'évaluation devient vite onéreux, que ce soit en temps ou en ressources financières. Pour pallier aux inconvénients de l'évaluation humaine, les chercheurs et utilisateurs de systèmes de TA aspirent à créer des mesures automatiques beaucoup plus pratiques et moins coûteuses. Nous verrons quelles ont été, au fil du temps, les solutions préconisées pour évaluer automatiquement la TA.

3.3.2.2 Évaluation automatique de la TA

Les années 90 ont vu la création et la prolifération de mesures automatiques à partir des WER et SER (issus de la RV) pour juger de la qualité de la traduction. Le principal motif invoqué pour utiliser des mesures automatiques est de parvenir à évaluer la TA plus facilement et à moindres coûts⁹⁵ (Tillmann, Vogel et al. 2000). Essentiellement, ces méthodes cherchent à comparer la traduction produite avec une traduction de référence. Le problème dans le domaine de la traduction est qu'il n'existe pas UNE bonne traduction qui peut servir de référence. Il faut donc utiliser plusieurs traductions de référence, ce qui demande malgré tout des ressources en traduction humaine. Ainsi, l'évolution des mesures automatiques va vers des méthodes d'apprentissage afin d'éviter de nouveau une dépendance aux ressources humaines. Nous tenterons d'inventorier les mesures automatiques selon leur appartenance à trois classes : (1) mesures statistiques, comparant une traduction à des références sans tenir compte d'aucune information d'ordre linguistique ; (2) mesures linguistiques, lorsque certains aspects

⁹⁵ Source (Tillmann et al. 2000 : p.55): «automatic and easy-to-use measure of the translation errors ».

linguistiques sont intégrés dans le calcul que ce soit d'ordre lexical, syntaxique ou sémantique ; et enfin (3) mesures d'apprentissage. Parfois, les mesures sont à la frontière entre ces catégories, comme nous le verrons par la suite.

3.3.2.2.1 *Première génération : WER, PER, BLEU*

Toutes les mesures de première génération se basent sur une comparaison entre une ou plusieurs références avec la traduction produite par le système de TA.

WER

La première à voir le jour est le WER, mesure appliquée couramment à la reconnaissance vocale (RV) (Niessen, Och et al. 2000). Dans le cas de la TA, on calcule le nombre de différences entre les mots de l'hypothèse (traduction) et ceux de la référence divisé par la longueur de la référence (Niessen, Och et al. 2000, Jurafsky et Martin 2009). Comme cette mesure compte le nombre d'erreurs, plus le score est élevé, moins la traduction évaluée est proche de la référence. Afin de mieux illustrer ces mesures, prenons un exemple. Quel est le score de WER pour la traduction « *Le duele la garganta* » par rapport à la référence unique « *Tiene la garganta cogida* » ? Etant donné que le WER se calcule en fonction du nombre d'édits par rapport au nombre de mots présents dans la référence ($WER = (S+I+D)/N$ où S= substitution, I = insertion, D = effacement et N= le nombre de mots dans la référence), il faut tout d'abord compter toutes les éditions nécessaires pour arriver de la TA à la référence :

Le	duele	la	garganta	
	tiene	la	garganta	cogida

Tableau 3.9 : Correspondance entre deux phrases pour WER

Nous comptons une substitution de « *duele* » par « *tiene* », une insertion de « *cogida* » et la suppression de « *le* », ce qui nous donne le calcul suivant :

$$WER = (3/4)*100=60\%$$

La difficulté par rapport à la RV, c'est qu'il est absolument impossible en traduction de définir une référence unique, qui serait LA bonne référence. C'est ainsi que la mesure **BLEU** (Papinemi, et al. ACL, 2002) a été proposée par l'équipe d'IBM.

BLEU

La mesure BLEU est basée sur le WER, mais calcule la proximité de la TA produite par rapport à une ou plusieurs traductions de référence. Le calcul se base sur le nombre de n-grammes qui sont présents dans la sortie du système par rapport à un ensemble de références.

L'uni-gramme permet de mesurer l'adéquation d'une traduction tandis que les n-grammes plus élevés servent à vérifier la fluidité des traductions (Papineni, Roukos et al. 2002)⁹⁶. Le score obtenu varie de 0 à 1, 1 étant le score le plus élevé. Il indique que la traduction évaluée est proche des références utilisées. Cette mesure a beaucoup contribué à l'amélioration des systèmes, parce qu'elle donne au développeur un moyen simple et rapide de comparer les différentes versions de leur système. Selon (Coughlin 2003), la corrélation entre BLEU et une évaluation humaine est assez bonne pour certaines fins, et même à partir d'une seule référence. Pour ces raisons, BLEU est l'une des mesures les plus utilisées. Cependant, comme nous le verrons, elle a été critiquée de plus en plus vivement ces dernières années.

Le calcul pour BLEU est le suivant (Papineni, Roukos et al. 2002, Koehn 2004, Blanchon et Boitet 2007) :

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Then,

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right).$$

Comme BLEU est une mesure entièrement basée sur la précision et qui ne prend pas en compte le rappel, elle comprend une pénalité de brièveté (BP) dont le rôle est de pénaliser les traductions trop courtes par rapport aux références données. Si la longueur de la traduction est équivalente à la plus longue des références (ou plus grande), BP= 1 ; autrement, cette pénalité se calcule de la manière suivante :

$$BP = \min(1, \text{longueur\_candidat} / \text{longueur\_référence})$$

Pour la deuxième partie de la formule, il faut calculer P_n, qui est la précision de n-grammes, c'est-à-dire le nombre de mots de la traduction (candidate) présents également dans les références, divisé par le nombre total de mots dans la traduction (P_n) pour chaque n-grammes (unigrammes, bigrammes, trigrammes, généralement jusqu'au quadrigrammes). La première étape consiste donc à compter le nombre de n-grammes correspondants.

⁹⁶Source (Papineni et al. 2002 : p.313) : "A translation using the same words (1-grams) as in the references tends to satisfy adequacy. The longer n-gram matches account for fluency".

Traduction	le	duele	la	garganta	
Référence 1		tiene	la	garganta	cogida
Référence 2	tiene	irritada	la	garganta	
Référence 3	le	duele	la	garganta	

Tableau 3.10 : Correspondance entre deux phrases pour BLEU

Prenons l'exemple du Tableau 3.10, pour lequel le nombre d'unigrammes présents dans la traduction et dans les références est de $4 / 4 = 1$, le nombre de bigrammes est $3/3=1$. Toutefois si nous excluons la dernière référence (Référence 3), la précision de n-grammes ne sera plus que de $1/3 = 0.33$ pour les bigrammes, puisque le seul bi-gramme commun entre la traduction et les références sera "la garganta" parmi les trois bigrammes présents. Dans notre exemple nous avons choisi de ne calculer que les unigrammes et bigrammes, ce qui fait que $n=2$.

Reprenons la formule entière :

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

Nous allons pouvoir déterminer la somme des précisions de n-grammes (P_n) pour l'exemple du Tableau 3.10 en nous basant sur le résultat obtenu pour les unigrammes ($4/4$) et pour les bigrammes ($3/3$). Ainsi, pour le premier cas,

$$BLEU = 1 * \exp(1/2 * \log 4/4 + 1/2 * \log 3/3) = 1 * \exp(1/2 * 0 + 1/2 * 0) = 1 * \exp(0) = 1 * 1 = 1^{97}$$

Cependant, si on applique BLEU à la même traduction, mais en faisant varier les trois références pour prendre uniquement celles fournies par des traducteurs/interprètes données dans le Tableau 3.11, le calcul est le suivant toujours avec $n=2$ et $BP=1$ (parce que le candidat et les références sont de longueur égale) :

Trad	le	duele	la	garganta	
Ref 1		tiene	la	garganta	cogida
Ref 2	tiene	irritada	la	garganta	
Ref 3		tiene	la	garganta	irritada

Tableau 3.11 : Correspondance entre deux phrases pour BLEU

La précision des uni-grammes est de 2 (la, garganta = 2 divisé par le total de mots du candidat (4) et la précision des bi-grammes = $1/3$ (la garganta) divisé par le nombre total de bi-grammes, soit :

$$BLEU = 1 * \exp(1/2 * \log 2/4 + 1/2 * \log 1/3) = 0.68^{98}$$

⁹⁷ Rappelons que « exp » est la fonction exponentielle et que « log » est la fonction réciproque de la fonction exponentielle. Nous avons utilisé les fonctions mathématiques EXP et LOG d'Excel pour effectuer le calcul.

⁹⁸ Idem que supra.

Ces exemples de calculs illustrent les inconvénients de cette mesure. Le premier est sa gourmandise en ressources parce qu'il faut pouvoir fournir au système un certain nombre de références humaines afin de s'approcher du jugement humain (Snover, Dorr et al. 2006). De plus, comme nous l'avons vu plus haut, la qualité de ces références influence beaucoup les résultats et le choix lui-même des références est crucial. Les résultats varieront en fonction du lot de références. Ainsi, ce choix doit être bien réfléchi et le coût de production de ces minimums trois ou quatre références demeure important.

Le deuxième inconvénient souvent relevé, c'est que BLEU ne peut pas prendre en considération les changements légitimes de mots ou d'ordre des mots, mais qu'il compte toutes les différences avec la référence comme des erreurs, que ce soit un changement dans l'ordre des mots ou l'usage de synonymes. De plus, lorsqu'une traduction est plus courte que son original, cela a une influence négative sur les scores de BLEU, du fait de sa pénalité de brièveté (BP, cf. formule). Selon (Hartley et Popescu-Belis 2004), la longueur des phrases, elle-même influence la fiabilité de BLEU. Il est donc certain que si on l'applique sur des phrases plus courtes que quatre mots, la forme classique de BLEU qui est calculé sur base des 4-grammes, le score sera automatiquement nul (Surcin, Hamon et al. 2005). Ainsi, à moins de choisir de réduire le nombre de n-grammes sur lesquelles elle se base, cette mesure n'est pas adaptée pour l'évaluation de phrases courtes, comme nous le verrons dans le chapitre 6.

Enfin, comme l'ont montré un certain nombre de recherches dans le domaine (Snover, Dorr et al. 2006, Callison-Burch, Fordyce et al. 2007, Owczarzak, van Genabith et al. 2007), pour certains types de corpus et pour certaines langues, les corrélations avec les mesures humaines sont assez faibles. C'est le cas pour une combinaison comme l'anglais et le tchèque, pour laquelle l'ordre des mots diffère beaucoup entre l'anglais assez rigide et le tchèque plutôt libre (Kos et Bojar 2009). (Callison-Burch 2006) décrit une contradiction entre les classements de systèmes de différents types (statistiques et linguistiques) au niveau des classements avec BLEU ou une mesure humaine. Cette observation nous intéressera au plus haut point et nous voudrions vérifier si pour nos données BLEU favorise les systèmes statistiques au détriment des systèmes linguistiques comme *MedSLT* (section 6.2.3).

Comme BLEU est fondé uniquement sur la précision et pas du tout sur le rappel (Culy and Riehemann 2003, Lavie, Sagae et al. 2004), très vite, d'autres mesures similaires ont vu le jour comme **NIST** (Doddington et al. HLT2002) qui est une variante de BLEU qui calcule la moyenne arithmétique de la précision des n-grammes par rapport à des traductions de référence. L'avantage de cette mesure par rapport à BLEU est qu'elle est moins sensible à la différence

de longueur. D'autres variantes existent comme **ORANGE** (Lin et Och, 2004a), **Rouge** (Lin et al, 2004b), **WNM** (Babych et al, 2004) et **GTM** (Melamed et al. 2003) que nous n'allons pas décrire ici faute de place.

3.3.2.2 *Deuxième génération : TER, METEOR*

Un exemple de mesure automatique de deuxième génération basée sur les références est le **Translation Edit Rate (TER)** qui a été créé par (Snover, Dorr et al. 2006). TER est une mesure de distance minimale d'édition dont le but est d'arriver avec le moins d'édicions possibles à au moins une des références, mais contrairement à BLEU elle est moins sensible au nombre de références fournies. Elle donne de très bonnes corrélations avec les mesures humaines parce que l'algorithme de comparaison est plus complet que pour un WER classique ou pour BLEU. Dans sa version améliorée, TERplus (TERp) (Snover, 2009), elle prend en compte les déplacements de séquences de mots lors du calcul du nombre d'insertions, omissions et de substitutions. Certes, avec cette mesure, le problème de la création de références subsiste, cependant elle répond au problème de rigidité de BLEU. La formule de TER est la suivante : $TER = \text{nombre d'édicions} / \text{la moyenne du nombre de mots dans la référence}$. La version TERp prend en plus en considération la racine des mots, les synonymes et des paraphrases générées automatiquement, ce qui lui permet d'intégrer des poids différents selon l'importance de l'édition faite afin d'approcher au plus près du jugement humain. Ainsi la substitution par un mot de la même racine ou par un synonyme reçoit une plus faible pénalité qu'un mot de racine différente. TERp utilise comme METEOR (classée dans les mesures linguistiques, cf. ci-dessous) le « *Porter stemming Algorithm* » (Porter 1980) et WordNet (Fellbaum 1998). Enfin, pour procéder à la substitution des séquences de mots, la mesure se base sur une comparaison avec une table contenant des paraphrases issues de toutes les recombinaisons possibles des sous-ensembles de phrases générées automatiquement. Enfin, TERp peut être paramétré selon l'évaluation que l'on veut faire en maximisant la correction avec le type d'évaluation voulu : une TER humaine (HTER) ou l'adéquation et la fluidité, ce qui fait que cette nouvelle mouture de TER pourrait être classée dans les mesures utilisant les méthodes d'apprentissage automatique (Sanders, Przybicki et al. 2011).

Afin d'illustrer la différence avec BLEU et WER, nous reprenons le même exemple que plus haut et ajoutons un exemple similaire. Pour calculer TER, il faut additionner le nombre de substitutions, d'insertions, de déplacements et d'effacements effectués et le diviser par le nombre moyen de mots dans la/les référence(s) :

$TER = SUB + INS + DEP + EFF / \text{Nb moyen de mots dans les références}$

Ref 1	tiene		la	garganta	cogida
Hypo 1	le	duele	la	garganta	
Editions 1	substitution	insertion	déplacement		effacement
Ref 2	tiene	dolor	de	garganta	
Hypo 2	le	duele	la	garganta	
Editions 2	substitution	substitution	substitution		

Tableau 3.12 : Alignements TER de l'hypothèse, de la référence et les éditions effectuées pour deux exemples

Pour le premier exemple le résultat est :

$$TER = 1 \text{ substitution} + 1 \text{ insertion} + 1 \text{ déplacement} + 1 \text{ effacement} / 4 = 4 / 4 = 1 ;$$

pour le second, nous avons :

$$TER = 3 \text{ substitutions} + 0 \text{ insertion} + 0 \text{ déplacement} + 0 \text{ effacement} / 4 = 3 / 4 = 0.75.$$

Il s'agit d'une mesure de taux d'erreurs allant de 0 à 1. Plus le score est bas, plus la qualité de la TA est élevée. Cette mesure pourrait aussi s'inscrire dans le cadre des mesures linguistiques, puisqu'elle intègre certains éléments de morphologie et de sémantique, mais puisqu'elle se base tout de même principalement sur la surface (les chaînes de caractères), et non sur la signification en profondeur en passant par une représentation sémantique, elle reste une mesure de ressemblance lexicale avec la référence.

Nous citerons également dans cette catégorie, la mesure *Maximum matching string* (MMS) de (Turian et al. 2003), qui est proche de la mesure *Translation Edit Rate* (TER) parce qu'elle permet à une chaîne de caractères d'être mise en correspondance qu'une seule fois et elle permet aussi la remise dans l'ordre des chaînes de caractères. Cependant, la MMS favorise clairement les correspondances longues, alors que la TER essaie de minimiser le nombre d'éditions entre la référence et l'hypothèse. Dans ce sens, TER semble plus adapté à notre contexte de TAP.

Les années 2005 ont vu l'apparition de mesures qualifiées de linguistiques parce qu'elles intègrent des éléments linguistiques qui pallient les inconvénients des mesures purement statistiques comme BLEU. La mesure « *Metric for Evaluation of Translation with explicit reordering* », plus communément connue par son abréviation **METEOR** (Banerjee and Lavie 2005) compte le nombre de correspondances exactes de mots entre la TA produite et la référence. Les mots qui ne correspondent pas sont transformés dans leur forme canonique et

mis en correspondance après un traitement morphologique. Cela permet de ne pas pénaliser l'emploi de synonymes ou de formes fléchies différentes de la traduction de référence. Cette mesure prend également en compte l'ordre des mots, en attribuant une pénalité pour chaque déplacement de mot nécessaire pour arriver à la référence. Pour illustrer le fonctionnement de cette mesure, nous allons calculer METEOR sur l'exemple du Tableau 3.13.

Ref 1	tiene	dolor	de	garganta
Hypo	le	duele	la	garganta
Editons	substitution	même racine	substitution	

Tableau 3.13 : Alignement METEOR de l'hypothèse, de la référence

Il faut tout d'abord calculer la précision (P), c'est à dire le nombre d'uni-grammes présents dans l'hypothèse et dans la traduction de référence (m) et le diviser par le nombre d'uni-grammes présents dans l'hypothèse (n-hypo).

$$P = m/n_{hypo} = 1/4 = 0.25.$$

Puis, il faut calculer le rappel (R), en divisant le nombre d'uni-grammes communs entre la référence et l'hypothèse par le nombre d'uni-grammes de la référence.

$$R = m/n_{ref} = 1/4 = 0.25.$$

Ensuite, on prend la moyenne harmonique en calculant

$$F = 10 * P * R / R + 9 * P = 10 * 0.25 * 0.25 / 0.25 + 9 * 0.25 = 0.25$$

Avec une pénalité fixée à $p=0.5$, nous obtenons le résultat suivant pour METEOR :

$$METEOR = F_{moyenne}(1-p) = 0.25(1-0.5) = 0.125$$

Cette mesure corrèlerait très fortement avec le jugement humain (Banerjee and Lavie 2005) et est utilisée comme référence dans les campagnes d'évaluation de la série *Workshop on Statistical Machine Translation* (Callison-Burch, Monz et al. 2009, Giménez and Màrquez 2010) ou dans (Owczarzak, Van Genabith et al. 2007).

D'autres mesures linguistiques plus expérimentales comme le **X-Score** (*predicted fluency score*) (Rajman et Hartley 2001) se basent sur l'analyse de la grammaticalité de la traduction produite, en comparant des informations morphologiques et syntaxiques de la traduction avec un corpus de référence annoté au niveau de la fluidité. Elles n'ont cependant pas été adoptées par de nombreux chercheurs parce qu'elles présentent l'inconvénient de nécessiter la production d'un corpus de référence annoté. Leur avantage est que seule la traduction est considérée par la mesure contrairement à BLEU ou METEOR. Dans sa version utilisée lors de la campagne

d'évaluation CESTA (Surcin, Hamon et al. 2005), la traduction fait l'objet d'un étiquetage grammatical (fonctions grammaticales). L'utilisation de cette mesure se fait en deux phases. Tout d'abord la phase d'entraînement lors de laquelle le modèle de grammaticalité est produit et ensuite la phase d'évaluation elle-même. Cette mesure peut être classée dans les mesures d'apprentissage automatique comme elle utilise des méthodes de ce type pour entraîner un modèle de prédiction linéaire. Les résultats obtenus lors de la campagne CESTA n'ont pas corrélé avec les jugements humains, et en plus les ressources nécessaires à l'application de cette mesure sont trop importantes par rapport aux résultats obtenus et dépendent trop de la qualité de l'étiqueteur employé. La conclusion finale de la campagne CESTA remettait en cause le fait que les mesures automatiques soient vraiment moins coûteuses qu'une évaluation humaine.

Giménez et Màrquez (2010) critiquent le fait que les mesures automatiques se contentent le plus fréquemment de prendre en compte des éléments lexicaux, et dans quelques cas certains éléments syntaxiques, mais encore trop peu d'éléments sémantiques. Giménez et Màrquez (2010) proposent la mesure UPC qui concatène plusieurs mesures classiques, et ajoute de la sémantique en se basant sur la correspondance des rôles sémantiques (Giménez et Màrquez 2010).

Kos et Bojar (2009) proposent une mesure de correspondance sémantique (*Semantic POS Overlapping* - SemPOS inspirée par les travaux de Giménez et Màrquez (2008) (Giménez et Màrquez 2008, Kos et Bojar 2009)).

3.3.2.2.3 *Troisième génération : mesures d'apprentissage et estimation de la qualité*

La dernière catégorie de mesures concerne les mesures d'apprentissage. Le but de ce type de mesure est de mettre un terme à la dépendance aux références humaines, en utilisant des corpus existants annotés. Un essai de ce type est le « *linear regression based model* » développé par (Albrecht et Hwa 2007) qui produit des pseudo-références à partir de traductions automatiques, ce qui permet de contourner le besoin de références humaines coûteuses. L'apprentissage peut se faire à différents niveaux et utilise des modèles du langage basés sur les dépendances (Albrecht et Hwa 2007). Les précurseurs dans la matière, (Kulesza et Shieber 2004) utilisent la méthode d'apprentissage des machines à vecteurs de support. Ces mesures ont le grand avantage de ne plus dépendre de ressources humaines.

L'approche de l'estimation qualité (Specia, Raj et al. 2010) entre également dans cette catégorie puisqu'il s'agit d'une tâche de prédiction de la qualité à partir d'attributs extraits des phrases sources et de leur traduction grâce à l'entraînement de modèles sur des données

d'entraînement importantes tout en évitant de produire des traductions de référence. Deux caractéristiques importantes pour évaluer la qualité de la traduction (section 3.3.2.1.1) sont l'adéquation et la fluidité. La première peut être estimée en calculant la « probabilité de la phrase et en comptant les distorsions (*phrase probability* et *distortion count*), la seconde grâce à des indicateurs de fluidité comme le degré de naturel et de grammaticalité d'une traduction (Specia, Hajlaoui et al. 2011).

Dans cette section de chapitre, nous avons tout d'abord fait le tour d'horizon des différentes mesures humaines et automatiques employées pour l'évaluation de la qualité externe de la TA. Dans le chapitre 5, nous utiliserons un certain nombre de ces mesures qui vont s'intégrer dans notre méthode d'évaluation EAGLES-ISO comme moyen d'évaluer les attributs qui découlent de notre modèle de qualité. Cependant, nous voulons tout d'abord compléter ce chapitre par une vue d'ensemble de l'évaluation des autres systèmes de TAP dans le domaine médical et dans les campagnes d'évaluation de TAP.

3.4 Vue d'ensemble de l'évaluation des autres systèmes de TAP

Dans le domaine de la TAP, la pratique d'évaluation peut se classer essentiellement en deux catégories. D'une part, une évaluation de type boîte transparente, laquelle se concentrera sur la qualité des composants de la TAP, à savoir la reconnaissance vocale, la traduction automatique et la synthèse vocale (TTS). D'autre part, une évaluation de type boîte noire concentrée sur la qualité à l'usage de l'ensemble du logiciel étudié et qui vise à répondre à la question suivante : « Le logiciel dans son ensemble aide-t-il l'utilisateur à accomplir une tâche précise ? ». Le premier type est ainsi plutôt orienté vers le développeur (*developer-based evaluation*), dont l'objectif principal est l'évaluation des fonctionnalités d'un système afin de continuer à le développer et à l'améliorer ; le deuxième type est lui axé sur les utilisateurs (*user-oriented evaluation*), notion couverte par les normes ISO/IEC 9126 et ISO/IEC 14598.

Même si certains travaux sont déjà cités tout au long de l'état de l'art ci-dessus, nous présentons dans la section suivante les différents modes d'évaluation choisis par les grandes campagnes d'évaluation comparative et par les systèmes de TAP que nous avons présentés dans la section 2.3. Nous allons commencer par la série de campagnes d'évaluation IWSLT, organisée à l'initiative du consortium C-STAR (Akiba, Federico et al. 2004) qui a lieu tous les ans depuis 2004.

3.4.1 Campagnes d'évaluation IWSLT

En résumé, la série de campagnes d'évaluation *International Workshop on Spoken Language Translation* (IWSLT) repose principalement sur les mesures classiques d'évaluation de la reconnaissance vocale (WER) et de la TA. Du côté des mesures automatiques, nous retrouvons BLEU, NIST, WER, PER, METEOR et GTM et du côté des mesures humaines, l'évaluation de l'adéquation et de la fluidité de la TA est le plus régulièrement utilisé (Akiba, Federico et al. 2004, Eck et Hori 2005, Boitet, Bey et al. 2006, Paul 2009, Paul, Federico et al. 2010, Federico, Bentivogli et al. 2011, Federico, Cettolo et al. 2012). Il est à noter que pour les campagnes 2005 (Eck et Hori 2005), une mesure humaine basée sur la conservation du sens a été introduite et qu'en 2009 (Paul 2009) et 2010 (Paul, Federico et al. 2010), la qualité de la TA était également mesurée par HTER (*Human translation edit rate*, cf. section 4.3.2.2). Dans la campagne 2011 (Federico, Bentivogli et al. 2011), l'évaluation humaine de la TA incluait un classement par ordre de préférence effectué sur un grand nombre d'évaluateurs grâce au « *crowdsourcing* » (cf. chapitre 6 pour une définition).

Dans les campagnes 2010 (Paul, Federico et al. 2010), 2013 (Paul, Federico et al. 2010) et 2014 (Cettolo, Niehues et al. 2014), la tâche repose sur la traduction des présentations TED talks⁹⁹, pour laquelle l'évaluation fait appel à HTER et en 2014 sur HTER en utilisant la plateforme de post-édition *Matecat*¹⁰⁰ (Cettolo, Niehues et al. 2014). La traduction automatique de Ted Talks (parole continue dans domaine variés) est assez éloignée de l'évaluation de système de TAP tel que MedSLT. Par contre, les campagnes 2009, 2010 et 2012 qui incluaient une tâche impliquant un dialogue dans le domaine des transports et tourisme sont plus proches. Dans ces campagnes, l'évaluation humaine classique (adéquation, fluidité et classement) était complétée par une évaluation de l'adéquation du dialogue dans son ensemble par rapport à l'information transmise (sur une échelle de 5 points) plutôt qu'une évaluation purement phrase par phrase. Cette mesure d'adéquation du dialogue se rapporte le plus à une qualité à l'usage, autrement, toutes les campagnes d'évaluation IWSLT reposent plutôt sur une approche de qualité externe. Ce choix est tout à fait logique puisque, dans ce cadre, le but est justement de faire avancer les technologies utilisées par les différentes composantes de la TAP. De ce fait, l'orientation est axée sur une évaluation de type développeur plutôt que sur une évaluation fondée sur l'utilisateur final.

⁹⁹ Site des conférences en ligne Ted Talks, , consulté le 17.04.2015.

¹⁰⁰ Site de Matecat, <https://www.matecat.com/features/>, consulté le 16.02.2016.

Nous allons maintenant étudier l'évaluation des logiciels développés dans le cadre du projet américain TRANSTAC (Sanders et al 2013) et de son précurseur Babylon.

3.4.2 NIST, MITRE et DARPA : de Babylon à TRANSTAC

L'une des premières campagnes d'évaluation axées sur la qualité à l'usage a été effectuée par Frederking et al. (2002), où le système de TAP anglais–croate *Tongues* a été évalué lors de son utilisation sur le terrain en Croatie afin d'étudier l'utilité de ce système grâce à des questionnaires et à l'analyse des dialogues obtenus, notamment au niveau des longueurs des dialogues et des erreurs survenues (Frederking, Black et al. 2002). L'évaluation conduite par le MITRE¹⁰¹ (Aberdeen, Oshika et al. 2005) est axée sur l'évaluation de l'efficacité du système qui permet au personnel médical de communiquer avec des patients non-anglophones. Elle comportait aussi une évaluation hors-ligne des résultats fournis par les composantes RV et TA, mais l'orientation principale était de type qualité à l'usage orientée vers les utilisateurs finaux, puisqu'une simulation d'utilisation du système développé pour le milieu médical a été conduite avec des patients standardisés et des professionnels de la médecine (PM). Les mesures retenues sont les classiques taux d'accomplissement de la tâche (dans le domaine médical, il s'agit de poser le bon diagnostic, en utilisant le système de TAP comme seul moyen de communication entre patient et médecin), le temps consacré à l'accomplissement de la tâche et aussi l'efficacité du dialogue (calculée en fonction du nombre de tours de parole par rapport à la longueur du dialogue ou le nombre d'interactions réussies ou au contraire le nombre de répétitions engendrées par un mauvais fonctionnement ou une mauvaise utilisation du système). Les mesures citées sont complétées par un questionnaire de satisfaction qui permet de collecter des informations plus subjectives. Les utilisateurs peuvent ainsi s'exprimer sur la qualité d'utilisation du système en répondant à des questions précises. Cette étude conclut que les mesures traditionnelles (WER, SER de la RV et adéquation/précision de la TA) ne sont pas toujours de bons indicateurs du succès de l'interaction, parce qu'elles ne tiennent pas compte de la capacité de l'homme à inférer l'information, notamment lorsque la motivation des utilisateurs est forte¹⁰².

¹⁰¹ Site du MITRE, ONG américaine dont l'objectif est d'effectuer de la recherche et développement financés par le gouvernement fédéral, actif dans la défense : <https://www.mitre.org> consulté le 15.04.2014.

¹⁰² “Like speech recognition accuracy, translation accuracy or adequacy might not be a good indicator of interaction success because traditional measures rarely take into account human ability to infer information, **especially when motivation is high**”, p. 7 du rapport technique Aberdeen, J., B. Oshika, S. Condon, L. Harper and J. Philips (2005). Evaluation of Speech-to-Speech Translation Systems, MITRE.

Dans le cadre du programme DARPA Babylon/CAST program, des évaluations avec de véritables médecins ont été organisées également pour le système du CMU qui permet un médecin et un patient de dialoguer en anglais et thaï (Schultz, Black et al. 2006).

Depuis, une méthodologie d'évaluation de logiciel a été proposée par le DARPA (Département de défense américain) appelée *SCORE : an experimental methodology: System, component and Operationally Relevant Evaluations*. Elle a été appliquée à l'évaluation de la TAP dans le cadre du projet américain TRANSTAC. Ce programme regroupe plusieurs groupes de recherche (notamment l'entreprise BBN (Stallard et al., 2007) avec le système de TAP Transtalk (Prasad, Natarajan et al. 2013), ou encore le CMU (Bach, Eck et al. 2007), IBM (Gao et al., 2006), SRI (Akbacak et al., 2009), et USC (Belvin et al., 2005)) afin de développer des logiciels de TAP notamment dans le domaine médical à l'usage des militaires sur le terrain. Selon (Sanders et al. 2013 : p.531), cette méthode englobe la performance technique quantitative et l'évaluation qualitative de l'utilité (dans le sens de qualité à l'usage – section 3.2 et section 3.3, p.69) en poursuivant deux objectifs principaux : d'une part, l'évaluation de l'utilité du système, grâce à l'estimation des capacités du système dans son ensemble (Weiss, Schlenoff et al. 2008, Weiss and Schlenoff 2011) et, d'autre part, l'évaluation des composantes du logiciel afin d'apprécier la performance de chacune d'elles séparément. Ces deux axes correspondent à la distinction entre évaluation interne et externe d'une part et qualité à l'usage d'autre part pour reprendre la terminologie des normes ISO. Il s'agit par conséquent d'une évaluation complémentaire qui, dans le domaine de la TAP, se résume le plus souvent exactement comme dans la citation ci-dessous. Il s'agira de prendre en compte, d'une part, l'adéquation de la traduction produite par la TAP (en incluant le résultat de la RV) et, d'autre part, l'évaluation de la qualité à l'usage : « *The TRANSTAC community is in agreement that the two aspects that best identify the ability of systems to meet the overall goal of seamless speech-to-speech communication are (1) the semantic adequacy (refers to having the meaning correct and being understandable) of the translations, leading to justified user confidence in the system's translations, and (2) the ability of military personnel and foreign language speakers to successfully carry out a task-oriented dialogue in a domain of known operation need under conditions that reasonably simulate us in the field.* » (Sanders, Weiss et al. 2013 : p.530).

Les évaluations par Weiss et Schlenoff (2011) et Sanders et al. (2013) comportent d'une part une évaluation des composantes (WER, BLEU) complétées de l'évaluation du taux de transfert des concepts ainsi qu'une évaluation humaine de l'adéquation sémantique des traductions (+3 : totalement adéquate à -3 totalement inadéquat). D'autre part, ils évaluent

également l'accomplissement d'un dialogue en simulant les conditions sur le terrain. Cette évaluation est complétée d'un questionnaire de satisfaction. Il ressort donc que dans le cadre de TRANSTAC, comme pour NIST et MITRE, l'évaluation s'articule sur plusieurs dimensions et niveaux¹⁰³ :

Premièrement, l'évaluation des composantes : WER pour la RV, pour la précision de la traduction (BLEU, TER (Condon, Parvaz et al. 2010), METOR, et évaluation humaine) et recours à un juge humain et WER pour la qualité de la synthèse vocale (SV) . Pour l'évaluation de la sortie finale, NIST et MITRE utilisent la mesure du taux de transfert des concepts de bas-niveau (*low-level concept transfer odds*) (Stallard, Choi et al. 2006).

Deuxièmement, la simulation de l'utilisation réelle (*real world scenario*) évaluée par l'accomplissement de la tâche (précision et efficacité en temps), le taux de transfert des concepts de haut niveau (*high-level concept transfer rate*), le nombre de tentatives nécessaires par succès (number of attempts per success) et le temps mis pour comprendre un concept (*time to retrieve a concept*).

Troisièmement, la satisfaction évaluée par des questionnaires post-tâche remplis anonymement (satisfaction sur une échelle) et des avis sur la performance, l'utilisabilité, et certains aspects liés à l'interface ou à l'interaction avec le système (interaction sans contact visuel, mobilité, interface, etc.).

Le successeur du programme TRANSTAC est le projet BOLT¹⁰⁴, dont l'objectif est d'aller au-delà des modèles statistiques et de proposer notamment une mesure sémantique d'évaluation afin de pouvoir mesurer la fidélité avec laquelle le sens a été transmis grâce à des méthode d'étiquetage sémantique utilisée à l'université de Colorado¹⁰⁵ et appliquée au système de BBN (Prasad, Kumar et al. 2012).

Dans cette section, nous nous sommes concentrée sur les grandes campagnes d'évaluation et sur les projets d'envergure (financés par DARPA). Nous allons compléter cette revue par les projets évoqués dans la section 2.3.

¹⁰³ Présentation intitulée : "Statistical Machine Translation and Speech-to-Speech Translation" par Bowen Zhou du IBM T. J. Watson Research Center, le 6 avril 2010 à l'Université de Columbia, consultée le 17.04.2015.

¹⁰⁴ Site de BOLT, <http://www.darpa.mil/program/broad-operational-language-translation>, consulté le 18.02.2016.

¹⁰⁵ Description du projet BOLT sur le site de l'Université de Columbia, Colorado, <http://www.cs.columbia.edu/speech/projects.cgi>, consulté le 24.04.2015.

3.4.3 Systèmes de TAP dans le médical : Nespole, S-minds, Transonics, Converser

Moins de publications sont accessibles concernant l'évaluation de ces systèmes de TAP dans le médical *Nespole* !, *S-minds*, *Transonics* et *Converser*. Pour *Nespole* ! (Mana, Burger et al. 2003), une expérience avec simulation de diagnostic en utilisant *Nespole* est montée afin d'évaluer la qualité d'utilisation par l'utilisateur final (*end-to end evaluation*) (Hamon, Fügen et al. 2009). Autrement, les évaluations ont plutôt été effectuées au niveau de la qualité externe en suivant le cadre de qualité proposé dans (Federico 2003) qui mènera vers la mise en place des campagnes IWSLT (Section 3.4.1).

Pour *S-minds* de *Fluential*, une évaluation a été effectuée portant sur la précision de la traduction et sur la satisfaction des utilisateurs (Ehsani, Kimzey et al. 2008). Aussi plus récemment dans le cadre du traitement du diabète de personnes non-anglophones dans (Soller, Chan et al. 2012). Les résultats de cette étude sont plutôt bons : la précision de la communication anglais-espagnol-anglais atteint entre 97 et 100%. La satisfaction des patients obtenait un score entre 4.7–4.9 de moyenne sur une échelle de Likert à cinq points). Cette étude conclut ainsi que la performance de *S-minds* est meilleure que les systèmes commerciaux autant dans des conditions optimales d'utilisation (silence environnement et peu d'hésitations) que dans un environnement bruyant avec un locuteur hésitant.

Pour *Transonics*, désormais appelé *Speechlinks*, (Shin, Georgiou et al. 2010) ont mis en place une étude de simulation d'utilisation en plus des mesures classiques de WER et BLEU pour la TA statistique, parce que l'objectif était de modéliser le comportement des utilisateurs, ce qui va bien évidemment au-delà d'une évaluation classique.

Enfin, pour *Converser* (Seligman et Dillinger 2006), il n'existe à ce jour pas de publication reprenant les résultats de l'étude de type qualité à l'usage annoncée dans (Seligman et Dillinger 2011).

3.5 Conclusion

Dans le présent chapitre, nous avons tout d'abord repris brièvement le cadre général de l'évaluation des logiciels selon les normes ISO/IEC 9126, ISO/IEC 14598 et SQuaRE – ISO 25000. Ces normes proposent un cadre d'évaluation que nous avons complété par l'interprétation proposée par le groupe de travail européen par EAGLES visant à suivre une méthode d'évaluation en sept étapes.

Nous avons ensuite expliqué la structure hiérarchique du modèle d'évaluation EAGLES-ISO, qui distingue l'évaluation de la qualité à l'usage (section 4.2) de celle de la qualité interne et externe (section 4.3) des logiciels. Pour chaque type de qualité, nous avons distingué et expliqué les caractéristiques et sous-caractéristiques amenées à constituer le modèle de qualité, qui est l'essence de la méthode d'évaluation décrite dans le chapitre suivant (méthodologie de la recherche) et dans le chapitre 5 (résultats). Ces caractéristiques doivent permettre à un évaluateur de constituer son propre modèle par rapport à un contexte d'utilisation particulier.

Nous avons ensuite passé en revue les différentes mesures appropriées pour mesurer les caractéristiques de qualité à l'usage et de qualité externe. Pour la partie concernant la qualité externe, nous avons distingués les mesures issues de la RV de celles issues de la TA. Pour l'évaluation de la précision de la TA, nous avons expliqué les principales mesures humaines et automatiques.

Enfin, nous avons achevé ce chapitre par une revue des pratiques d'évaluation des systèmes de TAP en général et plus particulièrement dans le domaine médical. Il ressort qu'en fonction du but de l'évaluation, celle-ci s'axe principalement sur la qualité externe des composantes de la TAP (comme c'est le cas pour les campagnes IWSLT puisque leur orientation est plutôt axée sur les développeurs) ou s'oriente vers une évaluation plus complexe et complète lorsque l'optique est de considérer l'utilisateur final. Dans ce cas, la qualité à l'usage complète la qualité externe des systèmes de TAP (Sanders, Weiss et al. 2013).

Au chapitre suivant, nous détaillerons notre méthodologie de recherche pour cette thèse.

4 Méthodologie de la recherche

L'objectif du présent chapitre est de décrire la méthodologie suivie dans le cadre de cette recherche. Dans l'introduction, nous contextualiserons nos travaux et présenterons les questions de recherche, qui seront étudiées successivement. En effet, de notre question de recherche générale que nous décrirons plus bas, découlent deux questions de recherche principales. Celles-ci sont, à leur tour, subdivisées en sous-questions. Pour chacune, nous expliquons quel est le procédé suivi pour y répondre en détaillant l'expérience, les participants, les données, les variables évaluées et les mesures employées.

Pour contextualiser notre recherche, il convient de revenir brièvement sur les observations exposées dans les précédents chapitres. Dans le chapitre 2, nous avons pu constater qu'il existe de nombreux systèmes de TAP dans le domaine médical. À ce stade, la question de l'évaluation de ces systèmes se pose. Nous avons aussi vu dans l'état de l'art de l'évaluation de la TAP (chapitre 3), que les méthodes classiques les plus fréquentes (IWSLT, section 3.4.1) mettent l'accent sur la **précision** (*accuracy*) du système, en étudiant la qualité des sorties des sous-tâches incluses dans la traduction automatique de la parole, c'est-à-dire de la qualité de la reconnaissance vocale et de la traduction automatique. L'avantage de cette méthode est de pouvoir recourir aux procédés existants pour évaluer la TA et la RV. Cependant, ce type d'évaluation se cantonne à une approche purement fonctionnelle et restreinte de l'évaluation, qui ne donne pas suffisamment d'informations sur la qualité effective d'un système TAP dans son ensemble par rapport à un autre.

De ces constats est née notre question de recherche générale : « Comment comparer des systèmes de TAP dans le domaine médical de manière fiable ? ». Pour y répondre, nous reformulons cette question trop générale, comme le préconisent (Saldanha et O'Brien 2014 : pp.16-17), pour obtenir une première question de recherche plus ciblée : « *Peut-on utiliser la méthode EAGLES-ISO pour évaluer les systèmes de TAP dans le domaine médical ?* » (**Q1_principale**).

Nous avons choisi cette méthode EAGLES-ISO pour deux raisons. Premièrement, dans le but d'inscrire notre évaluation dans une méthode que nous avons déjà utilisée auparavant (Starlander et Popescu-Belis 2002), mais également dans le cadre de l'enseignement de l'évaluation de logiciels d'aide à la traduction aux étudiants du cours de traduction assistée par ordinateur (cours de MA enseigné à la FTI) comme décrit dans (Starlander et Morado Vazquez 2013, Starlander 2015) avec l'interprétation de ces normes faites par le groupe de travail

EAGLES dans les années 1990. Deuxièmement, notre choix des normes ISO nous permet de poursuivre dans la direction indiquée par la thèse de Nikos Tsourakis (2013) qui propose également une interprétation des normes ISO appliquée à la traduction automatique de la parole, mais sous un angle différent, celui de l'interface et de la portabilité vers des supports mobiles. Ainsi, la méthode que nous proposons est fondée sur les travaux mentionnés plus haut et sur les normes de la série ISO 9126.

Q1_principale

L'objectif de **Q1_principale** est d'appliquer une méthode complète, directement inspirée des normes ISO/IEC 9126, ISO/IEC 14598 ou ISO 25000 *Software Product Quality Requirements and Evaluation* (SQuaRE) ainsi que de la méthode EAGLES (cf. chapitre 3) à deux systèmes de TAP dans le domaine médical. L'essence de cette méthode est de définir un **modèle de qualité** qui correspond à un ensemble de caractéristiques choisies en fonction du contexte d'utilisation (section 3.1). Elle doit par conséquent permettre d'aller au-delà de l'infime portion du modèle de qualité qui est couverte par les évaluations classiques décrites dans la section 3.4 (Azuma, 2001).

Pour appliquer notre méthode et ainsi répondre à **Q1_principale**, nous avons mis en place une expérience de simulation de diagnostic (**EXPQ1**) utilisant deux versions comparables de *MedSLT* (section 2.4.4.) que nous avons choisies comme candidates. **EXPQ1** suit un protocole (section 4.1.4) de type « *repeated measures design* » (Coolican 1996 : p.26, Lawal 2014 : p.697), dont le but est d'éviter d'introduire des biais. Pendant l'expérience **EXPQ1**, nos participants (section 4.1.3.1) utilisent tour à tour deux versions de *MedSLT* (V1 et V2), qui est, de ce fait, notre variable indépendante (Oates 2005 : p.129). Étant donné que notre question de recherche repose sur l'évaluation de la méthode EAGLES-ISO, nos variables dépendantes découlent de notre modèle de qualité qui sera décrit dans la section 4.1.1. En effet, les caractéristiques et sous-caractéristiques retenues dans notre modèle de qualité peuvent être reformulées comme des sous-questions de recherche (**Q1_S1 – Q1_S9**). Elles correspondent aux variables dépendantes étudiées au moyen d'une série de mesures présentées dans la section 4.1.2.5. La mise en place de l'expérience (**EXPQ1**) est présentée dans la section 4.1.2. Celle-ci détaillera les participants, le protocole, l'étude pilote, les données collectées et les mesures appliquées.

Le Tableau 4.1 énonce les neuf sous-questions (**Q1_S1 – Q1_S9**) qui découlent du modèle de qualité décrit dans la section 4.1.1.

N° de question et caractéristique	Questions de recherche de Q1_principale Testées par EXPQ1	Section
Q1_principale	Peut-on utiliser la méthode EAGLES-ISO pour évaluer les systèmes de TAP dans le domaine médical ?	
Qualité à l'usage		
Q1_S1	L'utilisation du système permet-elle au médecin de poser le bon diagnostic (=tâche accomplie) pour son patient ? <i>Cette question est reformulée de la manière suivante :</i> Quel est le taux d'accomplissement de la tâche ?	4.1.1
Q1_S2	Quel est le temps nécessaire pour accomplir la tâche effectuée à l'aide du logiciel ?	4.1.1
Q1_S3	Quelle version (V1 ou V2) donne la meilleure satisfaction aux utilisateurs ?	4.1.1
Q1_S4	L'un des deux systèmes évalués représente-t-il un plus grand danger ? <i>Cette question est reformulée de la manière suivante :</i> L'un des deux systèmes évalués génère-t-il plus d'erreurs d'interaction que l'autre ?	4.1.1
Qualité externe		
Capacité fonctionnelle	Quel système (V1 ou V2) est plus précis ?	
Q1_S5	Quel système est plus précis au niveau de la RV ?	4.1.2
Q1_S6	Quel système est plus précis au niveau de la rétro-translation ?	4.1.2
Q1_S7	Quel système est plus précis au niveau de la TA ?	4.1.2
Facilité d'utilisation	Quel système (V1 ou V2) est plus facile à utiliser ?	
Q1_S8	Le fonctionnement du système est-il plus simple à comprendre avec V1 ou V2 ?	4.1.2
Q1_S9	L'apprentissage du système est-il plus facile avec V1 ou V2 ?	4.1.2

Tableau 4.1 : Récapitulatif des questions de recherche de Q1_principale

Les résultats obtenus pour notre première question de recherche principale sont exposés dans le chapitre 5. Dans la partie récapitulative (section 5.4), nous arrivons à la conclusion que la méthode appliquée nous a effectivement permis d'évaluer les deux systèmes étudiés de manière à les départager. En revanche, lorsque nous procédons à l'examen critique de notre évaluation pour répondre à notre première question de recherche, nous constatons que la méthode EAGLES-ISO présente certaines faiblesses, notamment en termes de coûts et de temps investi.

Q2_principale

Ce constat nous a incitée à formuler une deuxième question de recherche: « *Peut-on proposer une nouvelle mesure pour la TAP qui serait équivalente à l'utilisation de la méthode EAGLES-ISO, mais plus rapide et plus économique à mettre en place que celle-ci ?* » (**Q2_principale**). De celle-ci découlent sept sous-questions (Tableau 4.2) qui nous amènent à

proposer la méthode TURKoise, apportant ainsi une contribution finale à notre question de recherche générale.

Q2_principale sera traitée en trois étapes : dans la première, nous proposons une échelle de qualité adaptée à la TAP (HMS) ; dans la deuxième, nous recherchons des mesures automatiques qui soient en corrélation avec celle-ci et, dans la troisième, nous appliquons l'échelle HMS dans un contexte de *crowdsourcing*.

Première étape - HMS

La première étape consiste à proposer une échelle de qualité adaptée à l'évaluation humaine de la TAP médicale (*Human medical scale* - HMS). De celle-ci découle la première sous-question (**Q2_S1**) : « *L'échelle HMS proposée est-elle plus adaptée que les échelles classiques d'évaluation de la fluidité et de l'adéquation ?* ».

Deuxième étape - HMS-auto

La deuxième étape consiste à étudier les différentes mesures automatiques les plus fréquentes (Section 3.3.2.2, BLEU, TER et METEOR) afin de déterminer *quelle mesure d'évaluation automatique corrèle le mieux avec notre méthode d'évaluation humaine avec l'échelle HMS* (**Q2_S2**). L'objectif ici est d'aboutir à une mesure automatique qui pourrait accélérer le processus d'évaluation et réduire l'effort requis par l'évaluation humaine. L'intégration d'une mesure automatique pourrait se traduire par un gain de temps et de coût non négligeable. Celle-ci évite aussi de devoir reposer sur la subjectivité du jugement humain. Cependant, comme nous le verrons avec les résultats obtenus pour l'expérience **EXPQ2_S2** mise en place, la corrélation des mesures automatiques étudiées avec notre mesure HMS est faible.

En nous fondant sur ces résultats, nous avons voulu rechercher les raisons de cette corrélation faible en étudiant si, fondamentalement, les mesures automatiques choisies étaient adaptées à notre contexte d'évaluation. Pour répondre à la sous-question de recherche **Q2_S3** - à savoir, « *Les mesures automatiques étudiées pour Q2_S2 sont-elles moins adaptées à l'évaluation de la TAP linguistique comme MedSLT ?* » - nous avons conçu une expérience complémentaire (**EXPQ2_S3**) qui consiste à comparer HMS et les mesures automatiques BLEU, WER et TER (section 3.3.2.2). Cette expérience implique deux systèmes de TAP de types différents : le système linguistique *MedSLT* (sans distinction de version, avec une partie des données collectées pour la question de recherche **Q1**) et le système statistique

GoogleTranslate (GT). Les résultats obtenus (section 5.2.3.2) confirment que la corrélation entre HMS et les mesures automatiques est très faible.

À partir de ce constat, nous avons choisi d'explorer une autre piste : recourir à l'évaluation humaine, mais tenter d'en éviter les écueils, notamment le coût important et le manque de fiabilité dû au défaut de cohérence entre évaluateurs (section 6.1.1 et 6.2.1), ceci en faisant appel au *crowdsourcing* (tâche effectuée par une foule, voir section 6.3). Ainsi, c'est à partir de la fusion de notre mesure HMS et du *crowdsourcing* que nous proposons la méthode TURKoïse.

Troisième étape - TURKoïse

Même sans avoir expliqué en détail en quoi consiste le *crowdsourcing* (section 6.3.1), et notamment le fonctionnement de la plateforme de *crowdsourcing* de *Amazon Mechanical Turk* (AMT)¹⁰⁶, nous pouvons exposer nos questions de recherche pour cette troisième étape de **Q2_principale**. Il s'agit ici d'un « *proof of concept* », qui vérifie si nous arrivons à des résultats comparables avec TURKoïse, en faisant évaluer nos systèmes par la masse, plutôt que par des évaluateurs experts (petits groupes) ou en employant des méthodes automatiques. Par conséquent, nous nous posons la question : « La méthode TURKoïse d'évaluation de la TAP représente-t-elle une alternative intéressante aux méthodes automatiques ou humaines classiques ? ». Pour répondre à cette question trop générale, nous la subdivisons en trois questions de recherche : **Q2_S4** : « *L'implémentation de TURKoïse est-elle faisable ?* », **Q2_S5** : « *Les travailleurs AMT recrutés sont-ils fiables ?* » et **Q2_S6** : « *Quelles sont les bénéfices du crowdsourcing en termes de compétitivité (coût et temps minimisés) ?* ». L'objectif de **Q2_S5** et **Q2_S6** est de répondre à la question trop générale : « Quels sont les bénéfices du *crowdsourcing* en termes de compétitivité (coûts, rapidité et facilité de mise en place) et de fiabilité ? ». Pour répondre à cette série de questions, nous testerons TURKoïse sur nos données issues de **Q1_principale** lors de l'expérience **EXPQ2_S4-S6**. La fiabilité de l'évaluation par des travailleurs AMT (**Q2_S5**) sera mesurée en termes de niveau de cohérence entre évaluateurs AMT (**Q2_S5_S1** : « *Les travailleurs AMT sont-ils cohérents ?* ») comparé à celui entre évaluateurs experts, (**Q2_S5_S2** : « *Les résultats obtenus avec TURKoïse sont-ils comparables à ceux obtenus par des experts avec HMS ?* ») mais aussi en termes de corrélation entre TURKoïse et une évaluation humaine classique de l'adéquation et de la fluidité (**Q2_S5_S3** : « *Les résultats obtenus avec TURKoïse sont-ils comparables aux résultats obtenus avec une*

¹⁰⁶ Page d'accueil du site AMT : <https://www.mturk.com/mturk/welcome>, consultée le 01.07.2015.

évaluation humaine classique de l'adéquation et de la fluidité ? ». Nous répondrons à la question **Q2_S6** : « Quelles sont les bénéfices du crowdsourcing en termes de compétitivité (coût et temps minimisés) ? », en comparant les coûts et le temps investi pour chaque type d'évaluation. Cette expérience **EXPQ2_S4-S6** a par ailleurs permis de préparer le terrain pour l'expérience **EXPQ2_S7**, qui doit répondre à la question générale : « Combien d'évaluateurs de la foule des travailleurs AMT faut-il recruter pour arriver à une évaluation de qualité correspondante à une évaluation effectuée par un petit groupe d'experts ? », que nous reformulons de la manière suivante : « Quelle est l'évolution de la cohérence entre travailleurs AMT lorsqu'on augmente leur nombre ? » (**Q2_S7**). Le Tableau 4.2 ci-dessous récapitule les sous-questions de recherche pour **Q2_principale**, avec en regard la section dans laquelle chacune est traitée en détail.

N° d'étape et expérience	N° de Q. et Sous-Question	Énoncé de la question de recherche	Section
Q2_principale		Peut-on proposer une nouvelle mesure pour la TAP qui serait équivalente à l'utilisation de la méthode EAGLES-ISO, mais plus rapide et plus économique à mettre en place que celle-ci ?	
<i>Étape 1 - HMS</i>			
EXPQ2_S1	Q2_S1	L'échelle HMS proposée est-elle plus adaptée que les échelles classiques d'évaluation de la fluidité et de l'adéquation ?	4.2.1.2
<i>Étape 2 - HMS-auto</i>			
EXPQ2_S2	Q2_S2	Quelle mesure d'évaluation automatique corrèle le mieux avec notre méthode d'évaluation humaine avec l'échelle HMS ?	4.2.2.1
EXPQ2_S3	Q2_S3	Les mesures automatiques étudiées pour Q2_S2 sont-elles moins adaptées à l'évaluation de la TAP linguistique comme <i>MedSLT</i> ?	4.2.2.2
<i>Étape 3 - TURKoise</i>			
Q_générale		La méthode TURKoise d'évaluation de la TAP représente-t-elle une alternative intéressante aux méthodes automatiques ou humaines classiques ? Cette question est subdivisée en Q2_S4 à Q2_S6	
EXPQ2_S4-S6	Q2_S4	L'implémentation de TURKoise est-elle faisable ?	4.2.3.2
	Q2_S4_S1	Quelle est la faisabilité du recrutement des travailleurs AMT pour notre tâche ?	
Q_générale		Quels sont les bénéfices du <i>crowdsourcing</i> en termes de compétitivité (coûts, rapidité et facilité de mise en place) et de fiabilité ?	
	Q2_S5	Les travailleurs AMT recrutés sont-ils fiables ? Cette question est subdivisée en Q2_S5_S1 à Q2_S5_S3	
	Q2_S5_S1	Les travailleurs AMT sont-ils cohérents ? (KAPPA, ICC, pourcentage d'accord)	
	Q2_S5_S2	Les résultats obtenus avec TURKoise sont-ils comparables à ceux obtenus par des experts avec HMS ?	
	Q2_S5_S3	Les résultats obtenus avec TURKoise sont-ils comparables aux résultats obtenus avec une évaluation humaine classique (adéquation et fluidité) ?	

	Q2_S6	Quelles sont les bénéfices du <i>crowdsourcing</i> en termes de compétitivité (coût et temps minimisés) ?	
Q_générale		Combien d'évaluateurs de la foule des travailleurs AMT faut-il recruter pour arriver à une évaluation de qualité correspondante à une évaluation effectuée par un petit groupe d'experts ?	
EXPQ2_S7	Q2_S7	Quelle est l'évolution de la cohérence entre travailleurs AMT lorsqu'on augmente leur nombre ?	4.2.3.3

Tableau 4.2 : Récapitulatif des questions de recherche de Q2_principale

Les résultats pour cette partie de la thèse seront repris étape par étape dans le chapitre 6. Dans la suite, nous consacrerons une section à chacune de nos questions principales en détaillant pour chacune les expériences successives mises en place pour y répondre. Ainsi, la section 4.1 correspond à la première question de recherche (**Q1_principale**) et la section 4.2 à la deuxième (**Q2_principale**).

4.1 Question de recherche 1 : Méthode d'évaluation EAGLES-ISO pour la TAP médicale

Comme nous venons de la voir, notre but est d'effectuer une évaluation complète fondée sur les normes ISO et la méthode EAGLES, afin de comparer deux versions différentes de notre système de traduction automatique de la parole dans le domaine médical *MedSLT* pour répondre à notre première question principale.

Notre contexte d'utilisation est celui d'une simulation d'utilisation. Il s'agit de permettre à un médecin de communiquer dans un environnement hospitalier avec un patient pour lui permettre de poser un diagnostic quant à la présence ou l'absence d'une angine à streptocoque. Ce contexte est délibérément très restreint, afin de pouvoir faire une évaluation comparative de deux versions de notre système de TAP dans le domaine médical *MedSLT*. Les utilisateurs dans notre contexte de simulation d'utilisation sont de véritables médecins de l'Hôpital de Dallas et des patients fictifs. Dans le cadre de ce travail, par conséquent, nous n'aborderons pas les problèmes plus vastes d'accessibilité (être capable d'utiliser un ordinateur, savoir lire ou avoir les mains libres) ni de convivialité de l'interface. Nous nous concentrerons exclusivement sur la comparaison de deux systèmes proches, dont la couverture est différente. Notre modèle de qualité correspond par conséquent à ce choix.

4.1.1 Modèle de qualité

Le schéma ci-dessous fait ressortir les principales caractéristiques de haute qualité retenues pour notre modèle de qualité :

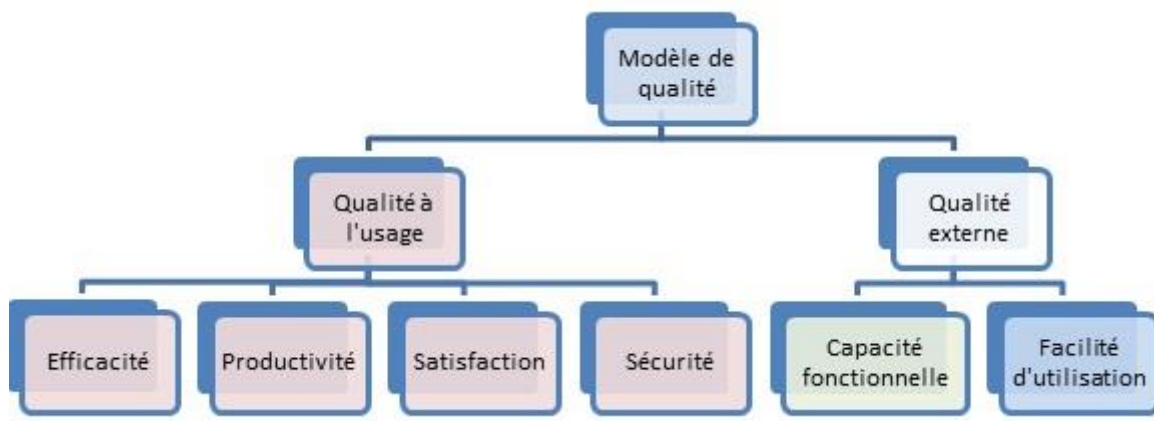


Figure 4.a : Modèle de qualité retenu pour notre contexte d'utilisation

Dans les sections 4.1.1.1 et 4.1.1.2, nous détaillerons notre modèle de qualité pour la qualité à l'usage et la qualité externe, en expliquant quelles sont les caractéristiques de qualité retenues et pourquoi. Nous détaillerons également comment les attributs finaux issus du modèle de qualité seront quantifiés. Dans la section 4.1.2, nous décrirons la mise en place de notre expérience.

4.1.1.1 Qualité à l'usage

S'agissant de la qualité à l'usage, nous retenons les quatre caractéristiques de la norme ISO 9126 : efficacité, productivité, satisfaction et sécurité. Dans le cas de *MedSLT* ou de la TAP dans le domaine médical, évaluer l'**efficacité** revient à vérifier si l'utilisation du système permet au médecin de poser le bon diagnostic pour son patient. Cette question générale est reformulée de la manière suivante : *Quel est le taux d'accomplissement de la tâche ? (Q1_S1)*. Dans notre contexte, cette caractéristique sera par conséquent évaluée en fonction du taux d'accomplissement de la tâche (résultats à la section 5.1.1.1), tâche qui en l'occurrence consiste à poser un diagnostic adéquat par rapport au scénario de simulation donné (voir protocole, section 4.1.2.2).

La **productivité** va de pair avec l'efficacité, puisqu'elle se mesure en termes de temps mis pour accomplir la tâche effectuée à l'aide du logiciel étudié. Cette caractéristique est importante dans le contexte de la TAP du fait de l'instantanéité de la tâche. En effet, celle-ci doit être accomplie dans un temps raisonnable pour que l'utilisation même du logiciel puisse être envisagée par les utilisateurs finaux. Nous arrivons ainsi à notre sous-question **Q1_S2** : *Quel est le temps mis pour accomplir la tâche effectuée à l'aide du logiciel ?*

La **satisfaction**, comme le soulignent (Sanders, Weiss et al. 2013) est une caractéristique fondamentale parce que si le logiciel n’atteint pas des niveaux suffisants, les utilisateurs ne seront pas incités à utiliser le logiciel en question¹⁰⁷. La caractéristique est d’ailleurs décomposée dans la nouvelle série SQuaRE (ISO/IEC 2014) en quatre sous-caractéristiques intéressantes à rappeler : **utilité, confiance, confort et plaisir**. Ce sont des caractéristiques qui pourront être évaluées par le biais de questionnaires de satisfaction. L’élément que nous mettrons en avant plus particulièrement dans notre cas est l’**utilité** du système. Notre question de recherche pour cette caractéristique devient par conséquent : *Quelle version (V1 ou V2) donne la meilleure satisfaction aux utilisateurs ? (Q1_S3)*

Pour ce qui est de la **sécurité**, dans le cadre de notre évaluation, il s’agira d’évaluer le risque en matière de dangers encourus par les utilisateurs dans le cadre de cette communication entre un patient et son médecin. L’un des deux systèmes évalués représente-il un plus grand danger ? Cette question trop générale est reformulée de la manière suivante, plus facile à traiter concrètement : *L’un des deux systèmes évalués génère-t-il plus d’erreurs d’interaction que l’autre ? (Q1_S4)*. Nous allons évaluer cette caractéristique par l’étude détaillée de l’échec des interactions entre les utilisateurs et le logiciel (section 5.3.4).

4.1.1.2 Qualité externe

Au niveau de la qualité externe, parmi les six caractéristiques de ISO/IEC 9126 (1991) qui servent à évaluer les sous-tâches : capacité fonctionnelle, fiabilité, facilité d’utilisation, rendement/productivité, maintenabilité et portabilité, nous retenons uniquement celles de capacité fonctionnelle et de facilité d’utilisation (Figure 4.b), comme nous allons l’expliquer par la suite.

¹⁰⁷ p. 542, “[...] even if a system has exceptional usability, if its product or service is of little value to Soldiers and Marines, they will not carry it in their backpacks at the expense of space that might be otherwise used to hold life-sustaining water that they value highly on patrol through the desert.” Sanders, G. A., B. A. Weiss, C. Schlenoff, M. P. Steves and S. Condon (2013). "Evaluation methodology and metrics employed to assess the TRANSTAC two-way, speech-to-speech translation systems." Computer Speech & Language 27(2): pp. 528-553.

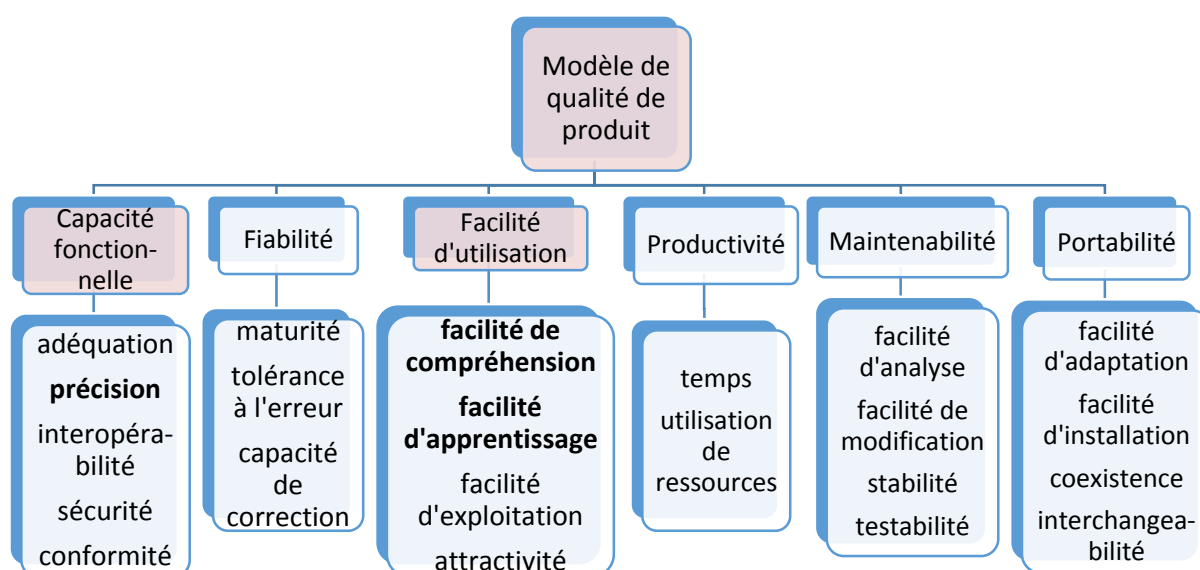


Figure 4.b : Caractéristiques de qualité externe de notre modèle de qualité

Précisons avant tout que les caractéristiques que nous avons retenues (en couleur dans la Figure 4.b ci-dessous) se décomposent en un certain nombre de sous-caractéristiques, que nous ne retiendrons pas toutes parce qu'elles ne correspondent pas toutes à notre contexte d'utilisation. De la même manière, nous n'avons pas retenu la caractéristique de fiabilité telle qu'elle est traitée par la caractéristique de sécurité mentionnée à la section précédente.

Nous évaluerons principalement la **capacité fonctionnelle**, dont nous retenons uniquement la sous-caractéristique de la **précision**. En effet, dans notre contexte, l'adéquation n'est pas une caractéristique importante dans la mesure où les deux versions à l'étude proposent les mêmes fonctionnalités et ne diffèrent qu'au niveau de la couverture des systèmes (chapitre 3). Par contre, dans le domaine de la TAP, la précision des sous-tâches qui la composent est une caractéristique primordiale, très largement utilisée lors de l'évaluation de logiciels de ce type, comme nous l'avons vu au chapitre 3. Nous allons par conséquent appliquer les mesures classiques du domaine afin d'évaluer la précision des sous-composantes de la TAP (section 5.2). Les composantes du système correspondent aux sous-tâches qu'impliquent *MedSLT* : la RV, la TA mais aussi la rétro-traduction, comme celle-ci occupe une place centrale dans l'interaction avec le logiciel. Nous n'évaluerons pas la synthèse vocale, puisque pour cette tâche nous utilisons Nuance¹⁰⁸ et que nous n'avons pas de moyen d'intervenir sur sa qualité. Comme cela a été décrit dans la section 2.4, le flux de travail de *MedSLT* implique que la qualité de la TA et de la RV, mais aussi de la rétro-traduction, vont influencer la qualité à l'usage du

¹⁰⁸ Site de Nuance pour développeur, <http://www.nuance.co.uk/for-developers/dragon/index.htm>, consulté le 21.04.2015.

logiciel en général. En effet, certaines erreurs de RV quantifiables par les mesures typiques de l'évaluation de la RV, comme la WER, n'auront que guère d'impact, voire aucun, sur la qualité à l'usage. Il est par conséquent intéressant de concevoir un cadre d'évaluation complet intégrant des éléments de qualité externe et de qualité à l'usage.

Des quatre sous-caractéristiques de la **facilité d'utilisation** (*usability*), nous ne retenons que la **facilité de compréhension** et la **facilité d'apprentissage** de l'utilisateur. La première caractéristique correspond à la sous-question **Q1_S8** : *Le fonctionnement du système est plus facile à comprendre pour V1 ou V2 ?* La deuxième caractéristique peut-être reformulée comme suit : *L'apprentissage du système est plus facile avec V1 ou V2 ? (Q1_S9).*

La **facilité d'apprentissage** est décrite comme pouvant se mesurer par le temps nécessaire pour apprendre à utiliser un logiciel. Il nous paraît par conséquent utile d'évaluer l'impact du module d'aide. Le rôle de celui-ci est de guider l'utilisateur vers la couverture du système en lui proposant des phrases couvertes (section 2.4.3). Nous allons par conséquent évaluer la capacité d'apprentissage des utilisateurs en fonction du recours au module d'aide (section 2.4.3) entre la première utilisation et la deuxième utilisation de chaque version (section 5.3.1).

Nous ne retenons pas la dernière sous-caractéristique de la facilité d'utilisation, c'est à dire la **facilité d'exploitation** (*operability*), qui traite plus particulièrement de l'interface en répondant à la question de savoir si les utilisateurs sont en mesure d'utiliser les boutons, les micros et l'interface en général pour interagir efficacement avec le système. Ces considérations sont importantes puisqu'elles auront une incidence sur la qualité de la RV notamment, mais dans notre contexte d'évaluation, nos systèmes ne diffèrent pas au niveau de l'interface. Cependant, avant d'envisager une utilisation véritable de ce type de systèmes dans un environnement hospitalier, une étude de cette caractéristique sera primordiale.

Nous ne retenons pas les caractéristiques de productivité, maintenabilité ou de portabilité. Nous allons exposer ci-après les principales raisons de ce choix.

Pour mesurer la **productivité** (ou l'efficacité du système), il faudrait pouvoir prendre comme point de comparaison la situation telle qu'elle se présente en l'absence de l'outil. En l'occurrence, il faudrait étudier comment a été effectuée la traduction des questions de diagnostic dans les hôpitaux jusqu'à présent pour avancer un élément de réponse (cf. Glenn Flores (2003) sur les erreurs de traduction dues à des interprètes peu formés ou non-

professionnels issus de l'entourage du patient). Ces considérations dépassent le contexte actuel de notre évaluation.

Enfin, les deux autres caractéristiques que nous n'avons pas retenues parce qu'elles ne s'appliquent pas à notre objectif sont la **maintenabilité** et la **portabilité**. La première concerne l'ensemble d'attributs portant sur l'effort nécessaire pour apporter des modifications au système. Cela pourrait être mesuré en termes de coût lors de l'adaptation à un autre domaine ou à une autre langue, mais cet aspect ne concerne pas notre contexte actuel.

La **portabilité** inclut l'ensemble d'attributs portant sur l'aptitude du logiciel à être transféré d'un environnement à l'autre (adaptation à un *Personal Digital Assistant* ou autre support mobile dans le contexte médical, par exemple). Cette caractéristique peut être importante dans certains contextes (Tsourakis 2013 : p.65), mais elle ne correspond pas au contexte d'évaluation présent.

Le Tableau 4.3 résume les questions de recherche et caractéristiques et sous-caractéristiques retenues qui constitueront les variables dépendantes de notre expérience. Ce tableau complète ainsi le Tableau 4.1 de l'introduction.

N° de question	Sous-questions de recherche	Variables dépendantes (Caractéristiques)	Variable indépendante
Q1_principale	Peut-on utiliser la méthode EAGLES-ISO pour évaluer les systèmes de TAP dans le domaine médical ?		V1 – V2
Qualité à l'usage			
Q1_S1	L'utilisation du système permet-elle au médecin de poser le bon diagnostic (=tâche accomplie) pour son patient ? <i>Cette question est reformulée de la manière suivante :</i> Quel est le taux d'accomplissement de la tâche ?	Efficacité	V1 – V2
Q1_S2	Quel est le temps mis pour accomplir la tâche effectuée à l'aide du logiciel ?	Productivité	V1 – V2
Q1_S3	Quelle version (V1 ou V2) donne la meilleure satisfaction aux utilisateurs ?	Satisfaction	V1 – V2
Q1_S4	L'un des deux systèmes évalués représente-t-il un plus grand danger ? <i>Cette question est reformulée ainsi :</i> L'un des deux systèmes évalués génère-t-il plus d'erreurs d'interaction que l'autre ?	Sécurité	V1 – V2

Qualité externe			
Capacité fonctionnelle	Quel système (V1 ou V2) est plus précis ?	Précision	V1 – V2
Q1_S5	Quel système est plus précis au niveau de la RV ?	Qualité de la RV	V1 – V2
Q1_S6	Quel système est plus précis au niveau de la rétro-translation ?	Qualité de la rétro-translation	V1 – V2
Q1_S7	Quel système est plus précis au niveau de la TA ?	Qualité de la TA (humaine et automatique)	V1 – V2
Facilité d'utilisation	Quel système (V1 ou V2) est plus facile à utiliser ?		
Q1_S8	Le fonctionnement du système est plus facile à comprendre pour V1 ou V2 ?	Facilité de compréhension	V1 – V2
Q1_S9	L'apprentissage du système est plus facile avec V1 ou V2 ?	Facilité d'apprentissage	V1 – V2

Tableau 4.3 : Traduction du modèle de qualité retenu en sous-questions de recherche, variables dépendantes et indépendante

Dans la section suivante, nous décrirons comment nous avons mis en place notre expérience de simulation, où il s'agit de s'approcher au plus près de la réalité pour comparer nos deux systèmes dans le contexte d'utilisation qui est le nôtre. Nous détaillerons les participants (section 4.1.2.1), le protocole (section 4.1.2.2), l'étude pilote (section 4.1.2.3), les données récoltées (section 4.1.2.4), et comment nous avons mesuré chaque variable (section 4.1.2.5).

4.1.2 Mise en place de l'expérience EXPQ1

L'**EXPQ1** vise à répondre à notre première question de recherche principale et à nos neuf sous-questions. Pour cette expérience, nous avons organisé une collecte de données avec des médecins et des patients standardisés à l'hôpital de Dallas en février 2008. La tâche consistait à simuler une entrevue médicale : le médecin doit diagnostiquer si le patient souffre ou non d'une infection bactérienne à la gorge, en utilisant à tour de rôle la version 1 (V1) et la version 2 (V2) du système bidirectionnel (voir section 2.4 pour une description détaillée). L'expérience suit un protocole minutieux que nous avons testé lors d'une phase pilote réalisée à Genève, en janvier 2008, et qui nous a permis de déterminer l'importance de suivre le protocole afin de contrôler l'effet de l'ordre d'utilisation. Celui-ci sera décrit dans la section 4.1.2.2. L'expérience nous a permis de collecter des données quantitatives et qualitatives, ce qui nous place dans une approche mixte (*mixed method approach*) (Saldanha et O'Brien 2014 : p.23).

4.1.2.1 Participants

Huit médecins du Dallas Children's Hospital ont participé à l'expérience parmi lesquels nous comptons trois médecins internes plus expérimentés et cinq jeunes internes de l'équipe du Prof. Glenn Flores. Contrairement aux expériences effectuées pour (Chatzichrisafis, Bouillon et al. 2006) et (Starlander 2007), il s'agissait de médecins praticiens et, même si leur niveau d'expérience n'était pas homogène, nous n'avons pas eu besoin de fournir d'exemples d'anamnèse pour le diagnostic des maux de gorge.

Les seize patients qui ont participé à l'expérience étaient tous des patients standardisés qui travaillent comme interprètes pour l'espagnol dans l'hôpital et qui ont par conséquent une certaine habitude du domaine. Cependant, afin de simuler correctement les symptômes d'une angine à streptocoque ou virale, les participants suivaient des scénarios prédéfinis.

Le choix des participants s'est fait par leur mise à disposition par notre partenaire, le Prof. Glenn Flores du Dallas Children's Hospital. Il s'agit par conséquent d'un échantillon auto-sélectionné (« *self-selected sample* »), découlant de la spécificité des participants nécessaires à l'expérience (Coolican 1996).

Au début de chaque session, les patients recevaient une feuille décrivant leurs symptômes. Nous avons prédéterminé huit scénarios de symptômes (annexe 5). L'expérience s'est déroulée en 64 sessions, de dix minutes chacune en moyenne (le temps de poser le diagnostic). Aucun participant n'avait utilisé le système auparavant. Chaque patient a utilisé les deux versions du système avec deux médecins différents. La moitié des patients ont commencé avec la version non-restreinte et l'autre moitié avec la version restreinte, en suivant le protocole décrit ci-dessous.

4.1.2.2 Protocole

Lors de l'élaboration du protocole fondé sur (Coolican 1996), nous avons dû prendre en compte différents variables décrites dans la section 4.1.3. Nous avons suivi le « *repeated measures design* » (Coolican 1996 : p.26), également appelé « *within-subject design* » (Saldanha et O'Brien 2014 : p.114). Celui consiste à soumettre un même groupe de participants à plusieurs conditions de la variable indépendante. Dans notre cas, il s'agit de faire utiliser par chaque groupe les deux systèmes comparés, à l'opposé du « *between-subject design* » qui impose qu'un groupe utilise uniquement V1 et l'autre V2.

De plus, le protocole final doit nous permettre de passer d'une version du système à l'autre sans introduire de biais dû à l'ordre d'utilisation (order effect, (Coolican, p. 27)), tout en

minimisant le temps de présence des participants. Nous avons alterné les systèmes et fonctionné en deux sessions parallèles qui ont eu lieu dans deux salles voisines. Lors de chaque session, le médecin voyait quatre patients différents, et chaque patient passait d'une salle à une autre, en commençant soit avec la version restreinte V1, soit avec la version moins restreinte V2, comme décrit pour la session 1 et 2 dans le Tableau 4.4 ci-dessous.

Salle 1	Salle 2
Session 1	Session 2
Doc. 1 - pat. 1 Version 1, scénario 1	Doc. 2 - pat. 2 Version 2, scénario 1
Doc. 1 - pat. 2 Version 1, scénario 2	Doc. 2 - pat. 1 Version 2, scénario 2
Doc. 1 - pat. 1 Version 2, scénario 3	Doc. 2 - pat. 2 Version 1, scénario 3
Doc. 1 - pat. 2 Version 2, scénario 4	Doc. 2 - pat. 1 Version 1, scénario 4
Pat. 2 remplit le questionnaire et quitte la salle	Pat. 1 remplit le questionnaire et quitte la salle
Doc. 1 - pat. 3 Version 1, scénario 5	Doc. 2 - pat. 4 Version 2, scénario 5
Doc. 1 - pat. 4 Version 1, scénario 6	Doc. 2 - pat. 3 Version 2, scénario 6
Doc. 1 - pat. 3 Version 2, scénario 7	Doc. 2 - pat. 4 Version 1, scénario 7
Doc. 1 - pat. 4 Version 2, scénario 8	Doc. 2 - pat. 3 Version 1, scénario 8
Pat. 4 remplit le questionnaire et quitte la salle	Pat. 3 remplit le questionnaire et quitte la salle
Doc. 1 remplit le questionnaire et quitte la salle	Doc. 2 remplit le questionnaire et quitte la salle
Temps total Docteurs : 1h20 ; patients : 40 min	

Tableau 4.4 : Extrait de protocole pour les sessions 1 et 2 (doc. = docteur, pat. = patient)

Nous avons effectué huit sessions de la même manière que dans le Tableau 4.4 ci-dessus. Lors de cette évaluation, nous nous sommes concentrés particulièrement sur les patients, puisque les deux versions diffèrent à ce niveau. Cependant, il est certain que la performance du patient dépend aussi de celle du médecin et vice-versa. Afin de contrôler au mieux cette variable, chaque patient a effectué une session avec deux médecins différents.

Afin de standardiser les patients, chaque patient recevait un scénario prédéfini. Lors d'une session, chaque médecin voyait quatre patients différents et était confronté à huit scénarios différents. Les scénarios (annexe 5) ont été établis en collaboration avec notre partenaire du *Dallas Childrens Hospital*, le Prof. Glenn Flores. Certains scénarios étaient plus simples que d'autres qui exigeaient des connaissances plus pointues. Cependant, chaque scénario a été suivi le même nombre de fois pour chaque version (V1 et V2).

Nous avons dû montrer aux utilisateurs comment utiliser les systèmes au niveau de l'interface et leur donner des informations sur le déroulement du test. Ces instructions concernaient aussi les types de phrases couvertes par chaque version du système et contenaient les règles de langage contrôlé (LC) à respecter. Ces règles, dont le rôle est de guider les utilisateurs vers la couverture de chaque système, sont exprimées dans la langue maternelle du

sujeto-paciente sous forme d'exemples, comme dans les tableaux 4.5 et 4.6 (voir les traductions entre parenthèse).

¿Cómo hablar con el sistema V1? (Comment parler au système V1)		
Puede usar solo frases incompletas o contestar sí / no (Vous ne pouvez utiliser que des phrases incomplètes ou répondre par oui/non)		
	Diga: (Dites)	No diga (Ne dites pas)
<u>Médico (médecin):</u> ¿Desde cuándo le duele la garganta? (Depuis quand avez-vous mal à la gorge ?)	Desde hace más de una semana (depuis plus d'une semaine)	me duele desde hace más de una semana (j'ai mal depuis plus d'une semaine)
<u>Médico (médecin):</u> ¿Tiene fiebre? (Avez-vous de la fièvre ?)	- Sí, un poco (oui, un peu) - Sí, más de 38 grados (oui, plus de 38 degrés) - No (non) - Sí (oui) - No sé (je ne sais pas)	- Sí, tengo un poco de fiebre (oui, j'ai un peu de fièvre) - Sí, tengo más de 38 grados (oui, j'ai plus de 38 degrés) - No, no tengo fiebre (non, je n'ai pas de fièvre) - Sí, tengo fiebre (oui, j'ai de la fièvre)

Tableau 4.5 : Instructions de LC pour V1

Pour V1, seules les phrases elliptiques sont autorisées. Il faut exprimer une seule idée par phrase. Aucune structure de phrase négative n'est prévue. Il faut impérativement répondre par « non ».

¿Cómo hablar con el sistema V2? (Comment parler au système V2)		
1. Expresa una idea por respuesta (Exprimez une idée par réponse)		
	Diga: (Dites)	No diga (Ne dites pas)
<u>Médico (médecin):</u> ¿Dónde le duele? (Où avez-vous mal?)	- Me duele la garganta (J'ai mal à la gorge) - Me duele la cabeza (J'ai mal à la tête)	Me duele la garganta y la cabeza (J'ai mal à la gorge et à la tête)
2. Puede usar (Vous pouvez utiliser)		
	frases enteras (phrases entières)	frases incompletas (phrases incomplètes)
<u>Médico (médecin):</u> ¿Desde cuándo le duele la garganta? (Depuis quand avez-vous mal à la gorge ?)	Me duele desde hace más de una semana (J'ai mal depuis plus d'une semaine)	Desde hace más de una semana (Depuis plus d'une semaine)
<u>Médico (médecin):</u> ¿Tiene fiebre? (Avez-vous de la fièvre)	- Sí, tengo mucha fiebre (Oui, j'ai beaucoup de fièvre) - No, tengo nauseas (Non, j'ai des nausées) (attention : il ne s'agit pas d'une phrase négative!)	- Sí, mucha fiebre (Oui, beaucoup de fièvre) - No, nauseas (Non, des nausées) - Sí (Oui) - No sé (Je ne sais pas)

3. No puede usar frases negativas (<i>Vous ne pouvez pas utiliser des phrases négatives</i>)		
Médico (<i>médecin</i>):	Diga (<i>Dites</i>) :	No diga (<i>Ne dites pas</i>) :
¿Tiene fiebre? (<i>Avez-vous de la fièvre ?</i>)	• No (<i>Non</i>)	• No tengo fiebre (<i>Non, je n'ai pas de fièvre</i>).

Tableau 4.6 : Instructions de LC pour V2

Pour V2, les patients peuvent répondre avec des phrases elliptiques et des phrases entières simples sans coordination, en se limitant toujours à une idée par phrase. Les phrases négatives sont interdites. Elles sont ambiguës en raison des phrases affirmatives précédées de "non", comme décrit au point 2 du Tableau 4.6. La réponse orale avec une virgule « *No, tengo nauseas* » (Non, j'ai des nausées) peut en effet signifier : « Non, j'ai des nausées » ou « Non, je n'ai pas de nausées » (sans la virgule). Comme la reconnaissance ne prend pas en compte l'intonation, il est difficile de distinguer les deux. Des tableaux qui précèdent, il apparaît clairement que les instructions étaient plus complexes pour V2 que pour V1.

Pour guider les utilisateurs vers la couverture, tous les participants avaient également la possibilité d'utiliser le module d'aide de l'interface (section 3.3).

4.1.2.3 Etude pilote

Comme conseillé dans (Saldanha et O'Brien 2014), nous avons tout d'abord effectué une étude pilote pour vérifier le bon déroulement de l'expérience. Celle-ci a eu lieu à Genève, UniMail, en janvier 2008, avec deux participantes hispanophones dans le rôle des patients et une collègue anglophone dans le rôle du médecin. Grâce à cette étude pilote, nous avons pu corriger quelques imprécisions dans les instructions et questionnaires fournis aux participants et vérifier que l'expérience pouvait effectivement se dérouler dans les temps prévus par le protocole.

4.1.2.4 Données

À l'issue de chaque session, les utilisateurs remplissaient un questionnaire de satisfaction. Cette expérience nous a aussi permis de collecter un corpus total de 2327 phrases transcrites : 1147 interactions pour la version restreinte (V1), dont 614 en anglais et 533 en espagnol, et 1180 pour la version moins restreinte (V2), dont 649 en anglais et 531 en espagnol, soit un total de 1263 interactions en anglais et de 1064 en espagnol. Ce corpus de données véridiques pourrait également être utilisé pour compléter la couverture de MedSLT (ou maintenant *BabelDr*¹⁰⁹) en vue d'une utilisation en situation réelle.

¹⁰⁹ Site du projet BabelDR, <http://babeldr.unige.ch/project-description/>, consulté le 01.03.2016.

Locuteur	Phrases transcrites
doctor	good morning my name is doctor calandro
doctor	good morning
patient	buenos días
doctor	how are you today
doctor	how are you feeling
doctor	how do you feel
doctor	what is bothering you
doctor	how are you
doctor	where is your pain
patient	me duele la garganta
doctor	how long have you had a sore throat
doctor	how long have you had a sore throat
patient	me duele desde hace dos días
doctor	how bad is the pain
doctor	what is the character of the pain
patient	fuerte
doctor	do you have a fever
doctor	do you have a fever
patient	sí yo he tenido fiebre
doctor	how high is your fever

Tableau 4.7 : Extrait brut d'un dialogue retranscrit, scénario 1, patient 1 et médecin 1

Dans le Tableau 4.6 ci-dessus, nous avons donné un extrait d'un dialogue pour illustrer la nature des phrases étudiées.

Dans la section suivante, nous allons décrire les mesures choisies pour notre analyse statistique.

4.1.2.5 Statistique : mesures employées

Comme nous l'avons vu dans le Tableau 4.3, nous avons retenu un certain nombre de variables. Nous allons maintenant examiner comment nous les quantifions.

Nous avons recours à des questionnaires pour évaluer certaines caractéristiques comme celle de la satisfaction (qualité à l'usage) des utilisateurs, mais aussi pour recueillir leur avis sur la facilité d'utilisation des logiciels (caractéristique de qualité externe). Ces questionnaires utilisent une échelle de type Likert de 5 points (1= le moins bon, et 5= le meilleur). Nous rapporterons les moyennes obtenues pour chaque question.

L'efficacité du logiciel est mesurée, d'une part, par le taux d'accomplissement de la tâche, mesuré en termes de pourcentage de diagnostics corrects, et, d'autre part, par la qualité de l'information sous la forme du rappel et de la précision des informations recueillies par les médecins. Nous quantifions également le pourcentage d'erreurs commises par les systèmes, les patients et les médecins.

La productivité est évaluée par le temps mis pour effectuer la tâche (poser un diagnostic en utilisant tour à tour V1 et V2 de MedSLT) et le nombre total d'interactions nécessaires pour y parvenir.

La sécurité, quant à elle, est évaluée en pourcentage d'interactions réussies et en pourcentage d'échecs des interactions.

Pour ce qui est de la qualité externe, pour la capacité fonctionnelle, nous utilisons les mesures classiques de la TAP (décrites au chapitre 3) pour quantifier la précision. S'agissant de l'évaluation de la facilité d'utilisation, outre le questionnaire, nous la quantifions en termes de pourcentage de phrases hors couverture pour la capacité de compréhension et en progression en temps entre la première et dernière utilisation du logiciel pour la capacité d'apprentissage.

Dans le Tableau 4.8 ci-dessous, nous reprenons les principales mesures choisies. Celles-ci sont des méthodes d'analyse statistique inférentielles et descriptives (Salkind 2012) selon la nature des données collectées. Pour toutes les variables dépendantes du tableau, la variable indépendante est V1 et V2. Les résultats de ces analyses statistiques sont décrits dans le chapitre 5.

Variables dépendantes (Caractéristiques)	Méthode de génération des données ¹¹⁰	Mesures choisies
Qualité à l'usage		
Satisfaction	Questionnaire de satisfaction Préférence entre V1 et V2	% par variable
Efficacité	Accomplissement de la tâche	% diagnostics corrects
	Qualité de l'information collectée	Rappel et précision
	Observation nombre d'erreurs systèmes	% d'erreurs système
	Observation nombre d'erreurs patients	% d'erreurs patients
	Observation nombre d'erreurs médecins	% d'erreurs médecins
Productivité	Observation du temps mis pour effectuer la tâche	Moyenne en minutes
	Observation du nombre d'interactions	Moyenne du nombre d'interactions par diagnostic
Sécurité	Observation du nombre d'interactions réussies	% d'inter. réussies (AC)
	Observation du nombre d'échecs des interactions	% d'échecs des interactions (NA)
Qualité externe		
Précision		
Qualité de la RV	Observation des erreurs de RV par mot	WER sur phrases couvertes
	Observation des erreurs de RV par phrase	SER sur phrases couvertes

¹¹⁰ Nous avons utilisé une méthode triangulaire (*method triangulation*, cf. Saldanha, G et S. O'Brien(2014), p.23 et Oates 2005, p.37) pour générer nos données afin d'étudier notre objet sous différentes perspectives. Il s'agit de deux types de génération de données (observation et questionnaire).

	Observation des phrases hors couverture	% de phrases hors couverture
	Observation des phrases acceptées	SemER
Qualité de la rétro-traduction	Observation de l'acceptabilité de la rétrotraduction	% de RetroTrad accepté / rejeté Binaire ¹¹¹
		Accord ¹¹² entre évaluateurs Dallas - kappa Fleiss
		Accord entre évaluateurs Traducteurs ¹¹³ - kappa Fleiss
TA - Évaluation humaine	Observation de la fluidité de la TA	Fluidité (échelle de 1 à 5)
		Accord entre évaluateurs Traducteurs - kappa Light
		Accord entre évaluateurs Traducteurs - ICC
	Observation de l'adéquation de la TA	Adéquation (échelle de 1 à 5)
		Accord entre évaluateurs Traducteurs - kappa Light
		Accord entre évaluateurs Traducteurs - ICC
TA - Évaluation automatique	Observation automatique de la qualité de la traduction	BLEU
Facilité d'utilisation		
Facilité de compréhension	Observation du nombre de phrases hors couverture	% de phrases hors couverture
Facilité d'apprentissage	Questionnaire	Échelle de Likert (1-5)
	Observation de la progression en temps entre 1ère et dernière utilisation	Progression moyenne en minutes

Tableau 4.8: Variables dépendantes assorties des mesures choisies

Comme le précise la recette en sept étapes d'EAGLES (section 3.1), il est important de définir les modalités d'interprétation des résultats obtenus avec la mesure choisie en déterminant une échelle de points pondérée selon l'importance accordée à l'attribut évalué en fonction du contexte d'utilisation. Dans la mesure du possible, nous avons suivi des indications de la littérature pour définir le seuil de tolérance à l'erreur afin de fixer une interprétation des résultats obtenus, objective et fondée. Ainsi, pour la qualité externe, les seuils définis dans le Tableau 4.8 sont inspirés de (Jurafsky et Martin 2009) pour la RV et de (Koehn 2010) pour la TA. En ce qui concerne les seuils d'interprétation kappa, nous avons suivi les indications de (Landis et Koch 1977). Pour le coefficient de corrélation intraclasse (ICC), nous nous fondons sur les travaux de (Cicchetti 1994). Nous avons défini les autres seuils selon notre expérience

¹¹¹ Les évaluateurs de Dallas effectuaient l'évaluation de la rétro-traduction en utilisant le système (en-système) pendant l'expérience EXPQ1.

¹¹² En bleu dans ce tableau, nous décrivons aussi les mesures de statistique inférentielle que nous avons appliquées pour mesurer la validité de nos données.

¹¹³ Les évaluateurs Traducteurs effectuaient l'évaluation de la rétro-traduction sans utiliser le système (hors-système).

du domaine. Nous n'avons pas appliqué de pondération spécifique, puisque nous avons gardé uniquement les caractéristiques importantes pour notre contexte d'utilisation.

Nous arrivons ainsi au Tableau 4.9, dans lequel nous avons ajouté à la mesure choisie la façon d'interpréter chaque échelon de l'échelle, ou la mesure obtenue, afin de pouvoir par la suite suivre cette grille pour attribuer un score unifié à chaque caractéristique et sous-caractéristique étudiée. Comme nous le verrons dans les sections 5.1, 5.2 et 5.3, dans lesquelles nous exposons les résultats, nous obtiendrons un score final (fondé sur un modèle de qualité motivé) pour chaque famille de caractéristiques afin de finalement parvenir à un total unique. L'application de cette méthode nous permettra d'obtenir un score final comparable pour chaque version du système étudié. Cette méthode et cette grille d'interprétation pourront être appliquées également à d'autres systèmes de TAP dans le domaine du médical.

4.1.3 Récapitulatif des variables pour Q1_principale : grille d'interprétation

Le Tableau 4.9 ci-dessous récapitule les variables dépendantes, la manière dont elles ont été mesurées et comment les résultats sont interprétés. Pour rappel, la variable indépendante pour l'EXPQ1 est V1 ou V2.

Caractéristiques et sous-caractéristiques (Variables dépendantes)	Evaluation effectuée (Chapitre 3)	Mesures choisies	Interprétation de la mesure	Points
Qualité à l'usage				
Efficacité (section 5.1.1)	Accomplissement de la tâche	% diagnostics justes	< 50%	-10
			>50% < 75%	0
			> 75%	10
	Qualité de l'information récoltée	Rappel	< 50%	-10
			>50% < 75%	0
			> 75%	10
		Précision	< 50%	-10
			> 50% < 75%	0
			> 75%	10

	Erreurs système	% d'erreurs	< 5% > 2% < 5% > 1% < 2%	-10 0 10
	Erreurs patients	% d'erreurs	< 10 % > 5 < 10% > 2,5 < 5%	-10 0 10
	Erreurs médecins	% d'erreurs	< 50 % > 25 < 50% < 25	-10 0 10
Productivité (section 5.1.2)	Temps mis pour effectuer la tâche	Moyenne en minutes	> 20 > 10 < 20 < 10	-10 0 10
	Nombre d'interactions	Moyenne du nombre d'interactions par diagnostic	> 50 > 30 < 50 < 30	-10 0 10
Satisfaction (section 5.1.3)	Préférence (en %)	< 50% > 50%		-10 10
Sécurité (section 5.1.4)	% d'inter. réussies (AC)		< 30 > 30 < 70 > 70	-10 0 10
	% d'échecs des interactions (NA)		> 70 > 50 < 70 < 50	-10 0 10
Qualité externe				
Précision				
Qualité de la RV (Jurafsky et Martin 2009) (section 5.2.1)	WER sur phrases couvertes	-> environ 15% ok	> 15% > 10 % < 15% < 10%	-10 0 10
	SER sur phrases couvertes	De 15% à 30% -> utilisable	> 30% > 15% < 30% < 15%	-10 0 10
	% de phrases hors couverture	Au-delà de 30% -> handicapant	> 30% > 15% < 30% < 15%	-10 0 10
	SemER	%	> 30% > 15% < 30% < 15%	-10 0 10
Qualité de la rétro- traduction (section 5.2.2)	Comparaison (binaire)	Binaire Accepté /rejeté En ligne et hors-ligne	< 70% acceptés > 70% < 80% > 80%	-10 0 10
(Landis et Koch 1977)	Évaluateurs Dallas	Kappa Fleiss	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
		% d'accord	< 50% > 50 < 80% > 80%	-10 0 10

	Évaluateurs Trad.	Kappa Fleiss	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
		% d'accord	< 50% > 60% > 80%	-10 0 10
TA - éval. humaine (section 5.2.3.1)	Fluidité échelle de 1 à 5 moyenne	Moyenne de 1-5	1-2 = mauvais 3 = neutre 4-5= bon	-10 0 10
(Landis et Koch 1977)	KAPPA Light	Entre 0 et 1	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
(Cicchetti, 1994)	ICC	Entre 0 et 1	< 0,70 (faible) > 0,70 < 0,79 (modéré) > 0,80 < 0,89 (fort)	-10 0 10
	Adéquation échelle de 1 à 5 moyenne	Moyenne de 1-5	1-2 = mauvais 3 = neutre 4-5= bon	-10 0 10
	KAPPA	Entre 0 et 1	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
	ICC	Entre 0 et 1	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
Evaluation automatique (section 5.2.3.2)	BLEU	Entre 0 et 1	< 0,40 < 0,60 > 0,60 < 0,80 > 0,80	-10 0 10
Facilité d'utilisation				
Facilité de compréhension (section 5.3.1)	Nombre de phrases hors couverture	Nombre de phrases hors couverture	> 60% > 40% < 60% < 40%	-10 0 10
Facilité d'apprentissage (section 5.3.2)	Evaluation qualitative par questionnaire	Likert 1-5	Entre 1-2 = mauvais >3 = neutre Entre 4-5= bon	-10 0 10
	Progression en temps entre 1ère et dernière utilisation	Progression moyenne en minutes	+ 1 à +5 -1 à -1,9 -2 à -5	-10 0 10
SCORE FINAL	Total de tous les points obtenus par chaque système V1 et V2 (V.I.)			

Tableau 4.9 : Variables dépendantes et interprétation des mesures choisies

Tout au long de l'application du modèle de qualité retenu, intégralement présenté dans le tableau ci-dessus, nous ferons le point pour chaque type d'évaluation (qualité à l'usage, qualité externe de précision ou de facilité d'utilisation) au moyen d'un tableau récapitulatif intermédiaire qui reprendra au fur et à mesure les principaux résultats. Tous les résultats obtenus seront décrits et analysés dans le chapitre 5.

Comme indiqué précédemment, nous nous sommes employée à exposer dans cette section 4.1, la démarche que nous avons suivie (modèle de qualité et expérience) pour répondre à **Q1_principale** et aux sous-questions qui en découlent (**Q1_S1-Q1_S9**). Les résultats obtenus dans ce cadre ont constitué la base de travail sur laquelle nous nous sommes appuyée pour mener à bien notre recherche sur notre deuxième question principale (**Q2_principale**) à laquelle est consacrée la section suivante.

4.2 Question de recherche 2 : TURKoïse

Nous avons exposé les sept sous-questions de recherche pour **Q2_principale** (*Peut-on proposer une nouvelle mesure pour la TAP qui serait équivalente à l'utilisation de la méthode EAGLES-ISO, mais plus rapide et plus économique à mettre en place que celle-ci ?*) dans l'introduction du présent chapitre (cf. Tableau 4.2). Dans cette section, nous allons préciser quelles sont nos variables et comment les mesurer grâce à quelles expériences. Tous les résultats concernant **Q2_principale** seront donnés au chapitre 6. Nous allons maintenant examiner comment nous comptons répondre à nos questions. Pour ce faire, nous procéderons en trois étapes.

4.2.1 Étape 1 : Evaluation humaine pour TAP médicale : échelle d'utilité à 4-points

4.2.1.1 Objectif

L'objectif de la première étape est de répondre à **Q2_S1** (*La méthode HMS proposée est-elle plus adaptée que les échelles classiques d'évaluation de la fluidité et de l'adéquation ?*). La nouvelle échelle *Human Medical Scale* (HMS) sera présentée à la section 6.1. Nous comparerons les résultats obtenus en faisant appliquer celle-ci par nos évaluateurs de Dallas aux résultats obtenus avec les échelles classiques de fluidité et d'adéquation (section 5.2.3.1). Notre comparaison portera sur deux points : la cohérence entre les évaluateurs et la capacité à départager V1 et V2. Dans la section suivante (4.2.1.2), nous présenterons l'expérience **EXPQ2_S1**, les participants, les données (4.2.1.2.1) et les mesures utilisées (4.2.1.2.2).

4.2.1.2 Mise en place de l'expérience EXPQ2_S1

Nos évaluateurs pour cette expérience sont les interprètes anglais-espagnol de l'Hôpital de Dallas qui avaient participé à la tâche de simulation en jouant le rôle des patients fictifs. Chaque interprète était rémunéré 50 dollars pour participer à l'expérience de simulation décrite dans section 4.1.2 (**EXPQ1**), mais aussi pour effectuer la tâche d'évaluation décrite à présent. Il devait également fournir une traduction de référence pour chaque phrase évaluée. Nous avons

demandé à chacun d'évaluer par courrier électronique une série aléatoire de phrases en utilisant l'échelle HMS. Les évaluateurs effectuaient l'évaluation dans les deux fichiers Excel (annexe 3). Chaque évaluateur obtenait un total de 190 phrases composées de 5 tranches de 16 phrases anglais-espagnol (selon la répartition indiquée ci-dessous dans le Tableau 4.10) et 5 tranches de 22 phrases espagnol-anglais.

Num phrases	1-16	17-32	33-48	49-64	65-80	81-96	97-112	113-128	129-144	145-160	161-176	177-192	193-208	209-229
Eval. N.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	2	3	4	5									
2		5	4	3	2	1								
3			1	2	3	4	5							
4				5	4	3	2	1						
5					1	2	3	4	5					
6						5	4	3	2	1				
7							1	2	3	4	5			
8								5	4	3	2	1		
9									1	2	3	4	5	
10										5	4	3	2	1
11	5										1	2	3	4
12	2	1										5	4	3
13	3	4	5										1	2
14	4	3	2	1										5

Tableau 4.10 : Répartition des 229 phrases (anglais-espagnol) entre les 14 évaluateurs de l'Hôpital de Dallas

Notre but était d'obtenir 5 évaluations au total si tous les évaluateurs sollicités répondaient, et de nous assurer d'avoir au moins 3 évaluations pour chaque phrase au cas où tout le monde ne répondait pas. Chaque interprète rémunéré n'a cependant pas pris le temps de participer à cette deuxième partie de la tâche (évaluation et traduction de référence). Ainsi, nous n'avons pu atteindre que 3 évaluations par phrase au lieu de 5. Nous donnons un court extrait de l'évaluation obtenue dans le Tableau 4.11 ci-dessous :

Original	Traduction	Eval.1	Eval. 2	Eval. 3
Are you allergic to any medicines?	¿Tiene alergia a alguna medicina?	CCOR	MEAN	MEAN
Are you having fever?	¿El dolor está aliviado cuando tiene fiebre?	DANG	DANG	DANG
Did you see a doctor this week?	¿Ha consultado un médico esta semana?	CCOR	CCOR	CCOR
Do you have a headache?	¿Tiene tos ayer?	DANG	NONS	DANG

Tableau 4.11 : Exemple d'application de l'échelle HMS

Il est à noter que les évaluateurs utilisaient les étiquettes CCOR, MEAN, NONS ou DANG que nous avons par la suite remplacées par une valeur numérique afin de procéder aux

calculs des moyennes. Nous détaillerons dans la section 6.1 comment nous avons procédé pour arriver à cette échelle HMS.

4.2.1.2.1 *Données*

Pour cette évaluation, nous avons ainsi utilisé un exemplaire unique de nos environ 2300 phrases au total pour chaque combinaison linguistique, c'est-à-dire que nous avons gardé uniquement la première occurrence d'une phrase et sa traduction et non les répétitions. Nous avons par conséquent tout d'abord conservé uniquement les types¹¹⁴. Ainsi, notre corpus de données évaluées s'élève à 222 phrases (types) pour la direction anglais-espagnol (annexe 3), et 305 phrases pour la direction espagnol-anglais. Ces phrases ont été évaluées comme décrit plus haut (par type), puis réintégrées dans leur contexte pour obtenir une évaluation pour chaque occurrence des phrases du corpus total. Une fois ces respectivement 222 et 305 phrases uniques (types) replacées dans leur contexte, nous arrivons au nombre de phrases total (répétition comprises) par système et langues décrit dans le Tableau 4.12 ci-dessous. Il faut également préciser que nous n'avons évalué que les phrases acceptées par les utilisateurs pendant la simulation de diagnostic.

Type de phrases	Nombre de phrases	Nb d'évaluations
anglais-espagnol - V1	358	3
anglais-espagnol - V2	390	3
espagnol-anglais - V1	335	3
espagnol-anglais - V2	364	3
Total phrases EXPQ2_S1	1147	

Tableau 4.12 : Total du nombre de phrases évaluées pour EXPQ2_S1

4.2.1.2.2 *Statistique : mesures employées*

Nous présenterons la moyenne obtenue, mais aussi la déviation standard pour les évaluateurs, ainsi que les résultats sous forme de mode pour clairement montrer laquelle des deux versions obtient de meilleurs résultats avec notre échelle. Nous comparerons ces résultats à ceux obtenus avec les mesures classiques de l'adéquation et de la fluidité (section 5.2.3.1). Nous mesurerons ensuite la cohérence de notre évaluation par l'échelle HMS grâce au kappa mais aussi à l'ICC et au pourcentage d'accord entre évaluateurs.

¹¹⁴ La notion de « type » est dérivée de celle de l'étude de corpus, qui est définie comme suit : « *Each different word in a corpus. Language is repetitive, which means that some words will appear in the corpus more than once. When calculating the number of types in the corpus, you do not count each word (tokens), you only count each different word.* », p. 233, in: Bowker, L. et J. Pearson (2002). *Working with specialized language: a practical guide to using corpora*, Routledge.

Le Tableau 4.13 ci-dessous récapitule les éléments essentiels de l'expérience **EXPQ2_S1** et indique les sections dans lesquelles les résultats sont présentés.

N° de question	N° d'exp.	Données employées	Mesures	Résultats
Q2_S1	EXPQ2_S1	3 évaluations de 1147 phrases ¹¹⁵ , par les évaluateurs experts de Dallas	Moyenne KAPPA ICC % accord	Section 6.1.2

Tableau 4.13: Tableau récapitulatif de l'expérience EXPQ2_S1

Nous passons à présent à la deuxième étape du processus suivi pour répondre à notre deuxième question de recherche. Cette étape consiste à rechercher une corrélation entre notre échelle HMS et des mesures automatiques pour les raisons que nous exposerons ultérieurement.

4.2.2 Étape 2 : Corrélation entre mesures automatiques et échelle HMS

4.2.2.1 Objectif

L'objectif de la **deuxième étape** est de répondre à notre question **Q2_S2** (*Quelle méthode d'évaluation automatique corrèle le mieux avec notre méthode humaine HMS ?*), nous procéderons tout d'abord à l'expérience (**EXPQ2_S2**) dans laquelle nous appliquerons les mesures automatiques classiques à nos données, puis nous calculerons la corrélation des résultats obtenus par ce biais avec ceux issus de l'évaluation humaine avec HMS (section 4.2.2.1). Ensuite, pour répondre à la question **Q2_S3** (*Les mesures automatiques étudiées pour Q2_S2 sont-elles moins adaptées à l'évaluation de la TAP linguistique comme MedSLT ?*) nous étudierons à quel point l'évaluation automatique de la TA (notamment BLEU) est adaptée à une évaluation phrase par phrase (**EXPQ2_S3**) afin de répondre à la question **Q2_S3**.

Nous alignant sur l'observation de (Jones, Gatewood et al. 2014), notre objectif ici n'est pas de démontrer que les mesures automatiques ne peuvent être utilisées pour prédire la qualité à l'usage de la TAP, mais d'étudier si, parmi les mesures automatiques existantes, il n'y en a pas une qui pourrait au moins servir de méthode de triage¹¹⁶. Pour cela, nous procéderons tout d'abord à l'expérience **EXPQ2_S2** que nous décrirons en détail dans la sous-section suivante

¹¹⁵ Tableau 4.12

¹¹⁶ "The appropriate observation is to note that the automatic score are not a reliable predictor of PASS/FAIL scores. The question is any of the chosen automatic metric a predictor of my own scale ?" Jones, D., P. Gatewood, M. Herzog et T. Marius (2014). Standardized ILR-Based and Task-Based Speech-to-Speech MT Evaluation. Workshop on Automatic and Manual Metrics for Operational Translation Evaluation - within LREC 2014, Keith J. Miller, Lucia Specia, Kim Harris, Stacey Bailey: 51-60.

(4.2.2.2). Ensuite, dans la section 4.2.2.3, nous détaillerons l'**EXPQ2_S3** qui doit répondre à **Q2_S3**.

4.2.2.2 Mise en place de l'expérience EXPQ2_S2

Dans un premier temps, notre objectif est de trouver quelles mesures automatiques corrélient le mieux avec nos mesures humaines axées sur l'utilité (ou dans certains cas, la dangerosité) des phrases produites. Afin d'étudier la corrélation entre les différentes mesures appliquées (humaines et automatiques), nous avons calculé à partir des données de notre évaluation du chapitre 5 le coefficient de Pearson comme dans (Estrella, King et al. 2007). La méthode utilisée dans cet article est celle de l'extrapolation des données (*bootstrapping technique*), le but étant de générer un grand nombre de points d'échantillon sur lesquels les corrélations sont ensuite calculées.

Nos évaluateurs pour cette expérience sont les interprètes de l'Hôpital de Dallas (voir description détaillée à la section 4.1.2.1), comme pour l'expérience précédente.

Les évaluations HMS étudiées dans cette expérience portent sur les mêmes phrases du corpus de Dallas que pour l'expérience précédente (**EXPQ2_S1**).

Les traductions de référence sont celles effectuées par les interprètes de l'hôpital de Dallas (cf. section 4.2.2.1) et (section 6.1.1). Nous avons collecté trois références par phrase.

Pour ce qui est des mesures, nous utiliserons le coefficient de corrélation de Pearson en comparant les évaluations HMS avec respectivement BLEU, NIST, la précision, le rappel, la F-mesure, le PER et le WER.

N° de question	N° d'exp.	Données employées	Mesures	Résultats
Q2_S2	EXPQ2_S2	3 évaluations de 1147 phrases ¹¹⁷ , par les évaluateurs experts de Dallas	Coefficient de corrélation de Pearson	Section 6.2.2

Tableau 4.14: Tableau récapitulatif de l'expérience EXPQ2_S2

Les résultats de cette expérience sont présentés dans la section indiquée dans le Tableau 4.14 ci-dessus. Nous allons maintenant décrire la deuxième expérience (**EXPQ2_S3**) effectuée dans le cadre de cette deuxième étape de notre deuxième question principale de recherche.

¹¹⁷ Tableau 4.12

4.2.2.3 Mise en place de l'expérience EXPQ2_S3

L'expérience que nous décrivons dans cette section correspond à notre sous-question **Q2_S3** : « *Les mesures automatiques étudiées pour Q2_S2 sont-elles moins adaptées à l'évaluation de la TAP linguistique comme MedSLT ?* ».

Nous examinerons la question de l'adaptation des mesures automatiques à l'évaluation de la TAP et plus particulièrement à la TAP de nature linguistique, en comparant cette fois-ci des données issues des deux versions de *MedSLT* avec un système statistique tel *Google Translate*¹¹⁸ (GT).

4.2.2.3.1 *Participants*

Nos participants à l'évaluation humaine se distinguent des évaluateurs des expériences précédentes. En effet, nous avons voulu intégrer une variable indépendante supplémentaire en formant deux groupes d'évaluateurs : les évaluateurs traducteurs et les évaluateurs non-traducteurs. Nous avons introduit cette distinction afin de pouvoir comparer la sévérité de chaque groupe, parce que, comme nous l'avons mentionné à la section 3.3.2, dans nos études précédentes, il se dégageait une tendance que nous voulions vérifier : les évaluateurs traducteurs sont-ils plus sévères que les évaluateurs non traducteurs à l'égard de la TAP ? En effet, il semble que les traducteurs avaient souvent du mal à se détacher de la forme syntaxique des phrases à évaluer pour se concentrer uniquement sur le sens. Nous avons par conséquent demandé à six nouveaux évaluateurs d'évaluer le sous-ensemble de phrases décrits ci-dessous (section 4.2.2.3.2). Trois sont traducteurs et trois sont des hispanophones maîtrisant l'anglais suffisamment pour évaluer les phrases sans être traducteurs.

4.2.2.3.2 *Données*

Le corpus d'évaluation HMS étudié dans cette expérience est un extrait de 222 phrases issues du corpus entier de Dallas. Il s'agit des phrases uniques (en éliminant les répétitions) énoncées par les patients et provenant autant de V1 que de V2. Pour pouvoir comparer avec les traductions de GT, nous avons fait passer les phrases originales collectées durant l'expérience décrite à la section 4.4.2 par l'interface de GT, mais sous forme écrite, sans passer par l'étape de RV, puisque nous voulons à ce stade évaluer la qualité de la TA et non celle de la RV.

¹¹⁸ Nous verrons au chapitre 6 pourquoi nous avons choisi GT. Ce choix est principalement justifié par l'utilisation fréquente de GT dans le milieu hospitalier, voir chapitre 2.

Type d'éval.	Type de phrases	Nombre de phrases	Groupe évaluateurs traducteurs	Groupe évaluateurs non-traducteurs
HMS	anglais-espagnol - MedSLT	222	3	3
	anglais-espagnol - GT	222	3	3
Nombre total de phrases pour EXPQ2_S3		444	Nb total d'éval.	6

Tableau 4.15 : Données de l'expérience EXPQ2_S3

Nous allons maintenant expliquer quelles sont les mesures que nous avons employées pour cette expérience **EXPQ2_S3**.

4.2.2.3.3 *Statistique : mesures employées*

Pour la partie expérience humaine, nous évaluerons l'accord entre évaluateurs par le kappa, le pourcentage d'accord et ICC.

Pour la partie de cette expérience qui concerne la corrélation, nous mesurerons le coefficient de corrélation de Pearson en comparant les évaluations HMS avec respectivement BLEU, NIST, la précision, le rappel, la F-mesure, le PER et le WER qui sont les mesures les plus communes (section 3.3.2.2).

N° de question	N° d'exp.	Données employées	Mesures	Résultats
Q2_S3	EXP Q2_S3	222 phrases anglais-espagnol extraites du corpus de Dallas traduites par MedSLT et GT (sans RV) → 3 évaluations d'experts (traducteurs) des 222 phrases → 3 évaluations de non-linguistes bilingues	KAPPA % accord Corrélation de Pearson HMS vs Mesures automatiques	Section 6.2.3

Tableau 4.16: Tableau récapitulatif de l'expérience EXPQ2_S3

Nous passons à présent à la troisième étape de notre deuxième question de recherche principale.

4.2.3 Étape 3 : TURKoise

4.2.3.1 Objectif

L'objectif de la troisième étape est de proposer une méthode d'évaluation de la TAP qui serait une alternative intéressante aux méthodes automatiques ou humaines classiques. Nous avons décomposé cette question générale en une série de sous-questions **Q2_S4** à **Q2_S7**. Celles-ci traitent respectivement de la faisabilité (**Q2_S4**), de la fiabilité (**Q2_S5**) et des bénéfices en temps et coûts (**Q2_S6**) de la méthode proposée. L'objet de **Q2_S7** est d'étudier

l'évolution de la cohérence entre évaluateurs en augmentant leur nombre (voir détails Tableau 4.2). Pour répondre aux questions de **Q2_principale**, nous avons mis en place deux expériences successives **EXPQ2_S4-S6** (section 4.2.3.2) et **EXPQ2_S7** (section 4.2.3.2), dont le but est de répondre à notre **Q2_principale**. De ces expériences et de la combinaison entre notre mesure humaine HMS et l'utilisation du *crowdsourcing* émerge notre proposition de méthode d'évaluation : *TURKoise*.

L'**EXPQ2_S4-S6** est, en quelque sorte, une étude pilote qui doit nous permettre d'étudier la faisabilité de *TURKoise*, en premier lieu, au niveau du recrutement des travailleurs AMT (**Q2_S4_S1**). Ensuite, nous nous intéresserons à la fiabilité des travailleurs AMT recrutés (**Q2_S5**) ainsi qu'à leur cohérence entre eux que nous mesurerons grâce au score kappa, ICC et au pourcentage d'accord entre évaluateurs (**Q2_S5_S1**). Puis, nous observerons si les résultats obtenus avec *TURKoise* sont comparables à ceux obtenus avec des évaluateurs experts (**Q2_S5_S2**). Enfin, nous évaluerons si les résultats obtenus avec *TURKoise* sont équivalents à ceux obtenus avec une évaluation humaine classique de l'adéquation et de la fluidité (**Q2_S5_S3**).

Notre expérience **EXPQ2_S7** visera à répondre à la sous-question de recherche : « *Quelle est l'évolution de la cohérence entre travailleurs AMT lorsqu'on augmente leur nombre ?* » (**Q2_S7_S7**). Pour ce faire, nous étudierons l'évolution de la cohérence entre évaluateurs (notamment par le biais du score kappa) selon leur nombre. C'est pour cette raison que nous sommes passée - pour effectuer les mêmes tâches que pour **EXPQ2_S4-S6** - d'un total de cinq évaluations par phrase à 8 évaluations au total. Ces données récoltées nous permettront de comparer les résultats obtenus avec successivement 3 évaluations, puis 5, puis 8, par rapport à un nombre plus réduit d'évaluations effectuées par des experts.

Nous commencerons par décrire la mise en place de l'expérience **EXPQ2_S4-S6**, puis nous continuerons avec l'**EXPQ2_S7**.

Cependant, avant de pouvoir expliquer la mise en place de nos expériences nous devons présenter le vocabulaire particulier du CS et de la plateforme de CS que nous avons employée, Amazon Mechanical Turk (AMT) : comme décrit plus en détail dans la section 6.3, les tâches imparties aux travailleurs sur AMT sont divisées en *Human Intelligence Tasks* (HIT). Ce sont des tâches restreintes que chaque travailleur AMT (*AMT worker*) pourra accepter d'effectuer en contrepartie d'une rémunération (voir section 6.3 pour de plus amples précisions).

4.2.3.2 Mise en place de l'expérience pilote : EXPQ2_S4-S6

La série de HITs construite couvre toutes les 222 phrases que nous voulions faire évaluer par les travailleurs AMT. Chaque HIT comportait ainsi 20 phrases à évaluer. Il s'agissait en effet de maintenir le volume de la tâche à un niveau raisonnable. Concrètement, nous avons tout d'abord transformé les données en format .csv pour ensuite pouvoir les intégrer dynamiquement dans le modèle html qui constitue la page web présentée aux AMT, grâce à un outil créé par notre collègue Johanna Gerlach de la FTI.

Pour répondre à notre question Q2_S4 (*L'implémentation de TURKoise est-elle faisable ?*) nous avons tout d'abord voulu expérimenter la faisabilité de recrutement d'un nombre suffisant de travailleurs AMT pour effectuer tous les HITs (Q2_S4_S1 : *Quelle est la faisabilité du recrutement des travailleurs AMT pour notre tâche ?*). Nous avons ainsi proposé la tâche sur la plateforme Amazon Mechanical Turk en février 2014. Dans la section des résultats, nous verrons que nous avons rapidement pu recruter suffisamment de travailleurs AMT pour arriver à 5 évaluations totales.

Il faut noter ici que les travailleurs AMT n'évaluent qu'une partie de ces 222 phrases sans savoir que celles-ci font partie d'un tout plus important. Il faut préciser également que c'est le groupe de travailleurs AMT qui effectue collectivement la tâche. Chacun peut accepter autant de HITs qu'il veut, une fois qu'il a été sélectionné comme étant qualifié pour effectuer la tâche. Par conséquent chacune des 222 phrases est évaluée par différents travailleurs, pour atteindre le nombre voulu d'évaluations totales.

Ensuite, nous avons demandé aux travailleurs AMT et au groupe d'experts d'évaluer les données issues de *MedSLT*, d'une part, en utilisant l'échelle classique de la fluidité et l'adéquation et, d'autre part, en appliquant notre échelle à 4-points axée sur l'utilité de la traduction (HMS).

4.2.3.2.1 Participants

Afin de recruter des travailleurs AMT qualifiés pour notre tâche (maîtrisant l'anglais et l'espagnol), nous avons tout d'abord publié un HIT contenant 20 phrases de référence, auparavant évaluées par les évaluateurs experts-maison, afin d'établir notre *gold standard*. Ce HIT a été rendu disponible à tous les travailleurs AMT, sans restriction autre qu'un prérequis sur leur historique de travail¹¹⁹. Tous les travailleurs AMT ayant obtenu un score minimum de

¹¹⁹ AMT permet de restreindre l'accès au HIT sur base de l'historique de chaque travailleur, qui sert d'indice de fiabilité. Cette restriction porte sur le nombre total de HITs complétés par le travailleur sur AMT, ainsi que sur le pourcentage de ces HITs qui ont été validés comme satisfaisants. Il est ainsi possible d'éviter les travailleurs peu scrupuleux. Eskenazi, M., G.-A. Levow,

75% de concordance avec le *gold standard* sur ce HIT de sélection ont obtenu une qualification (la plateforme AMT attribue des qualifications personnalisées qui permettent par la suite de donner ou non accès à des HITs). Au total, nous avons attribué les qualifications mentionnées à un ensemble de 11 travailleurs, dans les 24 premières heures après publication de la tâche. Seuls les travailleurs AMT qualifiés ont alors eu accès aux trois « véritables » tâches, qui correspondaient à 29 HITS au total.

Nous avons fidélisé les travailleurs AMT qui avaient effectué la tâche de sélection avec succès grâce à un bonus de 0,20 \$, qui paraît être une somme dérisoire mais qui, au vu des tarifs appliqués généralement, correspond à une « coquette somme ». Le sujet de la rétribution des travailleurs de CS mérite d’être débattu et a fait d’ailleurs l’objet d’un article sur l’éthique du recours généralisé au CS depuis 2010 (Fort, Adda et al. 2011).

À part les travailleurs AMT, nous avons également recruté des évaluateurs experts, c’est-à-dire dans ce cas des collègues traducteurs ou des étudiants de MA de la FTI. Il faut préciser que nous bénéficions d’un contexte particulièrement propice au recrutement d’évaluateurs bilingues. Ce n’est pas le cas de tous les chercheurs ou de toute personne chargée d’effectuer une évaluation de ce type. Nous avons ainsi pu recruter notre relativement petit groupe d’évaluateurs (traducteurs et non-traducteurs) assez rapidement, puisqu’ils ont pu répondre favorablement à la demande en moins de 24 heures. Le corpus est toujours constitué de 222 phrases extraites du corpus généré par l’expérience de la première question de recherche. Cependant, ces 222 phrases regroupent des phrases issues de MedSLT V1 et V2 sans distinction.

Évaluateurs	Sous-groupe	Nb. d'évaluations	Nb. d'évaluateurs	Nb. de phrases
Groupe Experts - maison	Traducteurs	5	5	222
	Non-traducteurs	3	3	222
Groupe Experts- de l'hôpital de Dallas	Interprètes de l'H. de Dallas	3	14 au total	190
Travailleurs AMT	EXPQ2_S4-S6	5	11/20 AMT recrutés	222
	EXPQ2_S7	3	3/13 AMT invités mail	222

Tableau 4.17 : Récapitulatif des groupes d'évaluateurs et du nombre d'évaluations

Nous allons à présent examiner plus en détail les types de tâche à effectuer par nos participants.

4.2.3.2.2 *Protocole*

Les travailleurs AMT évaluaient nos phrases sous forme de HIT comme mentionné plus haut. Ces HITs étaient constitués des trois tâches à effectuer en ligne sur AMT mais par petits groupes de phrases :

1. L'évaluation monolingue de la fluidité des traductions anglais-espagnol avec l'échelle classique de 5-points (section 3.3.2.1.1)
2. L'évaluation bilingue de l'adéquation des traductions anglais-espagnol avec l'échelle classique de 5-points (section 3.3.2.1.1)
3. L'évaluation bilingue des phrases en utilisant notre échelle HSM à 4 points (section 6.)

Nos évaluateurs experts ont effectué la même évaluation à l'aide d'une simple feuille de calcul Excel (voir annexe 11 pour un exemple). Les évaluateurs hispanophones (non-traducteurs) ont d'abord évalué la fluidité, en lisant uniquement la traduction espagnole sans avoir recours à l'original. Ensuite, les évaluateurs bilingues ont évalué l'adéquation.

4.2.3.2.3 *Statistique : mesures employées*

Nous avons fait évaluer nos phrases avec l'échelle HMS mais aussi avec l'échelle classique de fluidité et d'adéquation.

Dans notre étude, nous mesurerons la fiabilité des travailleurs AMT sélectionnés par rapport à leur cohérence (kappa et pourcentage d'accord, voir section 3.3.2.1.2) mais également en comparant leurs résultats avec les résultats obtenus par un groupe de contrôle d'experts.

N° de question	N° d'expérience	Données employées	Mesures	Résultats
Q2_S4 Q2_S5 Q2_S6	EXPQ2_S4-S6	222 phrases de MedSLT (V1 et V2 confondu) Evaluation HMS, fluidité et adéquation	kappa % accord	Section 6.3.2

Tableau 4.18: Tableau récapitulatif de l'expérience EXPQ2_S4-S6

Le Tableau 4.18, ci-dessus, reprend les informations essentielles de l'expérience **EXPQ2_S4-S6** et indique dans quelle section seront exposés les résultats de celle-ci.

4.2.3.3 *Mise en place de l'expérience EXPQ2_S7*

L'expérience **EXPQ2_S7** est la même que l'expérience **EXPQ2_S4-S6** décrite plus haut, sauf que nous avons augmenté le nombre de travailleurs AMT afin d'arriver à un total de huit évaluations par phrase. Afin d'augmenter le nombre de travailleurs AMT, nous avons contacté les travailleurs qui avaient été sélectionnés lors de la phase qualificative, mais qui n'avaient pas

encore participé à l'expérience et nous avons également contacté les précédents travailleurs en leur proposant un bonus afin de les fidéliser et de les inciter à participer à cette deuxième expérience.

Nous évaluons l'évolution de la cohérence en termes de kappa et ICC sur 3, puis 5, puis 8 évaluations, comme indiqué dans le Tableau 4.19.

N° de question	N° d'expérience	Données employées	Mesures	Résultats
Q2_S7	EXP_Q2_S7	Comparaison entre 3 évaluations AMT, 5 éval. AMT et 8 éval. AMT	kappa % accord	Section 6.3.3

Tableau 4.19: Tableau récapitulatif de l'expérience EXPQ2_S7

Tous les résultats pour **Q2_S7** se trouvent dans le chapitre 6, section 6.3.3.

4.2.4 Récapitulatif des variables pour Q2_principale

Nous reprenons ici les données clés concernant la deuxième question de recherche principale.

Pour la première étape, la variable indépendante est V1 et V2 de *MedSLT* et les variables dépendantes sont la qualité et la cohérence.

Pour la deuxième étape, la variable indépendante est *MedSLT* et *GT* et la variable dépendante, la corrélation entre HMS et les mesures automatiques étudiées.

Pour la troisième étape, la variable indépendante est l'identité de l'évaluateur (travailleur AMT ou expert). Les variables dépendantes sont la rapidité de réponse, la qualité, la cohérence, le temps et les coûts (**EXPQ2_S4-S6**). Pour **EXPQ2_S7**, la variable indépendante est le nombre d'évaluations (3, 5, puis 8). La variable dépendante est la cohérence.

Le Tableau 4.20 récapitule toutes les variables, les données utilisées ainsi que les mesures employées pour chaque expérience de **Q2_principale**.

N° de question	N° d'exp.	Données employées	Variable dépendante	Mesures	Variable indép.
Q2_S1	EXP_Q2_S1	3 évaluations de 1147 phrases ¹²⁰ , par les évaluateurs experts de Dallas	Qualité (HMS) Cohérence	Moyenne KAPPA ICC % accord	V1 ou V2
Q2_S2	EXP_Q2_S2		Corrélation	Coefficient de Pearson	MedSLT ou GT
Q2_S3	EXP_Q2_S3	222 phrases anglais-espagnol extraites du corpus de Dallas traduites par MedSLT et GT → 3 éval. de traducteurs → 3 éval. de non-linguistes	Qualité Cohérence Corrélation HMS vs mesures auto	Moyenne KAPPA % accord Coefficient de Pearson	
Q2_S4_S1	EXP_Q2_S4-S6	222 phrases de MedSLT (V1 et V2 confondues)	Rapidité de réponse Qualité (HMS, fluidité / adéquation)	Temps Moyenne	Travailleur AMT ou évaluateur expert
Q2_S5_S1			Cohérence	Kappa % accord	
Q2_S5_S2			Cohérence TURKoise et évaluateurs experts HMS	Kappa % accord	
Q2_S5_S3			Cohérence TURKoise et adéquation / fluidité	Kappa % accord	
Q2_S6			Temps et coûts	Heures, jours et dollars	
Q2_S7	EXP_Q2_S7	Comparaison entre 3 évaluations AMT, 5 éval. AMT et 8 éval. AMT	Cohérence	Kappa % accord	Nombre d'éval.

Tableau 4.20 : Tableau récapitulatif des expériences effectuées pour Q2_principale

¹²⁰ Voir Tableau 4.12

4.3 Conclusion

Dans ce chapitre, nous avons décrit la méthodologie suivie pour répondre à notre question de recherche générale : comment comparer des systèmes de TAP dans le domaine médical de manière fiable ? Ce questionnement nous a conduit à notre première question de recherche principale (**Q1_principale**, section 4.1 : *Peut-on utiliser la méthode EAGLES-ISO pour évaluer les systèmes de TAP dans le domaine médical ?*), que nous avons traitée sous forme de neuf sous-questions qui correspondent au modèle de qualité défini dans la section 4.1.1. Le modèle de qualité retenu comporte des caractéristiques de qualité à l'usage (efficacité, productivité, satisfaction et sécurité) et de qualité externe (capacité fonctionnelle et facilité d'utilisation). Dans la section 4.1.2, nous avons décrit l'expérience de simulation de diagnostic utilisant deux versions comparables de *MedSLT* (**EXPQ1**). Celle-ci nous a permis de collecter les données pour répondre à **Q1_principale**. Elles nous serviront également pour la deuxième question de recherche principale (**Q2_principale** : *Peut-on proposer une nouvelle mesure pour la TAP qui serait équivalente à l'utilisation de la méthode EAGLES-ISO, mais plus rapide et plus économique à mettre en place que celle-ci ?*). Dans la section 4.1.3, nous avons repris toutes les variables dépendantes et indépendantes pour la première question de recherche principale. Le Tableau 4.9 constitue un outil important de notre méthode EAGLES-ISO puisqu'il correspond à notre grille d'interprétation des résultats pour cette partie de la thèse (chapitre 5).

Pour répondre à **Q2_principale** (section 4.2), nous avons procédé en trois étapes. La première étape (section 4.2.1) consiste à proposer une échelle humaine d'évaluation adaptée à la TAP du domaine médicale (HMS) et à la tester sur notre corpus de phrases issu d'**EXPQ1**. La deuxième étape (section 4.2.2) vise à rechercher des mesures automatiques qui corrèlent avec l'échelle HMS. La troisième étape (section 4.2.3) a pour objectif de proposer la méthode d'évaluation TURKoise, qui combine l'échelle HMS et le *crowdsourcing*. Chacune de ces étapes comporte une série de sous-questions auxquelles nous répondrons grâce aux expériences que nous avons présentées dans ce chapitre et qui sont reprises dans la section récapitulative 4.2.4.

Le schéma ci-dessous (Figure 4.c) permet de visualiser les principaux points de notre méthodologie. Il reprend nos questions de recherche, les expériences et les variables dépendantes et indépendantes de toute la thèse.

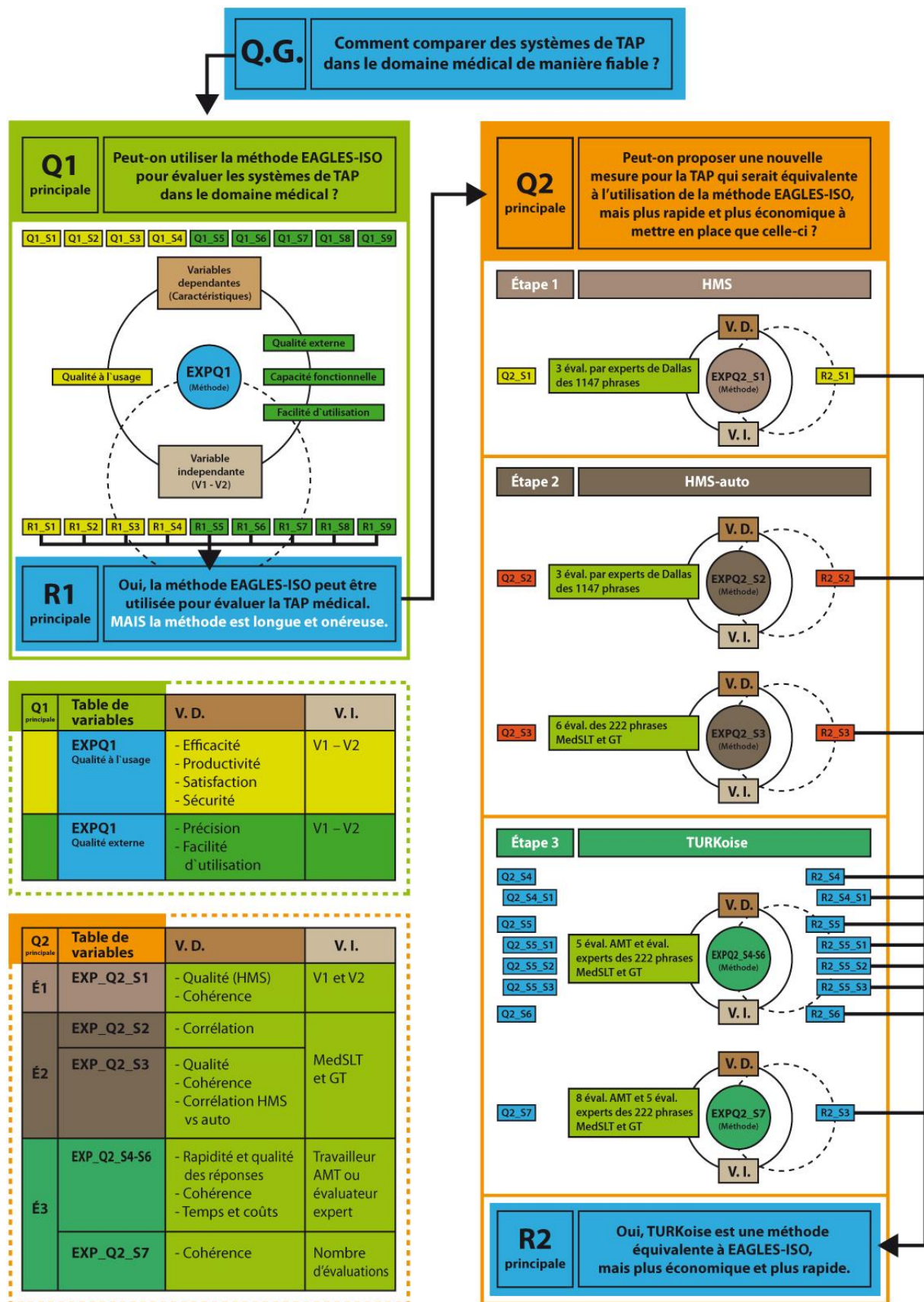


Figure 4.c: Schéma récapitulatif de notre méthode de recherche

Dans le chapitre suivant, nous présenterons les résultats obtenus pour **Q1_principale**, pour laquelle nous avons soumis les deux versions de *MedSLT* à la méthode d'évaluation EAGLES-ISO.

5 Résultats de Q1_principale : évaluation EAGLES-ISO

Dans ce chapitre, nous donnerons les résultats concernant notre première question de recherche principale : *Peut-on évaluer les systèmes de TAP dans le domaine médical avec la méthode EAGLES-ISO ? (Q1_principale)*.

Pour répondre à cette question, nous avons appliqué la méthode décrite dans la section 4.1 afin de comparer les deux versions V1 et V2 de *MedSLT*. Celle-ci repose fondamentalement sur la définition d'un modèle de qualité adapté au contexte d'utilisation futur des logiciels comparés.

Le modèle de qualité comporte un certain nombre de caractéristiques de qualité à l'usage et de caractéristiques de qualité externe. Du choix de ces caractéristiques découlent toutes les variables dépendantes de l'expérience **EXPQ1** que nous avons menée pour répondre à **Q1_principale**. Pour rappel, nous reprenons le tableau 4.3 de la section 4.1.1.2, ci-dessous :

N° de question	Sous-questions de recherche	Variables dépendantes (Caractéristiques)	Variable indépendante
Q1_principale	Peut-on utiliser la méthode EAGLES-ISO pour évaluer les systèmes de TAP dans le domaine médical ?		V1 – V2
Qualité à l'usage			
Q1_S1	L'utilisation du système permet-elle au médecin de poser le bon diagnostic (=tâche accomplie) pour son patient ? <i>Cette question est reformulée de la manière suivante :</i> Quel est le taux d'accomplissement de la tâche ?	Efficacité	V1 – V2
Q1_S2	Quel est le temps mis pour accomplir la tâche effectuée à l'aide du logiciel ?	Productivité	V1 – V2
Q1_S3	Quelle version (V1 ou V2) donne la meilleure satisfaction aux utilisateurs ?	Satisfaction	V1 – V2
Q1_S4	L'un des deux systèmes évalués représente-t-il un plus grand danger ? <i>Cette question est reformulée ainsi :</i> L'un des deux systèmes évalués génère-t-il plus d'erreurs d'interaction que l'autre ?	Sécurité	V1 – V2

Qualité externe			
Capacité fonctionnelle	Quel système (V1 ou V2) est plus précis ?	Précision	V1 – V2
Q1_S5	Quel système est plus précis au niveau de la RV ?	Qualité de la RV	V1 – V2
Q1_S6	Quel système est plus précis au niveau de la rétro-translation ?	Qualité de la rétro-translation	V1 – V2
Q1_S7	Quel système est plus précis au niveau de la TA ?	Qualité de la TA (humaine et automatique)	V1 – V2
Facilité d'utilisation	Quel système (V1 ou V2) est plus facile à utiliser ?		
Q1_S8	Le fonctionnement du système est plus facile à comprendre pour V1 ou V2 ?	Facilité de compréhension	V1 – V2
Q1_S9	L'apprentissage du système est plus facile avec V1 ou V2 ?	Facilité d'apprentissage	V1 – V2

Tableau 5.1 : Reprise du tableau 4.3 du modèle de qualité et des variables d'EXPQ1

Dans le Tableau 5.1, chaque question de recherche et caractéristique évaluée est reprise. L'organisation du présent chapitre suit notre modèle de qualité. Ainsi, nous allons donner les résultats obtenus en commençant par détailler ceux concernant la qualité à l'usage (section 5.1) avant de décrire les résultats obtenus pour la qualité externe (capacité fonctionnelle, section 5.2, et facilité d'utilisation, section 5.3). Chacune de ces sections sera terminée par un tableau récapitulatif reprenant les principaux résultats. Nous clorons ce chapitre par une discussion sur la méthode EAGLES-ISO.

5.1 Qualité à l'usage

Pour rappel, les caractéristiques de qualité à l'usage retenues à la section 4.1 étaient : efficacité, productivité, satisfaction et sécurité. Nous exposerons les résultats pour chaque caractéristique de qualité retenue, avant de reprendre tous les résultats dans un tableau récapitulatif, conformément à la méthode décrite au chapitre 3 et à la section 4.1. Pour faciliter la lecture, nous reprenons le tableau d'interprétation des résultats pour la qualité à l'usage ci-dessous :

Caractéristiques et sous-caractéristiques (Variables dépendantes)	Evaluation effectuée (Chapitre 3)	Mesures choisies	Interprétation de la mesure	Points
Qualité à l'usage				
Efficacité (section 5.1.1)	Accomplissement de la tâche	% diagnostics justes	< 50% >50% < 75% > 75%	-10 0 10
	Qualité de l'information récoltée	Rappel	< 50% >50% < 75% > 75%	-10 0 10
		Précision	< 50% > 50% < 75% > 75%	-10 0 10
	Erreurs système	% d'erreurs	< 5% > 2% < 5% > 1% < 2%	-10 0 10
	Erreurs patients	% d'erreurs	< 10 % > 5 < 10% > 2,5 < 5%	-10 0 10
	Erreurs médecins	% d'erreurs	< 50 % >25 < 50% < 25	-10 0 10
Productivité (section 5.1.2)	Temps mis pour effectuer la tâche	Moyenne en minutes	> 20 > 10 < 20 < 10	-10 0 10
	Nombre d'interactions	Moyenne du nombre d'interactions par diagnostic	> 50 > 30 < 50 < 30	-10 0 10
Satisfaction (section 5.1.3)	Préférence (en %)	< 50% > 50%		-10 10
Sécurité (section 5.1.4)	% d'inter. réussies (AC)		< 30 > 30 < 70 > 70	-10 0 10
	% d'échecs des interactions (NA)		> 70 > 50 < 70 < 50	-10 0 10

Tableau 5.2 : Reprise du tableau d'interprétation de la section 4.1.2.5, Tableau 4.9

Nous commencerons par les résultats concernant l'efficacité des systèmes étudiés, pour suivre l'ordre indiqué dans le tableau ci-dessus.

5.1.1 Efficacité (Q1_S1)

Comme nous l'avons vu dans la section 4.1.1.1, l'accomplissement de la tâche est une façon classique d'évaluer l'efficacité d'un logiciel. Dans notre contexte d'utilisation, celle-ci se résume à la capacité de communiquer par le biais du système de TAP afin d'établir le bon diagnostic. Nous allons tout d'abord évaluer l'accomplissement de la tâche, puis nous évaluerons son efficacité en regard du rappel et de la précision de l'information récoltée avec

chaque système. Pour ce faire, nous mesurerons la proportion des objectifs atteints puisque c'est grâce aux informations exactes collectées que la tâche pourra être effectuée avec succès.

5.1.1.1 Accomplissement de la tâche

Dans cette section, nous considérons que la tâche a été accomplie avec succès lorsque le médecin parvient à poser le diagnostic correspondant aux symptômes décrits dans les scénarios prédéterminés. Le Tableau 5.3 nous montre que V2 obtient de meilleurs résultats (62.5%) que V1 (43.75%) pour l'accomplissement de la tâche.

	V1	V2
Aucun diagnostic	18,8	12,5
Diagnostic exact	43,8	62,5
Diagnostic inexact	37,5	25,0

Tableau 5.3 : Accomplissement de la tâche en pourcentage de diagnostics posés

Ce résultat pourrait indiquer qu'il est plus difficile pour un médecin d'obtenir les informations nécessaires d'un patient lorsque celui-ci est limité à des réponses de type oui/non ou elliptiques. Nous verrons avec l'analyse en profondeur des informations récoltées avec chaque version du système (section 5.1.1.2) si cette observation se confirme. Le fort pourcentage de cas où le dialogue n'a pas abouti à un diagnostic, autour de 15% en moyenne, est dû principalement au fait que certains médecins étaient réticents à poser un diagnostic sans avoir au préalable effectué un test de détection du streptocoque ou au moins un examen médical, même si les instructions allaient dans ce sens (Tableau 5.4).

	Méd. 1	Méd. 2	Médecin 3	Méd. 4	Méd. 5	Méd. 6	Méd. 7	Méd. 8
Type Expérience	résident 1 an	résident 1 an	expérimenté 7 ans	exp. inconnu	résident 3 ans	résident 2 ans	résident 2 ans	exp. 7 ans
strep 1	ps	1	1	1	1	1	ps	1
strep 2	0	0	1	ps	0	0	0	0
strep 3	0	0	ps	0	ps	ps	1	0
strep 4	1	1	1	0	1	0	1	1
virale 5	1	0	0	1	1	1	0	ps
virale 6	1	1	0	1	1	1	1	1
virale 7	1	ps	1	1	0	1	0	ps
virale 8	1	1	0	1	1	0	1	ps
Taux de succès	62,5	50	50	62,5	62,5	50	50	37,5 ¹²¹

Tableau 5.4 : Diagnostic par médecin et par système – vert = V1, bleu = V2, 1=correct, 0=incorrect, ps =pas sûr, strep = angine à streptocoque, virale = angine virale

¹²¹ Ce médecin, bien qu'expérimenté, ne voulait pas se prononcer sans examen clinique à l'appui.

En outre, l'explication fournie par le Prof. Glenn Flores est que certains médecins internes (avec en moyenne seulement une à trois années d'expérience) ne sont simplement pas capables de poser le bon diagnostic même avec les signes et symptômes classiques de l'angine à streptocoque¹²². Ainsi, nous avons décidé de confronter erreurs de diagnostic et informations recueillies pour voir d'où provenaient les erreurs. Il ressort que certaines erreurs de diagnostic s'expliquent par la différence de difficulté des scénarios prédéterminés avec notre conseiller et partenaire médical, le Prof. Glenn Flores. Dans deux des huit scénarios (scénarios 2 et 3, en rouge dans le Tableau 5.4), il était particulièrement difficile de trancher entre une angine à streptocoque et une angine virale au regard des symptômes décrits. Le bon diagnostic a été posé seulement pour 2 cas sur 16. Cependant, même en enlevant les deux scénarios difficiles, la différence entre V1 et V2 s'accroît encore en faveur de V2, comme le montre la figure suivante :

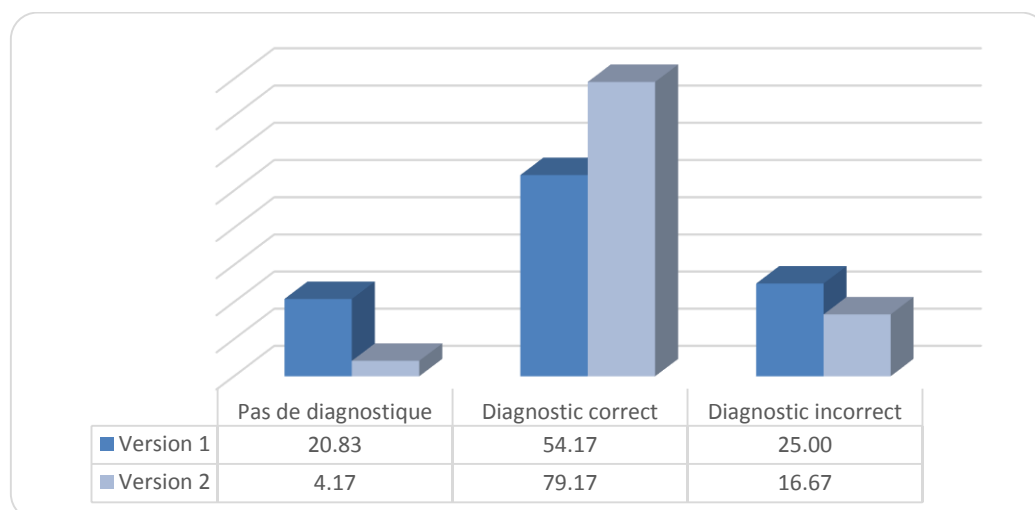


Figure 5.a : Accomplissement de la tâche en éliminant les deux scénarios difficiles.

Pourquoi la V2 obtient-elle de meilleurs résultats que la V1 dans l'accomplissement de la tâche alors que les difficultés de scénarios étaient les mêmes pour les deux versions ? Afin de compléter l'étude de l'efficacité des deux versions V1 et V2, nous avons procédé à une analyse complémentaire des données récoltées, qui nous permettra d'évaluer l'efficacité de la tâche au regard du nombre d'informations importantes récoltées grâce à l'utilisation de chaque système.

¹²² Citation du courrier électronique reçu le 09.06.2008 de Prof. Glenn Flores : “Unfortunately, I believe what we're seeing here is that some physicians (especially residents) are unable to make the diagnosis, even with some classic signs and symptoms”.

5.1.1.2 Efficacité de la tâche

Pendant chaque entrevue avec un patient, les médecins devaient noter les informations fournies par les patients afin de remplir une fiche d'information contenant les sept catégories suivantes : degré de la douleur, durée des symptômes, énumération des symptômes, contact avec une personne malade, visite médicale récente, test de diagnostic pratiqué récemment et allergie aux antibiotiques.

En nous basant sur les notes prises par les médecins pendant la simulation de diagnostic, nous avons constaté que certaines informations ont été difficiles à obtenir. Ainsi, les informations sur la présence d'éruption cutanée ou de maux de ventre n'ont souvent pas été récoltées, comme le montre l'exemple du Tableau 5.5 ci-dessous. Il s'agit des notes prises par le médecin 5 lors du diagnostic du patient 9 qui suivait le scénario 1.

Données cliniques de référence	Cas 1 - info du scénario 1	Médecin 5 - Patient 9
Degré de douleur	++	1
Durée du symptôme	Deux jours	1
Symptômes	Fièvre	1
	39- 39.5 C°	1
	Maux de tête (++)	1
	Ganglions gonflés (++)	1
	Éruption cutanée rouge (+)	0
	Maux de ventre	0
Contact avec personne malade	Oui, il y a deux jours	0
Récente visite médicale	non	1
Test de diagnostic pratiqué	non	1
Allergies aux antibiotiques	aucune	0

Tableau 5.5 : Extrait traduit du tableau de collecte d'information (0= info non-recueillie, 1= info recueillie)

Nous avons tout d'abord repris toutes les transcriptions des dialogues et rempli chaque case d'un tableau dont le Tableau 5.5 ci-dessus est un extrait traduit en français, puis nous avons choisi quatre mesures afin de quantifier cette étude. Tout d'abord, en nous basant sur les informations données par le Prof. Glenn Flores, nous avons compté combien d'informations sur le total préconisé ont bien été recueillies par les médecins grâce à l'utilisation de *MedSLT*. Cette mesure correspond au rappel des informations collectées. Nous avons donc attribué un point si l'information a pu être recueillie correctement et avons donné 0 point si l'information n'a pas été demandée par le médecin ou qu'elle n'a pas pu être recueillie. Nous nous basons sur le script original du scénario tel que fourni par le Prof. Glenn Flores. Il faut noter que les médecins qui ont participé à l'expérience ont bien sûr posé également d'autres questions que

les points importants retenus par notre spécialiste. Il faut savoir aussi que, pour trancher entre une angine à streptocoque ou virale, tous les 15-18 éléments d'informations (selon le scénario) n'étaient pas indispensables et que d'autres informations étaient implicites. Parmi ces informations, certaines pouvaient aider à guider le médecin vers la bonne réponse ou au contraire induire le médecin en erreur si le patient, ne sachant pas que répondre sur la base des réponses prévues dans les scénarios, fournissait des informations contradictoires au lieu de répondre « *no sé* » (je ne sais pas). Ainsi, dans cette étude nous nous sommes strictement limités aux informations des scénarios de référence, afin de garantir une meilleure cohérence dans les résultats. Pour compléter la mesure du rappel, nous avons ajouté la précision, où cette fois nous calculons le taux d'informations pertinentes (uniquement celles figurant dans les scénarios de départ) par rapport au nombre total de questions posées par le médecin lors de l'entrevue médicale. La différence entre le nombre d'informations récoltées et le nombre total de questions s'explique aussi par le fait que ces dialogues incluent aussi des actes de dialogues introductifs tels que « Qu'est-ce qui vous amène ? », « Bonjour, je m'appelle X. », etc. Ces tours de paroles sont hors couverture, ce qui explique qu'ils sont surtout présents dans les premières sessions et qu'ils disparaissent rapidement au fur et à mesure que le médecin apprend la couverture du système parce qu'ils rallongent inutilement le dialogue.

Pour compléter cette étude, nous avons différencié trois catégories qui expliquent pourquoi les informations n'ont pas pu être collectées. Nous avons ainsi calculé (1) le taux de mauvaises informations récoltées à cause d'une faille du système. Pour chaque scénario, nous avons compté les informations qui n'ont pas pu être récoltées à cause du système pour différentes raisons comme : phrase hors couverture, ou erreur de RV ou de TA qui a engendré l'abandon de l'utilisateur, calculé sur le nombre total d'informations récoltées.

Nous avons aussi comptabilisé (2) toutes les réponses erronées données par les patients et qui, suivant leur gravité, ont pu induire le médecin en erreur. Nous avons divisé ce nombre par le nombre total d'informations présentes dans le scénario afin d'obtenir le taux de précision de ces dialogues. On pourrait envisager une pondération selon la gravité de l'erreur, puisque certaines informations sont cruciales pour déterminer un diagnostic alors que d'autres sont plus secondaires, mais nous avons choisi de ne pas le faire pour faciliter le calcul.

Enfin, nous avons répertorié (3) les cas où le médecin n'a simplement pas posé la question prévue par le scénario. Ce problème ne se serait pas posé dans le cas de l'utilisation des systèmes avec de véritables patients. Cette réflexion touche à un des inconvénients d'utiliser des patients standardisés. Cependant, les avantages de cette méthode sont multiples, comme

nous l'avons vu dans la section 4.1.2.1. Nous avons regroupé ces trois mesures dans la Figure 5.b afin de mieux tirer des conclusions par rapport à la différence entre les deux versions.

Cette étude nous a permis tout d'abord de créer un tableau récapitulatif de toutes les informations recueillies (annexe 7) et de détecter clairement au moins quatre cas où l'erreur de diagnostic est très probablement due à des réponses erronées des patients sur des points cruciaux comme un test de détection de streptocoque positif alors que ce n'est pas précisé sur le scénario, ou encore la présence d'une forte fièvre alors que ce n'est pas prévu dans le scénario. Dans d'autres cas, on peut mettre en cause le manque de précision du scénario. Par exemple, comme lorsque l'intensité de la douleur est représentée par le signe + par rapport à ++, ce signe a été interprété de diverses manières par les patients variant de « *un poco de dolor* » (un peu de douleur) à « *fuerte* » (une douleur forte) par rapport à « *muy fuerte* » (une douleur très forte) lorsqu'il y avait ++.

À partir de ces informations détaillées, nous avons pu produire la Figure 5.b qui nous montre qu'au niveau du rappel et de la précision, les deux systèmes sont équivalents avec un rappel de 56% pour V1 et de 54,2% pour V2 et une précision de 39,6% pour les deux.

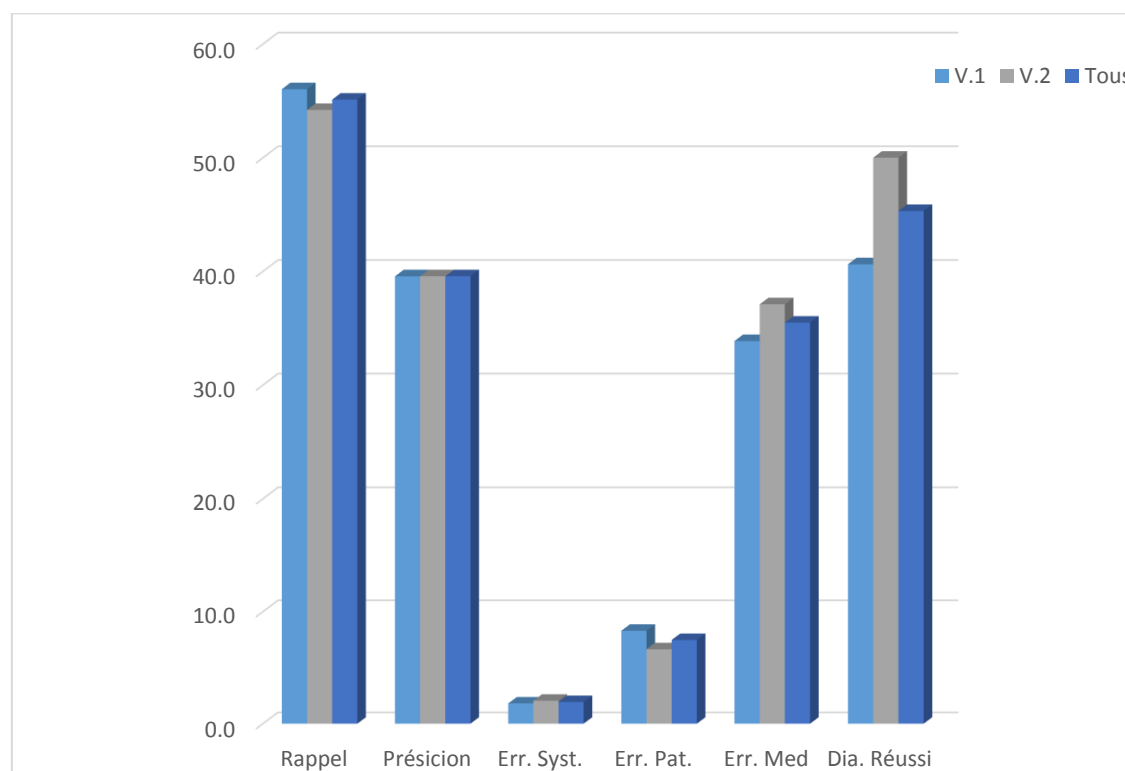


Figure 5.b : Récapitulation des mesures pour chaque version

Une différence apparaît au niveau du taux d'erreurs commises par les patients et les médecins. Il ressort, en effet, que les patients ont commis plus d'erreurs avec la version 1 (8,3 %

V1 vs 6,6 % V2, cf. colonne Err. Pat.) et que les médecins ont omis plus de questions avec la version 2 (V1 33,9 % et V2 37.1 %, cf. colonne Err.Med). Il est intéressant de noter que le taux d'erreurs système est sensiblement le même pour les deux versions (V1 : 1,8 et V2 : 2.1).

L'analyse du résultat de cette étude scénario par scénario (cf. Figure 5.c) montre que les scénarios 2 et 3 ont obtenu de nettement moins bons résultats. Pour le premier scénario, le rappel et la précision sont plus bas par rapport aux suivants. On retrouve la même baisse au scénario 5, que lorsque nous observerons le temps mis pour chaque scénario et le nombre d'interactions (Figure 5.d, section suivante). On note une progression tout au long des scénarios, cependant la moyenne n'est pas très élevée et revient à des niveaux proches du scénario 1 pour le rappel. Par contre, la précision s'est bien améliorée. Dans la version 2, elle est en quasi constante augmentation avec juste une chute pour le scénario 5. Cette tendance montre que les utilisateurs apprennent bien la couverture de la version 2 (voir approfondissement à la section 5.3.1). Pour la version 1, la progression de la précision est plus en dents de scie, et elle chute pour les scénarios 3 et 7. Le rappel fluctue davantage, mais de manière générale, ces informations semblent se recouper avec la Figure 5.d (section suivante) indiquant le temps mis en moyenne pour effectuer un diagnostic.

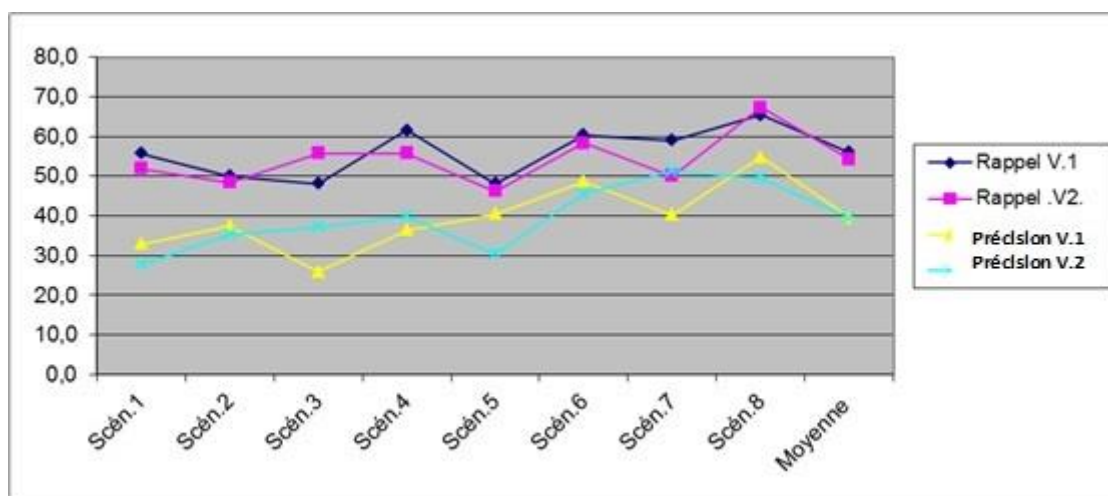


Figure 5.c : Rappel et précision par version et par scénario

En conclusion, dans cette section sur l'efficacité des systèmes et pour répondre à notre question **Q1_S1**, nous avons mené une étude approfondie de l'accomplissement de la tâche eu égard au diagnostic final mais aussi de la qualité de l'information recueillie. Cela nous a permis de constater que, si V2 ressort premier au niveau de l'accomplissement de la tâche, les performances des deux systèmes sont en fait très proches au niveau du rappel et de la précision des informations collectées. Pour ce qui est des différents types d'erreurs observées, les

différences apparaissent pour le taux d'erreurs du patient (plus élevé pour V1) et pour le taux d'erreurs du médecin (plus élevé pour V2). V1 et V2 présentent le même taux d'erreur système. Cependant, ces résultats n'expliquent pas pourquoi V2 obtient le plus fort taux d'accomplissement de la tâche. La seule observation en faveur de V2 semble être que les utilisateurs apprennent plus vite la couverture de ce système. Nous suivrons cette piste, en étudiant plus en détails dans la section 5.3.1 qui concerne la facilité d'apprentissage des systèmes étudiés.

Nous aborderons la productivité des deux versions en évaluant quelle proportion des objectifs de la tâche ont été atteints. Ces analyses complémentaires de nos données nous permettront de faire le parallèle entre l'accomplissement de la tâche et le temps mis ainsi que le nombre moyen d'interactions nécessaires pour effectuer le diagnostic de chaque scénario avec les deux versions.

5.1.2 Productivité : Temps nécessaire pour effectuer la tâche (Q1_S2)

La mesure du temps nous donne des informations pour la caractéristique de la productivité. Lorsqu'on étudie uniquement le temps moyen pour tous les scénarios, les résultats pour V1 et V2 sont proches, puisqu'il est de 7,88 minutes pour V1 et de 8,25 pour V2 (Tableau 5.6).

	Nombre d'interactions		Temps moyen en minutes	
	V1	V2	V1	V2
Scénario 1	41	48	14	11
Scénario 2	39	50	7	12
Scénario 3	36	35	6	10
Scénario 4	44	35	9	6
Scénario 5	28	50	6	9
Scénario 6	27	28	5	7
Scénario 7	41	25	10	5
Scénario 8	31	33	6	6
Moyenne	36	38	7,88	8,25

Tableau 5.6 : Temps et nombre d'interactions par scénario et par système

La Figure 5.d illustre bien que malgré la moyenne très proche entre les systèmes et les scénarios, de fortes variations entre les deux versions apparaissent pour certains scénarios (notamment 2, 5 et 7) s'agissant du nombre d'interactions requis. Nous essaierons d'expliquer d'où viennent ces différences, la raison n'est pas la difficulté du diagnostic puisque, dans cette liste, seul le scénario 2 fait partie des scénarios difficiles. Notons que le nombre moyen

d'interactions est très proche pour les deux systèmes avec respectivement 36 (V1) et 38 (V2). Cependant, de nettes différences apparaissent dans le haut de la Figure 5.d.

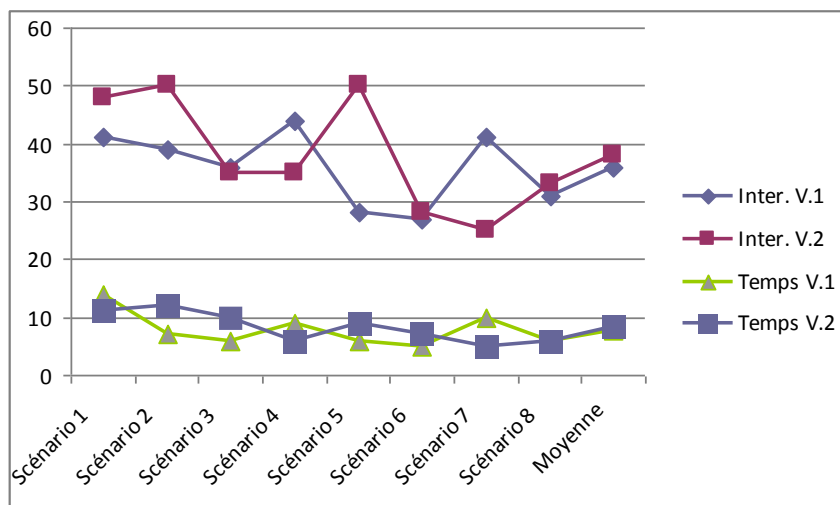


Figure 5.d : Comparaison entre temps et nombre d'interactions moyens pour chaque scénario avec V1 et V2

Pour le scénario 7, le diagnostic avec V2 a nécessité en moyenne moins d'interactions qu'avec V1. La différence s'explique par un incident technique qui a fait que le système s'est retrouvé sans courant et a coupé l'interview en plein milieu. Pour le scénario 5, c'est le contraire : V2 obtient un nombre supérieur d'interactions et un temps plus long. On constate, en reprenant le dialogue le plus long effectué avec ce scénario, que le patient a essayé 7 fois de répondre à la question sur la présence de fièvre en employant une phrase négative (non-couverte), 11 fois à la question « Depuis quand avez-vous de la fièvre » et 5 fois pour indiquer que dans son entourage souffrait de maux de gorge. Toutes ces tentatives de réponse de ce patient sont hors couverture, d'où le nombre élevé d'interactions. Le patient en question dans ce scénario persévère dans des réponses hors couverture et en général répond de manière erronée. Cependant, il se pourrait que ce ne soit pas anodin que ce patient ait éprouvé plus de difficultés avec la V2, puisque étant moins restreinte, elle fait oublier aux utilisateurs que les phrases négatives ne sont pas permises conformément aux instructions données au début de chaque session (section 4.1.2.2). Enfin, pour le même scénario, toujours avec la V2, deux autres dialogues ont été particulièrement longs, l'un d'eux parce que le médecin a posé plus de questions et que le patient a donné plusieurs réponses en dehors du scénario défini.

Le scénario 2 est l'un des deux scénarios les plus difficiles avec le scénario 3 pour lequel on ne constate pas une différence au niveau du nombre moyen d'interactions. Cependant, dans ces deux cas, le temps moyen est nettement supérieur avec V2, où de 7 minutes pour V1, nous

passons à 12 minutes pour V2 pour le scénario 2 et de 6 minutes avec V1 et 10 minutes pour V2 pour le scénario 3. Pourquoi la difficulté de ces deux scénarios ne ressort-elle pas avec la V1 ? Il est intéressant de noter qu'au niveau du nombre d'interactions, le scénario 3, pourtant classé difficile, a nécessité exactement le même nombre d'interactions avec V1 et V2, mais pour une durée différente. Ainsi, il semble que la difficulté des scénarios n'ait pas de répercussions directes sur la durée de l'entrevue, alors qu'elle a un net impact sur le taux d'accomplissement de la tâche.

Récapitulons les résultats obtenus pour la caractéristique de productivité afin de répondre à **Q1_S2** : les deux systèmes obtiennent des résultats pratiquement identiques avec respectivement 36 interactions pour V1 et 38 pour V2. Au niveau du temps moyen par scénario, V1 obtient 7,88 et V2 8,25 minutes. Les différences apparaissent uniquement lorsqu'on étudie les détails des scénarios.

Nous allons maintenant voir les résultats obtenus pour la caractéristique de la satisfaction des utilisateurs.

5.1.3 Satisfaction (Q1_S3)

La satisfaction des utilisateurs a pu être évaluée grâce aux questionnaires remplis par chaque participant à l'issue des sessions de l'expérience EXPQ1 décrite dans la section 4.1.2. Chacun a utilisé autant les deux versions du système avant de remplir un questionnaire de satisfaction, où il devait répondre à une série de questions. Pour quantifier la satisfaction des utilisateurs, nous avons procédé à une évaluation par questionnaire en nous fondant sur la littérature décrite dans la section 3.2.4.

Le questionnaire destiné aux patients était constitué de cinq parties (annexe 9), rédigées en espagnol. La première partie contenait les questions générales sur le participant (âge, sexe, pays d'origine et expérience préalable avec un système de TAP). La deuxième partie consistait à choisir entre les deux versions en retenant la déclaration adéquate : « Je préfère MedSLT V1 ou MedSLT V2 ». Cette dernière était précédée d'une brève description de chaque système pour rappel et suivie d'un espace libre pour étayer le choix effectué. Dans la troisième partie, nous abordions des questions concernant *MedSLT* en général. Les réponses à ces questions se trouvent dans l'annexe 10.

Le Tableau 5.7 reprend les résultats de la question « Quel système préférez-vous ? », adressée à toutes les catégories d'utilisateurs. Il montre qu'autant les médecins que les patients

ont clairement préféré la version moins restreinte de *MedSLT* (V2) : 68,75 % des patients et 62,5 % des médecins ont choisi la version V2.

	V1	V2	Aucune préférence
Patients	18,75	68,75	12,5
Médecins	12,5	62,5	25

Tableau 5.7 : Taux de préférence par système pour chaque type d'utilisateur

Selon les utilisateurs interrogés, la principale justification de ce choix est que V2 permet une communication plus libre entre les utilisateurs : le dialogue est plus facile, puisque tous les types de phrases sont permis. Ainsi, l'utilisateur n'a pas besoin de réfléchir à la façon de formuler les réponses. Remarquons cependant que, pour les médecins, la distinction entre les deux versions était plus difficile à faire (25% d'indécis), sans doute parce que, pour les médecins, la couverture des deux versions est la même. Le médecin est seulement concerné indirectement dans les cas où l'un des deux systèmes empêcherait le patient de répondre. Le seul utilisateur médecin qui s'est clairement prononcé en faveur de la version plus restreinte arguait que c'est sans doute plus facile d'utiliser la version restreinte avec des personnes « moins éduquées », parce que ce type de système réduirait le risque d'erreur. Du côté des patients, l'argument qui revient le plus souvent concerne la plus grande flexibilité de la V2. La réponse à **Q1_S3** est par conséquent que les utilisateurs sont clairement plus satisfaits de V2 en général. Cependant, ce plébiscite n'est pas confirmé par les autres réponses du questionnaire que nous verrons dans la section concernant la facilité d'utilisation des systèmes (section 5.3).

Afin de compléter l'analyse de la qualité à l'usage, nous évaluerons à présent la dernière caractéristique de ce type : la sécurité. Nous avons choisi d'étudier celle-ci sous l'angle de la fréquence des erreurs. Pour ce faire, nous avons procédé à une analyse détaillée de tous les tours de paroles collectés pendant l'expérience de Dallas.

5.1.4 Sécurité : étude de l'échec des interactions (Q1_S4)

Dans cette section, nous présentons l'analyse de tous les tours de paroles comme dans (Starlander, Bouillon et al. 2008) afin d'évaluer le degré de sécurité apporté par les deux versions étudiées. Nous tenons compte de toutes les interactions des utilisateurs avec le système afin de mieux comprendre la raison de l'échec d'une interaction. Notre but principal est de comparer le pourcentage d'erreurs effectuées avec chaque version. En effet, un système présentant un grand nombre d'erreurs d'interactions pourra être considéré comme peu sûr. L'étude des interactions nous renseignera aussi sur la fiabilité du système. Nous avons ainsi

décidé d'étudier cette caractéristique simultanément à la sécurité, puisque ces deux caractéristiques sont très fortement liées (cf. section 4.1.1.2).

Nous avons classé toutes les interactions en huit catégories différentes. Nous distinguons les phrases acceptées après reconnaissance (AC) de celles reprises de l'aide (AC_help). Parmi les phrases qui n'ont pas été acceptées par l'utilisateur, nous distinguons celles dont la cause est une mauvaise reconnaissance (NA_BR) de celles sans traduction (NA_NT). Ensuite, nous avons annoté les erreurs d'interactions : lorsqu'il s'agissait d'une phrase acceptée par erreur, nous l'avons annotée comme « AC_IE » et lorsque c'était une phrase qui aurait dû être acceptée parce qu'elle était correcte, nous l'avons marquée comme « NA_IE ». Enfin, la catégorie NA comprend les interactions non-acceptées pour une raison indéterminée. Nous avons également compté les phrases que les utilisateurs ont fait écouter plusieurs fois à leur interlocuteur (REP). Cette méthode est principalement utilisée avec la version restreinte V1 par les médecins (environ 2%) parce que cela leur permettait de redonner au système le contexte perdu d'une réponse elliptique. En effet, comme décrit dans la section 4.3.3 et dans (Bouillon, Rayner et al. 2007), le traitement de l'ellipse requiert la conservation du contexte de la phrase précédente afin de replacer les informations omises. Le Tableau 5.8 reprend les huit catégories différentes.

Catégorie	Num.	Explication
AC	1	Interaction acceptée
AC_help	2	Phrase issue de l'aide acceptée par l'utilisateur
AC_IE	3	Acceptée par erreur
REP		Répétition
NA	4	Interaction rejetée sans raison apparente
NA_BR	5	Rejetée pour cause de mauvaise RV
NA_NT	6	Rejetée pour cause d'absence de traduction
NA_IE	7	Rejetée par erreur
NA_help	8	Phrase issue de l'aide et rejetée

Tableau 5.8 : Huit catégories de classement des interactions

Le Tableau 5.9 résume les résultats obtenus lors de cette analyse, en faisant la distinction entre les versions de *MedSLT* mais aussi entre les patients et les médecins.

Utilisateur	Médecin	Médecin	Patient	Patient
Système	V1	V2	V1	V2
AC	42,83	42,68	59,10	59,13
AC_help	15,31	17,26	6,00	6,40
AC_IE	0,65	0,77	0,75	3,58
REP	2,44	1,54	0,19	0,00
Total AC	61,24	62,25	66,04	69,11
NA	2,93	1,85	1,69	0,56
NA_BR	29,15	31,28	25,89	25,61
NA_NT	3,09	3,08	2,81	1,13
NA_IE	2,93	0,92	1,50	0,75
NA_help	0,65	0,62	2,06	2,82
Total NA	38,76	37,75	33,96	30,89

Tableau 5.9 : Analyse des interactions des utilisateurs avec V1 et V2, en pourcentage

Pour l'heure, ce qui nous intéresse particulièrement dans ce tableau concerne le taux de phrases qui n'ont pas été acceptées (NA, NA_BT, NA_NT ou NA_help) à cause d'une erreur. Au total, on compte moins de phrases rejetées pour V2 que ce soit par les médecins (37,75% V2 contre 38,76 V1) ou par les patients (30,89% V2 contre 33,96 V1). Si l'on étudie l'origine des erreurs, on constate qu'en ce qui concerne les patients, le taux de phrases rejetées en raison d'une mauvaise reconnaissance (NA_BR) est équivalent entre les deux versions, avec 25,89% pour V1 et 25,61% pour V2. Pour le médecin, ce taux est légèrement plus élevé pour V2 (31,28%) que pour V1 (29,15%).

Les différences sont plus nettes pour les phrases rejetées à cause d'une erreur de traduction (NA_NT). V2 ne compte que 1,13% de phrases non-acceptées pour cause d'absence de traduction alors que, pour V1, ce chiffre atteint 2,81%. Cette différence s'explique par la dépendance de V1 à la résolution d'ellipse. En effet, pour pouvoir bien traduire les phrases elliptiques, le contexte de la phrase originale est indispensable pour V1. Dans certains cas, il n'était plus accessible au système, d'où le nombre important de phrases REP pour V1. Inversement, pour V2, il n'est pratiquement jamais fait recours à la répétition de phrase (0,19% pour les médecins et 0% pour les patients).

Au vu des résultats ci-dessus, la réponse à **Q1_S4** est que V1 génère légèrement plus d'erreurs d'interactions en général qui poussent les utilisateurs à rejeter les phrases. Par conséquent, le pourcentage de phrases acceptées est plus élevé pour V2 (pour les patients 69,11% V2 contre 62,25 V1, pour les médecins, 62,25% V2 contre 61,24 V1). La différence n'est pas assez significative pour être répercutée dans le score final, comme nous le verrons à la section 5.1.5.

La seule grande différence entre V1 et V2 concerne le taux d'interactions acceptées par erreur (AC_IE), qui est particulièrement élevé pour V2, puisque les patients atteignent 3.58% d'erreurs alors que pour V1 ce chiffre reste en dessous de 1%. La plupart de ces erreurs d'interaction proviennent du non-respect du langage contrôlé, qui excluait les formes négatives dans les phrases entières puisque celles-ci n'étaient pas couvertes par le système. Par exemple, avec V2, pour exprimer la négation, il fallait simplement utiliser « *no* » au lieu d'une phrase entière comme « *no tengo* » (« je n'ai pas »). Ainsi, les phrases négatives engendraient systématiquement des erreurs de reconnaissance ou de traduction. Les exemples ci-dessous illustrent comment un utilisateur, malgré la rétro-traduction correcte, pouvait facilement être induit en erreur et être conduit à commettre des erreurs d'interaction en acceptant des phrases mal reconnues qui, à la fin de la chaîne, produisent des traductions dangereuses. Ce problème ne concernait pas la version V1, puisque les phrases entières n'y sont pas autorisées.

	Exemple 1	Exemple 2
Question	Do you have a cough → ¿Tiene tos?	Does it hurt? → Tiene dolor / le duele
Transcription	no, no tengo tos (non, je n'ai pas de toux)	no muy fuerte (non, pas fort)
Reconnaissance	no tengo tos (non, j'ai de la toux)	no muy fuerte (non, très forte)
Rétro-traduction	tengo tos (j'ai de la toux)	el dolor es muy fuerte (la douleur est très forte)
Traduction	No, I cough (non, je tousse)	The pain is very severe (la douleur est très forte)

Tableau 5.10 : Exemple de formes négatives mal reconnues, produisant potentiellement des erreurs d'interaction

Dans l'exemple 1 du Tableau 5.10, le patient a dit la phrase non permise « *no, no tengo tos* ». En outre, ces exemples illustrent bien l'utilité de l'affichage de la rétro-traduction pour permettre aux utilisateurs de se prémunir contre d'éventuelles erreurs de traitement du système. Rappelons en effet que, dans le système *MedSLT*, nous présentons à l'utilisateur la rétro-traduction au lieu de lui donner le résultat de la reconnaissance vocale brute. L'avantage de la rétro-traduction est que le résultat final est plus proche du résultat réel. En effet, souvent le produit de la RV peut contenir des erreurs et néanmoins déboucher sur de bons résultats. Nous étudierons la rétro-traduction dans la partie consacrée à la qualité externe puisqu'elle peut être considérée comme une sous-tâche de *MedSLT* au même titre que la RV et la TA.

Dans la section suivante, nous récapitulerons les principaux résultats pour la qualité à l'usage.

5.1.5 Bilan et tableau récapitulatif des résultats de la qualité à l'usage

Le bilan de l'évaluation de la qualité à l'usage est le suivant : l'évaluation de l'**efficacité** nous permet de conclure que le taux d'accomplissement de la tâche est plus élevé pour V2 (62,5%) que pour la V1 (43.75%). Au niveau du rappel et de la précision des informations récoltées, les deux systèmes sont équivalents avec un rappel de 56% pour V1 légèrement plus élevé que V2 (de 54,2%) et une précision de 39,6% pour les deux.

Par contre, au niveau de la **productivité**, c'est-à-dire du temps mis pour effectuer le diagnostic et du nombre moyen d'interactions nécessaires, les deux versions arrivent au coude à coude avec un temps moyen respectif de 7,88 minutes pour V1 et de 8,25 minutes pour V2 et un nombre moyen d'interactions pour tous les scénarios de 36 pour V1 et 38 pour V2. Nous avons aussi analysé, au long de cette section, les détails des variations entre les scénarios.

Pour l'évaluation de la **satisfaction**, V2 arrive clairement en tête avec 68,7% de patients et 62,5% des médecins qui classent celle-ci comme étant leur version préférée. Les participants à l'expérience mettent en avant que V2 permet de communiquer plus librement.

Dans la dernière partie de cette section, nous avons évalué la **sécurité** des deux systèmes en analysant en détail les interactions afin de déceler si l'une ou l'autre version était plus sûre. L'analyse portait sur le nombre d'erreurs et de phrases acceptées par les utilisateurs, et c'est encore une fois la version V2 qui remporte le plus haut taux de phrases acceptées avec, pour V1, 62.3% pour les médecins et 66,04% pour les patients et pour la V2 62.24% pour les médecins et 69.11% pour les patients. Cependant, c'est aussi avec la V2 que les utilisateurs ont commis le plus d'erreurs d'interaction.

Nous reprenons les principaux résultats dans notre tableau de correspondance entre caractéristiques (variables), mesures et résultats, comme nous l'avons décrit dans le Tableau 4.9 de la section 4.1.3. Nous avons reporté les principaux résultats obtenus dans le Tableau 5.11 ci-dessous et interprété ceux-ci selon l'échelle prévue afin d'obtenir un score final unique pour chaque système.

Objet de l'évaluation	Mesures choisies et type	Résultats bruts		Interprétation Score final	
		V1	V2	V1	V2
Satisfaction (section 5.1.3)					
Préférence (en %)	Patients	18,8 %	68,8 %	-10	10
(Tableau 5.7)	Médecins	12,5 %	62,5 %	-10	10
Efficacité (section 5.1.1)					
Accomplissement de la tâche (Tableau 5.3)	% diagnostics correctes	43,8 %	62,5 %	-10	0
Qualité information récoltée	Rappel	56,0 %	54,2%	0	0
	Précision	39,6 %	39,6 %	-10	-10
Erreurs systèmes	% d'erreurs	1,8 %	2,1 %	10	0
Erreurs patients	% d'erreurs	8,3 %	6,6 %	0	0
Erreurs médecins	% d'erreurs	33,9 %	37,1 %	0	0
Productivité (section 5.1.2)					
Temps mis pour effectuer la tâche	Moyenne en minutes	7,88	8.25	10	10
Nombre d'interactions	Moyenne du nombre d'interactions par diagnostic	35	38	0	0
Sécurité (Tableau 5.9)					
% d'échecs des interact. (NA)	Médecins	38,8 %	37,8 %	10	10
	Patients	34 %	30,9 %	10	10
TOTAL Qualité à l'usage				0	40

Tableau 5.11: Principaux résultats pour la qualité à l'usage et interprétation

Ainsi, pour la qualité à l'usage, le score obtenu pour la version 1 est de 0, alors que la version 2 obtient 40 points au total. Cette différence de score provient principalement d'une meilleure **efficacité**, mesurée au moyen du taux d'accomplissement de la tâche (62,5% pour V2 par rapport à 43,8% pour V1), mais aussi du fait que V2 est plébiscité par les utilisateurs dans le questionnaire de **satisfaction**. Au niveau de la **productivité** et de la **sécurité**, les deux systèmes obtiennent des scores similaires. Il est à noter qu'aucune des deux versions n'obtient de points pour la **sécurité** selon les seuils que nous avons établis à la section 4.1.3. Nous poursuivrons à présent notre évaluation avec la qualité externe.

5.2 Qualité externe : capacité fonctionnelle de la RV, la rétro-translation et la TA

Dans cette section, nous traiterons de la capacité fonctionnelle, et plus particulièrement de la sous-caractéristique de la **précision** (Figure 5.e). La facilité d'utilisation fera l'objet de la section 5.3.

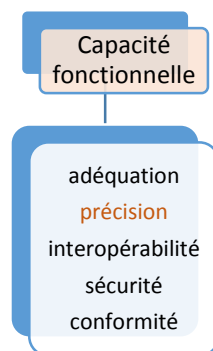


Figure 5.e : Première caractéristique externe retenue - capacité fonctionnelle

Nous présenterons les résultats de la précision en les divisant par rapport à la sous-tâche à laquelle elle se rapporte. Pour rappel, nous reprenons ci-dessous la partie du tableau de notre modèle de qualité qui concerne la sous-caractéristique de la précision.

Qualité externe : capacité fonctionnelle				
Précision				
Qualité de la RV (Jurafsky et Martin 2009)	WER sur phrases couvertes	-> environ 15% ok	> 15% > 10 % < 15% < 10%	-10 0 10
	SER sur phrases couvertes	De 15% à 30% -> utilisable	> 30% > 15% < 30% < 15%	-10 0 10
	% de phrases hors couverture	Au-delà de 30% -> handicapant	> 30% > 15% < 30% < 15%	-10 0 10
	SemER	%	> 30% > 15% < 30% < 15%	-10 0 10
Qualité de la rétro- traduction (Section 5.4.2) (Landis et Koch 1977)	Comparaison (binaire)	Binaire Accepté /rejeté	< 70% acceptés > 70% < 80% > 80%	-10 0 10
	Évaluateurs Dallas	Kappa Fleiss	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
		% d'accord	< 50% > 50 < 80% > 80%	-10 0 10
	Évaluateurs Trad.	Kappa Fleiss	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
		% d'accord	< 50% > 60% > 80%	-10 0 10

TA - éval. humaine	Fluidité échelle de 1 à 5 moyenne	Moyenne de 1-5	1-2 = mauvais 3 = neutre 4-5= bon	-10 0 10
(Landis et Koch 1977)	KAPPA Light	Entre 0 et 1	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
(Cicchetti, 1994)	ICC	Entre 0 et 1	< 0,70 (faible) > 0,70 < 0,79 (modéré) > 0,80 < 0,89 (fort)	-10 0 10
	Adéquation échelle de 1 à 5 moyenne	Moyenne de 1-5	1-2 = mauvais 3 = neutre 4-5= bon	-10 0 10
	KAPPA	Entre 0 et 1	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
	ICC	Entre 0 et 1	< 0,20 (très faible) > 0,20 < 0,40 (faible) > 0,40 < 0,60 (modéré) > 0,60 (fort)	-20 -10 0 10
Evaluation automatique	BLEU	Entre 0 et 1	< 0,40 < 0,60 > 0,60 < 0,80 > 0,80	-10 0 10

Tableau 5.12 : Évaluation de la capacité fonctionnelle (extrait du Tableau 4.9)

Nous commencerons cette étude par l'évaluation de la précision de la RV (section 5.2.1), puis nous évaluerons la rétro-traduction (section 5.2.2), celle-ci ayant un impact significatif sur l'acceptation ou le rejet des phrases. Enfin, dans la section 5.2.3, nous évaluerons la précision de la TA des phrases qui ont été acceptées par les utilisateurs, afin de comparer la qualité des deux systèmes V1 et V2.

5.2.1 Reconnaissance vocale (Q1_S5)

Comme nous l'avons mentionné dans la section 3.3.1, cela semble désormais acquis que l'évaluation d'un système de TAP par les mesures classiques de précision par le taux d'erreur au mot (WER) et à la phrase (SER) ne reflète pas réellement la qualité d'un système de TAP. C'est pourquoi nous mesurerons également la qualité de la reconnaissance par d'autres critères et notamment le taux d'erreur sémantique, en nous basant sur le jugement de l'utilisateur, qui, avant de faire traduire une phrase peut l'accepter par le biais de la rétro-traduction. À la lumière des résultats obtenus dans l'évaluation de cette sous-tâche, nous intégrerons les résultats obtenus pour la RV et ceux obtenus lors de l'évaluation de la rétro-traduction et de la TA dans le tableau récapitulatif de la section 5.2.4.

5.2.1.1 WER et SER

Nous avons calculé le WER et le SER selon la méthode décrite dans (Jurafsky et Martin, 2009) et dans la section 4.1 de ce travail. Nos calculs reposent sur une transcription de nos données orales. Le Tableau 5.13 fournit les résultats pour tout le corpus de l'expérience EXPQ1, tandis que le Tableau 5.14 donne le WER et SER uniquement pour les phrases en couverture (COV).

Selon le Tableau 5.13, la version moins restreinte (V2) obtient uniquement de meilleurs résultats pour le WER des utilisateurs médecins, avec un taux de 31,1 % contre 32,5 % pour V1. Tous les autres résultats sont meilleurs pour V1 que pour V2. On par conséquent conclure, qu'il y a moins d'erreurs de reconnaissance avec la version V1 pour tout le corpus sans distinction entre phrases couvertes et non couvertes.

	V1 - tout	V2- tout
WER médecins	32,5 %	31,1 %
WER patients	20,1 %	27,5 %
SER médecins	58,4 %	62,0 %
SER patients	31,1 %	37,6 %

Tableau 5.13 : Taux d'erreur au mot (WER) et à la phrase (SER) pour tout le corpus

Nous verrons si ce résultat se confirme lorsqu'on étudie uniquement les phrases couvertes par le système (COV). En effet, il ressort du Tableau 5.14 que la RV est meilleure avec V1 également pour les phrases en couverture. Il est à noter que de manière général les taux d'erreur sont moins élevées que dans le Tableau 5.13.

	V1 - COV	V2- COV
WER médecins	15,3 %	17,4 %
WER patients	10,0 %	10,4 %
SER médecins	40,7 %	46,6 %
SER patients	9,6 %	14,0%

Tableau 5.14 : Taux d'erreur au mot (WER) et à la phrase (SER) pour les phrases couvertes par les systèmes V1 et V2

Le Tableau 5.15 montre le taux de phrases hors couverture pour chaque système. Dans V1 et V2, il y a plus de phrases hors couverture pour les médecins que pour les patients. Au niveau des patients, il est intéressant de noter que le taux de phrases hors couverture est plus élevé pour V2 que pour V1 alors que la couverture de V2 est plus étendue. Par contre, pour les médecins, c'est l'inverse : 34, 2% des phrases sont hors couverture pour V1 par rapport à 29,8% pour V2.

	V1 en nb	V1 en %	V2 en nb	V2 en %
COV médecins	322/490	65,7 %	367/523	70,1 %
COV patients	363/488	74,3 %	343/481	71,3 %
HORSCOV médecins	168/490	34,2 %	156/523	29,8 %
HORSCOV patients	125/488	25,6 %	138/481	28,6 %

Tableau 5.15 : Pourcentage de phrases compris dans la couverture par système

La grande différence au niveau du WER entre patients et médecins, s'explique par la nature des phrases générées par chaque type d'utilisateur lors du dialogue entre médecin et patient. Pour les deux systèmes, les phrases n'excèdent jamais 17 mots avec une moyenne d'environ 6 mots chez les patients, et une grande majorité de phrases résolument courtes chez les patients.

	V1	V2
Phrases entières	16.1 %	45.0 %
Ellipses	52.8 %	25.8 %
dont "un poco"	14 %	13.2 %
dont "mucho"	7.2 %	5.9 %
Oui/ Non	31.1 %	29.2 %

Tableau 5.16 : Types de phrases employées par les patients dans V1 et V2.

Comme on peut le voir dans le Tableau 5.16, ci-dessus, et dans la Figure 5.f, ci-dessous, les patients génèrent environ 85% de phrases courtes dont 52.8% sont des ellipses et 31.1% des phrases oui/non, et malgré les instructions, on peut déplorer la présence de 16.1% de phrases entières dans la version restreinte du système (V1) qui n'est pas conçue pour les traiter.

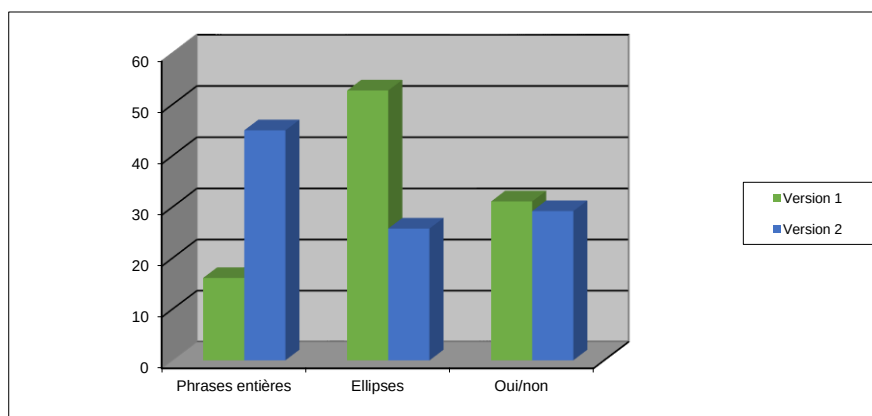


Figure 5.f : Types de phrases employées par les patients dans V1 et V2

Les réponses oui/non sont presque toujours bien reconnues. En revanche, nous avons mis les phrases « *un poco* » et « *mucho* » en exergue, parce que ce sont typiquement des mots difficiles à reconnaître, et on peut constater que leur fréquence est légèrement plus forte dans

la version restreinte (V1). De nombreuses occurrences de « *poco* » ou « *un poco* » ont été mal reconnues dans les deux systèmes. Cela dit, dans les données collectées avec V1, le nombre de réponses oui/non est plus élevé. Ce qui explique le SER très peu élevé pour V1 au niveau des patients (9,6%, voir Tableau 5.14).

S'agissant du système V1, les erreurs subsistantes pour les phrases en couverture sont principalement dues à des omissions, substitutions ou des erreurs de temps, comme illustré par le Tableau 5.17 ci-dessous.

Anglais - médecins	V1	V2
Omissions	Omissions des articles « a » ou « an »	Idem
Substitutions	« the » par « this »	Idem
	pluriel par un singulier : « antibiotic » au lieu de « antibiotics » ou « medecine » au lieu de « medecines ».	Idem
Erreurs de temps	« have you visited » au lieu de « have you been to visit », ou « have you taken » au lieu de « are you taking », ou encore « did you » au lieu de « do you »	Idem
Espagnol - patients	V1	V2
Omissions	Omissions de mots courts comme « en », « un », « no », « mucho », « dos », « a la », « hace » ;	Idem
Substitutions	de pluriel par le singulier : « meses » au lieu de « mes »	Idem
Erreurs de temps	Pas de verbes conjugués dans cette version.	« que ha tenido » au lieu de « que tiene », « tomo » au lieu de « tomé », « ha empezado » au lieu de « empenzó »

Tableau 5.17 : Types d'erreur de reconnaissance vocale par langue et par système.

Il est également intéressant de noter que, sur les phrases contenant des erreurs de reconnaissance au mot, il s'agit souvent d'omissions et de substitutions comme celles décrites dans le Tableau 5.17, c'est-à-dire la perte d'article ou de mots courts qui, dans la plupart de ces cas, ne posent pas de problème pour la suite du traitement avec *MedSLT*. Cette observation confirme le besoin d'évaluer la qualité de la RV non seulement sous l'angle purement quantitatif mais aussi au niveau de l'utilité des phrases reconnues. Nous comparerons maintenant ces résultats avec le taux d'erreur sémantique, qui, comme nous l'avons décrit dans la section 3.3.1, reflète sans doute mieux la véritable qualité de la RV, puisque cette mesure nous renseigne sur l'utilité de la RV.

5.2.1.2 Taux d'erreur sémantique (SemER)

Comme nous l'avons vu dans la section 2.4, le système *MedSLT* propose à l'utilisateur la rétro-traduction au lieu de lui donner le résultat de la reconnaissance vocale brute. L'avantage de la rétro-traduction est que le résultat final est plus proche du résultat réel. Nous avons donc un moyen de calculer le taux d'erreur sémantique de la RV, eu égard aux phrases acceptées directement par les utilisateurs. Il existe également une méthode semi-automatique appliquée dans (Bouillon, Rayner et al. 2011), qui consiste à vérifier la bonne formation de l'interlangue. Si celle-ci est correcte, la chaîne entière de traitement (reconnaissance vocale, rétro-traduction et traduction) a de fortes chances d'être correcte.

Les résultats ci-dessous se basent sur l'étude manuelle des interactions. Le Tableau 5.18 présente le pourcentage de phrases rejetées par les utilisateurs à cause d'une mauvaise reconnaissance pour V1 et V2.

	Restreint (V1)	Non-restreint (V2)
SemER (médecins)	29,15 %	31,28 %
SemER (patients)	25,89 %	25,61 %

Tableau 5.18 : Pourcentage de phrases rejetées à cause d'une mauvaise reconnaissance

Pour les médecins, le taux de SemER est légèrement meilleur avec V1 qu'avec V2. Par contre, étonnamment, pour les patients, ce taux est quasiment identique pour V1 (25,89%) et V2 (25,61%). Dans le système V1, les phrases rejetées étaient soit hors couverture (utilisation de phrases entières à la place d'ellipses), soit elles étaient constituées de mots courts difficiles à reconnaître.

La technologie de reconnaissance linguistique de notre système explique pourquoi, malgré un WER d'environ 30% pour les médecins (V1 et V2) et de 20% pour V1 et 27,5% pour V2 (voir Tableau 5.13) pour les patients, sur notre corpus entier contenant les phrases hors couverture et en couverture, le pourcentage de phrases acceptées est pour tous les utilisateurs et systèmes entre 70% et 75%. Cela montre bien que, la non-reconnaissance de certains mots n'a aucune conséquence sur le bon traitement de la phrase, ce que la rétro-traduction reflète, en réintroduisant notamment les articles qui étaient perdus lors de la RV (voir Tableau 5.17). D'autres erreurs de WER comme l'omission du "no" de la négation ou la différence de temps peuvent au contraire avoir des conséquences plus graves, comme nous le verrons dans la section 5.2.3 concernant la qualité de la traduction.

Le SemER a permis de faire un lien entre les mesures traditionnelles de la RV et l'utilité d'un système. Il existe aussi un lien très fort entre la SemER et la qualité de la traduction, puisque le SemER se base sur la révision de la rétro-traduction, dont le but est d'éliminer les phrases qui posent problème et qui déboucheraient sur des traductions erronées.

L'évaluation de la sous-tâche de la RV donne un léger avantage à V1 quant aux mesures classiques de la WER et du SER, cependant, nous avons vu que la mesure plus appropriée pour véritablement évaluer l'utilité de la RV est la SemER. Celle-ci dépend dans le cas de *MedSLT* de la qualité de la rétro-traduction puisque l'utilisateur du système se base sur celle-ci pour accepter ou rejeter une phrase, au lieu de se baser sur le résultat brut de la RV qui ne lui est même pas présenté dans l'interface. Au niveau du SemER, pour le patient, il n'y a pas de différence notable entre V1 et V2. Par contre, pour les médecins, c'est V1 qui arrive en tête.

Précision			V1	V2	V1	V2
RV (section 5.2.1)	WER sur phrases couvertes	Médecins	15,3 %	17,4 %	-10	-10
		Patients	10,0 %	10,4 %	0	0
	SER sur phrases couvertes	Médecins	40,7 %	46,6 %	-10	-10
		Patients	9,6 %	14,0 %	10	10
	% de phrases hors couverture	Médecins	34,2 %	29,8 %	-10	0
		Patients	25,6 %	28,6 %	0	0
SemER	Médecins	29,2 %	31,3 %	0	-10	
	Patients	25,9 %	25,6 %	0	0	
SOUS-TOTAL					-20	-20

Tableau 5.19 : Récapitulatif pour la précision de la RV

Par conséquent, comme résumé dans le Tableau 5.19, et pour répondre à **Q1_S5** : la précision de la RV est dans l'ensemble équivalente pour V1 et V2. Nous étudierons à présent la qualité de la rétro-traduction.

5.2.2 Rétro-traduction (Q1_S6)

Comme nous venons de le mentionner, la rétro-traduction (Rétrotra) permet de vérifier la qualité de la RV, mais elle va au-delà de cette étape puisque l'utilité de la phrase traitée en dépend complètement (section 2.4). En effet, elle est le moyen de fournir une sécurité accrue aux utilisateurs. Ainsi le rôle de la rétro-traduction est d'assurer la fiabilité du système. Sur le marché, certains systèmes de TAP dans le domaine médical ont choisi de ne pas présenter de rétro-traduction afin d'accélérer l'interaction, comme c'est le cas pour S-minds (Ehsani, Kimzey et al. 2008). Ce système commercial de TAP montre qu'il est possible de proposer un système performant de ce type dans le domaine médical sans pour autant proposer un « garde-fou » aux utilisateurs. Comment savoir dès lors que cette rétro-traduction apporte quelque chose au niveau

de l'utilité ou si, au contraire, elle fait perdre du temps aux utilisateurs qui doivent lire et comprendre en quoi la rétro-traduction présentée correspond à ce qu'ils ont dit ?

(Tsourakis, Lisowska et al. 2010) proposent d'évaluer l'effort demandé à l'utilisateur pour traiter (lire, comprendre et accepter) les rétro-traductions dans le cas de *MedSLT* et la production des paraphrases dans le cas de l'application du calendrier de gestion de réunions, décrite dans l'article susmentionné. L'objectif est, en résumé, d'évaluer la qualité de la rétro-traduction en se basant sur le temps de lecture et de réflexion avant d'accepter ou de rejeter une rétro-traduction jugée non-conforme à la phrase dite à l'origine en suivant un certain nombre de scénarios prédéfinis. Cette étude sur le traitement cognitif de la rétro-traduction pourrait contribuer à l'améliorer. En effet, si le rôle de la *Rétrotra* est de donner la possibilité à l'utilisateur de vérifier que la production du système est conforme à ce qu'il a dit, nous avons pu constater dans la section concernant la sécurité (section 5.1.4) que certaines erreurs de traduction (phrases rejetées pour cause de mauvaise reconnaissance ou d'absence de traduction ou phrases acceptées à tort) résultaient probablement de la qualité discutable de la rétro-traduction.

Pourquoi la rétro-traduction n'a-t-elle pas été efficace pour éviter des erreurs comme celles contenues dans le Tableau 5.10 (section 5.1.4) ? Dans ces cas, les phrases acceptées n'ont soit pas produit de traduction, soit donné une traduction fautive, alors que la rétro-traduction aurait dû mettre l'utilisateur en garde et lui faire rejeter la phrase à moins qu'il n'ait pas vu l'erreur en parcourant la rétro-traduction sur l'écran. Dans de nombreux autres cas, c'est au contraire une rétro-traduction a priori tout à fait correcte qui a été rejetée.

Nous avons demandé à un petit groupe d'hispanophones natifs de vérifier la qualité de la rétro-traduction espagnol-espagnol afin d'avoir des informations sur la qualité de ces rétro-traductions et de la compréhension de celles-ci, en évaluant l'équivalence entre la phrase d'origine et la rétro-traduction produite par le système. Lors de cette étude, nous pourrions également observer si les utilisateurs de *MedSLT* lors de l'expérience de Dallas (EXPQ1, décrite section 4.1.2.1) sont cohérents avec ce groupe d'évaluateurs qui a effectué cette tâche en dehors de l'EXPQ1 (hors expérience) et qui a eu tout loisir de lire tranquillement les phrases à évaluer. Il s'agit d'une tâche binaire : après avoir lu la question en anglais et la réponse en espagnol, l'évaluateur doit lire la rétro-traduction et l'accepter comme étant équivalente ou non, comme dans l'extrait du Tableau 5.20, ci-dessous :

Question - contexte	Transcription	Rétro-traduction	Acc.
how long have you had a sore throat?	me duele desde hace dos días	me duele desde hace dos días	Oui
what is the character of the pain?	fuerte	el dolor es severo	Non
do you have a fever?	sí yo he tenido fiebre	fiebre ha empezado siempre	Non
do you have any pain?	sí en la garganta	me duele la garganta	Oui

Tableau 5.20 : Extrait de l'évaluation binaire de la rétro-traduction

Comme illustré par le Tableau 5.20, la qualité de la rétro-traduction et celle de la traduction sont très étroitement liées, dans la mesure où la phrase reconnue fait l'objet d'une représentation sémantique source (espagnol), puis elle passe par l'interlangue et elle est retransformée en représentation cible (espagnol à nouveau dans le cas de la rétro-traduction) avant d'être générée en espagnol.

En analysant en détail le résultat de l'évaluation, il apparaît premièrement que les évaluateurs hors expérience confirment le rejet de certaines rétro-traductions, qui ont également été refusées par les utilisateurs lors de l'expérience, alors même que les développeurs les considèrent comme correctes (voir catégorie NA_IE, dans Tableau 5.18 ci-dessous) et, deuxièmement, que certains cas intéressants révèlent la possibilité d'améliorer et de corriger la rétro-traduction afin d'améliorer la qualité à l'usage du système.

Dans seulement 32 sur 241 des phrases, les deux groupes d'évaluateurs sont en désaccord. Voici dans le tableau suivant des exemples de désaccord soit entre les utilisateurs et les évaluateurs, soit entre le développeur et les évaluateurs.

Contexte : Question	Reconnaissance vocale brute	Rétro-traduction	Acc. En/Hors
V1			
are you allergic to any medicine	sí penicilina	la penicilina me da alergias	NA Non
for how long have you had your sore throat	desde hace los días	me duele la garganta desde hace días	NA_IE Non
have you seen a doctor recently	sí viernes pasado	he consultado un médico el viernes pasado	NA Oui
do you have a sore throat	sí un poco	me duele un poco la garganta	NA Oui
V2			
what is the type of pain	el dolor es muy fuerte	el dolor es muy fuerte	NA Oui
have you been to another doctor	No	No	NA Oui

Tableau 5.21 : Exemples de désaccord entre utilisateurs en EXPQ1 et hors EXPQ1, en bleu V1, en blanc V2.

Le bas du Tableau 5.22 montre que toutes les phrases qui avaient été acceptées par erreur (catégorie AC_IE) n'auraient en effet pas été acceptées, si l'évaluation avait eu lieu hors-ligne, après utilisation, comme c'est le cas pour l'évaluation actuelle de la rétro-translation. Cela indique donc que, sur le moment, lors d'une utilisation réelle des systèmes, une mauvaise lecture ou compréhension étaient possibles. Ces erreurs proviennent-elles d'un manque de clarté des rétro-translations ou est-ce simplement une question de manque de temps et de lecture globale trop rapide ?

Question du médecin pour contexte	Réponse du patient	Reconnaissance brute	Rétro-translation espagnol-espagnol
Phrases rejetées lors de l'utilisation (NA BR) de V1 et confirmées incorrectes par évaluation hors-ligne			
have you been in contact with anyone who has a sore throat	sí hace un mes	sí hace mes	he estado en contacto con alguien que tiene el dolor de garganta hace meses
for how long have you had your sore throat	desde hace menos de	desde hace mes	me duele la garganta desde hace meses
have you seen a doctor for your sore throat	no todavía no	no entiendo	no entiendo
what is your temperature	treinta y nueve grados	sí nueve grados	tengo nueve grados de fiebre
Phrases rejetées lors de l'utilisation (NA) de V2 et confirmées incorrectes par évaluation hors-ligne			
have you had fever?	mi cabeza se siente caliente	ví la persona recientemente	he estado recientemente en contacto con la persona
have you been in contact with anyone recently who has a sore throat	creo que no	lo tengo	presento este síntoma
for how long have you had your sore throat	no estoy segura	no desde hace hora	me duele la garganta desde hace horas
how long have you had the pain	desde hace tres o cuatro días	desde hace treinta y ocho días	me duele desde hace treinta y ocho días
have you had a fever	sí he tenido un poco	sí disminuye un poco	este síntoma disminuye un poco
Phrases rejetées doublement, hors couverture pour cause de négation			
how long have you had a stomachache	no me duele	no me duele	me duele
what is the character of the pain	no muy fuerte	no muy fuerte	el dolor es muy fuerte
are you taking any medicine for the pain	no he tomado medicina	no he tomado medicinas	he tomado medicina
have you had a fever	no no tengo fiebre		presento siempre este síntoma

Tableau 5.22 : Exemples de phrases AC_IE lors de l'utilisation réelle, rejetées par les évaluateurs hispanophones hors expérience.

Si nous analysons la source d'erreur dans le traitement total effectué par *MedSLT* tel qu'il est décrit plus haut dans cette section, dans de nombreux cas, la qualité de la production finale était compromise par l'absence ou la présence d'un article dès l'étape de la RV. Cependant,

comme la représentation sémantique (cf. section 2.4.1) ne représente pas les articles, c'est au moment de la génération que ceux-ci sont réintroduits. Afin d'améliorer la qualité de la rétro-traduction et aussi de la traduction vers une langue cible, il faudra s'assurer au moment de l'étape de la génération que les articles introduits correspondent bien à la phrase originale.

Il est intéressant d'observer qu'autant les évaluateurs hors EXPQ1 que les utilisateurs de l'expérience n'ont pas aimé la façon de retourner les phrases du Tableau 5.23 et auraient préféré une rétro-traduction plus littérale : « *estoy alérgico / alérgica a la penicilina* » (« Je suis allergique à la pénicilline »), mais nous avons exprès évité cette formulation à cause de la difficulté apportée par l'introduction de la marque du genre.

	Contexte : question	Reconnaissance brute	Rétro-traduction
V1	are you allergic to any antibiotic?	sí a la penicilina	la penicilina me da alergias
V2	are you allergic to any antibiotics?	estoy alérgica al penicilina	la penicilina me da alergias

Tableau 5.23 : Exemple de formulation impersonnelle, rejetés par les utilisateurs et évaluateurs

Le type d'erreur décrit dans le Tableau 5.24 est propre à la version V2 puisque qu'elle ne concerne que la version moins restreinte. Du fait que cette version couvre des phrases entières, elle traite des verbes conjugués :

	Contexte : question	Reconnaissance brute	Rétro-traduction
V2	what were the results?	el resultado fue positivo	el resultado ha sido positivo
	are you allergic to any medications?	Sí la amoxicilina	la amoxicilina me ha dado alergias

Tableau 5.24 : Exemple de décalage de temps pour V2

Dans ces cas, les rétro-traductions jugées mauvaises comportaient une erreur de temps.

Nous avons pu constater que, dans l'ensemble, à part quelques exemples décrits dans le Tableau 5.21, plus haut, les évaluateurs hors-ligne sont en accord avec les utilisateurs lors de l'expérience, ce qui tendrait à montrer que cette rétro-traduction est bien comprise et qu'elle apporte véritablement une sécurité aux utilisateurs. Cependant, comme l'impact de cette rétro-traduction sur la qualité à l'usage est important, nous avons aussi montré comment elle pourrait encore être améliorée. Enfin, comme il s'agit d'une évaluation humaine, nous présentons dans le Tableau 5.25, ci-dessous, le taux d'accord entre évaluateurs et le kappa (cf. section 4.1.2.5).

	Méthode	Résultat	Inférence/Sujets		Inférence/Sujets	
			ErrStd	95% I.C.	ErrStd	95% I.C.
V1	Kappa de Fleiss	0,55	0,06	0,43 à 0,67	0,08	0,40 à 0,70
V2	Kappa de Fleiss	0,73	0,04	0,64 à 0,81	0,05	0,63 à 0,83
V1	% d'accord	0,87	0,02	0,83 à 0,91	0,02	0,83 à 0,91
V2	% d'accord	0,89	0,02	0,85 à 0,92	0,02	0,85 à 0,933

Tableau 5.25 : Pourcentage d'accord entre évaluateurs et kappa pour V1 et V2 pour l'évaluation de la rétro-translation

Il ressort que, pour cette évaluation humaine binaire, l'accord entre les évaluateurs est bon. Il s'élève entre 0,87% et 0,89% d'accord pour les deux versions et un taux kappa de Fleiss de 0,66 pour tout le corpus évalué. Par ailleurs, nous pouvons noter une nette différence entre la version 1 et la version 2, qui n'apparaît pas avec le simple taux d'accord entre évaluateurs. En effet, le taux kappa de Fleiss est de 0,55 pour V1 et de 0,73 pour V2, ce qui revient sur l'échelle d'interprétation des résultats (cf. section 5.2.1) à un accord élevé pour V1, et encore plus fort pour V2. Une explication plausible serait que, pour V2, à cause des erreurs flagrantes de négation et de temps, l'opinion des évaluateurs est plus clairement unanime.

La différence entre les deux versions V1 et V2 n'est pas très forte dans cette analyse détaillée des interactions et de la rétro-translation. En effet, même si V2 obtient un pourcentage légèrement plus élevé (69,11%) de phrases acceptées que V1 (66,04%), ce résultat doit être nuancé par le fait que V2 compte aussi un plus grand nombre d'erreurs d'interaction, et si l'on retranche ces erreurs d'interactions, nous arrivons à des résultats sensiblement identiques avec 65,29% pour V1 et 65,43% pour V2. De l'analyse de l'origine de ces erreurs dans la qualité de la rétro-translation, il ressort que certaines erreurs spécifiques à V2 augmentent d'autant ces erreurs dans ces rétro-translations (temps, et négation). Cependant, il est intéressant de noter qu'au niveau de l'évaluation décrite ci-dessus, les deux groupes (utilisateurs lors de l'expérience et évaluateurs hors expérience) se rejoignent, tendant à prouver que la qualité de la rétro-translation en elle-même n'est pas en cause mais qu'une lecture trop rapide est l'explication la plus plausible des erreurs constatées dans les sections antérieures. Certaines améliorations pourront toutefois être apportées, mais il semble qu'il serait judicieux de maintenir la présentation de la rétro-translation aux utilisateurs. Néanmoins, il faudrait vérifier avec une étude similaire à celle effectuée par (Tsourakis, Lisowska et al. 2010) si la présentation de la rétro-translation représente une perte de temps trop conséquente par rapport à son bénéfice.

Pour répondre à **Q1_S6**, nous reprenons les principaux résultats dans Tableau 5.26, ci-dessous. Il en ressort qu’au niveau de la précision de la rétro-translation, V1 et V2 arrivent à égalité avec chacun 10 points.

Rétro-translation	Comparaison (binaire)	En pourcentage de phrases acceptées				
(Section 5.2.2)	Évaluateurs Dallas	En-expérience (proche SemER)	66,0 %	69,1 %	-10	-10
	Évaluateurs traducteurs	Hors-expérience	83,3 %	72,4 %	10	0
(Tableau 5.25)		Kappa Fleiss	0,55	0,73	0	10
		% d’accord	0,87	0,89	10	10
SOUS-TOTAL					10	10

Tableau 5.26 : Récapitulatif de la précision de la rétro-translation

5.2.3 Traduction automatique (Q1_S7)

L’évaluation de la traduction automatique est un domaine pour lequel de nombreuses sources existent (cf. section 3.3.2). Nous utiliserons par conséquent les mesures classiques pour cette sous-composante de la TAP. Nous commencerons par présenter les résultats de l’évaluation humaine de la TA pour finir avec l’évaluation automatique de la TA, achevant cette section par un bilan récapitulatif des résultats obtenus pour la qualité externe des sous-composantes de la TAP.

5.2.3.1 Évaluation humaine

Comme nous l’avons vu dans le chapitre 3, la précision de la TA est évaluée le plus classiquement par la fluidité (*fluency*) et l’adéquation (*adequacy*). Dans le tableau ci-dessous, nous reprenons les principaux résultats obtenus. De manière générale, nous pouvons constater que les scores obtenus pour V1 et V2 sont très proches au niveau de l’évaluation humaine sur une échelle de 1 à 5 (voir Tableau 5.27, ci-dessous).

	V. 1	Moy.DS	Kappa de Light ¹²³	ICC ¹²⁴	V. 2	Moy. DS	Kappa Light	ICC
Fluidité-ES (patients)	4,05	0,41	0.13	0.81	4.07	0.42	0.14	0.81 125
Fluidité-EN (médecins)	4,74	0,14	0.32	0.86	4.56	0.21	0.34	0.90
Adéquation EN-ES	4,82	0,04	0.32	0.78	4.83	0.06	0.33	0.81
Adéquation ES-EN	4,63	0,23	0.22	0.76	4.44	0.26	0.33	0.88

Tableau 5.27 : Fluidité et adéquation pour V1 et V2

En ce qui concerne la fluidité, la moyenne atteint respectivement 4,05 pour V1 et 4,07 pour V2 pour l’espagnol (patients) et 4,74 pour V1 et 4.56 pour V2 pour l’anglais. Nous obtenons ainsi une très faible différence, non significative, qui par ailleurs ne sera pas répercutée dans nos scores finaux (voir tableau récapitulatif section 5.2.4), la différence étant trop faible.

Le Tableau 5.27 montre que l’adéquation des traductions anglais vers l’espagnol obtient respectivement les scores 4,82 pour V1 et 4,83 pour V2. Pour les traductions espagnol vers l’anglais, ces scores s’élèvent à 4,63 pour V1 et 4,44 pour V2. Concernant ce dernier score, la différence entre les deux versions est un peu plus marquée avec +0,19 de différence en faveur de V1.

La déviation standard (DS) dénote de la variabilité entre les scores donnés par chaque évaluateur pour chaque phrase évaluée, cependant, ici encore, nous n’observons pas de différence notable entre V1 et V2.

Le **coefficient de corrélation intraclasse** (ICC, section 3.3.2.1.2) est mis en évidence dans le Tableau 5.27 parce qu’il est légèrement meilleur pour la V2, en cohérence avec les scores obtenus pour le kappa de Light (section 3.3.2.1.2) également légèrement plus élevé pour V2. Il semblerait donc que les évaluateurs sont légèrement plus cohérents pour V2 que pour V1. De manière générale pour la fluidité comme pour l’adéquation, il faut modérer ces résultats très bas, obtenus pour le kappa de Light par la sensibilité de kappa à la prévalence (Gwet 2012: p.44-45, p.49), comme expliqué à la section 3.3.2.1.2. Dans le Tableau 5.27, il apparaît très clairement que la valeur ICC est bien plus représentative du véritable accord entre nos évaluateurs. Pour notre objectif qui est d’évaluer la fluidité et l’adéquation, ainsi que pour le

¹²³ “Light’s generalized Kappa Coefficient : averaging all pairwise Cohen-Type Kappa coefficients.”, p.49 Gwet, K. L. (2012). Handbook of inter-rater reliability. Wiley Encyclopedia of Clinical Trials, Advanced Analytics LCC

¹²⁴ ICC par défaut : “Two-way mixed” : des évaluateurs différents sur un même ensemble de phrases à évaluer.

¹²⁵ Intervalle de confiance 95% (0.777<0.809<0.838).

chiffre de kappa, nous retenons les mesures de moyenne de fluidité et d'adéquation, le KAPPA et l'ICC. Ces mesures sont reportées dans notre tableau récapitulatif ci-dessous.

TA Évaluation humaine	Fluidité échelle de 1 à 5 (moyenne)	Médecins Patients	4.74	4.56	10	10
			4.05	4.07	10	10
	KAPPA Light	Médecins Patients	0.323	0.339	-10	-10
			0.132	0.140	-10	-10
	ICC	Médecins Patients	0.859	0.897	10	10
			0.811	0.809	10	10
	Adéquation échelle de 1 à 5 (moyenne)	Médecins Patients	4.82	4.83	10	10
			4.63	4.44	10	10
	KAPPA Light	Médecins Patients	0.322	0.333	-10	-10
			0.219	0.326	-10	-10
	ICC	Médecins Patients	0.778	0.808	0	10
			0.758	0.884	0	10
SOUS-TOTAL					20	40

Tableau 5.28 : Récapitulatif de la précision de la TA – évaluation humaine

Il ressort du Tableau 5.28 que V2 obtient un meilleur score final (40 points) que V1 (20 points) pour l'évaluation de la TA par les mesures humaines classiques (fluidité et adéquation). Cependant, avant de pouvoir répondre à **Q1_S7**, nous verrons les résultats de l'évaluation automatique de la TA.

5.2.3.2 Evaluation automatique

Parmi les mesures automatiques décrites dans le chapitre 3, nous avons retenu un certain nombre de mesures classiques comme le WER, le PER (Position Independent Error Rate), la précision, le rappel ainsi que la F-mesure et enfin les grands classiques BLEU et NIST. Afin de calculer ces résultats, nous avons demandé aux traducteurs de Dallas de nous fournir trois traductions de référence. Cependant, nous avons repris dans le tableau récapitulatif final uniquement les scores obtenus avec BLEU, qui est la mesure automatique la plus fréquemment utilisée.

Mesures	Ang-Esp V1	Ang-Esp V2	Esp-Ang V1	Esp-Ang V2
BLEU	0,55	0,31	0,54	0,42
NIST	5,74	4,58	5,91	5,39
WER	0,38	0,52	0,36	0,41
PER	0,34	0,47	0,31	0,36
Précision	0,73	0,65	0,77	0,76
Rappel	0,79	0,68	0,73	0,69
F-mesure	0,76	0,67	0,75	0,71

Tableau 5.29 : Résultats de l'évaluation automatique

Le Tableau 5.25 montre tout d'abord que pour les deux versions, les résultats sont faibles et très similaires. Cependant, pour toutes les mesures, V1 obtient de meilleurs résultats que V2. Ce qui rejoint les résultats de l'évaluation humaine décrite dans la section ci-dessus de la fluidité et de l'adéquation. Ils sont aussi cohérents avec ceux de la RV (WER, SER et SemER), décrits dans la section 5.2.1. Cependant, nous retenons également que ces scores sont assez bas, ce qui nous pousse à étudier dans le chapitre 6 pourquoi il en est ainsi. Nous reprenons tous les principaux résultats de la section 5.2 sur la qualité externe des sous-tâches de la TAP dans la section suivante.

5.2.4 Bilan et tableau récapitulatif : Précision de la RV, RetroTra, TA

Le Tableau 5.30 reprend les principaux résultats concernant la capacité fonctionnelle des systèmes au niveau de la précision. Nous analyserons les résultats par sous-tâche ci-dessous.

Qualité externe			Résultats		Scores	
Précision			V1	V2	V1	V2
RV (section 5.2.1)	WER sur phrases couvertes	Médecins	15,3 %	17,4 %	-10	-10
		Patients	10,0 %	10,4 %	0	0
	SER sur phrases couvertes	Médecins	40,7 %	46,6 %	-10	-10
		Patients	9,6 %	14,0 %	10	10
	% de phrases hors couverture	Médecins	34,2 %	29,8 %	-10	0
		Patients	25,6 %	28,6 %	0	0
	SemER	Médecins	29,2 %	31,3 %	0	-10
		Patients	25,9 %	25,6 %	0	0
SOUS-TOTAL					-20	-20
Rétro-translation	Comparaison (binaire)	En pourcentage de phrases acceptées				
(Section 5.2.2)	Évaluateurs Dallas	En-expérience (proche SemER)	66,0 %	69,1 %	-10	-10
	Évaluateurs traducteurs	Hors-expérience	83,3 %	72,4 %	10	0
(Tableau 5.25)		Kappa Fleiss	0,55	0,73	0	10
		% d'accord	0,87	0,89	10	10
SOUS-TOTAL					10	10

TA Évaluation humaine	Fluidité échelle de 1 à 5 (moyenne)	Médecins	4.74	4.56	10	10
		Patients	4.05	4.07	10	10
	KAPPA Light	Médecins	0.323	0.339	-10	-10
		Patients	0.132	0.140	-10	-10
	ICC	Médecins	0.859	0.897	10	10
		Patients	0.811	0.809	10	10
	Adéquation échelle de 1 à 5 (moyenne)	Médecins	4.82	4.83	10	10
		Patients	4.63	4.44	10	10
	KAPPA Light	Médecins	0.322	0.333	-10	-10
		Patients	0.219	0.326	-10	-10
	ICC	Médecins	0.778	0.808	0	10
		Patients	0.758	0.884	0	10
SOUS-TOTAL					20	40
TA Évaluation Automatique	BLEU	Médecins	0.55	0.31	0	-10
		Patients	0.54	0.42	0	0
SOUS-TOTAL					0	-10
TOTAL					10	20

Tableau 5.30 : Récapitulation des scores de précision de la RV, la RétroTra et la TA

On peut voir dans le Tableau 5.30 ci-dessus qu'au niveau de la RV, V1 et V2 obtiennent au final le même score négatif de -20. Le taux d'erreur au mot est assez proche pour V1 et V2, cependant, V1 arrive au seuil de 10% de WER, ce qui explique l'obtention de 10 points. V1 comptabilise plus de phrases hors couverture, ce qui explique les - 10 points obtenus. Au niveau du taux d'erreur sémantique, V2 perd de nouveau 10 points pour avoir franchi le seuil de 30 % de SemER.

Au niveau de la précision de la rétro-translation, V1 et V2 arrivent à ex-aequo avec chacun 10 points.

Si nous observons les scores obtenus pour la précision de la traduction automatique, nous pouvons constater que V1 et V2 sont très proches au niveau de la fluidité et de l'adéquation des TA. Ce qui les différencie, c'est le niveau d'accord entre évaluateurs qui est plus élevé pour V2 lorsqu'on considère le coefficient de corrélation intraclasse (ICC) qui est plus adapté à notre cas (cf. section 3.3.2.1.2.). Ainsi, V2 obtient un net avantage avec 40 points contre 20 points pour V1 pour la précision de la TA, évaluée manuellement. Pour l'évaluation automatique de la précision de la TA, le résultat est inversé puisque V1 obtient un score plus élevé pour BLEU qui se traduit par 0 point de notre échelle contre -10 pour V2.

Par conséquent, le bilan pour la caractéristique capacité fonctionnelle au niveau de la précision de la RV, rétro-translation et TA est favorable à V2, qui obtient un total de 20 points (selon l'échelle établie à la section 4.1.3) pour 10 points obtenus par V1.

5.3 Qualité externe : Facilité d'utilisation

Dans la section 5.2, nous avons traité de la caractéristique de capacité fonctionnelle et, plus particulièrement de la précision des résultats obtenus par les sous-tâches des deux systèmes à l'étude (V1 et V2). Cette caractéristique est très communément évaluée, souvent même exclusivement, au détriment d'autres caractéristiques. Dans cette nouvelle section, nous ajouterons une autre caractéristique de la qualité externe qui est importante à nos yeux et surtout dans notre contexte de TAP destiné au milieu médical : la facilité d'utilisation. Comme décrit dans la section 4.2.1.3, nous avons retenu uniquement les deux sous-caractéristiques de la facilité d'utilisation : la facilité de compréhension (section 5.3.1) et la facilité d'apprentissage (section 5.3.2).

5.3.1 Facilité de compréhension (Q1_S8)

La capacité de compréhension pour un système de TAP pourrait être mesurée par rapport à la compréhension du fonctionnement de l'interface. Cependant, cette étude empièterait sur la caractéristique d'opérabilité et le design d'interface, qui ne sont pas nos propos dans le cadre de cette thèse. Alors que, dans notre cas, la compréhension de la couverture du système est un point central puisque les deux versions étudiées varient en fonction de celle-ci. Par conséquent, pour répondre à **Q1_S8** (*Le fonctionnement du système est plus facile à comprendre pour V1 ou V2 ?*), nous comparerons la facilité de compréhension de V1 et V2 en étudiant le nombre de phrases hors couverture.

Le tableau ci-dessous indique le taux de phrase hors couverture (OOC) par système et par type d'utilisateur (médecin ou patient).

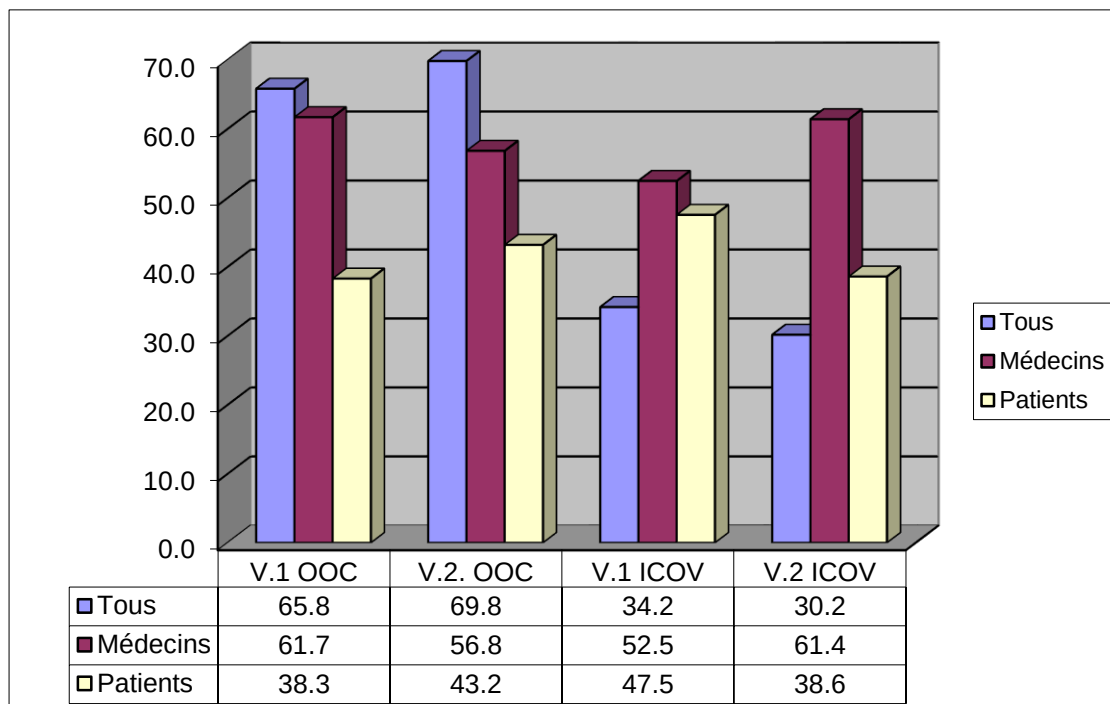


Tableau 5.31 : Taux de phrases hors couverture parmi les phrases mal reconnues

Les résultats du Tableau 5.31 indiquent que de manière générale (Tous), V2 compte moins de phrases hors couverture (69,8 % de phrases OOC) que V1 (65,8 % de phrases OOC). Ils nous montrent aussi que les médecins ont buté sur des phrases hors couverture bien plus souvent (61.7% pour V1 et 56.8% pour V2) que les patients (38.3% pour V1 et 43.2% pour V2). Un examen précis de ces phrases hors couverture nous a permis de constater que ce sont toujours plus ou moins les mêmes phrases qui sont hors couverture chez les médecins que ce soit pour V1 ou V2. Il s'agit principalement des phrases d'accueil (Tableau 5.32). Elles disparaissent vite à mesure que les médecins s'habituent à la couverture des systèmes, après avoir toutefois été répétées plusieurs fois. Le module d'aide oriente les médecins vers les phrases d'ouverture prévue par les systèmes : « où avez-vous mal ? ».

Phrases d'ouverture (hors couverture) tentées par les médecins
What brings you into the office today?
What brings you to the clinic today?
What are your symptoms?
How are you today?
How are you feeling?
How do you feel?
What is bothering you

Tableau 5.32 : Exemples de phrases hors couverture

Au niveau des patients, le pourcentage de phrases hors couverture est plus élevé pour V2 (43,2%) que pour V1 (38,3%). Ce qui paraît étonnant puisque la couverture de V2 est plus étendue que celle de V1.

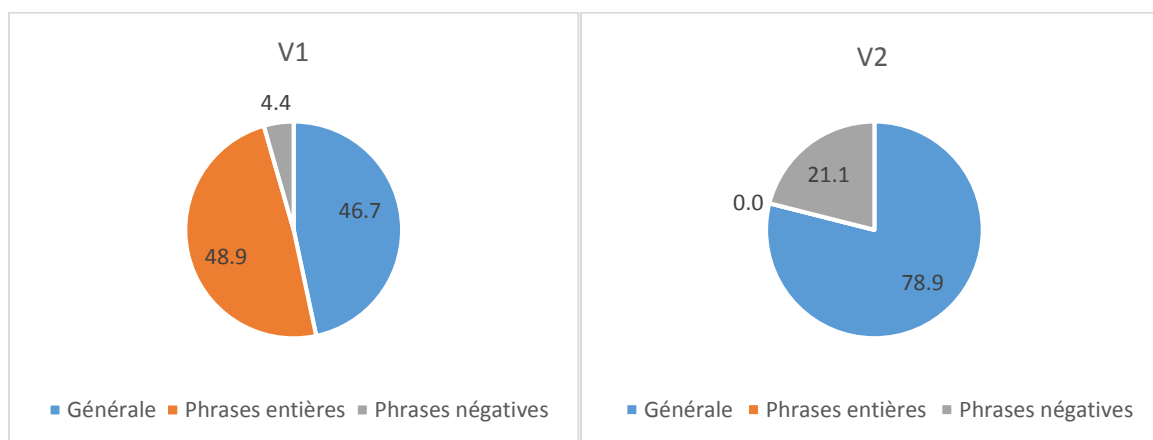


Figure 5.g : Type de phrase hors couverture par système pour les patients

Après classification par type de phrase, on constate dans la Figure 5.h que pour V2, les phrases négatives (exclues du LC indiqué, voir section 4.1.2.2) comptent pour une part importante des phrases hors couverture. Pour V1, la plus grande source de phrases hors couverture est lorsque les utilisateurs de V1 limitée aux ellipses utilisent des phrases entières, qui sont pourtant traitées que par V2. Ainsi, malgré le module d'aide certains utilisateurs s'obstinent à utiliser des phrases entières alors que le système ne comprenait que des phrases elliptiques, ce qui indique probablement que les utilisateurs en question n'avaient pas totalement compris la couverture des deux versions, malgré les instructions données au début de chaque session d'expériences et malgré le module d'aide (voir section suivante).

De ces résultats, nous pouvons conclure que la réponse à Q1_S8 est que le fonctionnement de V2 est plus facile à comprendre que celui de V1, puisque les participants à l'expérience de manière générale comprennent mieux la couverture de V2. Il ressort aussi de cette étude que les participants n'ont pas très bien suivi les instructions de LC concernant le type de phrases couvertes par les systèmes (section 4.1.2.2). Cependant, il est à noter que, lorsque l'utilisateur tente une phrase hors de la couverture du système, il y a presque toujours de l'aide pour réparer l'erreur. Du moins, lorsque il s'agit de phrases à reformuler légèrement, ce qui est le cas pour des phrases elliptiques par rapport à des phrases entières. En effet, lorsqu'une phrase est complètement hors couverture, par exemple si elle porte sur un autre domaine non couvert par le domaine étudié (maux de gorge), le module d'aide ne pourra pas proposer une phrase

alternative proche mais pourra guider l'utilisateur vers un nouveau départ. Le module d'aide est l'objet de la section suivante qui compare la facilité d'apprentissage de chaque système.

5.3.2 Facilité d'apprentissage (Q1_S9)

L'apprentissage du système consiste principalement en deux principaux éléments : le maniement de l'interface et l'apprentissage de la couverture du système. Ceux-ci ont tous deux fait l'objet d'instructions précises communiquées aux utilisateurs avant le début de l'expérience (annexe 4). Le maniement des deux systèmes V1 et V2 est le même. Il suffit de retenir qu'il faut cliquer sur le bouton « *Push to talk* » ou « *Pulsa y habla* » avant de parler, puis vérifier la rétro-traduction pour ensuite cliquer sur le bouton traduction si la rétro-traduction donne satisfaction à l'utilisateur. Pour évaluer la facilité d'apprentissage des deux systèmes et l'utilité des informations fournies (instructions sur la façon d'utiliser MedSLT V1/V2, et instructions de LC) pour aider à cet apprentissage, nous avons introduit quatre questions sur la facilité d'utilisation dans le questionnaire (Tableau 5.33) soumis à nos participants lors d'EPXQ1 (section 4.2.2).

1	Il est facile d'apprendre à utiliser MedSLT 1 .
2	Les instructions sur la façon d'utiliser MedSLT 1 sont claires.
3	Les instructions sur la façon de formuler les réponses sont très utiles.
4	La fenêtre d'aide est utile pour apprendre à formuler les réponses.

Tableau 5.33 : Extrait du questionnaire pour V1 concernant la facilité d'utilisation

La Figure 5.h donne les résultats pour les questions sur la facilité d'apprentissage données dans le Tableau 5.33. Les réponses des patients sont identiques (3.88) pour la V1 et la V2 pour la question 1 (Q1 : Il est facile d'apprendre à utiliser MedSLT V1 / V2).

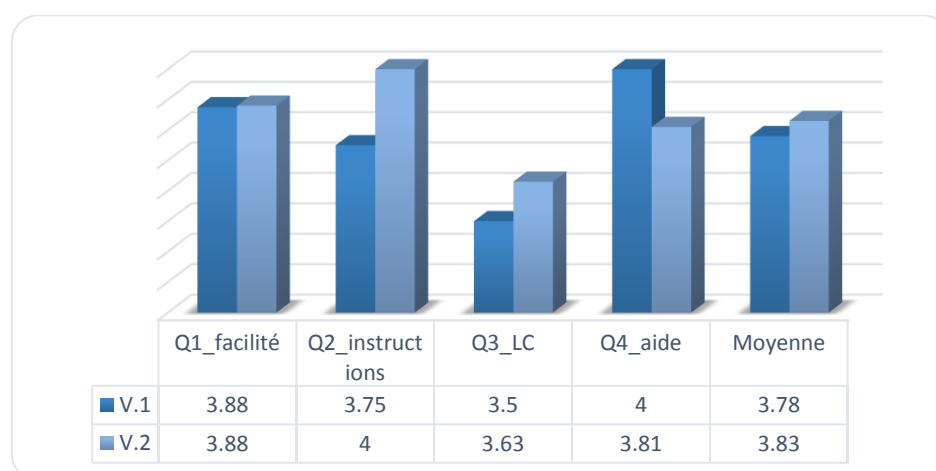


Figure 5.h : Comparaison de l'opinion des patients sur V1 et V2

En revanche, l'opinion sur les instructions fournies tant pour le maniement du système (Q.2 : Les instructions sur la façon d'utiliser MedSLT V1/V2 sont claires.) que pour le langage contrôlé (Q.3 : Les instructions sur la façon de formuler les réponses sont très utiles.) diffère beaucoup. Pour Q.2, le score moyen est de 3,75 sur 5 pour V1 et de 4 sur 5 pour V2. L'utilité du LC (Q.3) obtient des scores particulièrement bas. Il est étonnant que V2 obtienne un score plus élevé que V1 (3.63 vs. 3.5), alors qu'on pourrait considérer que des règles sur la formulation des phrases sont moins utiles pour V2 puisqu'une plus grande variété de phrases est permise. Cependant, il est possible que ce résultat vienne du fait que les règles pour V2 sont plus complexes (3 règles et de nombreux exemples) que pour V1 (cf. p. 150 et 151). Ce résultat est peut-être aussi dû à des instructions pas assez claires, notamment sur l'interdiction d'utiliser les formes négatives.

Pour la question 4, concernant l'aide (« La fenêtre de l'aide a été utile pour apprendre à formuler les réponses »), le système V1 obtient un meilleur score, ce qui semble indiquer que le patient a dû avoir recours plus souvent à l'aide pour arriver à formuler des phrases conformes à V1. Nous incluons ces résultats dans le tableau récapitulatif des résultats de la section 5.3.3.

Ainsi une première part de réponse à la **Q1_S9** provient de ce questionnaire. Nous allons maintenant étudier les résultats obtenus pour notre part quantitative à cette question pour laquelle nous nous concentrons sur l'étude de l'apprentissage de la couverture du système grâce au module d'aide de *MedSLT*.

Comme nous l'avons vu, la couverture du système varie entre V1 et V2, et de manière générale, elle est plus difficile à comprendre (section 5.3.1). Il faut souligner que le module d'aide (section 2.4.3) joue un rôle important pour guider l'utilisateur vers la couverture adéquate. Nous avons par conséquent évalué la capacité d'apprentissage des utilisateurs à l'aune du recours à l'aide entre la première utilisation et la dernière utilisation de V1 et V2.

Utilisateur	Médecin	Médecin	Patient	Patient
Système	V1	V2	V1	V2
AC	42,83	42,68	59,10	59,13
AC_help	15,31	17,26	6,00	6,40
AC_IE	0,65	0,77	0,75	3,58
REP	2,44	1,54	0,19	0,00
Total AC	61,24	62,25	66,04	69,11

Tableau 5.34 : Extrait du Tableau 5.9

En nous référant au Tableau 5.34, une première constatation que nous avons pu faire est que les médecins utilisent le module d'aide de *MedSLT* en moyenne deux fois plus souvent (V1

15,31%, et V2 17,26%) que les patients (V1 6% et V2 6,40%). Cette tendance s'explique par le fait que les médecins sont les initiateurs du dialogue. Ils doivent donc, dans un premier temps, trouver les questions couvertes par le système auxquelles les patients doivent « seulement » trouver une façon de répondre. La tâche est plus difficile pour les médecins, mais fort heureusement cette population peut être formée à utiliser le système. Pour la même raison, on constate que le pourcentage de phrases acceptées tout de suite est en moyenne nettement inférieur pour les médecins que pour les patients.

Dans la majorité des cas de mauvaise reconnaissance ou d'interaction sans succès, l'utilisateur trouve de l'aide directement et clique sur une phrase proposée ou il reformule en suivant l'exemple donné. Certains médecins utilisent l'aide beaucoup plus fréquemment que d'autres. Pour certains, une véritable stratégie d'utilisation de l'aide se met en place : notamment avec des phrases longues comme « *Have you been in contact with anyone who has a sore throat?* », le médecin ne dit qu'une partie de la phrase (« *anyone with a sore throat?* ») sachant que l'aide va lui proposer la phrase correcte et entière pour obtenir "*have you been in contact with anyone who has a sore throat*" ou "*for how long*" pour obtenir "*how long have you had your sore throat*".

D'autres médecins ont eu recours directement à l'aide en évitant de parler, notamment ceux qui avaient tendance à ne pas parler assez fort (malgré le micro) et qui n'obtenaient pas de bons résultats avec la RV. Dans ces cas-là, il est plus facile de piocher directement une question parmi celles proposées par le module d'aide.

Dans les études (Starlander, Bouillon et al. 2005, Chatzichrisafis, Bouillon et al. 2006), l'impact de l'aide sur l'apprentissage a été évalué en quantifiant le recours à l'aide par chaque participant à l'expérience, mais aussi par l'étude visant à vérifier que l'aide était utile aux utilisateurs en leur proposant un secours adéquat. En effet, selon ces études, l'aide ne semble pas avoir été toujours efficace pour éviter les confusions des utilisateurs, notamment lorsque, comme c'est le cas dans cette étude actuellement, les deux systèmes étudiés se distinguent par leur couverture. En effet, l'étude de 2005 montre que le module d'aide permet d'éviter aux utilisateurs de rester « coincés » sur une même phrase qui ne fonctionne pas en la répétant en le guidant vers une phrase similaire qui est couverte par le système (p. 193, (Starlander, Bouillon et al. 2005). Malgré certains défauts du module d'aide, celui-ci reste indispensable pour guider l'utilisateur au fur et à mesure de l'utilisation des systèmes, comme nous avons pu le constater lors de la collecte de données de mars 2006 avec des étudiants en médecine (5^{ème} et 6^{ème} année). En effet, certains utilisateurs s'adaptent très vite à la couverture de notre

système, en utilisant toujours la même stratégie de diagnostic et les mêmes phrases parce qu'ils ont constaté qu'elles donnaient de bons résultats (Chatzichrisafis, Bouillon et al. 2006). Nous constatons le même type de comportement dans le cadre de la comparaison entre V1 et V2 : le plus souvent les utilisateurs médecin ont très vite mis sur pied une sorte de routine, une fois que certaines phrases fonctionnaient bien, ils les ont utilisées et ainsi augmenté la rapidité du diagnostic, et il semble que les patients font de même. Nous nous proposons ici de vérifier cette évolution afin de répondre à **Q1_S9** (*L'apprentissage du système est plus facile avec V1 ou V2 ?*).

5.3.2.1 Médecin : Facilité d'apprentissage

Le tableau ci-dessous montre la progression entre la première et la quatrième utilisation en temps mis en moyenne pour poser le diagnostic.

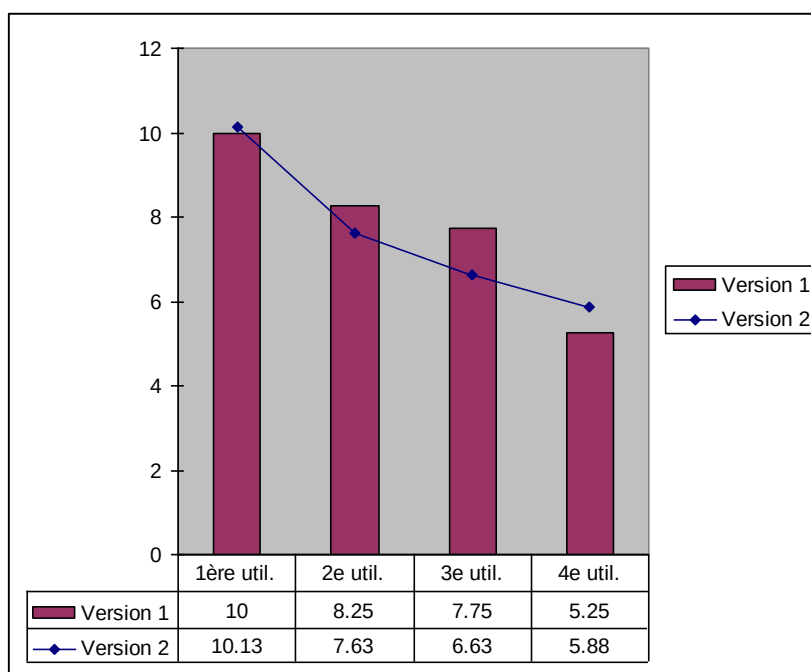


Tableau 5.35 : Evolution entre première et quatrième utilisation du médecin

Cette progression est nette et s'accroît pour la V1, moins restreinte, en passant de 10 minutes en moyenne à 5,25 minutes, ce qui représente une amélioration de 4,75 minutes, pour une amélioration de 4,25 minutes pour V2. Cette étude permet de prouver qu'il y a bien une nette progression entre la première et la dernière utilisation grâce à l'apprentissage progressif de l'utilisateur médecin. Nous verrons à présent ce qu'il en est pour les patients.

5.3.2.2 Patient : Facilité d'apprentissage

Dans le tableau ci-dessous, qui montre le temps moyen par diagnostic, la progression entre la première et la deuxième utilisation pour chaque système est nette. Cette différence s'accroît encore pour V2. L'utilisateur s'adapte plus vite au système.

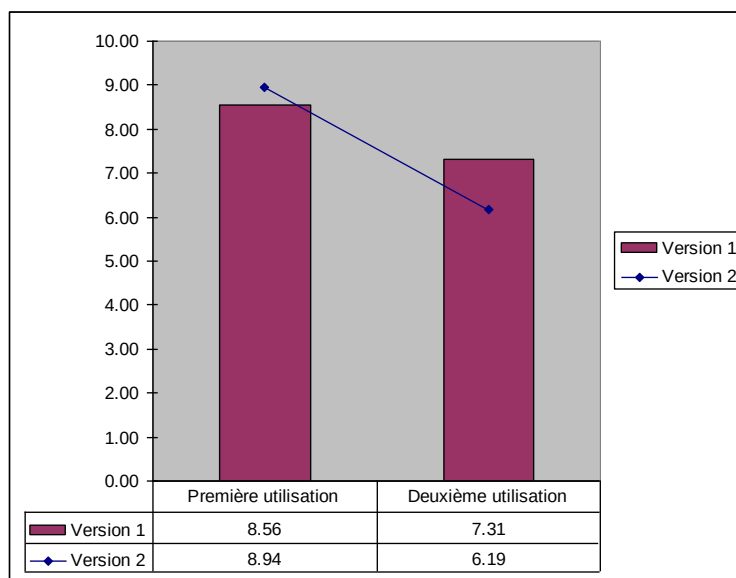


Tableau 5.36 : Evolution entre première et deuxième utilisation du patient

Les patients n'ont utilisé chaque version que deux fois, alors que les médecins ont utilisé chacune pour quatre diagnostics différents (cf. section 4.1.2.2). Néanmoins, une nette progression apparaît pour les deux versions, pour V1, celle-ci est de déjà de 1,29 minutes et pour la V2 la progression est encore plus nette puisqu'elle est de 2,75 minutes. Cette progression moyenne s'observe clairement patient par patient avec une baisse considérable pour certains scénarios, mais pour d'autres patients (patients 7, 8 et 12) cette progression est inversée. La progression la plus fulgurante intervient pour les patients 13 et 1, qui passent d'une durée de diagnostic respectivement de plus de 20 minutes ou 16 minutes à environ 7-8 minutes. On peut conclure que l'apprentissage des systèmes en général s'effectue rapidement, mais que celui-ci est encore plus rapide pour V2 que pour V1. Ainsi la réponse à Q1_S9 est que l'apprentissage du système est plus facile avec V2.

5.3.3 Bilan et tableau récapitulatif de la qualité externe : facilité d'utilisation

Le Tableau 5.37 reprend les principaux résultats concernant la facilité d'utilisation. Nous reprenons tout d'abord l'évaluation qualitative faite à l'aide du questionnaire sur la facilité d'utilisation en général pour obtenir le ressenti des utilisateurs. 3,78 et 3,83 correspondent à la

moyenne des scores obtenus pour 4 questions ayant trait à la facilité d'utilisation. Les résultats pour V1 et V2 sont quasiment identiques.

Pour ce qui est de la facilité de compréhension, une différence apparaît pour les utilisateurs médecins en faveur de V2, mais au niveau des patients la différence n'est pas importante et paradoxalement en défaveur de V2, qui est pourtant moins restreinte. V1 obtient - 10 points pour 0 point à V2.

La facilité d'apprentissage a été mesurée en progression de temps mis pour arriver à un diagnostic entre la première et la dernière utilisation du système. La progression est plus nette avec V2 ce qui lui vaut 10 points supplémentaires. Ainsi, au total, pour la qualité externe le score final pour V1 est de 0 point contre 20 points pour V2.

Qualité externe			Résultats		Scores	
Facilité d'utilisation			V1	V2	V1	V2
Facilité de compréhension (couverture)	Nombre de phrases hors couverture	Médecins (EN-ES)	61,7 %	56,8 %	-10	0
		Patients (ES-EN)	38,3 %	43,2 %	0	0
Facilité d'apprentissage	Evaluation par questionnaire Likert 1-5	moyenne des 4 questions, cf. Figure 5.h)	3,78	3,83	10	10
	Progression en temps entre 1ère et dernière utilisation	Minutes en moins pour diagnostic	1,29	2,75	0	10
TOTAL					0	20

Tableau 5.37: Récapitulation des scores obtenus pour la facilité d'utilisation

Nous reprendrons à présent tous les résultats obtenus pour l'évaluation de V1 et V2 avec la méthode ISO-EAGLES afin de tirer les conclusions pour nos systèmes. Nous allons voir lequel des deux systèmes V1 et V2 obtient le plus de points au score final, ce qui nous permettra de répondre à notre première question de recherche principale: *Peut-on utiliser la méthode EAGLES-ISO pour évaluer les systèmes de TAP dans le domaine médical ?*

5.4 Conclusion : méta-évaluation de la méthode EAGLES-ISO

Nous avons appliqué la méthode complète EAGLES-ISO à deux systèmes comparables V1 et V2. Nous avons tout d'abord défini un modèle de qualité comprenant des caractéristiques de qualité à l'usage (*quality in use*) et de qualité externe. L'avantage de cette méthode est que nous obtenons un score final comparable. Nous reprenons tous les scores finaux pour la qualité à l'usage et la qualité externe dans le tableau ci-dessous.

	Caractéristique	V1	V2
Qualité à l'usage			
	Satisfaction	-20	20
	Efficacité	-10	-10
	Productivité	10	10
	Sécurité	20	20
Sous-total		0	40
Qualité externe			
	Précision de RV	-20	-20
	Précision de Rétrotra	10	10
	Précision de TA (humaine)	20	40
	Précision de TA (auto)	0	-10
	Capacité fonctionnelle	10	20
	Facilité d'utilisation	0	20
Sous-total		10	40
TOTAL FINAL		10	80

Tableau 5.38 : Score final pour V1 et V2

Il ressort de ce tableau récapitulatif que V2 obtient 80 points alors que le score final pour V1 se monte à 10. V2 arrive donc très clairement en tête, notamment grâce aux meilleurs scores obtenus pour la qualité à l'usage. V1 obtient un total de 0 en additionnant les résultats pour chaque caractéristique de la qualité à l'usage, tandis que V2 arrive à un score de 40 points.

Pour la qualité externe, au niveau de la capacité fonctionnelle, V2 arrive en tête avec 20 contre 10. La différence entre les deux systèmes s'accroît au niveau de la facilité d'utilisation, V2 obtient 20 points contre 0 point pour V1. C'est dans la facilité d'utilisation que les deux versions se distinguent le plus. Notre méthode d'évaluation EAGLES-ISO a été efficace pour le montrer.

Nous pouvons constater que cette évaluation notamment de la qualité à l'usage est très informative puisqu'elle permet de donner une idée du comportement du logiciel dans son utilisation future, cependant c'est également une méthode très coûteuse et longue à mettre en place.

L'évaluation de la qualité externe au niveau de la capacité fonctionnelle est souvent employée parce qu'elle utilise des mesures classiques appliquées de manière quasi systématique dans le domaine de la TAP mais aussi pour l'évaluation des sous-composantes ou sous tâches de celle-ci. Dans la section 5.2, nous avons pu appliquer de manière équitable ces mesures. Pour toutes les sous-composantes, la différence entre V1 et V2 n'est pas vraiment significative ce qui est répercuté dans le score obtenu final assez proche de 10 pour V1 et de 20 pour V2. Comme le montre le Tableau 5.38, au niveau de la RV et de la rétro-translation, les deux systèmes arrivaient à exæquo (-20 pour la RV et 10 pour la rétro-translation). En ce qui concerne

la TA, V2 arrive en tête uniquement parce que les évaluateurs étaient plus cohérents pour V2 que pour V1. En effet, il faut souligner qu'au niveau de la fluidité et de l'adéquation de la TA V1 et V2 obtiennent le même nombre de points (10 points partout, voir Tableau 5.30).

Lors de l'évaluation de la qualité de la TA en matière d'adéquation et de fluidité, qui sont des mesures humaines, nous avons rencontré le problème classique du manque d'accord entre évaluateurs, si l'on en croit les très faibles scores obtenus pour le kappa. Cela nous a permis de confirmer que le kappa n'est sans doute pas le meilleur indice de cohérence entre évaluateurs pour les raisons décrites dans la section 3.3.2.1.2 ; ce qui est confirmé par les résultats obtenus dans la section 5.2.3.1. De cette partie de l'évaluation, l'un de nos principaux résultats de recherche est que des mesures humaines comme l'adéquation et la fluidité ne semblent pas tout à fait adaptées à la TAP, du moins pas pour départager deux systèmes proches.

En ce qui concerne l'évaluation automatique, les scores pour BLEU sont très bas. Selon les études critiques de BLEU (cf. section 3.3.2.2.1), ces niveaux très bas pourraient s'expliquer par le fait que cette mesure classique n'est pas adaptée à la TAP du domaine médical, où les phrases sont très courtes et idiomatiques (distantes des traductions de référence plus littérales, voir chapitre 6).

De manière générale, il serait intéressant d'appliquer la méthode proposée dans ce chapitre à d'autres systèmes plus éloignés pour déterminer si elle donne également des résultats fiables et informatifs. Cependant, la principale limite est le temps et le coût investis dans cette méthode. Nous avons en effet effectué une évaluation complète (qualité à l'usage et externe) de nos deux systèmes, et les résultats globaux obtenus sont certes intéressants, satisfaisants et fiables, mais ils ne remplissent pas tous les critères de qualité d'une évaluation, tels qu'énoncés par Koehn (2010) et King (2007). Selon ces auteurs, les principales caractéristiques de qualité d'une bonne évaluation sont : la fiabilité et la reproductibilité (l'évaluation peut être répétée et reproduite par d'autres) et la validité (l'évaluation doit correspondre à ce qui devait être évalué, pour ce faire, il faut bien définir le but de l'évaluation et l'objet de celle-ci avant de procéder à celle-ci). Une autre caractéristique importante est la qualité de l'exécution : la méthode doit être bien définie. Enfin, une évaluation de qualité doit fournir des résultats exploitables, ce qui sera assuré à condition de savoir comment interpréter les résultats obtenus. Pour tous ces critères, la méthode que nous avons proposée dans le présent chapitre est conforme aux bons usages décrits dans la littérature. Cependant, en ce qui concerne les deux critères mis en avant par (Koehn 2010) pour l'évaluation de la TA, à savoir la possibilité de reproduire une évaluation rapidement

et pour un faible coût, notre méthode d'évaluation obtiendrait de mauvais résultats étant donné l'investissement que ce type d'évaluation représente.

Nous arrivons ainsi à la conclusion qu'il serait intéressant de pouvoir évaluer ou plus exactement prédire la qualité à l'usage en se concentrant sur certains éléments de la qualité externe, comme l'écrit King (2007 : p.14) à propos des travaux de EAGLES et ISLE dans la citation suivante :

*“The particular slant on this assumption [concerning a quality chain] in EAGLES and ISLE work is that by looking at a combination of user's needs and system performance in terms of the quality characteristics, we can construct specialized models and thus, **on the basis of an evaluation of external quality, go a long way towards predicting quality in use for a specific user**”.*

Dans le chapitre suivant, nous poursuivrons notre objectif en proposant une méthode d'évaluation permettant de répondre aux critiques formulées ci-dessus. Nous proposons une méthode appelée TURKoise. Il s'agit d'une évaluation externe mais qui intègre des éléments de qualité d'usage utilisant le *crowdsourcing* pour évaluer l'utilité du produit final : la TA issue de la TAP en utilisant le corpus assemblé lors de l'expérience EXPQ1. Dans le chapitre 6, nous décrirons comment nous sommes arrivés à la méthode TURKoise pour répondre à notre deuxième question de recherche principale : *Peut-on proposer une nouvelle mesure pour la TAP qui serait équivalente à l'utilisation de la méthode EAGLES-ISO, mais plus rapide et plus économique à mettre en place que celle-ci ?*

L'objectif de cette question est d'arriver, au bout du cheminement que nous allons suivre dans le chapitre suivant, à une méthode qui sera applicable plus rapidement mais qui reflètera aussi la qualité à l'usage des systèmes étudiés.

6 Résultats de Q2_principale : méthodologie alternative d'évaluation de la TAP (TURKoise)

Dans le chapitre précédent, nous avons appliqué l'évaluation EAGLES-ISO à nos deux versions de MedSLT (V1 et V2) pour répondre à notre première question de recherche qui était de savoir si cette méthode est adaptée à l'évaluation de la TAP du domaine médical. Nous sommes arrivés à la conclusion que cette méthode est informative et permet bien de départager deux systèmes proches, mais qu'elle est très coûteuse en temps et en argent.

Par conséquent, nous proposons dans ce chapitre de répondre à notre deuxième question de recherche qui consiste à rechercher une méthode d'évaluation alternative qui serait plus rapide à mettre en place et moins coûteuse, tout en étant axée sur la qualité à l'usage du système de TAP étudié.

Ainsi, dans la suite de ce chapitre, nous procédons en trois étapes, comme indiqué dans le chapitre de méthodologie (chapitre 4). Dans la première étape, afin de répondre à **Q2_S1**, nous proposons une nouvelle mesure de la qualité de la TAP : une échelle à quatre points spécialement adaptée à l'évaluation humaine de la TAP médicale (*Human Medical Scale* - HMS). Cette échelle sera appliquée à notre corpus par les évaluateurs de Dallas dans le cadre de notre expérience **EXPQ2_S1**.

Dans la deuxième étape, cette mesure étant manuelle, nous aimerions pouvoir trouver une mesure automatique qui corrèlerait suffisamment avec HMS pour automatiser, au moins en partie, l'évaluation. C'est pour cette raison que nous étudierons les différentes mesures automatiques les plus utilisées (BLEU3, BLEU4, WER et TER, voir section 3.3.2.2) afin de déterminer laquelle corrèle le mieux avec l'échelle HMS (**EXPQ2_S2**). En effet, si un certain degré de corrélation peut être établi, celle-ci pourrait servir à faire un premier tri en donnant une indication sur la qualité des TA produites par un système de TAP du médical. L'expérience **EXPQ2_S3** approfondit l'étude des mesures automatiques en étudiant si les mesures choisies pour **EXPQ2_S2** sont véritablement adaptées à l'évaluation de la TAP linguistique.

Enfin, dans la troisième étape, nous verrons qu'il semble difficile pour l'heure de se passer entièrement d'évaluation humaine lorsque nous voulons arriver à une évaluation « juste » (au niveau de la qualité de l'usage) des systèmes de TAP linguistiques comme *MedSLT*. Nous proposons donc la mesure **TURKoise** afin de fournir une alternative aux évaluations humaine et automatique classiques. L'objectif est de recourir à des évaluateurs humains tout en nous assurant de garantir à notre évaluation une certaine efficacité en temps et

coût (« *cost and time effectiveness* »), mais également de rendre l'évaluation humaine plus objective et fiable en obtenant des scores élevés au niveau de l'accord entre les évaluateurs (« *inter-rater agreement* »). **TURKoise** est une méthode d'évaluation de qualité externe, mais qui intègre des éléments de qualité à l'usage afin de concentrer l'évaluation sur l'utilité (ou l'exploitabilité) du produit final (la TA issue de la TAP). Cette méthode s'appuie sur le *crowdsourcing* : afin d'appliquer une échelle humaine spécialement définie pour l'évaluation de la TAP dans le domaine médical à une plus grande échelle que lors d'une évaluation classique effectuée par un petit ensemble d'experts (**EXPQ2_S4-S6**, décrite section 4.2.3.2).

Ce chapitre est structuré comme suit : dans la section 6.1, nous traiterons de l'étape 1. Pour ce faire, nous décrirons, dans la section 6.1.1, en quoi consiste notre échelle à quatre points pour l'évaluation humaine de la TAP médicale (HMS). Ensuite, nous présentons les résultats pour la question de recherche **Q2_S1**. Pour y répondre, nous avons demandé à nos évaluateurs de l'hôpital de Dallas (cf. section 4.2.1.2) d'appliquer l'échelle HMS (section 6.1.2). Nous présenterons la moyenne obtenue pour V1 et V2 et nous comparerons ces résultats avec ceux obtenus en utilisant l'évaluation classique d'adéquation et fluidité (6.1.2.1). Dans la section 6.1.2.2, nous étudierons la cohérence des évaluateurs et nous clorons la première étape en répondant à **Q2_S1** dans la section 6.1.3.

Dans la section 6.2, nous traiterons de la deuxième étape dont l'objectif est d'étudier les mesures automatiques en comparaison avec l'échelle HMS. Dans la section 6.2.1, nous exposerons notre objectif et la problématique des mesures automatiques pour l'évaluation de la TAP. Dans la section 6.2.2, nous donnerons les résultats de l'expérience **EXPQ2_S2**, et dans la section 6.2.3. ceux pour l'expérience **EXPQ2_S3**. La section 6.2.4 récapitulera les principaux résultats de la deuxième étape.

La section 6.3 est consacrée à la troisième étape (TURKoise). Dans un premier temps, nous introduirons le *crowdsourcing* et ses domaines d'utilisation (section 6.3.1). Dans la section 6.3.2, nous présenterons les résultats de l'expérience **EXPQ2_S4-S6** dont l'objectif est de répondre à notre série de sous-questions énoncés à la section 4.2.3.2. La section 6.3.3 fera l'objet de l'expérience **EXPQ2_S7** qui vise à étudier l'évolution de la cohérence entre évaluateurs CS lorsque leur nombre augmente. La section 6.3.4 clora l'étape 3 en résumant les principaux résultats obtenus pour celle-ci.

6.1 Etape 1 : Evaluation humaine pour TAP médicale (HMS) par évaluateurs experts (EXPQ2_S1)

6.1.1 Échelle *Human Medical Scale* - HMS

Il existe déjà de nombreuses mesures issues de l'évaluation empirique de logiciels, cependant, il n'est pas rare de vouloir proposer sa propre mesure pour coller au plus près des besoins d'une évaluation particulière. C'est notre cas, et nous proposons de suivre la méthode en quatre étapes proposée dans le cas d'une création d'une nouvelle méthode de mesure par (Desharnais et Abran 2012) :

1. Déterminer l'objectif de la mesure (*measurement objective*)
2. Caractériser le concept (et sous-concepts) à quantifier
3. Créer la méthode (modéliser les relations entre les sous-concepts)
4. Définir les règles numériques de la mesure.

Cette méthode peut aussi être utilisée pour améliorer les « mesures de base » proposées par les normes ISO ou comme la fluidité et l'adéquation (voir chapitre 5). Selon (Abran 2010) : « *some base quantities in ISO 9126 are difficult to relate to a measurable concept and have no base unit* ». Elle peut par conséquent pallier ce manque, en permettant si besoin est d'ajouter des mesures adaptées.

Dans notre évaluation, nous voulons nous concentrer sur la qualité de la traduction au niveau du contenu du message. Nous proposons donc une échelle fondée uniquement sur l'adéquation des traductions au niveau du sens. Pour reprendre la méthode de (Desharnais et Abran 2012) ci-dessus, notre objectif est de vérifier si le sens de la phrase originale est préservé, et non pas sa forme. En effet, les cas typiques d'erreurs comme l'omission d'un article, l'insertion, l'ordre des mots ou la transformation d'une phrase personnelle comme « *Are you allergic to antibiotics* » (Êtes-vous allergique aux antibiotiques ?) en une forme qui évite la marque du masculin ou du féminin comme « *Tiene alergia a alguno antibiotico* » (Avez-vous des allergies à des antibiotiques) n'affecte pas l'utilité de notre système.

Pour ces mêmes raisons, nous avons aussi choisi de nous démarquer de l'échelle à trois points (*good-ok-bad*) appliquée précédemment pour l'évaluation manuelle de *MedSLT* (Chatzichrisafis, Bouillon et al. 2006, Rayner, Bouillon et al. 2006). Cette échelle a en effet montré ses limites pour l'évaluation de la TAP. Nous avons pu observer que, dans ces évaluations, les juges évaluaient essentiellement des éléments purement linguistiques (comme

l'omission d'un article) qui n'ont pourtant que peu d'impact sur la transmission du sens de l'énoncé.

Différents aspects ont influencé la définition de notre nouvelle échelle. D'une part, les discussions avec des médecins et d'autres chercheurs du domaine lors d'un atelier dans le cadre de la conférence MT Summit à Manchester (Rayner, Bouillon et al. 2008) ont permis de cerner l'objectif de l'évaluation. D'autre part, nous nous sommes inspiré de l'atelier NYC (Somers et Lovel 2006), et notamment des articles de (Somers et Sugita 2003) qui proposaient déjà une échelle de sept points allant d'utile à inutile. Nous avons également considéré les travaux de (Boitet, Bey et al. 2006), qui suggèrent la définition d'une échelle d'évaluation humaine qui vise l'utilisation finale de la traduction¹²⁶. Plus récemment, dans le cadre de TRANSTAC, Weiss et al. utilisent aussi une échelle de Likert pour évaluer l'adéquation sémantique des traductions obtenues (cf. section 3.4.2) (Weiss et Schlenoff 2011, Sanders, Weiss et al. 2013).

Nous avons par conséquent choisi l'échelle d'évaluation suivante comme moyen d'évaluation de l'adéquation d'une traduction automatique, clairement axée sur l'adéquation du sens :

- La traduction est complètement correcte. Tout le sens de la source est présent dans la phrase cible (CCOR - 4 points).
- La traduction n'est pas complètement correcte. Le sens est légèrement différent, sans présenter un danger de mauvaise communication entre le médecin et le patient (MEAN - 3 points).
- Cette traduction ne fait pas de sens, c'est du non-sens. Cette traduction n'est pas correcte dans la langue cible (NONS - 2 points).
- Cette traduction est incorrecte et le sens de la phrase cible est très différent de la phrase source. C'est un faux-sens, dangereux pour la communication entre le médecin et le patient (DANG - 1 point).

Les deux catégories NONS et DANG peuvent paraître similaires, toutefois elles sont bien distinctes. En effet, les phrases de catégorie NONS étant beaucoup plus faciles à identifier par les utilisateurs finaux que les phrases DANG, elles sont moins dangereuses. Les phrases

¹²⁶ « A suggestion, then, would be to suppress the classical fluency and adequacy measures as we know them since 4-5 years or so, and to invent and introduce usage-related measures for deployed systems. », p.8 in: Boitet, C., Y. Bey, M. Tomokio, W. Cao et H. Blanchon (2006). IWSLT-06: Experiments with Commercial MT Systems and Lessons from Subjective Evaluations. International Workshop on Spoken Language Translation (IWSLT-06), Kyoto, Japan.

évaluées NONS sont tellement déformées qu'elles ne veulent plus rien dire, tandis que les phrases classées comme DANG ressemblent à des phrases tout à fait correctes mais qui sont en fait des non-sens. C'est ce type de phrase qui représente un véritable danger pour l'utilisateur final qu'est le patient, avec comme conséquence des erreurs potentielles de diagnostic.

6.1.2 Résultats de l'expérience EXPQ2_S1

Nous avons appliqué cette échelle aux traductions des interactions des médecins (anglais - espagnol), ainsi qu'aux traductions des réponses de patients (espagnol - anglais). Il s'agit du même corpus que celui utilisé au chapitre 5. Nous avons réintégré les respectivement 222 et 305 types dans le corpus complet (voir section 4.2.1.2.1). Pour chaque phrase évaluée, nous avons récolté trois évaluations, ce qui nous a permis de vérifier le niveau d'accord entre les évaluateurs (6.1.2.2). Dans la section suivante, nous donnons les moyennes obtenues par chaque version de *MedSLT*.

6.1.2.1 Application de HMS

Dans le Tableau 6.1, ci-dessous, on voit que la version V1 obtient des résultats légèrement meilleurs au niveau de la moyenne.

	Ang-Esp V1	Ang-Esp V2	Esp-Ang V1	Esp-Ang V2
Moyenne	3,64	3,59	3,74	3,57
Erreur-type	0,02	0,02	0,03	0,04
Médiane	3,67	3,67	4	4
Mode	4	4	4	4
Écart-type	0,46	0,49	0,54	0,75
Variance de l'échantillon	0,21	0,24	0,30	0,57
Nombre d'échantillons	358	390	335	364
Niveau de confiance (95.0%)	0,05	0,05	0,06	0,08

Tableau 6.1 : Moyenne des scores obtenus sur une échelle de 1-4 (1=DANG, 4=CCOR), sur toutes les phrases du corpus (avec répétition)

Pour la direction anglais vers l'espagnol, la différence est très faible (de +0,05 en faveur de V1). Celle-ci est par conséquent non-significative. Dans la direction espagnol vers l'anglais, la différence est un peu plus marquée avec +0,16 en faveur de V1. Il est intéressant de noter que l'écart type pour V2 est particulièrement élevé (0.7522 en moyenne) tout comme sa variance. Cela semble montrer que, pour V2, il y a une plus grande différenciation entre les

phrases¹²⁷. Nous verrons dans le Tableau 6.2, plus bas, qu'il y a en effet plus de phrases des catégories NONS et DANG, comparé à V1 (9% des phrases dans ces catégories pour V2 contre seulement environ 4% pour V1).

Le Tableau 6.2 indique le mode (c'est-à-dire la classe d'une distribution statistique qui présente la fréquence la plus élevée,¹²⁸ ce qui équivaut à la tendance centrale) par système et pour chaque catégorie. Dans cette perspective de « la majorité gagne », la version plus restreinte V1 obtient le pourcentage le plus élevé de phrases totalement correctes (catégorie 4) pour les deux sens de traduction, tandis que V2 obtient le plus grand pourcentage de phrases jugées comme dangereuses.

Catégorie	Anglais - Espagnol V1	Anglais - Espagnol V2	Espagnol - Anglais V1	Espagnol - Anglais. V2
4	70.7%	68.0%	80.9%	73.1%
3	24.0%	25.8%	14.8%	17.4%
2	2.2%	2.2%	0.9%	2.0%
1	2.3%	3.5%	3.2%	7.1%

Tableau 6.2 : Evaluation de la qualité de la traduction par catégorie

Les résultats obtenus avec HMS sont assez proches des résultats obtenus à la section 5.2.3.1, lorsque nous avons évalué V1 et V2 avec les mesures classiques de fluidité et d'adéquation. Les moyennes obtenues pour la fluidité (échelle de 1-5) de l'espagnol pour V1 étaient de 4,05 et de 4,07 pour V2, pour l'anglais, V1 obtenait 4,74 et V2 4,56. L'écart entre V1 et V2 est également très faible pour les scores de l'adéquation. La direction anglais-espagnol obtenait 4,82 pour V1 et 4,83 pour V2. Concernant les traductions de l'espagnol vers l'anglais, ces scores s'élèvent à 4,63 pour V1 et 4,44 pour V2. Pour ce dernier score, la différence entre les deux versions est un peu plus marquée avec +0,19 de différence en faveur de V1. De ce point de vue, la nouvelle échelle HMS ne permet pas de distinguer plus clairement les deux versions étudiées que lorsque nous utilisons les mesures classiques. Ce résultat semble montrer que les phrases obtenues avec l'une et l'autre version sont similaires. Cependant, afin d'étudier pourquoi il en est ainsi, nous avons analysé en détail les réponses obtenues.

Le Tableau 6.3 donne un extrait du tableau de l'évaluation (annexe 6) effectuée par le groupe d'évaluateurs de Dallas en appliquant de l'échelle HMS. Il en ressort que l'évaluation montre une grande variation entre les évaluateurs. Dans la plupart des cas, les avis diffèrent entre deux points proches de l'échelle (CCOR et MEAN ou DANG et NONS) mais rien que

¹²⁷ « The larger the standard deviation, the more spread out the values are, and the more different they are from one another », p.48, Salkind, N. J. (2012). *Statistics for people who (think they) hate Statistics: Excel 2010 Edition*, Sage.

¹²⁸ Définition du Petit Robert, version 2012.

dans le cours extrait ci-dessous, il apparaît que l'accord entre les évaluateurs est loin d'être parfait.

Source	Cible	Eval. 1	Eval. 2	Eval. 3
do you have any stomachache	el estómago le ha dolido	CCOR	DANG	CCOR
are you allergic to antibiotics	los antibióticos le dan alergia	CCOR	NONS	MEAN
are you better	el dolor disminuye	DANG	MEAN	DANG
are you congested	tiene la nariz tapada	MEAN	CCOR	MEAN
are your glands swollen	tiene los ganglios inflamados	DANG	CCOR	CCOR
are you having fever	el dolor está aliviado cuando tiene fiebre	DANG	DANG	DANG
are you vomiting	vomita	CCOR	MEAN	MEAN
did the doctor do a rapid strep test	le han realizado la prueba rápida por estreptococo	CCOR	CCOR	MEAN
did you have a throat culture	le han realizado un cultivo de la garganta	CCOR	CCOR	CCOR
did you have any test	le han realizado un examen	CCOR	CCOR	CCOR
do you have a headache	tiene tos ayer	DANG	NONS	NONS

Tableau 6.3 : Extrait de l'évaluation par les évaluateurs de Dallas

C'est pour cette raison, que nous allons évaluer dans la section suivante la cohérence des évaluateurs experts.

6.1.2.2 Cohérence des évaluateurs experts

Comme évoqué dans la section 3.3.2.1.2, en présence d'évaluations humaines, il est indispensable d'évaluer si les évaluateurs sont d'accord les uns avec les autres, en calculant un score d'accord entre évaluateurs. Pour rappel, nous avons choisi d'appliquer le kappa de Fleiss¹²⁹ (Fleiss 1971).

Dans notre cas, nous avons au minimum trois évaluations et quatre catégories (CCOR -4, MEAN -3, NONS-2, et DANG-1), ce qui influence beaucoup les résultats, que voici :

Kappa de Fleiss	V1	V2
4 catégories ANG-ESP - Médecins	0,18	0,10
4 catégories ESP-ANG - Patients	0,41	0,49
2 catégories ANG-ESP - Médecins	0,27	0,24
2 catégories ESP-ANG - Patients	0,62	0,71

Tableau 6.4 : Estimation du kappa de Fleiss avec Agreestat

Il ressort du Tableau 6.4 que pour les deux versions (V1 et V2) de l'anglais vers l'espagnol, l'accord entre les juges est très faible (interprété d'après (Landis & Koch, 1977)).

¹²⁹ Calculé ici, avec l'outil Agreestat, http://www.agreestat.com/agreestat_excel.html, Kilem L. Gwet, Ph.D.

L'accord pour les patients semble plus élevé pour V2 que pour V1. Une analyse plus approfondie des phrases évaluées révèle que les cas où les juges sont le moins d'accord sont ceux où une phrase est évaluée CCOR ou MEAN. En multipliant le nombre de catégories, l'effet de dispersion des évaluateurs s'accroît. Nous avons par conséquent également calculé l'accord entre évaluateurs en regroupant les catégories d'après leur finalité (traduction utilisable vs inutilisable). Nous avons donc fusionné dans une même catégorie CCOR et MEAN et dans une autre catégorie NONS et DANG. Dans ce cas, comme on peut le voir dans les rangées inférieures du Tableau 6.4, l'accord est fort, notamment pour les traductions de l'espagnol vers l'anglais. Cela montre que l'échelle serait plus facile à appliquer si elle ne comportait que deux catégories, mais nous perdrons en granularité.

Afin de compléter cette partie sur l'accord entre évaluateurs, nous présentons dans le Tableau 6.5 le pourcentage d'accord parfait entre les trois évaluations, comme dans (Hamon, Fügen et al. 2009). Ces résultats rejoignent ceux du Tableau 6.4 pour indiquer que l'accord est plus élevé dans le sens Espagnol-Anglais. Par contre, pour V2, les résultats ne concordent pas avec le Tableau 6.4, puisque dans celui-ci, V2 obtenait un meilleur kappa, alors qu'en termes d'accord parfait, V2 présente un score moins élevé que V1.

Accord entre les trois évaluateurs	V1	V2
4 catégories (4-1) - ENG-SPA	47%	38%
4 catégories (4-1) - SPA-ENG	70%	66%
2 catégories : (4-3) et (2-1) fusionnés - ENG-SPA	90%	88%
2 catégories : (4-3) et (2-1) fusionnés - SPA-ENG	96%	93%

Tableau 6.5 : Pourcentage d'accord entre les trois évaluations

Il sera intéressant d'étudier le comportement de la cohérence entre évaluateurs lorsque cette échelle sera utilisée par un plus grand nombre d'évaluateurs, ce qui sera fait grâce à l'introduction du *crowdsourcing* (section 6.3).

Dans la section suivante, nous tirerons les conclusions des résultats obtenus afin de répondre à **Q2_S1**.

6.1.3 Conclusion de l'étape 1 : Q2_S1

En résumé, lorsque l'on considère uniquement la moyenne des évaluations humaines données par nos traducteurs, les deux systèmes arrivent ex-æquo. Si on nuance ces résultats, en considérant le pourcentage de phrases classées dans chacune des quatre catégories qui constituent notre échelle humaine, la V1 obtient assez clairement de meilleurs résultats. Nous

constatons que l'échelle HMS permet de mieux trancher entre les deux systèmes V1 et V2 que les mesures classiques de fluidité et adéquation (section 5.2.3.1). Cependant, comme toujours avec des mesures humaines composées d'une échelle de plus de 2 degrés (Graham, Baldwin et al. 2013), l'accord entre évaluateurs n'est pas élevé. Le degré d'accord entre évaluateurs oscille entre 0,41 et 0,49, ce qui correspond sur l'échelle d'interprétation de (Landis et Koch 1977) seulement à un accord modéré. Cette constatation justifie pourquoi, afin d'éviter la subjectivité des mesures humaines, les chercheurs ont souvent recours à des mesures automatiques.

Nous étudierons à présent à quel point les mesures humaines et automatiques corrélient dans le but de trouver quelle mesure automatique pourrait aller dans le même sens que notre échelle HMS.

Notre objectif est maintenant de tester les mesures automatiques classiques afin de trouver lesquels corrélient le mieux avec notre mesure humaine axée sur l'utilité (HMS). Dans la section suivante, nous exposons les résultats de l'analyse de la corrélation entre les deux types d'évaluation (automatique et HMS).

6.2 Etape 2 : Corrélation entre mesures automatiques et échelle HMS

Dans cette deuxième étape de la question principale 2 (**Q2_principale**), nous présenterons tout d'abord (section 6.2.1) les résultats de **EXPQ2_S2** qui consiste à étudier quelle mesure automatique corréle le mieux avec notre échelle HMS. Ensuite, à la section 6.2.2, nous donnerons les résultats pour notre expérience **EXP2Q2_S3** qui répond à la question **Q2_S3** (*Les mesures automatiques étudiées pour Q2_S2 favorisent-elles la TA statistique et sont-elles moins adaptées à l'évaluation de la TAP linguistique comme MedSLT ?*). Dans la section 6.2.2.1, nous présenterons les résultats obtenus avec HMS sur notre corpus de phrases traduites par *MedSLT* et *GT* et dans la section 6.2.2.2, ceux obtenus avec les mesures automatiques retenues pour cette expérience (WER, BLEU, TER). Enfin, nous arriverons aux résultats concernant la corrélation entre HMS et les mesures automatiques (section 6.2.2.3). Dans la section 6.2.3, nous terminerons cette deuxième étape par une conclusion sur les réponses obtenues à nos questions Q2_S2 et Q2_S3.

6.2.1 Mesures automatiques : critiques

Les expériences décrites dans (Blatz, Fitzgerald et al. 2004) suggèrent que les mesures automatiques actuellement disponibles obtiennent une corrélation insuffisante au niveau de la phrase. Ces mesures fonctionnent mieux sur un corpus entier (Kulesza et Shieber 2004). De

plus, (Turian, Shen et al. 2003) observaient déjà que la performance des mesures automatiques classiques (BLEU) est mauvaise sur des phrases courtes, ce qui est le cas pour nos données.

De plus, comme il s'agit de production orale, de nombreuses variantes d'une même phrase sont possibles. Dans le cas de *MedSLT*, l'architecture par interlangue (voir chapitre 2) fait que plusieurs phrases en LS, différentes syntaxiquement mais porteuses du même sens, seront traduites de la même façon en LC, perdant certaines nuances (entraînant quelques fois un appauvrissement) mais gagnant en cohérence.

Phrases LS	Traduction LC
(1) For how long have you had your sore throat?	(A) ¿Cuánto tiempo le ha dolido la garganta?
(2) Have you had pains for a long time?	
(3) For how long time does your throat hurt?	

Tableau 6.6 : Traduction unique pour multiples phrases LS

Comment peut-on utiliser des mesures automatiques sans pénaliser un système sur ce type de variations dues au passage par une interlangue ? Si la mesure se contente de compter les différences au niveau de la surface comme BLEU ou NIST, partant de la phrase source (A), considérant la traduction (2) et utilisant les phrases (1) et (3) comme traduction de référence, le score BLEU pour la phrase (2) sera très bas, parce qu'au niveau des mots, les phrases (1) et (3) sont très différentes de la phrase (2) alors que du point de vue du sens elles sont similaires.

Le choix des traductions de référence est de ce fait déterminant pour la fiabilité des scores d'évaluation automatique. Différentes pistes pour arriver à une traduction de référence sont envisageables en utilisant comme étalon-or, soit :

- la traduction attendue par le développeur,
- la traduction préférée des juges humains, mais il faudra tout d'abord trouver un accord,
- la traduction produite par un traducteur professionnel, cependant, il faudra en juger la qualité,
- la traduction issue d'un « phrase book » bilingue, l'avantage étant que ces traductions ont normalement déjà été évaluées avant publication (exemple : (Bongiovanni 2005)).

Nous verrons à la section 6.2.3 comment nous avons procédé dans le cadre de ce travail.

Pour contourner la difficulté de choisir des traductions de référence, une alternative serait d'établir une mesure automatique qui ne se base pas sur la phrase en surface, mais sur le sens. Celle-ci se baserait, pour évaluer la qualité de la traduction, non plus sur des mesures de surface telles que BLEU ou WER.

Dans l'état de l'art, le standard de comparaison communément admis pour les mesures automatiques est l'évaluation humaine (Dybkjaer, Hemsén et al. 2007, Koehn 2010). On constate que la plupart des études cherchent à évaluer une méthode d'évaluation par rapport à une autre, et notamment les mesures automatiques entre elles. L'objectif est de trouver une mesure automatique qui corrèle avec le jugement humain. Notre but est l'inverse : trouver une mesure automatique qui corrèle avec l'échelle HMS afin d'accélérer le processus d'évaluation humaine.

Dès lors, comment choisir au mieux des mesures automatiques qui pourraient corrélérer avec HMS ? Un critère de choix pour déterminer quelle mesure est meilleure peut simplement consister à retenir la méthode la plus utilisée. Dans ce cas, BLEU arrive en tête, toutes catégories confondues. En effet, lorsque l'on considère les études menées pendant cette décennie, depuis son apparition en 2002, BLEU s'est très vite imposé. L'étude la plus courante dans l'état de l'art consiste à vérifier la corrélation entre un certain ensemble de mesures automatiques - dont BLEU fait quasiment toujours partie - et les mesures humaines le plus couramment admises, à savoir, d'une part l'adéquation et la fluidité, et d'autre part, une classification binaire des systèmes évalués. Cette dernière évaluation est simple, puisqu'il s'agit pour des évaluateurs humains de classer deux systèmes selon leur préférence au niveau de la traduction produite. Nous avons par conséquent principalement retenu BLEU, mais dans un premier temps nous présentons les résultats également pour les mesures automatiques décrites à la section 3.3.2.2.

6.2.2 Résultats de l'expérience EXPQ2_S2 : Corrélation entre mesures automatiques et échelle HMS

Afin d'étudier la corrélation entre les différentes mesures appliquées (humaines et automatiques), nous avons calculé le coefficient de Pearson comme dans (Estrella, King et al. 2007) sur les données d'EXPQ1. La méthode utilisée dans cet article est celle de l'extrapolation des données (*bootstrapping technique*) dans le but de générer un grand nombre de points d'échantillon sur lesquels les corrélations sont ensuite calculées. Le Tableau 6.7 indique les mesures de corrélation entre notre évaluation humaine et différents scores automatiques

appliquées à nos données. Les traductions de référence sont celles effectuées par les interprètes de l'hôpital de Dallas (voir section 4.2.2.2).

Mesures	Ang-Esp V1	Ang-Esp V2	Esp-Ang V1	Esp-Ang V2
HMS vs. BLEU	0,25	0,23	0,40	0,29
HMS vs. NIST	0,21	0,23	0,35	0,32
HMS vs. Précision	0,25	0,32	0,44	0,36
HMS vs. Rappel	0,09	0,15	0,49	0,50
HMS vs. F-mesure	0,19	0,24	0,48	0,47
HMS vs. PER	-0,26	-0,31	-0,49	-0,41
HMS vs. WER	-0,24	-0,29	-0,50	-0,42

Tableau 6.7 : Corrélation entre les différentes mesures automatiques et HMS.

Malgré la faible corrélation pour les deux versions du système, V1 pour les patients obtient les scores les plus élevés concernant la corrélation entre HMS et BLEU. Pour les deux versions, au niveau des médecins, la corrélation obtenue est négligeable, ce qui suggère que les mesures automatiques, dans ce cas, devraient être utilisées avec prudence pour effectuer un premier tri.

En réalité, la corrélation entre les mesures automatiques et HMS est vraiment faible. Nous avons évoqué plus haut un certain nombre de raisons qui pourraient expliquer nos résultats. Cependant, une des pistes serait que l'évaluation automatique basée sur des références (comme BLEU et toutes les mesures étudiées) n'est pas adaptée à l'évaluation de la TAP dans le médical, surtout si elle est de nature linguistique. Nous allons par conséquent étudier la question de l'adaptation des mesures automatiques afin de répondre à **Q2_S3** (*Les mesures automatiques étudiées pour Q2_S2 sont-elles moins adaptées à l'évaluation de la TAP linguistique comme MedSLT ?*) en comparant *MedSLT* avec un système statistique *Google Translate*, au lieu de comparer V1 et V2.

6.2.3 Résultats de l'expérience EXPQ2_S3 : *MedSLT* comparé à *GT*

Dans cette section, nous comparerons les résultats obtenus pour *MedSLT* et *Google Translate* (GT) tout d'abord avec HMS, puis avec les mesures automatiques BLEU, TER et WER.

Nous avons choisi GT comme système SMT pour différentes raisons : premièrement, c'est le logiciel le plus commun et accessible à tous, deuxièmement, il existe sur Internet de nombreux sites de diagnostic bilingue espagnol-anglais, ce qui lui donne une bonne base de corpus d'entraînement adapté à notre domaine. La dernière raison est que GT est souvent utilisé dans la pratique hospitalière (voir chapitre 2). Ainsi notre choix de *GT*, que nous ne pouvons

certaines pas entraîner sur des corpus spécifiques, est néanmoins assez équitable dans la mesure où les données dont *Google* dispose pour entraîner son modèle du langage comportent au moins un ensemble important de données adaptées. Une autre possibilité aurait été de choisir Moses (Koehn, Hoang et al. 2007) et de l’entraîner sur nos données comme cela a été fait avec *MedSLT* pour d’autres combinaisons linguistiques comme le français-anglais et l’anglais-japonais. Cependant, comme décrit dans (Rayner, Estrella et Bouillon, 2011), le SMT ainsi obtenu grâce à des méthodes d’extrapolation (*bootstrapping*) sur notre petit corpus d’origine ne donne pas de meilleurs résultats que notre système linguistique avec lequel il était comparé. De ce fait, les recherches dans ce sens n’ont pas été étendues à l’espagnol-français. Notre meilleure solution est donc de choisir GT comme système de comparaison. Enfin, puisqu’il s’agit d’une comparaison entre un SMT et un système linguistique, comme annoncé en conclusion du chapitre précédent, pour maximiser la taille de notre corpus, nous avons regroupé toutes les données issues de l’expérience de Dallas en un seul corpus qui est constitué des phrases uniques de V1 et du V2 confondues.

Cette étude est justifiée par le fait qu’une mesure doit être fiable, reproductible (Banerjee et Lavie, 2005, et ISO standards) mais elle ne doit pas être influencée par la technologie sous-jacente. Ainsi une mesure ne devrait pas avantager un type de technologie. Ce point a été considérablement discuté parce que, comme le montrent plusieurs recherches et notamment les grandes campagnes d’évaluation en TA et TAP (Callison-Burch et al 2006), certaines mesures automatiques comme BLEU sont plus favorables aux systèmes de traduction automatique statistique, donnant systématiquement un moins bon score aux systèmes de traduction automatique linguistique alors que les juges humains donnaient une meilleure évaluation à ces derniers ¹³⁰(Koehn et Monz, 2006).

Nous commencerons par l’évaluation humaine par HMS, puis nous poursuivrons par l’évaluation automatique des deux systèmes *MedSLT* et *GT*.

6.2.3.1 Évaluation humaine (HMS)

Dans cette section, nous utilisons notre échelle HMS (section 6.1). Pour rappel, celle-ci va de 0 (DANG) à 4 (CCOC), 4 étant le score maximum. Afin d’augmenter le nombre d’évaluateurs, nous avons demandé à 3 personnes supplémentaires d’évaluer les phrases issues de *MedSLT* et de *GT*. Il s’agit cette fois de non-traducteurs. L’idée est de comparer le degré de

¹³⁰ « The results of last year’s workshop further suggested that Bleu systematically underestimated the quality of rule-based machine translation systems », in : (Koehn et Monz, 2006).

sévérité entre ces deux groupes d'évaluateurs ainsi que de voir quelle est leur cohérence. En effet, nous avons pu constater tout au début de nos recherches sur *MedSLT* que les traducteurs avaient tendance à être plus sévères que d'autres évaluateurs médecins ou ingénieurs (voir section 4.2.2.3.1). Comme dans la section 6.1.2 (Tableau 6.1), nous pouvons constater dans le Tableau 6.8 que lorsqu'on prend en compte uniquement la moyenne des résultats, les chiffres sont très proches entre *MedSLT* et GT. Le premier obtient cependant un score total (les 6 évaluateurs) plus élevé que le second. En outre, on peut remarquer que les évaluateurs non-traducteurs sont en effet moins sévères que les évaluateurs traducteurs.

	MedSLT	GT
Traducteurs	3,40	3,43
Non-traducteurs	3,62	3,46
Tous	3,51	3,44

Tableau 6.8 : Evaluation humaine, comparaison entre *MedSLT* et GT

Comme dans la section 6.1.2.1, nous compléterons ces résultats par une étude catégorique par catégorie, en ne considérant que le mode, c'est-à-dire en adoptant une perspective « de la majorité gagne ». Pour les deux groupes d'évaluateurs, *MedSLT* obtient de meilleurs résultats avec 68% contre 65,8% de phrases totalement correctes (CCOR, 4) pour les traducteurs et même 76,1% contre 70,7% pour les non-traducteurs (Tableau 6.9). De ce Tableau 6.9, il ressort aussi clairement que les non-traducteurs donnent de meilleurs scores que les évaluateurs traducteurs que ce soit pour *MedSLT* ou GT, en franchissant (même largement pour *MedSLT*) la barre des 70% de phrases classées comme CCOR.

Catégorie	MedSLT Traducteurs	GT Traducteurs	MedSLT Non-traducteurs	GT Non-traducteurs
4 = CCOR	68,0%	65,8%	76,1%	70,7%
3 = MEAN	25,2%	20,7%	16,7%	15,3%
2 = NONS	2,3%	6,3%	0,0%	2,3%
1 = DANG	4,5%	3,2%	3,2%	2,7%
Aucun accord	Néant	4,1%	4,05%	9,0%

Tableau 6.9 : Evaluation humaine, comparaison entre *MedSLT* et GT par catégories

En outre, si l'on compare ce tableau avec le Tableau 6.2 qui présente les résultats équivalents pour la comparaison entre V1 et V2, les résultats obtenus ici pour *MedSLT* sont équivalents à la moyenne des deux résultats pour V1 et V2, et les proportions de phrases placées dans chaque catégorie sont également similaires. Au niveau de l'évaluation HMS, *MedSLT* obtient de meilleurs résultats que GT. Nous verrons dans la section suivante si les résultats obtenus par les mesures automatiques sont cohérentes avec ce résultat.

Comme il s'agit d'une évaluation humaine, nous analyserons ces résultats par rapport à l'accord entre évaluateurs avec la méthode du kappa de Fleiss et avec le pourcentage d'accord selon la méthode de (Hamon, Fügen et al. 2009), qui est plus facile à interpréter.

	MedSLT	GT
Traducteurs	0,18	0,37
Non-traducteurs	0,10	0,26

Tableau 6.10 : Kappa de Fleiss pour MedSLT et GT pour différents groupes d'évaluateurs

Selon le Tableau 6.10, l'accord parfait entre les évaluateurs est bien plus élevé pour les phrases produites par GT que par *MedSLT*, même si ces chiffres restent peu élevés.

	MedSLT	GT
Tous les 6 évaluateurs	18,5%	27,9%
Traducteurs (3)	33,8%	49,5%
Non-traducteurs (3)	41,9%	46,4%

Tableau 6.11 : Pourcentage d'accord entre évaluateurs pour MedSLT et GT

Les scores du Tableau 6.11 s'alignent sur les scores kappa, indiquant un accord plus élevé pour l'évaluation de *GT*, ceci pour les deux groupes d'évaluateurs. Nous observons cependant que les non-traducteurs atteignent des niveaux d'accord similaires pour les deux systèmes, alors que les traducteurs sont nettement plus souvent d'accord sur les catégories attribuées aux traductions de *GT* (49,5% contre 33,8%).

Par rapport à la différence entre *MedSLT* et *GT*, une explication plausible pourrait être que justement, comme notre échelle laisse de côté l'évaluation de la fluidité des traductions, en se concentrant uniquement sur l'adéquation au niveau du sens, les évaluateurs se trouvent plus souvent d'accord sur la médiocrité des traductions de *GT* qui sont bien moins idiomatiques que les phrases produites par *MedSLT*, mais pour lesquels le sens était maintenu. Afin de vérifier cela, nous avons choisi d'ajouter une évaluation humaine binaire, sous la forme d'une évaluation de préférence entre les deux systèmes (*ranking*). Dans cette évaluation, *MedSLT* est clairement préféré dans 61,1% des cas aux dépens de *GT* qui n'obtient la préférence des évaluateurs que dans 34,5% des cas. Il faut aussi noter que, pour cette tâche d'évaluation binaire, le kappa de Fleiss s'élève à 0.5564 ce qui est clairement plus élevé que pour notre échelle à quatre points (Tableau 6.4). Ce résultat rejoint le kappa pour notre autre évaluation binaire de la rétro-translation (Tableau 5.25).

En résumé, les principaux résultats pour cette section consacrée à l'évaluation humaine HMS de *MedSLT* et *GT* sont que les chiffres obtenus au niveau de la moyenne sont très proches,

cependant, le premier système obtient un score total plus élevé que le second. Les résultats sont plus clairement en faveur de *MedSLT* lorsqu'on retient le mode de chaque évaluation pour classer chaque phrase dans une catégorie (Tableau 6.9). En outre, on peut remarquer que les évaluateurs non-traducteurs sont en effet moins sévères que les évaluateurs traducteurs. Nous retenons, enfin, qu'avec l'évaluation HMS, *MedSLT* obtient de meilleurs résultats que *GT*. Nous verrons dans la section suivante si les résultats obtenus par les mesures automatiques étudiées sont cohérentes avec ce résultat.

6.2.3.2 Evaluation automatique : BLEU, TER, WER

Nous avons décidé de ne retenir pour cette évaluation automatique qu'un certain nombre de mesures classiques : taux d'erreurs au niveau des mots (WER) et BLEU, afin de les comparer avec une mesure plus nouvelle : le taux d'erreurs de traduction (TER). Celui-ci est moins sensible que BLEU à l'ordre des mots, ce qui pourrait être un avantage dans notre contexte où les traductions ne sont pas littérales. Nous aurions aimé ajouter à cette comparaison METEOR - qui inclut une composante sémantique en faisant recours à Wordnet (section 3.3.2.2.2) - mais les ressources nécessaires pour l'espagnol n'étaient pas disponibles au moment de cette étude.

Etant donné que toutes ces mesures sont dépendantes des références sur lesquelles elles se basent pour calculer leur score, nous avons utilisé différentes séries de références. Premièrement, *ref_dev* qui correspond à la référence donnée par le développeur du système, deuxièmement, *ref_trad* qui correspond à la traduction référence fournie par les traducteurs de Dallas, et enfin, les deux types de référence mélangés (*ref_mixtes*). Dans le Tableau 6.12, ci-dessous, nous pouvons observer que les résultats des deux systèmes sont proches quand on utilise les références des traducteurs (cf. colonnes 4 et 5) et aussi quand on choisit les références mixtes (cf. colonnes 6 et 7). La seule différence claire apparaît quand on utilise la référence du développeur (colonnes 2 et 3). Dans la colonne *MedSLT ref_dev*, le score pour BLEU est élevé puisqu'il compare les traductions effectuées avec des traductions de références proches (*ref_dev*). Ces résultats confirment la seconde évaluation humaine, l'évaluation binaire de la section précédente.

Mesures	MedSLT Ref_dev	GT Ref_dev	MedSLT Ref_trad	GT Ref_trad	MedSLT Toutes_ref	GT Toutes_ref
WER	0,12	0,80	0,59	0,67	0,55	0,58
TER	0,10	0,67	0,53	0,65	0,65	0,66
BLEU	0,84	0,17	0,35	0,33	0,35	0,33

Tableau 6.12 : Résultats des mesures automatiques

Nous avons déjà utilisé ce corpus de référence dans une évaluation précédente (Rayner et al, 2009), mais selon les résultats obtenus dans ce travail, il est clairement plus équitable pour les deux systèmes d'utiliser un mélange des deux types de référence. Cette étude pointe, comme exposé dans la section 5.2.2, qu'il est primordial de bien choisir son ensemble de références. Une autre conclusion à tirer de ce tableau est que ce serait un avantage de ne pas reposer sur des références humaines, ainsi les nouvelles mesures automatiques proposées récemment (décrites dans la section 5.3), qui se fondent sur l'apprentissage, sont intéressantes. Cette évolution est à suivre attentivement, même si dans notre cas, nous ne disposons sans doute pas d'assez de données.

Afin de mieux comprendre les résultats obtenus avec nos mesures automatiques, en utilisant les références mixtes (colonnes 6 et 7), nous avons décidé de vérifier nos résultats phrase par phrase. Dans le Tableau 6.13, ci-dessous, nous donnons un extrait de cette évaluation.

Source	Cible	BLEU4	Hum.	TER	WER	BLEU2
Are you coughing?	¿Tiene tos?	0	4.0	0.0	0.0	1
Do you have a cough?	¿Tiene tos?	0	4.0	0.0	0.0	1
Do you have a fever?	¿Tiene fiebre?	0	4.0	60.0	75.0	0
Have you vomited?	¿Ha vomitado?	0	4.0	1	1	1

Tableau 6.13 : Extrait de l'évaluation phrase par phrase

Ce court extrait montre immédiatement que les résultats bruts notamment pour BLEU ne peuvent être utilisés tels quels pour *MedSLT*. En effet, on constate que pour toutes ces phrases, le score pour BLEU4 est de zéro. BLEU4 correspond à la version couramment utilisée de BLEU qui calcule la précision des n-grammes sur les unigrammes, les bigrammes, les trigrammes et les quadrigrammes. Ceci est problématique avec notre corpus, puisque comme nous l'avons mentionné plus tôt, nous sommes en présence de nombreuses phrases très courtes. En effet, 20% des phrases du corpus, c'est-à-dire 46/222, comptent moins de quatre mots, contre 14%

pour GT, ce qui explique les nombreux scores de zéro pour *MedSLT* alors que les évaluateurs humains donnaient la meilleure note (4). Ces cas représentent 10% de notre corpus, ce qui fausse tout de même beaucoup les résultats totaux pour cette évaluation automatique. Il est clair que les mesures comme BLEU ont été développées pour la TA où les phrases sont nettement plus longues. Ainsi, l'une des caractéristiques de la TAP, à savoir la brièveté des phrases traitées, semble incompatible avec BLEU4.

Une deuxième caractéristique des traductions produites par *MedSLT* influence les résultats obtenus avec les scores automatiques : *MedSLT* produit des traductions moins littérales et bien plus idiomatiques et fluides que GT, ce qui est d'ailleurs l'une des forces du système.

Cependant, les mesures automatiques qui se contentent de comparer les phrases de surface avec des références sans aller jusqu'au sens, comme c'est le cas pour WER, BLEU et même TER, pénalisent ce type de traduction. Comment évaluer de manière juste des traductions comme celles-ci : *MedSLT* traduit « *Do you have a rash* » par « *Tiene una erupción cutánea* » (« Avez-vous une irruption cutanée »), mais dans toutes les références humaines produites par les traducteurs de Dallas nous retrouvons au lieu de cette formulation générique, un régionalisme: « *Tiene sarpullido* » ou une version plus familière comme « *Tiene un picor* » (« Est-ce que ça gratte? ») ou encore une variation : « *Tiene urticaria* ». Un autre exemple de ce type, déjà mentionné dans cette thèse est la traduction pour « *What are you allergic to?* » que nous avons choisi de traduire par « *Qué le da alergias?* » (« Qu'est-ce qui vous donne des allergies ? »). Cette solution nous permet tout simplement d'éviter les problèmes d'ambiguïtés de genre (*alérgica* (fem.) / *alérgico* (masc.)). Nos traducteurs qui ont fourni les références n'avaient pas besoin de tenir compte de ce type de contrainte, que le développeur doit, lui, résoudre. Dans ces cas, seule une mesure sémantique peut apporter une évaluation équitable. Seule une mesure capable de tenir compte de synonymes et de régionalismes serait capable de détecter que ces phrases sont équivalentes, même si les n-grammes qui les composent sont différents. Ces observations montrent que les scores de BLEU dans le Tableau 6.12 ne reflètent sans doute pas la véritable qualité de la traduction, et par conséquent ne fournissent pas d'indication valable sur l'utilité du système. Nous verrons si la corrélation entre les mesures automatiques et humaines confirme ce sentiment.

6.2.3.3 Corrélation entre mesures automatiques et humaines – *MedSLT* et GT

Nous poursuivons l'étude de la section 6.2.2, mais sur *MedSLT* et GT et non plus sur V1 et V2 de *MedSLT*. En outre, nous avons ajouté les variantes BLEU-3 (fondé sur les trigrammes)

et BLEU-2 (fondé sur les bigrammes), après avoir observé que le BLEU classique (BLEU4, fondé sur les quadrigrammes) n'est sûrement pas adapté à nos données parce qu'elles contiennent trop de phrases de moins de 4 mots. Nous pourrions ainsi comparer ces résultats avec ceux obtenus dans la section 6.2.2.

Correlation	MedSLT	GT
BLEU-4vs H	0,13	0,26
BLEU-3 vs H	0,21	0,29
BLEU-2 vs H	0,33	0,22
TER vs H	-0,30	-0,21
WER vs H	-0,49	-0,26

Tableau 6.14 : Corrélation Pearson entre les différentes mesures

Le Tableau 6.14 montre que les mesures BLEU-2 et WER obtiennent les meilleures corrélations et non TER, comme nous l'attendions lorsque nous avons ajouté cette mesure par rapport à ses propriétés (section 3.3.2.2.2). Il est intéressant de noter également que la corrélation pour *GT* est bien moins bonne que pour *MedSLT*, mais que par contre, BLEU (3-grammes et 4-grammes) corrèlent mieux avec les jugements humains pour *GT*. BLEU-2 et WER obtiennent les meilleures corrélations pour *MedSLT*. Contre toute attente, TER obtient des résultats très proches des autres mesures classiques de comparaison à une référence, ainsi elle ne sera pas retenue. Nous poursuivrons par conséquent nos recherches vers une mesure automatique adaptée à la TAP de type *MedSLT* dans la direction de BLEU-2 et WER. Une piste à explorer serait de coupler par exemple BLEU-2 à une autre mesure, soit automatique nouvelle comme RTE (Padó, Galley et al. 2009) ou humaine. Dans l'état actuel, la solution automatique exploitée dans cette section ne permet pas de filtrer les TA issues de la TAP qui présentent un danger pour la communication entre le médecin et le patient. De plus, la piste exploitée ici demande de créer des traductions de référence, une tâche qui n'est pas anodine et qui peut se révéler coûteuse, comme nous l'avons vu dans les sections 6.1.2 et 6.2.2.

6.2.4 Conclusion de l'étape 2 : Q2_S2 et Q2_S3

Nous avons choisi d'investiguer l'utilisation de mesures automatiques pour évaluer la composante de TA de notre système, afin de voir si celles-ci permettent d'obtenir une indication de la qualité à l'usage des traductions telle que mesurés avec notre échelle HMS. Nous avons dans un premier temps appliqué la mesure HMS sur nos données du chapitre 5, en faisant la différence entre V1 et V2 puis, suite à la constatation que cette mesure ne permet pas de départager ces deux versions très proches, nous avons décidé d'étendre l'application de HMS

en comparant *MedSLT* en général (sur un corpus de phrases issus de V1 et V2) à Google Translate.

Notre objectif est de donner les moyens aux utilisateurs de déterminer assez vite quel système de TAP leur est plus utile, ou, plus particulièrement pour le domaine médical, lequel représente le moins de danger. Cet objectif n'est atteint ni avec les mesures automatiques étudiées, ni totalement avec notre échelle HMS qui est sujette aux problèmes usuels des évaluations humaines qui s'appuient sur un petit groupe d'évaluateurs : le manque de cohérence et le coût important en termes de temps et d'argent pour la mettre en place.

Du côté des mesures automatiques, les résultats obtenus confirment les critiques vues dans la section 6.2 sur la très répandue mesure BLEU. Celles-ci nous avaient encouragée à tester d'autres mesures automatiques plus récentes comme TER et METEOR. Cependant, force est de constater que pour TER notamment, la corrélation avec HMS est plus faible que celle obtenue avec BLEU2.

Plusieurs pistes s'offrent à nous à ce stade : les nouvelles mesures automatiques basées sur l'estimation de la qualité de la TA (Specia, Raj et al. 2010, Specia, Hajlaoui et al. 2011), sur la qualité sémantique (Kos et Bojar 2009, Padó, Galley et al. 2009) ou encore sur l'analyse de l'erreur de la TAP (Condon, Parvaz et al. 2010) ou de la TA (Lommel, Popović et al. 2014). Cependant, selon les critères de (Koehn 2010) sur l'évaluation, l'analyse des erreurs ne satisfait pas non plus au critère de rapidité et de coût d'application peu élevé, puisque l'annotation se fait pour l'instant principalement manuellement, ou, dans le meilleur des cas, de manière semi-automatique (Popović, Arčan et al. 2015).

Une dernière piste est de recourir à l'évaluation humaine, mais de tenter d'en éviter les écueils, notamment le coût important et le manque de fiabilité dû au manque d'accord entre annotateurs (cf. section 6.1.1 et 6.2.1), ceci en faisant appel au *crowdsourcing*. C'est cette piste que nous nous proposons maintenant d'exploiter comme exposé dans l'introduction à ce chapitre et dans le chapitre 4.

6.3 Étape 3 : crowdsourcing pour évaluation TAP- TURKoise

En explorant la possibilité d'utiliser des mesures automatiques pour arriver à un résultat comparable à ceux obtenus avec notre échelle HMS (section 6.1.1), nous avons constaté qu'aucune mesure automatique ne corrèle suffisamment et que, par ailleurs, celle-ci ne sont pas adaptées à l'évaluation de la TAP médicale linguistique. Par conséquent, nous suggérons une alternative qui repose sur le *crowdsourcing* afin de proposer une évaluation humaine (plus

adaptée à notre contexte) mais qui serait débarrassée de ces principaux travers (manque de cohérence, coût et temps).

6.3.1 Crowdsourcing

Commençons par une définition de ce concept. Comme nous l'avons vu dans le chapitre 1 de cette thèse, *crowdsourcing* est un anglicisme communément employé au lieu du terme « **externalisation ouverte**¹³¹ », proposé par le Grand Dictionnaire québécois et le bureau de terminologie de l'administration fédérale suisse, et du terme « **production participative** », préconisé par les autorités terminologiques françaises et publié dans le journal officiel¹³² (section 4).

Le néologisme *crowdsourcing* (CS) a été créé en 2006 par Jeff Howe dans la revue *Wired*. à partir du mot *outsourcing*, qui consiste à externaliser des tâches en sous-traitance, en ajoutant le concept de foule (*crowd*, cf. sur ce concept le livre « *The wisdom of crowds* » (Surowiecki 2005)) afin de refléter la possibilité d'exploiter la créativité, l'intelligence et le savoir-faire d'un grand nombre d'internautes (Zhao et Zhu 2014).

(Eskenazi, Levow et al. 2013) donnent deux définition du CS : la première d'ordre opérationnel et la seconde d'ordre fonctionnel. La première peut être résumée de la façon suivante : 1. une tâche doit être effectuée, 2. il y a moyen de convaincre des non-experts de l'effectuer et 3. cette tâche est annoncée par un appel ouvert. La deuxième définition repose sur les quatre principales caractéristiques de la « foule intelligente » selon (Surowiecki 2005) : 1. la diversité d'opinions, 2. l'indépendance de chaque opinion, 3. la décentralisation de l'information (aucun membre de la foule n'a besoin d'avoir accès à toutes les informations pertinentes pour effectuer la tâche) et 4. les diverses opinions des membres de la foule peuvent être mises ensemble pour former une solution collaborative. Ces caractéristiques permettent à la foule d'arriver à une solution proche de celle formulée usuellement par des experts, d'où l'intérêt pour notre recherche et pour la recherche de manière générale.

Dans notre cadre de travail, nous considérons ce concept de CS depuis le même angle que l'un des acteurs principaux dans ce domaine : *Amazon*. Le site Service Web d'Amazon¹³³ donne

¹³¹ Définition du terme « externalisation ouverte » issue de Termdat <https://www.termdat.bk.admin.ch/Search/Search>, (site terminologique de l'administration suisse), et du Grand dictionnaire <http://www.granddictionnaire.com/> consultés le 28 juillet 2014 : Pratique qui consiste pour une organisation à externaliser une activité, par l'entremise d'un site Web, en faisant appel à la créativité, à l'intelligence et au savoir-faire de la communauté des internautes, pour créer du contenu, développer une idée, résoudre un problème ou réaliser un projet innovant, et ce, à moindre coût.

¹³² Définition du terme « production participative » issue du site : <http://www.culture.fr/franceterme/>, consulté le 28 juillet 2014 : Mode de réalisation d'un projet ou d'un produit faisant appel aux contributions d'un grand nombre de personnes, généralement des internautes.

¹³³ Site Amazon Web Services, Amazon Mechanical Turk : <https://aws.amazon.com/fr/mturk/details/>, consulté le 08.05.2014.

la description suivante de son service de *crowdsourcing* : il s'agit d'un accès facilité, grâce à la mise à disposition d'une interface web, à une communauté virtuelle de travailleurs. Cet outil permet de proposer à des travailleurs à distance des tâches à effectuer. L'intérêt est la possibilité de proposer à des humains des tâches encore difficiles à accomplir par des machines.

Le CS est désormais utilisé dans de nombreux domaines, et notamment dans le domaine des sciences computationnelles mais également dans le commerce (p. 425, (Zhao et Zhu 2014). En TAL, l'essor de cette méthode remonte à 2009, avec l'application à l'annotation (Sheng, Provost et al. 2008) , à l'étiquetage (Jurcicek et al, 2011), puis à l'utilisation de CS pour la collecte de données orales (Callison-Burch et Dredze 2010) et enfin tout récemment l'application à l'évaluation de logiciels (Sherief, Jiang et al. 2014) pour ne donner que quelques exemples.

L'idée d'étendre l'emploi des AMT à l'évaluation de la TA date de (Callison-Burch, Monz et al. 2009) et a été intégrée dans les campagnes d'évaluation IWSLT depuis 2010 (Paul et al, 2010 et Federico et al; 2012), comme nous l'avons vu dans la section 3.4.1.

Le principe de base de cette sous-branche du CS, également appelé « microworking » d'où le nom d'une des plateformes existantes autres que AMT, Microworkers.com¹³⁴, est de proposer des petites tâches à une communauté de « travailleurs » (dans le cadre d'AMT : on utilise « workers » ou « turkers », dans le cadre de MiniFreelance.com¹³⁵, on utilise « freelancer », et enfin, dans le cadre de Crowdfunder.com¹³⁶, c'est le terme « contributor » qui est utilisé). Nous allons suivre le « jargon » de AMT puisque c'est la plateforme que nous avons choisi d'utiliser, étant la plateforme la plus utilisée (selon (Eskenazi, Levow et al. 2013) AMT est utilisé par 53,5% des recherches prises en compte dans leur étude). Ainsi, comme mentionné dans la section 4.2.3.1, les tâches à effectuer par les travailleurs sont découpées en sous-tâches appelées « *human intelligence tasks* » (HITS). Des études comme celle de (Kittur, Smus et al. 2011) citée par (Eskenazi, Levow et al. 2013) dans la section 2.5.6 du livre ont montré que « *breaking large tasks up into several smaller subtasks will optimize accuracy* » (Eskenazi, Levow et al. 2013).

Nous avons choisi d'utiliser le CS parce que la plateforme AMT nous permet de faire appel à une multitude de personnes pour évaluer nos traductions à un moindre coût. Ceci nous

« Mechanical Turk donne accès à une communauté virtuelle de travailleurs qui sont disponibles pour vous aider à réaliser vos objectifs. Un ensemble robuste d'API et d'outils de ligne de commande vous permet de répartir, par programmation, les tâches qui nécessitent une intelligence humaine sur une main-d'œuvre mondiale à la demande.»

¹³⁴ Site de Microworkers, <https://microworkers.com/>, accédé le 01.07.2015.

¹³⁵ Site de Minifreelance, <http://minifreelance.com/>, accédé le 01.07.2015.

¹³⁶ Site de Crowdfunder, <http://www.crowdfunder.com/>, accédé le 01.07.2015.

donne le potentiel de multiplier le nombre d'évaluateurs par rapport à l'évaluation classique faite avec des experts tout en approchant la qualité de l'évaluation faite par les experts selon (Eskenazi, Levow et al. 2013) : « [...] *it is believed that the aggregate opinion of many nonexperts will approach the quality of the opinion of an expert. It is also believed that the use of nonexperts in this manner will be less onerous and more rapid than the use of experts* ».

C'est exactement ce que nous recherchons en recourant à AMT. En effet, le CS nous évite par exemple de consacrer du temps et de l'argent à la création de traductions de référence. D'autre part, dans notre contexte d'utilisation particulier, les personnes travaillant pour AMT (*turkers* ou travailleur AMT) sont des utilisateurs potentiels des logiciels de TAP, ils sont plus proches de la population d'utilisateurs potentiels des logiciels à l'étude.

Nous avons vu une série d'avantage des AMT et du CS en général, cependant, il existe aussi un certain nombre de réserves qui font hésiter (ou ont fait hésiter) les chercheurs à les utiliser dans un premier temps, notamment la fiabilité. En effet, il faut avoir un moyen de contrôler la qualité du travail effectué par des personnes inconnues. Il est probable, que parmi les travailleurs, il y ait des personnes qui ne compléteront pas les tâches de manière satisfaisante, que ce soit pour des raisons de compétence ou de motivation. Ainsi il est indispensable d'inclure un contrôle de qualité d'une forme ou d'une autre (le plus souvent en comparant les résultats d'un prétest effectué par le travailleur avec une référence fiable - « *goldstandard* » - fournie par des experts) afin de « *weed out the work of poor workers (who have good intentions, but who furnish work that is not of good quality) and malicious workers (those who randomly enter answers or automated bots.* » (Eskenazi, Levow et al. 2013).

Une question qui découle de cette réserve est : « La somme payée a-t-elle une influence positive ou négative sur la qualité du travail effectué ? ». (Eskenazi, Levow et al. 2013) fait référence à plusieurs études sur cette question et cite notamment (Marge et al 2010)¹³⁷ qui confirment notre impression que ce n'est pas forcément le travailleur le mieux payé qui travaille le mieux, sans doute à cause de l'appât du gain (en moyenne environ 70% des donneurs d'ordre - ou « *requesters* » en jargon AMT - proposent une rétribution de 0,05 dollar ou moins par HIT). Celui-ci ne doit pas être le seul moteur du travailleur, mais peut en effet être un bon moyen de fidéliser les travailleurs, comme nous le verrons dans notre expérience ci-dessous.

En résumé, le CS pour notre cas de figure se traduit par la possibilité de recruter facilement des évaluateurs afin d'atteindre une masse critique qui compenserait le manque de

¹³⁷ « higher pay led to more errors » *ibid.*

fiabilité ou de cohérence entre les évaluateurs humains experts. En outre, puisqu'il s'agit d'un service Internet, il est synonyme de rapidité. Un autre avantage probable est que les évaluateurs du *crowdsourcing* présentent un profil plus proche de l'utilisateur final des systèmes de TAP à l'étude que les évaluateurs experts, qui sont la plupart du temps des experts linguistiques (académique), des chercheurs, des traducteurs ou encore des étudiants. Enfin, par rapport à l'évaluation automatique classique (qui repose sur la ressemblance à des traductions de références, BLEU, NIST, WER, METEOR, TER...), le fait de recourir à des évaluateurs du *crowdsourcing* permet d'éviter de créer des références (inconvenients décrits dans section 6.2.3).

6.3.2 Résultats de l'expérience EXPQ2_S4-S6

Dans cette section, nous donnons les résultats de l'expérience **EXPQ2_S4-S6**, décrite dans la section 4.2.3.2, dont le but est de répondre à notre question : *La méthode d'évaluation de la TAP TURKoïse représente-t-elle une alternative intéressante aux méthodes automatiques ou humaines classiques ?*

Pour rappel, cette question générale est composée de plusieurs sous-questions de recherche. Nous nous pencherons tout d'abord sur **Q2_S4** (*L'implémentation de TURKoïse est-elle faisable ?*) et sa sous-question **Q2_S4_S1** (*Est-ce faisable de recruter des travailleurs AMT pour notre tâche ?*). En effet, pour effectuer notre tâche, nous devons recruter des travailleurs AMT bilingues, ce qui pourrait être un problème. La réponse à cette sous-question est présentée dans la section 6.3.2.1.

Ensuite, nous aborderons la question de la fiabilité des travailleurs recrutés (**Q2_S5**). Cette question est subdivisée en trois sous-questions, auxquelles nous répondrons dans la section 6.3.2.2.

Enfin, dans la section 6.3.2.3, nous répondrons à **Q2_S6**, qui porte sur les bénéfices de TURKoïse en termes d'économies de temps et d'argent.

6.3.2.1 Q2_S4 : Faisabilité du recrutement sur AMT

Nous avons eu suffisamment de participants pour l'accomplissement de la tâche de notre étude pilote (**EXPQ2_S4-S6**), et ce dans des délais courts et malgré le fait que nous recherchions des bilingues anglais-espagnol. En effet, en moins de 24 heures, nous avons attribué des qualifications à 20 travailleurs qui correspondaient à nos prérequis. Il faut se référer

à la section 4.2.3.2.1 pour voir les modalités de recrutement des travailleurs AMT. En l'espace de trois jours, tous les HITS (correspondant à cinq évaluations au total) avaient été accomplis.

Pour la deuxième phase de l'étude (EXPQ2_S7), qui visait à augmenter le nombre d'instances, seulement trois des 13 participants invités ont répondu favorablement à la requête en l'espace d'une semaine. Nous aurions pu rouvrir la tâche de sélection afin de recruter de nouveaux travailleurs AMT qualifiés mais nous avons préféré nous limiter à huit évaluations au lieu des dix prévues initialement, en continuant de travailler avec les travailleurs AMT déjà sélectionnés.

Par conséquent, nous pouvons répondre à **Q2_S4_S1** en affirmant qu'il est possible de recruter des travailleurs AMT même pour une tâche demandant certaines compétences spécifiques, dans notre cas le fait d'être suffisamment bilingue (anglais-espagnol) pour pouvoir évaluer les phrases des HITS.

6.3.2.2 Q2_S5 : Fiabilité des travailleurs AMT

L'objectif de la présente section est de répondre à la question suivante : une fois qu'une sélection en vue de notre tâche a été effectuée, les travailleurs AMT recrutés sont-ils fiables ? Tout d'abord, nous avons pu observer un indice de fiabilité que nous n'avions pas prévu : plusieurs de nos travailleurs AMT nous ont contactée par courrier électronique pour nous poser des questions relatives à la tâche d'évaluation, ce qui indique qu'ils ont effectué la tâche avec sérieux. Par ailleurs, chaque travailleur AMT a effectué plusieurs tâches (HIT) ; ainsi, parmi les travailleurs AMT sélectionnés c'est un sous-ensemble qui a effectué toutes les tâches proposées. Cette fidélité peut être interprétée comme un nouveau gage de fiabilité.

Outre ces remarques, dans notre étude, nous mesurerons la fiabilité des travailleurs AMT sélectionnés par rapport à leur cohérence (**Q2_S5_S1**, kappa et ICC) mais également en comparant leurs résultats avec les résultats obtenus par un groupe de contrôle d'experts (**Q2_S5_S2**). Enfin, nous analyserons si les résultats obtenus avec TURKoise sont comparables aux résultats obtenus avec une évaluation humaine classique de l'adéquation et de la fluidité (**Q2_S5_S3**).

Nous présenterons les résultats obtenus pour ces trois questions de la manière suivante : nous comparerons tout d'abord les résultats obtenus pour les mesures classiques (fluidité et adéquation), puis ceux obtenus avec HMS d'une part avec les travailleurs AMT (TURKoise) et d'autre part avec les évaluateurs experts (HMS). Pour chaque type d'évaluation, les résultats seront accompagnés des mesures de cohérence (pourcentage d'accord unanime et kappa de

Fleiss). Dans les tableaux de résultats des sections suivantes, la deuxième colonne présente les résultats obtenus par nos experts-maison et la troisième colonne ceux des travailleurs AMT.

Fluidité

Le Tableau 6.15 ci-dessous montre une nouvelle fois (section 5.2.3.1) la difficulté et la subjectivité de l'évaluation de la fluidité. En effet, le pourcentage de phrases pour lesquelles une majorité d'évaluateurs sont d'accord entre eux n'excède pas 84% pour les experts- maison (traducteurs) et 79% pour les travailleurs AMT, le kappa de Fleiss est très bas tout en étant légèrement meilleur pour les experts-maison.

Fluidité	Experts-maison (5)	Travailleurs AMT (5)
Accord unanime	21%	22%
4 en accord	32%	29%
3 en accord	31%	28%
Majorité d'accords	84%	79%
Kappa de Fleiss	0.174	0.164

Tableau 6.15 : Fluidité pour expert- maison comparés aux travailleurs AMT

Le meilleur accord entre évaluateurs est obtenu pour les phrase assortie du meilleur score de fluidité (5 sur 5). Cette observation s'applique aux deux types d'évaluateurs (expert-maison et travailleurs AMT). Nous verrons plus loin que l'évaluation de la fluidité ressort clairement comme étant la plus subjective ou aléatoire, au vu de la variation observée dans les réponses et le manque d'accord entre évaluateurs. L'utilisation du CS pour ce type de tâche pourrait être bénéfique étant donné que la multiplication des évaluateurs permet de compenser le manque d'accord (section 3.3.2.1.2).

Adéquation

Le Tableau 6.16 présente les résultats pour l'adéquation. Les résultats divergent clairement entre le pourcentage de majorité d'accord et le kappa de Fleiss : pour le premier, les évaluateurs experts obtiennent 93% et les travailleurs AMT 86%, alors que ces derniers obtiennent un meilleur kappa de Fleiss (0.236) que les experts-maison (0.121).

Adéquation	Expert-maison	Évaluateurs AMT
Accord unanime	26%	36%
4 en accord	40%	27%
3 en accord	27%	24%
Majorité d'accords	93%	86%
Kappa de Fleiss	0.121	0.236

Tableau 6.16 : Adéquation pour experts maison comparés aux travailleurs AMT

Si l'on se fie au score kappa, ces résultats semblent encourager l'utilisation de l'AMT. Cependant, ces résultats doivent être nuancés par le fait que le choix du kappa de Fleiss est discutable (section 3.3.2.1.2). Nous aimerions appliquer ici l'ICC (Intraclass Correlation Coefficient) pour compléter les résultats en distinguant bien la situation de « *fully crossed design* », qui s'applique aux évaluateurs-maison puisque chaque évaluateur a traité toutes les phrases, par opposition aux AMT et interprètes de Dallas, dont chacun n'a évalué qu'une partie du corpus total. Nous passons maintenant aux résultats pour HMS (évaluateurs experts) et TURKoïse (HMS avec travailleurs AMT).

Échelle HMS par des experts : Traducteurs vs. Non traducteurs

Comme nous l'avions annoncé à la section 4.2.2.3.1, nous avons décidé de compléter l'évaluation avec des évaluateurs non-traducteurs, comme mentionné plus haut, afin de vérifier la tendance observée selon laquelle les évaluateurs traducteurs ont tendance à être plus sévères que les non-traducteurs (Chatzichrisafis, Bouillon et al. 2006). Le Tableau 6.17, ci-dessous, montre qu'il est difficile d'obtenir un accord entre les 6 évaluateurs (18.5%).

Évaluateurs experts	% d'accord unanime
Les 6 évaluateurs	18.5%
Traducteurs (3)	33.8%
Non-traducteurs (3)	41.9%

Tableau 6.17 : Pourcentage d'accord unanime entre évaluateurs

Il ressort ce tableau que les évaluateurs non-traducteurs sont plus cohérents que le groupe de traducteurs. Cela suggère que, contrairement à la pratique courante, les traducteurs, qui sont souvent choisis pour cette tâche parce qu'ils sont compétents dans les deux langues évaluées, ne sont pas forcément les évaluateurs les plus représentatifs de la qualité réelle en termes de qualité à l'usage. Souvent, ils se focalisent trop sur la forme - style et syntaxe - même si l'échelle

HMS doit normalement éviter ce biais, observé également par (Costa-Jussà, Farrús et al. 2012) dans le cadre des comparaisons entre la TA linguistique et statistique¹³⁸.

Dans la section suivante, nous comparerons ces résultats avec ceux obtenus par les travailleurs AMT, c'est-à-dire ce que nous avons appelé TURKoïse.

TURKoïse

Dans cette section, nous décrirons les résultats obtenus pour les trois tâches en comparant les différents groupes d'évaluateurs (experts-maison et travailleurs AMT) lorsqu'ils appliquent HMS.

Les résultats pour TURKoïse (Tableau 6.18) confirment que les travailleurs AMT obtiennent des meilleurs niveaux d'accord que ceux obtenus par les experts-maison lors de l'évaluation de l'adéquation (32% contre 15% pour les experts-maison). Encore une fois, le kappa de Fleiss pour les travailleurs AMT, avec 0,232, est faible selon la grille d'interprétation de (Landis & Koch, 1977). Cependant, ce résultat reste plus élevé que celui obtenu par les évaluateurs experts-maison (0,199).

TURKoïse	Experts-maison	Travailleurs AMT
Accord unanime	15%	32%
4 en accord	35%	26%
3 en accord	42%	37%
Majorité d'accords	92%	95%
Kappa de Fleiss	0.199	0.232

Tableau 6.18 : TURKoïse

Dans la plupart des cas, les scores attribués par les évaluateurs ne diffèrent que d'un point. Ils sont par conséquent d'accord avec la qualité générale des phrases évaluées.

¹³⁸ p. 266 : "These results led us to two considerations: in the first place, humans rely most on syntactic and semantic errors than in orthographic, morphological errors and lexical errors. Second, the percentage of cases is directly proportional to the depth of the language structure: orthographic level is related to the most superficial structure, while syntactic level is related to the deepest one, so that human evaluators use to penalise strongly those errors committed at deeper levels", Costa-Jussà, M. R., M. Farrús, J. B. Mariño et J. A. R. Fonollosa (2012). "Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems." *Computational Informatics* 31(2): 245–270.

%	CCOR	MEAN	NONS	DANG	Total
Nombre total d'éval.	735	301	28	44	1108
Unanime	31%	1%	0.0%	0.0%	32%
4 d'accord	20%	5%	0.5%	1.4%	26%
3 d'accord	22%	14%	0.0%	1.8%	37%

Tableau 6.19 : Accord par catégorie pour TURKoise.

Il ressort du Tableau 6.19 ci-dessus que l'accord est plus élevé pour la catégorie CCOR que pour NONS ou DANG. Cette observation rejoint celle faite au moment de l'évaluation de l'adéquation par les experts, selon laquelle l'accord était plus élevé pour les phrases obtenant les meilleurs scores. Il est probable que si nous regroupions les catégories CCOR et MEAN d'une part et les catégories NONS et DANG d'autre part, pour arriver à une échelle binaire, nous obtiendrions un accord entre évaluateurs plus élevé (comme observé dans le chapitre 5, section 5.2.3.1).

De la série de résultats présentés jusqu'ici (évaluations de la fluidité, de l'adéquation et HMS effectuées par les experts et les travailleurs AMT), nous arrivons aux conclusions suivantes concernant **Q2_S5** : premièrement (réponse à **Q2_S5_S1**), les travailleurs AMT atteignent des niveaux de cohérence comparables à ceux des experts, voire meilleurs ; deuxièmement, les résultats obtenus avec TURKoise sont comparables à ceux obtenus par les experts en appliquant HMS ; troisièmement, les résultats obtenus avec TURKoise (c'est-à-dire avec HMS) sont comparables aux résultats obtenus avec l'évaluation classique de l'adéquation et de la fluidité. Les niveaux de cohérence ne sont pas les mêmes, mais cela provient également tout simplement du fait qu'une échelle compte 4 points (HMS) et que les autres en comptent 5, ce qui rend la comparaison directe impossible.

Nous présenterons maintenant les résultats concernant Q2_S6, concernant la compétitivité de la méthode proposée.

6.3.2.3 Q2_S6 : Compétitivité au niveau du temps et des coûts

Pour le groupe d'évaluateurs maison (collègues ou traducteurs FTI) et pour le groupe de non-traducteurs (amis bilingues), chaque évaluateur a évalué les 222 phrases. Aucun coût n'a été généré par ces deux groupes qui travaillaient sur base de bénévolat. D'autre part, selon la charge de travail respective des participants auxquels nous avons fait appel, tous ont répondu dans une période allant d'une heure à 24 heures.

Pour le deuxième groupe (interprètes de Dallas, cf. section 4.2.1.2), l'évaluation s'est faite par email en découpant le corpus de phrases total en série de 190 phrases afin d'obtenir au

moins trois évaluations par phrase. Cette manière de procéder pourrait s'apparenter à une tâche de CS (avant l'heure, puisqu'elle a été effectuée en 2008). En effet, nous ne connaissons pas ces interprètes et, surtout, chaque évaluateur n'a pas évalué toutes les phrases (comme les travailleurs AMT). Nous demandions cependant une tâche bien plus longue à effectuer qu'un HIT d'AMT, puisque les interprètes devaient également nous fournir une traduction de référence pour chaque phrase (c'est ainsi que nous avons récolté nos phrases de références utilisées pour l'évaluation automatique, voir section 5.2.3.2). La comparaison avec AMT s'arrête là, parce que cette tâche nous a pris bien plus de temps et d'effort que la mise en place des tâches sur AMT.

Si nous n'avons malheureusement pas conservé d'indication précise du temps nécessaire, la récolte des données de Dallas s'est poursuivie sur plus de deux semaines selon la correspondance de mail effectuée à cette époque. Enfin, il faut noter que cette tâche était indirectement payée comme ce sont les mêmes interprètes qui ont participé à l'expérience EXPQ1. Chaque interprète rémunéré n'a cependant pas pris le temps de participer à cette deuxième partie de la tâche (évaluation et traduction de référence).

Évaluateurs	Sous-groupe	Nb. d'éval.	Nb. d'évaluateurs	Nb. de phr.	Temps	Coûts
Groupe Experts - maison	Traducteurs	5	5	222	< 24 h	Gratuit
	Non-traducteurs	3	3	222	< 24 h	Gratuit
Groupe Experts- de l'hôpital de Dallas	Interprètes de l'H. Dallas	3	14 au total	190	> 2 semaines pour la tâche	14*50\$ =700\$ au total (pour l'ensemble des tâches)
Travailleurs AMT	EXPQ2_S4-S6	5	11/20 AMT recrutés	222	recrutement < 3jours tâche	Env. 50\$
	EXPQ2_S7	3	3/13 AMT invités mail	222	une semaine	Env. 50 \$

Tableau 6.20 : Tableau récapitulatif du temps et des coûts

Le Tableau 6.20 reprend les principales informations quant aux temps mis et aux coûts engendrés pour les évaluations d'experts et pour les évaluations AMT (TURKoise). Il ressort de ce tableau qu'en ce qui concerne la rapidité du recrutement, les différences ne sont pas très grandes. En effet, dans les deux cas, 24 heures ont suffi pour trouver assez d'évaluateurs. Cependant, il faut bien souligner que notre contexte est particulier, puisque nous avons la chance de compter de nombreuses personnes bilingues dans notre entourage (traducteurs mais

aussi non-traducteurs). Ainsi, il est intéressant de pouvoir s'appuyer sur une plateforme d'évaluateurs accessibles facilement et rapidement, même lorsqu'il s'agit de bilingues. Il faut noter que la combinaison linguistique étudiée joue en notre faveur. Il faudrait tester si le recrutement est faisable pour d'autres paires de langues.

En ce qui concerne les coûts engendrés, les travailleurs AMT présentent bien sûr un coût plus élevé que les évaluateurs bénévoles que nous avons recrutés. Cependant, si l'on compare avec les coûts engendrés par les évaluateurs de Dallas (50 dollars par personne pour la participation à l'expérience, la création de traductions de référence et l'évaluation des phrases, ce qui fait au total 700 dollars), il apparaît très nettement que cela revient moins cher de passer par une plateforme comme AMT (100 dollars au total pour les deux expériences).

Nous pouvons par conséquent conclure en répondant à **Q2_S6** par l'affirmative : les bénéfices du *crowdsourcing* en termes de compétitivité (coûts et temps réduits) sont réels.

6.3.3 Résultats de l'expérience EXPQ2_S7

Lorsqu'on a la possibilité de collaborer avec des travailleurs AMT, aisément disponibles en grand nombre et peu chers (du moins en comparaison avec les évaluateurs experts-maison, la FTI étant elle-même une source tarissable), la tentation de multiplier le nombre d'évaluateurs pour compenser l'inhérente incohérence de l'évaluation humaine est grande. Dans cette section, nous voulons par conséquent évaluer le nombre d'évaluations nécessaires avant d'arriver à un seuil.

Dans le tableau ci-dessous, nous présentons les résultats obtenus lors de la deuxième phase d'AMT en indiquant les scores pour 3, 5 et 8 instances, ceci pour les trois échelles étudiées. A titre de comparaison, nous incluons le kappa obtenu pour ces mêmes évaluations avec les cinq évaluateurs experts-maison.

Nombre d'éval totales	Fluidité	Adéquation	TURKoise
3 instances. AMT	-0,05	0,14	0,18
5 instances. AMT	0,16	0,24	0,23
8 instances. AMT	0,13	0,23	0,23
5-eval. Experts-maison	0,17	0,12	0,20

Tableau 6.21 : Évolution du kappa selon le nombre d'évaluations

Il est intéressant de constater que le kappa le plus bas corresponde à l'évaluation de la fluidité (cf. section 6.3.2.2.1).

De manière plus générale, le tableau montre que TURKoise obtient des scores encourageants, les niveaux pour les travailleurs AMT étant comparables ou même meilleurs dès trois instances. Il est intéressant d'observer que les kappa plafonnent à partir de cinq instances d'évaluation et que l'ajout de trois évaluateurs supplémentaires n'améliore pas le score d'accord. D'autre part, pour une mesure comme la fluidité, trois instances ne sont clairement pas suffisantes : les travailleurs AMT sont en léger désaccord selon (Landis & Koch, 1977).

Pour TURKoise, il semble possible de conclure que, dès cinq instances, les travailleurs AMT obtiennent des scores comparables à ceux obtenus par les évaluateurs experts, ce qui confirme (Eskenazi, Levow et al. 2013).

La réponse à **Q2_S7** est par conséquent qu'il n'est pas nécessaire de multiplier le nombre d'évaluateurs, puisque la cohérence pour les évaluations effectuées par les travailleurs AMT semble plafonner à partir de cinq instances.

6.3.4 Conclusion de l'étape 3 : TURKoise

Les résultats de l'étape 3 nous permettent de répondre par l'affirmative à notre question : *La méthode d'évaluation de la TAP TURKoise représente-t-elle une alternative intéressante aux méthodes automatiques ou humaines classiques ?*

En effet, nous avons vu dans la section 6.3.2.1 que le recrutement d'un important nombre d'évaluateurs bilingues (20 personnes qualifiées recrutées en moins de 24h) est tout à fait faisable (**Q2_S4** et **Q2_S4_S1**).

Dans la section 6.2.2.2, concernant la fiabilité des travailleurs AMT recrutés (**Q2_S5**), nous avons conclu que, sur les points étudiés (cohérence (**Q2_S5_S1**), résultats comparables à ceux des experts (**Q2_S5_S2**) et aux mesures classiques (**Q2_S5_S3**)), les résultats obtenus par les travailleurs AMT sont équivalents, voire meilleurs que ceux obtenus par nos plus petits groupes d'évaluateurs experts. L'expérience **EXPQ2_S4-S6** prouve bien qu'il est important, en présence d'évaluation humaine, d'arriver à une masse critique d'évaluateurs pour arriver à des taux de cohérence satisfaisants. L'avantage des plateformes de *crowdsourcing* est de faciliter la mise en place d'évaluation à plus grande échelle qu'avec des groupes d'experts, de manière plus rapide et à moindres frais.

Cependant, l'**EXPQ2_S7** nous a également montré qu'il n'est pas nécessaire de multiplier sans fin le nombre d'évaluateurs, puisque en comparaison avec les évaluateurs

experts, nous arrivons à un plafond pour la cohérence (mesurée avec kappa de Fleiss) à partir de 5 évaluations issues de différents évaluateurs. Nous verrons dans la conclusion générale qu’il serait souhaitable d’approfondir ces résultats en appliquant, comme nous l’avons fait pour le chapitre 5, d’autres mesures de cohérence telles que l’ICC.

Cependant, pour terminer ce bilan de l’étape 3, nous pouvons conclure que les résultats pour TURKoïse, et de manière générale pour l’évaluation par le biais du *crowdsourcing*, sont plutôt encourageants.

6.4 Conclusion pour Q2_principale

Dans la première étape, nous avons proposé une nouvelle évaluation humaine fondée sur l’utilité finale de la TAP (*Human Medical Scale*, HMS), que nous avons ensuite appliquée à nos versions V1 et V2, comme dans le chapitre 5. Les résultats obtenus montrent que l’échelle HMS permet de mieux trancher entre les deux systèmes V1 et V2 que les mesures classiques de fluidité et d’adéquation (section 5.2.3.1). Cependant, nous avons constaté également que le niveau de cohérence entre les évaluateurs est assez bas. Ce qui nous a incitée à rechercher une évaluation automatique qui pourrait corrélérer avec notre échelle afin de moins dépendre de la subjectivité de l’évaluation humaine.

Dans la deuxième étape (section 6.2), nous avons ainsi appliqué, dans un premier temps, les mesures automatiques classiques à V1 et V2 afin de trouver celle qui corrèle le mieux avec l’échelle HMS. Nous avons observé que les résultats et les corrélations étaient très faibles. Cela nous a poussée, dans un deuxième temps, à étendre cette étude à des données provenant de la TA statistique de GT afin de vérifier si les mesures automatiques classiques sont moins adaptées à l’évaluation de la TAP linguistique. La conclusion de cette étape était qu’en effet, la mesure classique BLEU (et notamment BLEU4) n’est pas adaptée à notre contexte de TAP du domaine médical, d’une part en raison de la brièveté de nos phrases et d’autre part, de leur nature (phrases issues d’un traitement linguistique par interlangue). Ainsi, les résultats obtenus confirment les critiques exposées dans la section 6.2 sur la très répandue mesure BLEU. En outre, ces résultats nous ont encouragée à tester d’autres mesures automatiques plus récentes comme TER et METEOR, sans obtenir de meilleures corrélations avec HMS. TER obtient une corrélation avec HMS plus faible que celle obtenue avec BLEU2. À partir de ces résultats, nous avons réorienté nos recherches vers une évaluation humaine qui corrigerait les travers de HMS. C’est ainsi que nous avons proposé TURKoïse (échelle HMS combinée au CS) afin de permettre l’évaluation équitable, rapide et peu coûteuse des systèmes de TAP dans le domaine médical.

Notre troisième étape nous a permis de conclure que la méthode TURKoise proposée est plus économique et plus rapide que la méthode EAGLES-ISO dans le chapitre précédent, tout en reflétant bien la qualité à l'usage de la TA produite par la TAP du domaine médical. En général, le *crowdsourcing* semble être une alternative à la fois efficace et peu onéreuse effectuée par de petits groupes d'experts. Ainsi, même en augmentant le nombre de travailleurs, les coûts restent très compétitifs.

Nous pouvons ainsi conclure que TURKoise pourrait avantageusement remplacer les mesures classiques de la fluidité et de l'adéquation pour l'évaluation de la TAP dans le domaine médical. Enfin, pour ce qui est de l'évaluation de la composante TA de la TAP, il est intéressant de noter que TURKoise pourrait apporter une réponse plus équitable que les mesures automatiques lorsqu'il s'agit de comparer des systèmes linguistiques et statistiques.

En résumé, nous avons effectué un cheminement en trois étapes avant d'arriver à la méthode TURKoise et montrer qu'il est possible de faire une évaluation manuelle efficace et peu coûteuse grâce au *crowdsourcing*. Nous pouvons ainsi affirmer que la méthode d'évaluation de la TAP TURKoise représente bien une alternative intéressante aux méthodes automatiques ou humaines classiques.

7 Conclusion générale

Le sujet de la présente thèse a beaucoup évolué au fil du temps. Nous sommes partie d'un point de vue de développeur –comme nous étions impliquée dans le développement de *MedSLT* (système de traduction automatique de la parole –TAP– dans le domaine médical), notamment pour la version bidirectionnelle anglais-espagnol que nous avons utilisée dans les expériences menées dans le cadre de ce travail– pour arriver à un point de vue axé sur l'utilisateur final. En partant de cette expérience de développement, nous avons décidé de faire évoluer ce travail vers l'évaluation des logiciels de TAP et même vers la méta-évaluation des méthodes d'évaluation.

7.1 Principaux résultats

Comme décrit dans le chapitre de méthodologie (chapitre 4), la thèse s'articule autour de deux volets qui répondent chacun à une question de recherche principale.

Notre première contribution, avec ce travail, est la mise sur pied d'une évaluation complète de deux versions de *MedSLT*, en appliquant la méthode EAGLES-ISO.

Dans ce **premier volet**, nous avons procédé en deux étapes. Dans la première, nous avons défini le modèle de qualité sur lequel repose notre méthode. Celui-ci comporte principalement des caractéristiques de qualité à l'usage mais il est complété par deux sous-caractéristiques de la caractéristique de capacité fonctionnelle, à savoir la précision et la facilité d'utilisation (section 4.1.1). Dans la deuxième étape, nous avons appliqué cette méthode en mettant en place une expérience de simulation de l'utilisation des logiciels en milieu hospitalier (**EXPQ1**). Celle-ci consistait à faire utiliser tour à tour les deux versions de *MedSLT* par de véritables médecins et des patients standardisés (section 4.1.2.1) afin de simuler le diagnostic d'une angine à streptocoques ou virale (section 4.1.2.2). Cette expérience a eu lieu en février 2008, au *Dallas Childrens' Hospital*, grâce à la collaboration avec le Prof. Glenn Flores. Il nous a permis d'effectuer des tests dans ses locaux avec la participation de médecins de son service et nous a mis à disposition un groupe de patients standardisés.

Les résultats de cette expérience sont exposés dans le chapitre 5. En résumé, la méthode EAGLES-ISO nous a bien permis d'évaluer équitablement les deux systèmes en parvenant à les départager, malgré leur fort degré de similitude. De manière générale, V2, qui comporte une couverture plus large a obtenu les meilleurs scores.

En ce qui concerne les caractéristiques de **qualité à l'usage**, V2 obtient de meilleurs scores que V1 pour l'efficacité, mesurée grâce au taux d'accomplissement de la tâche et

également par le rappel et la précision des informations collectées grâce à l'utilisation de *MedSLT* V1 et V2.

En matière de **qualité externe**, V1 et V2 arrivent à égalité pour la capacité fonctionnelle en termes de précision des sous-tâches qui composent les deux systèmes : RV, rétro-translation et TA. Cependant, V2 arrive clairement en tête lorsqu'on évalue la facilité d'utilisation des systèmes.

Le bilan que nous pouvons tirer de cette comparaison entre nos deux systèmes très proches est que c'est principalement la qualité à l'usage qui les distingue. Au niveau de la précision, la couverture des systèmes ne semble pas avoir un impact direct.

Par ailleurs, nous pouvons conclure que la méthode EAGLES-ISO employée nous a permis de départager ces deux systèmes, ce que nous n'aurions pas pu faire, si nous avions uniquement étudié leur capacité fonctionnelle comme c'est le cas la plupart du temps dans l'évaluation de la TAP (section 3.4).

Cependant, cette méthode a été longue et coûteuse à mettre en place. En effet, elle est le fruit d'un travail de longue haleine qui n'est pas vraiment réalisable dans d'autres contextes que le contexte académique.

Ainsi, le **deuxième volet** de ce travail, exposé dans le chapitre 6, a consisté à proposer une méthode alternative à EAGLES-ISO, qui serait également axée sur la qualité à l'usage mais pouvant être exécutée plus rapidement et à moindres frais. Nous sommes parvenue à notre contribution finale avec la méthode TURKoise après un cheminement en trois étapes.

La **première étape** consistait à proposer une évaluation humaine adaptée à la TAP dans le domaine médical grâce à l'échelle *Human Medical Scale* (HMS) que nous avons élaborée dans la section 6.1.1. L'objectif de cette échelle est de parvenir à évaluer l'utilité de la TA produite par un système de TAP, et ainsi évaluer la qualité à l'usage d'un système, tout en évaluant le produit fini obtenu en l'utilisant. Ensuite, afin de tester l'efficacité de HMS, nous l'avons appliquée sur nos données issues de l'expérience **EXPQ1**.

Lors de la **deuxième étape**, nous avons évalué les mesures automatiques qui pourraient corrélérer avec l'échelle proposée afin d'accélérer le processus d'évaluation. En effet, l'échelle HMS n'est pas dépourvue de l'inconvénient principal de l'évaluation humaine : la difficulté d'atteindre un accord entre évaluateurs. Par ailleurs, il est clair que l'évaluation humaine, comme celle proposée, est plus longue à mettre en place qu'une évaluation automatique (à condition d'avoir des traductions de référence à disposition, voir section 6.2.3.2). Elle est

surtout moins facile à répéter. Les résultats pour cette expérience **EXPQ2_S2** montrent que la corrélation entre les mesures automatiques étudiées et HMS est très faible.

Nous avons ainsi poussé les recherches plus en avant (**EXPQ2_S3**) en testant les mesures automatiques et HMS sur deux systèmes d’approches distinctes : *MedSLT* et *Google Translate*, pour les raisons évoquées à la section 6.2.3.3, à savoir si les mesures appliquées sont vraiment adaptées à la TAP linguistique ou s’il existe un biais en faveur de la TA statistique. Les principaux résultats obtenus sont que BLEU n’est clairement pas adapté à notre contexte d’utilisation, puisque dans sa version classique les scores reposent sur un calcul de quadrigrammes. Or, une large portion de notre corpus consiste en des phrases de moins de quatre mots. La mesure TER, que nous pensions plus adaptée à notre contexte, n’obtient pas une meilleure corrélation que les autres mesures étudiées. BLEU2 obtient la meilleure corrélation et pourrait être envisagé pour faire un premier tri automatique des TAP, à condition d’avoir la possibilité de créer facilement des traductions de référence.

Cependant, le bilan de l’expérience **EXPQ2_S3** reste très mitigé. Celle-ci a surtout contribué à confirmer que ces mesures automatiques ne semblent pas être très bien adaptées à notre contexte d’évaluation.

De ce constat nous nous sommes réorientée vers l’évaluation humaine mais avec l’objectif d’éviter ses écueils (coûts importants et manque de fiabilité dû au manque d’accord entre évaluateurs) en faisant appel au *crowdsourcing* (CS). Nous avons étudié cette piste lors de notre **troisième étape**, concernant TURKoïse. Cette méthode repose sur deux éléments centraux : l’échelle HMS et le CS.

Nous avons tout d’abord étudié les possibilités qu’offrent le CS mais également les pièges et inconvénients (section 6.3.1). Nous sommes arrivée à la conclusion qu’avant de pouvoir proposer une méthode qui repose sur le CS, il faut tester ses principales failles potentielles : la fiabilité mais aussi la faisabilité de ce type d’évaluation en ayant recours à une plateforme de *crowdsourcing* comme *Amazon mechanical turk* (AMT) (**EXPQ2_S4-S6**, section 4.2.3.2).

Les principaux résultats de cette troisième étape sont décrits dans la section 6.3. Ils peuvent être résumés ainsi : tout d’abord, la faisabilité du recrutement de travailleurs AMT pour une tâche nécessitant des compétences linguistiques en deux langues a été prouvée dans la section 6.3.2.1. Ensuite, concernant la fiabilité d’une évaluation par les travailleurs AMT, nous arrivons à la conclusion que les résultats obtenus par les évaluateurs AMT sont comparables à ceux obtenus par notre groupe d’évaluateurs experts, composé d’une part d’évaluateurs

traducteurs et d'autre part d'évaluateurs non-traducteurs. Nous arrivons à cette conclusion après comparaison du niveau de cohérence des évaluateurs lorsqu'ils appliquent, d'une part, les mesures classiques d'évaluation humaine de la TA (fluidité et adéquation) et, d'autre part, en utilisant l'échelle HMS.

La dernière expérience menée avec TURKoïse (**EXPQ2_S7**) a été d'explorer combien d'évaluations sont nécessaires pour arriver à une stabilisation du degré de cohérence des évaluateurs AMT. Nous avons montré qu'à partir d'un nombre relativement petit, de cinq évaluations, nous atteignons un seuil acceptable pour TURKoïse. Il n'est par conséquent pas nécessaire de multiplier le nombre d'évaluations, ce qui permet de circonscrire les dépenses engendrées par l'évaluation sur AMT (section 6.3.2.3).

Nous sommes ainsi arrivée au terme de notre cheminement, en réponse à notre deuxième question de recherche en proposant la méthode TURKoïse comme alternative viable et compétitive à la méthode EAGLES-ISO pour l'évaluation de la TAP dans le domaine médical.

7.2 Contributions

Au-delà de la principale contribution évoquée plus haut, nos apports au domaine de l'évaluation de la TAP dans le domaine médical sont les suivants :

- la proposition d'un modèle de qualité transposable à d'autres évaluations de systèmes de TAP de ce type ;
- la réalisation d'une étude à large échelle comprenant des tests en milieu hospitalier afin d'appliquer la méthode EAGLES-ISO dans des conditions les plus proches de l'utilisation finale réelle des systèmes comparés ;
- la confirmation de certains faits concernant l'évaluation de la TAP, souvent évoqués sans être explicitement confirmés, notamment, le fait que les mesures automatiques classiques d'évaluation de la TA (et par conséquent de la TAP) ne sont pas véritablement adaptées à l'évaluation de la TA et de la TAP linguistiques. Nous avons aussi montré qu'il est discutable de choisir des traducteurs (interprètes ou autre professionnels de la langue) comme évaluateurs, comme c'est le plus souvent le cas dans le cadre de l'évaluation de la précision de la TA ou TAP, puisqu'ils s'avèrent plus sévères que les évaluateurs bilingues non-linguistes ;
- la définition de l'échelle *Human Medical Scale* (HMS), adaptée à l'évaluation de la TAP dans le domaine médical. Le fondement de celle-ci est de se concentrer sur le

maintien du sens des phrases évaluées au lieu de se focaliser sur leur bonne formation (grammaticale ou stylistique). En outre, son objectif est de pouvoir évaluer autant la TA linguistique que statistique, sans introduire de biais en faveur de l'une ou de l'autre technologie employées par les systèmes évalués ;

- enfin, l'élaboration de la méthode TURKoise pour l'évaluation de la TAP médicale. Du fait qu'elle se fonde sur HMS, celle-ci est plus équitable du point de vue de la technologie employée (statistique ou linguistique) et surtout, elle est axée sur la qualité à l'usage du produit fini. Cette méthode permet d'effectuer plus rapidement et à moindre coût une évaluation informative quant à la qualité à l'usage des systèmes évalués.

Le dernier point évoqué répond à une des principales limites de notre méthode EAGLES-ISO que nous évoquerons dans la section suivante.

7.3 Limites

Comme mentionné plus haut, la méthode EAGLES-ISO, bien que donnant un résultat probant, est bien trop longue à mettre en place et trop coûteuse, ce qui l'empêche d'être utilisée notamment dans le cadre de grandes campagnes d'évaluation de TAP (comme l'atelier annuel IWSLT). Elle pourrait, en revanche, compléter des études effectuées par le DARPA, puisqu'elle va dans le sens de leur méthode d'évaluation SCORE (section 3.4.2).

D'autre part, concernant l'échelle HMS que nous proposons, son efficacité a été testée uniquement sur deux systèmes très proches et avec des critères qui ne sont pas toujours faciles à interpréter. En effet, un résultat secondaire de ce travail est la constatation que la mesure classique de cohérence entre évaluateurs (kappa) n'est pas la mesure la plus adaptée et qu'elle est difficile à interpréter. Celle-ci doit être utilisée avec une grande prudence en veillant à utiliser la variante appropriée. Nous avons ainsi commencé à utiliser le coefficient de corrélation intraclasse (ICC – *Intraclass Correlation Coefficient*; Shrout et Fleiss, 1979) qui nous a donné des résultats plus cohérents notamment avec le pourcentage d'accord entre évaluateurs. Il serait souhaitable d'utiliser ICC de manière plus extensive, et de manière systématique pour toutes les évaluations de cohérence au lieu de la mesure kappa.

Une autre limite de ce travail est de n'avoir pas exploité toutes les possibilités du corpus constitué lors de l'**EXPQ1**. Il serait par ailleurs intéressant de le préparer afin de le mettre à

disposition des chercheurs du domaine. Il pourrait ainsi directement être exploité notamment par l'héritier de *MedSLT*, le projet *BabelDr*¹³⁹.

Enfin, comme mentionné dans le chapitre 5, à propos des résultats obtenus pour **EXPQ1** (section 5.1.1.2), l'utilisation de patients standardisés, tout en présentant les avantages décrits dans la section 4.1.2 (standardisation des symptômes, possibilité de contrôler les résultats obtenus, non-mise en danger de véritables patients lors de l'étude), engendre certains biais. Son principal inconvénient est la nécessité de fournir des scénarios de simulation de diagnostic (qui ne sont pas toujours bien suivis par les participants à l'expérience) et de ne pas forcément bien refléter la réalité des besoins des utilisateurs finaux (notamment le niveau d'éducation et la capacité d'utiliser une interface sur un ordinateur).

7.4 Perspectives

Pour le premier volet de ce travail, comme évoqué ci-dessus, ce serait intéressant de réutiliser le modèle de qualité élaboré pour la TAP dans le domaine médical et d'arriver à une expérience impliquant de préférence de véritables patients. Bien que cette évaluation soit plus compliquée à mettre en place, elle présenterait les trois avantages suivants : (1) pouvoir utiliser une version modernisée de *MedSLT* (comme *BabelDr*), (2) éviter l'utilisation de scénarios pour la simulation de diagnostic, (3) répéter également TURKoise sur les données générées afin de comparer directement les deux méthodes.

Pour le second volet, nous aimerions tester la méthode TURKoise sur une plus grande variété de systèmes de TAP du médical, notamment sur des systèmes statistiques afin de vérifier sa viabilité, son utilité et sa compétitivité pour d'autres systèmes de TAP.

Nous pourrions aussi contribuer au développement de nouvelles applications de TAP en mettant à disposition le corpus collecté lors d'**EXPQ1**.

Enfin, il serait envisageable d'étudier d'autres plateformes de *crowdsourcing*, plus facilement accessibles en Europe que l'AMT (restreint aux USA depuis quelque temps), comme *Crowdfunder.com*¹⁴⁰, ou *MiniFreelance.com*¹⁴¹, afin de vérifier si nous obtenons des résultats comparables par rapport à la compétitivité, la fiabilité et la faisabilité de TURKoise.

¹³⁹ Site du projet BabelDR, <http://babeldr.unige.ch/project-description/>, consulté le 01.03.2016.

¹⁴⁰ Site de Crowdfunder, <http://www.crowdfunder.com/>, accédé le 01.07.2015.

¹⁴¹ Site de Minifreelance, <http://minifreelance.com/>, accédé le 01.07.2015.

8 Bibliographie

Abeillé, A. et P. Blache (2000). Grammaires et analyseurs syntaxiques. Ingénierie des langues. J.-M. Pierrel. Paris, Hermès Science Publications: 51-76.

Aberdeen, J., B. Oshika, S. Condon, L. Harper et J. Philips (2005). Evaluation of Speech-to-Speech Translation Systems, MITRE.

Abran, A. (2010). Software metrics and software metrology, John Wiley & Sons.

Agency, U. S. D. A. R. P. (2008). DARPA: 50 Years of Bridging the Gap, Faircount LLC.

Akiba, Y., M. Federico, N. Kando, H. Nakaiwa, M. Paul et J. i. Tsujii (2004). Overview of the IWSLT04 Evaluation Campaign. International Workshop on Spoken Language Translation IWSLT04, Kyoto, Japan.

Albrecht, J. et R. Hwa (2007). Regression for sentence-level MT evaluation with pseudo references. ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS.

Arnold, D. (2003). Why translation is difficult for computers. Computers and Translation: A translator's guide. H. Somers. Amsterdam; Philadelphia PA, John Benjamins: 119-142.

Arnold, D., L. Balkan, R. L. Humphreys, S. Meijer et L. Sadler, Eds. (1994). Machine Translation : An Introductory Guide. London, Blackwell-NCC.

Avramidis, E. et P. Koehn (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation. ACL 2008. Columbus, Ohio.

Bach, N., M. Eck, P. Charoenpornasawat, T. Köhler, S. Stüker, T. Nguyen, R. Hsiao, A. Waibel, S. Vogel et T. Schultz (2007). The CMU TransTac 2007 eyes-free and hands-free two-way speech-to-speech translation system. Proceedings of the IWSLT.

Bach, N., M. Eck, P. Charoenpornasawat, T. Kohler, S. Stüker, T. Nguyen, R. Hsiao, A. Waibel, S. Vogel, T. Schultz et A. W. Black (2007). The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System. IWSLT 2007, Trento, Italy.

Belz, A. (2009). "That's Nice... What Can You Do With It?" Computational Linguistics 35(1): 111-118.

Blanchon, H. (2004). Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de TAO de l'écrit et de l'oral — Une bataille contre le bruit, l'ambiguïté et

le manque de contexte Habilitation à Diriger des Recherches, Université Joseph Fourier - Grenoble 1.

Blanchon, H. et C. Boitet (2007). "Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche." Traitement Automatique des Langues (TAL) **48**(1): 33-65.

Blatz, J., E. Fitzgerald, G. Foster, S. Grandrabur, C. Goutte, A. Kulesza, A. Sanchis et N. Ueffing (2004). *Confidence Estimation for Machine Translation*. Baltimore, MD, John Hopkins

Boitet, C., Y. Bey, M. Tomokio, W. Cao et H. Blanchon (2006). IWSLT-06: Experiments with Commercial MT Systems and Lessons from Subjective Evaluations. International Workshop on Spoken Language Translation (IWSLT-06), Kyoto, Japan.

Bongiovanni, G. L. (2005). Medical Spanish. New-York, McGraw-Hill.

Bouillon, P. et A. Clas (1993). La traductique. Montréal, Presses de l'Université de Montréal.

Bouillon, P., M. Rayner, N. Chatzichrisafis, B. A. Hockey, M. Santaholma, M. Starlander, H. Isahara, K. Kanzaki et Y. Nakao (2005). A generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation, 30-31, May, 2005, Budapest, Hungary. Tenth Conference of the European Association of Machine Translation, Budapest, Hungary.

Bouillon, P., M. Rayner, J. Gerlach et P. Estrella (2011). "Pour une interlangue utile en traduction automatique de la parole dans des domaines limités." Traitement Automatique des Langues (TAL) **52**(1).

Bouillon, P., M. Rayner, B. Novellas Vall, M. Starlander, M. Santaholma, Y. Nakao et N. Chatzichrisafis (2006). "Une grammaire partagée multi-tâche pour le traitement de la parole : application aux langues romanes." Revue TAL **47**(3): 155-173.

Bouillon, P., M. Rayner, M. Starlander et M. Santaholma (2007). Les ellipses dans un système de Traduction Automatique de la Parole. Traitement Automatique des Langues Naturelles, Toulouse, France.

Bowen, Z. (2013). "Statistical Machine Translation for Speech: A Perspective on Structures, Learning, and Decoding." Proceedings of the IEEE **101**(5): 1180-1202.

Bowker, L. et J. Pearson (2002). Working with specialized language: a practical guide to using corpora, Routledge.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, R. L. Mercer et P. S. Roossin (1988). A Statistical Approach to Language Translation. Computational Linguistics, Budapest, Hungary.

Callison-Burch, C. et M. Dredze (2010). Creating speech and language data with Amazon's Mechanical Turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Los Angeles, California, Association for Computational Linguistics: 1-12.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz et J. Schroeder (2007). (Meta-) evaluation of machine translation. Second Workshop on Statistical Machine Translation StatMT '07, Prague, Czech Republic, Association for Computational Linguistics.

Callison-Burch, C., P. Koehn, C. Monz, K. Peterson, M. Przybocki et O. F. Zaidan (2010). Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. ACL 2010 Fifth Workshop on Statistical Machine Translation, Uppsala, Sweden.

Callison-Burch, C., C. Monz, P. Koehn et J. Schroeder (2009). Findings of the 2009 Workshop on Statistical Machine Translation. Fourth Workshop on Statistical Machine Translation, Athens, Greece.

Chatzichrisafis, N., P. Bouillon, M. Rayner, M. Santaholma et M. Starlander (2006). Evaluating Task Performance for a Unidirectional Controlled Language Medical Speech Translation System. First International Workshop on Medical Speech Translation, HLT-NAACL, New York, ACL.

Chow, Y., M. Dunham, O. Kimball, M. Krasner, G. Kubala, J. Makhoul, P. Price, S. Roucos et R. Schwartz (1987). BYBLOS: The BBN continuous speech recognition system. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.

Cicchetti, D. V. (1994). "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." Psychological assessment 6(4): 284.

Condon, S., D. Parvaz, J. Aberdeen, C. Doran, A. Freeman et M. Awad (2010). Evaluation of machine translation errors in English and Iraqi Arabic. LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation. Valetta, Malta: 729-735.

Coolican, H. (1996). Research methods and statistics in psychology, Hodder & Stoughton Educational.

Costa-Jussà, M. R., M. Farrús, J. B. Mariño et J. A. R. Fonollosa (2012). "Study and comparison of rule-based and statistical Catalan–Spanish machine translation systems." Computational Informatics **31**(2): 245–270.

Cuvilliez, C. et G. Tiger (2014). Système de traduction numérique, rapport de synthèse. Genève, Direction générale, Gestion des projets & qualité, HUG.

Dabbadie, M., A. Hartley, M. King, K. Miller, W. M. El Hadi, A. Popescu-Belis, F. Reeder et M. Vanni (2002). A Hands-On Study of the Reliability and Coherence of Evaluation Metrics. Workshop at the LREC 2002 Conference: Machine Translation Evaluation: Human Evaluators Meet Automatec Metrics, Las Palmas, Canary Islands, Spain.

Desharnais, J.-M. et A. Abran (2012). "Software Measurement Methods: An Analysis of Two Designs." Journal of Software Engineering and Applications **Vol.05No.10**: 13.

Deville, G. (2001). Corpus-based Sublanguage Modelling for NLP Applications: A Tentative Methodology. International Colloquium on Trends in Special Language and language Technology, Bruxelles.

Dillinger, M. et M. Seligman (2006). Converser (TM): Highly Interactive Speech-to-Speech Translation for Healthcare. Workshop on Medical Speech Translation at HLT-NACCL, New-York, New-York.

Dorr, B. J., J. Olive, J. MacCary et C. Christianson (2011). Chapter 5: Machine Translation Evaluation and Optimization. Handbook of Natural Language Processing and Machine Translation. J. Olive, J. MacCary et C. Christianson. New-York, Springer Science + Business Media: 745-843.

Dybkjaer, L., H. Hemsén et W. Minker (2007). Evaluation of text and speech systems, Springer Science & Business Media.

Eck, M. et C. Hori (2005). Overview of the IWSLT 2005 Evaluation Campaign. Pittsburgh, USA.

Eck, M., I. Lane, Y. Zhang et A. Waibel (2010). JIBBIGO: Speech-to-speech Translation on Mobile Devices. Spoken Language Technology Workshop (SLT), 2010 IEEE.

Ehsani, F., R. Frederking, M. Rayner et P. Bouillon (2008). Spoken Language Translation. New Trends in Speech Based Interactive Systems. F. Chen et K. Jokinen, Springer Verlag.

Ehsani, F., J. Kinzey, D. Master, K. Lesea et H. Park (2006). Speech to speech Translation for Medical Triage in Korean. Workshop on Medical Speech Translation at HLT-NACCL, New-York, New-York.

Eisele, A., C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Herrmann et Y. Chen (2008). "Hybrid architectures for multi-engine machine translation." Proceedings of Translating and the Computer **30**.

Eskenazi, M., G.-A. Levow, H. Meng, G. Parent et D. Suendermann (2013). Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment, John Wiley & Sons.

Estrella, P., M. King et A. Popescu-Belis (2007). A New Method for the Study of Correlations between MT Evaluation Metrics. 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07), Skövde, Sweden.

Ettelaie, E., S. Gandhe, P. Georgiou, R. S. Belvin, K. Knight, D. Marcu, S. Narayanan et D. Traum (2005). Transonics: A Practical Speech-to-Speech Translator for English-Farsi Medical Dialogs. 43rd Annual Meeting Meeting of the Association for Computational Linguistics: ACL-05 Interactive Poster and Demonstration Sessions, Ann Arbor, MI.

Federico, M., L. Bentivogli, M. Paul et S. Stuker (2011). Overview of the IWSLT 2011 evaluation campaign, International Workshop on Spoken Language Translation 2011. California, USA.: 11-27.

Federico, M., M. Cettolo, L. Bentivogli, M. Paul et S. Stüker (2012). Overview of the IWSLT 2012 evaluation campaign. 9th International Workshop on Spoken Language Translation Proceedings IWSLT-2012. Hong Kong: 12-33.

Fleiss, J. L. (1971). "Measuring nominal scale agreement among many raters." Psychological Bulletin **76**: 378-382.

Flores, G., B. Laws, S. Mayo, B. Zuckerman, M. Abreu, L. Medina et E. Hardt (2003). "Errors in Medical Interpretation and Their Potential Clinical Consequences in Pediatric Encounters." Pediatrics **111**(1): 6-14.

Fort, K., G. Adda et K. B. Cohen (2011). "Amazon mechanical turk: Gold mine or coal mine?" Computational Linguistics 37(2): 413-420.

Frederking, R., A. W. Black, R. D. Brown, J. Moody et E. Steinbrecher (2002). Field Testing the Tongues Speech-to-Speech Machine Translation System. Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Gran Canaria, Spain.

Freitag, M., J. Wuebker, S. Peitz, H. Ney, M. Huck, A. Birch, N. Durrani, P. Koehn, M. Mediani et I. Slawik (2014). "Combined Spoken Language Translation." Proc. of the Int. Workshop on Spoken Language Translation (IWSLT), South Lake Tahoe, CA, USA.

Giménez, J. et L. Màrquez (2008). The UPC Participation at the Metrics MATR Challenge 2008. Metrics MATR Workshop at AMTA'08 Machine Translation, Waikiki, Hawaii.

Giménez, J. et L. Màrquez (2010). "Linguistic measures for automatic machine translation evaluation." Machine translation 24(3-4): 209-240.

Graham, Y., T. Baldwin, A. Moffat et J. Zobel (2013). "Continuous measurement scales in human evaluation of machine translation." Proc. 7th Linguistic Annotation Wkshp. & Interoperability with Discourse: 33-41.

Gwet, K. L. (2012). Handbook of inter-rater reliability. Wiley Encyclopedia of Clinical Trials, Advanced Analytics LCC

Hamon, O., C. Fügen, D. Mostefa, V. Arranz, M. Kolss, A. Waibel et K. Choukri (2009). End-to-end evaluation in simultaneous translation. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, Association for Computational Linguistics.

Hartley, A. et A. Popescu-Belis (2004). Évaluation des systèmes de traduction automatique. Évaluation des systèmes de traitement de l'information. S. Chaudiron. Paris, Hermès: p.311-335.

Haton, J.-P., C. Cerisara, D. Fohr, Y. Laprie et K. Smaïli (2006). Reconnaissance automatique de la parole: Du signal à son interprétation. Paris, Dunod.

Hockey, B. A., O. Lemon, E. Campana, L. Hiatt, G. Aist, J. Hieronymus, A. Gruenstein et J. Dowding (2003). Targeted help for spoken dialogue systems: Intelligent feedback improves naïve users' performance. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary.

Hovy, E., M. King et A. Popescu-Belis (2002). "Principles of Context-Based Machine Translation Evaluation." Machine Translation **17**(1): 43-75.

Hyman, P. (2014). Speech-to-Speech translations Stutter, But Researchers Seem ellifluous future. Communication of the ACM. **57**.

Isabelle, P., M. Dymetman et E. Macklovitch (1988). CRITTER: a Translation System for Agricultural Market Reports. 12th International Conference on Computational Linguistics, Budapest.

ISO/IEC (2001). ISO/IEC 9126-1 (E) - Software Engineering - Product Quality - Part 1: Quality Model. Geneva, International Organization for Standardization / International Electrotechnical Commission: 25.

ISO/IEC (2003). ISO/IEC 9126-2 (E) Software engineering - Product quality - Part 2: External metrics. Geneva, International Organization for Standardization / International Electrotechnical Commission: 86.

ISO/IEC (2004). ISO/IEC 9126-4 (E) Software engineering - Product quality - Part 4: Quality in use metrics. Geneva, International Organization for Standardization / International Electrotechnical Commission: 59.

ISO/IEC (2011). ISO/IEC 25010:2011 : Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models.

ISO/IEC (2014). ISO/IEC 25000:2014(en) Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE. Geneva, International Organization for Standardization / International Electrotechnical Commission.

Jones, D., P. Gatewood, M. Herzog et T. Marius (2014). Standardized ILR-Based and Task-Based Speech-to-Speech MT Evaluation. Workshop on Automatic and Manual Metrics for Operational Translation Evaluation - within LREC 2014, Keith J. Miller, Lucia Specia, Kim Harris, Stacey Bailey: 51-60.

Jones, K. S. et J. R. Galliers (1995). Evaluating natural language processing systems: An analysis and review, Springer Science & Business Media.

Jurafsky, D. et J. H. Martin (2000). HMMS and Speech Recognition. Speech and Language Processing: An Introduction to Natural Language Processing, Computational

Linguistics, and Speech Recognition. D. Jurafsky et J. H. Martin. Upper Saddle River, NJ, Prentice Hall: 235-284.

Jurafsky, D. D. et J. H. Martin (2009). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.

Kaliyadan, F. et S. Gopinathan Pillai (2010). "The use of Google language tools as an interpretation aid in cross-cultural doctor–patient interaction: a pilot study." Informatics in Primary Care **18**(2): 141-143.

Khanna, R. R., L. S. Karliner, M. Eck, E. Vittinghoff, C. J. Koenig et M. C. Fang (2011). "Performance of an online translation tool when applied to patient educational material." Journal of Hospital Medicine **6**(9): 519-525.

King, M. (2007). General Principles of User-oriented Evaluation. Springer, Springer: 125-161.

Kittur, A., B. Smus, S. Khamkar et R. E. Kraut (2011). CrowdForge: crowdsourcing complex work. Proceedings of the 24th annual ACM symposium on User interface software and technology. Santa Barbara, California, USA, ACM: 43-52.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. EMNLP 2004.

Koehn, P. (2009). A web-based interactive computer aided translation tool. Proceedings of the ACL-IJCNLP 2009 Software Demonstrations. Suntec, Singapore, Association for Computational Linguistics: 17-20.

Koehn, P. (2010). Evaluation. Statistical machine translation, Cambridge University Press.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin et E. Herbst (2007). Moses: open source toolkit for statistical machine translation. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Prague, Czech Republic, Association for Computational Linguistics: 177-180.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, Ond, #345, e. Bojar, A. Constantin et E. Herbst (2007). Moses: open source toolkit for statistical machine translation. Proceedings of the 45th Annual

Meeting of the ACL on Interactive Poster and Demonstration Sessions. Prague, Czech Republic, Association for Computational Linguistics: 177-180.

Kos, K. et O. Bojar (2009). "Evaluation of machine translation metrics for czech as the target language." The Prague Bulletin of Mathematical Linguistics **92**: 135-148.

Kulesza, A. et S. Shieber (2004). A Learning Approach to improving Sentece-Level MT Evaluation. Tenth Conference on Theoretical and Methodological issues in Machine Translation, Baltimore.

Labaka, G., C. España-Bonet, L. Màrquez et K. Sarasola (2014). "A hybrid machine translation architecture guided by syntax." Machine Translation **28**(2): 91-125.

Lamel, L. et J.-L. Gauvain (2003). Speech Recognition. The Oxford Handbook of Computational Linguistics. R. Mitkov. Oxford, Oxford University Press: 305-322.

Landis, J. R. et G. G. Koch (1977). "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers." Biometrics: 363-374.

Lavie, A., L. Levin, T. Schultz, A. Waibel , C. Langley, B. Han, A. Tribble, D. Gates, D. Wallace et K. Peterson (2001). Domain Portability in Speech-to-Speech Translation. HLT-2001 Human Language Technology Conference, San Diego, CA.

Lavie, A., F. Metze, R. Cattoni et E. Costantini (2002). A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System. Speech-to-Speech Translation Workshop at the 40th Annual Meeting of the Association of Computational Linguistics, Philadelphia, PA.

Lawal, B. (2014). Repeated Measures Design. Applied Statistical Methods in Agriculture, Health and Life Sciences, Springer International Publishing: 697-718.

Lee, Y.-S., D. J. Sinder et C. J. Weinstein (2002). "Interlingua-based English-Korean Two-way Speech Translation of Doctor-Patient Dialogues with CCLINC." Machine Translation **17**(3): 213-243.

Levin, L., D. Gates, A. Lavie et A. Waibel (1998). An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. International Conference on Spoken Language Processing (ICSLP), Sydney, Australia.

Levin, L., A. Lavie, M. Woszczyna, D. Gates, M. Gavalda, D. Koll et A. Waibel (2000). "The JANUS-III Translation System: Speech-to Speech Translation in Multiple Domains." Machine Translation 15: 3-25.

Lewis, W. D. (2015). Skype Translator: Breaking Down Language and Hearing Barriers. Proceedings of Translating and the Computer (TC37). Londres.

Lommel, A., M. Popović et A. Burchardt (2014). Assessing Inter-Annotator Agreement for Translation Error Annotation. Automatic and Manual Metrics for Operational Translation Evaluation Workshop Programme.

Mana, N., S. Burger, R. Cattoni, L. Besacier, V. MacLaren, J. McDonough et F. Metze (2003). The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains. Eurospeech, Geneva, Switzerland.

Matsuda, S., H. Xinhui, Y. Shiga, H. Kashioka, C. Hori, K. Yasuda, H. Okuma, M. Uchiyama, E. Sumita, H. Kawai et S. Nakamura (2013). Multilingual Speech-to-Speech Translation System: VoiceTra. Mobile Data Management (MDM), 2013 IEEE 14th International Conference on.

Miller, K. et M. Vanni (2005). Inter-rater agreement measures, and the refinement of metrics in the PLATO MT evaluation paradigm. MT Summit X, Phuket, Thailand.

Nakamura, S. (2009). "Overcoming the Language Barrier with Speech Translation Technology." Quarterly review(31).

Narayanan, S., S. Ananthakrishnan, R. S. Belvin, E. Ettelaie, S. Gandhe, S. Ganjavi, P. Georgiou, C. M. Hein, S. Kadambe, K. Knight, D. Marcu, H. E. Neely, N. Srinivasamurthy, D. Traum et D. Wang (2004). The Transonics Spoken Dialogue Translator: An aid for English-Persian Doctor-Patient interviews. American Association for Artificial Intelligence (AAAI) Fall Symposium, Menlo Park, California.

Niessen, S., F. J. Och, G. Leusch et H. Ney (2000). An evaluation tool for machine translation: Fast evaluation for machine translation research. 2nd LREC, Athens, Greece.

Nirenburg, S. (1993). L'interlangue et le traitement du sens dans les systèmes de traduction automatique. La traductique. P. Bouillon et A. Clas. Montréal, Presse Universitaire de l'Université de Montréal: 91-108.

Nuance (2003). Nuance Grammar Developer's guide, version 8.5, Nuance.

- Oates, B. J. (2005). Researching information systems and computing. Londres, Sage.
- Owczarzak, K., J. van Genabith et A. Way (2007). "Evaluating machine translation with LFG dependencies." Machine Translation **21**(2): 95-119.
- Padó, S., M. Galley, D. Jurafsky et C. Manning (2009). Robust machine translation evaluation with entailment features. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1. Suntec, Singapore, Association for Computational Linguistics: 297-305.
- Papineni, K., S. Roukos, T. Ward et Z. Wei-Jing (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, USA.
- Patil, S. et P. Davies (2014) "Use of Google Translate in medical communication: evaluation of accuracy." BMJ **349**.
- Paul, M. (2009). Overview of the IWSLT 2009 Evaluation Campaign. IWSLT 2009, Tokyo, Japan.
- Paul, M., M. Federico et S. Stuker (2010). Overview of the IWSLT 2010 Evaluation Campaign. International Workshop on Spoken Language Translation 2010. Paris, France.
- Pierce, J. R. et J. B. Carroll (1966). Language and Machines - Computers in Translation and Linguistics (ALPAC report). Washington DC, Automatic Language Processing Advisory Committee.
- Popescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. TAL (Traitement Automatique de la Langue), vol. 48, n. 1: 67-91.
- Popescu-Belis, A. (2007). "Le rôle des métriques d'évaluation dans le processus de recherche en TAL." Traitement automatique des langues (TAL) **48**(1): 67-91.
- Popović, M., M. Arčan, E. Avramidis, A. Burchardt et A. Lommel (2015). Poor man's lemmatisation for automatic error classification. Eighteenth Annual Conference of the European Association for Machine Translation (EAMT 15). Antalya, Turkey: 105-112.
- Porter, M. F. (1980). "An algorithm for suffix stripping." Program **14**(3): 130-137.

Prasad, R., R. Kumar, S. Ananthakrishnan, W. Chen, S. Hewavitharana, M. E. Roy, F. Choi, A. Challenner, E. Kan, A. Neelakantan et P. Natarajan (2012). Active error detection and resolution for speech-to-speech translation: 150-157.

Precoda, K., H. Franco, A. Dost, M. Frandsen, J. Fry, A. Kathol, C. Richey, S. Riehemann, D. Vergyri, J. Zheng et C. Culy (2004). Limited-Domain Speech-to-Speech Translation between English and Pashto. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.

Rajman, M. et T. Hartley (2001). Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. Proceedings of MT Summit VIII: 4th ISLE Workshop on MT Evaluation.

Randhawa, G., M. Ferreyra, R. Ahmed, O. Ezzat et K. Pottie (2013). "Using machine translation in clinical practice." Canadian Family Physician 59(4): 382-383.

Rayner, E., P. S. Estrella et P. Bouillon (2011). Bootstrapping A Statistical Speech Translator From A Rule-Based One. Proceedings of the 2nd International Workshop On Free/Open-Source Rule-Based Machine Translation.

Rayner, M., P. Bouillon, J. Brotanek, G. Flores, S. Halimi, B. A. Hockey, H. Isahara, K. Kanzaki, E. Kron, Y. Nakao, M. Santaholma, M. Starlander et N. Tsourakis (2008). The MEDSLT 2008 system. Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications (Coling 2008), Manchester, England.

Rayner, M., P. Bouillon, N. Chatzichrisafis, M. Santaholma et M. Starlander (2006). MedSLT: A Limited-Domain Unidirectional Grammar-Based Medical Speech Translator. First International Workshop on Medical Speech Translation, HLT-NAACL, New-York, Omnipress Inc.

Rayner, M., P. Bouillon, G. Flores, F. Ehsani, M. Starlander, B. A. Hockey, J. Brotanek et L. Biewald (2008). A Small-Vocabulary Shared Task for Medical Speech Translation. Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications (Coling 2008), Manchester.

Rayner, M., P. Bouillon et B. A. Hockey (2006). Introduction. Putting Linguistics into Speech Recognition. M. Rayner, P. Bouillon et B. A. Hockey. Sanford, California, CSLI: 1-22.

Rayner, M., P. Bouillon, B. A. Hockey, N. Chatzichrisafis et M. Starlander (2004). Comparing Rule-Based and Statistical Approaches to Speech Understanding in a Limited

Domain Speech Translation System. 10th International Conference on Theoretical and Methodological Issues in Machine Translation, Baltimore, MD.

Rayner, M., P. Bouillon, V. Van Dalsem III, H. Isahara, K. Kanzaki et B. A. Hockey (2003). A Limited-Domain English to Japanese Medical Speech Translator Built Using Regulus 2. Association for Computational Linguistics, Sapporo, Japan.

Rayner, M., D. Carter, P. Bouillon, V. Digalakis et M. Wirén (2000). The Spoken Language Translator. Cambridge, Cambridge University Press.

Rayner, M., B. A. Hockey et P. Bouillon (2006). Putting Linguistics into Speech Recognition. Stanford, California, Stanford University Center for the Study of Language and Information.

Rayner, M., B. A. Hockey et J. Dowding (2002). Grammar Specialisation Meets Language Modelling. 7th International Conference on Spoken Language Processing, Denver, Colorado.

Saldanha, G. et S. O'Brien (2014). Research methodologies in translation studies. Londres et New-York, Routledge.

Salkind, N. J. (2012). Statistics for people who (think they) hate Statistics: Excel 2010 Edition, Sage.

Sanders, G., B. Weiss, C. Schlenoff, M. Steves et S. Condon (2013). "Evaluation methodology and metrics employed to assess the TRANSTAC two-way, speech to speech translation systems." Computer Speech and Language 27(2): 528-553.

Sanders, G. A., B. A. Weiss, C. Schlenoff, M. P. Steves et S. Condon (2013). "Evaluation methodology and metrics employed to assess the TRANSTAC two-way, speech-to-speech translation systems." Computer Speech & Language 27(2): 528-553.

Sarich, A. (2004). Development and Fielding of Phraselator Phrase Translation System. 26th Translation and the Coputer Conference (ASLIB), London.

Schultz, T., D. Alexander, A. W. Black, K. Peterson, S. Suebvisai et A. Waibel (2004). A Thai speech translation system for medical dialogs. Human Language Technologies (HLT), Boston, MA.

Schultz, T. et A. W. Black (2006). Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs. 31st International Conference on Acoustics, Speech, and Signal Processing (ICASSP2006), Toulouse, France.

Schultz, T., A. W. Black, S. Vogel et M. Woszczyna (2006). "Flexible speech translation systems." Audio, Speech, and Language Processing, IEEE Transactions on **14**(2): 403-411.

Seligman, M. et M. Dillinger (2006). Usability Issues in an Interactive Speech-to-Speech Translation System for Healthcare. Workshop on Medical Speech Translation at HLT-NACCL, New-York, New-York.

Seligman, M. et M. Dillinger (2008). Rapid Portability among Domains in an Interactive Spoken Language Translation System. Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications (SLT4MED). Manchester, UK: 40-47.

Seligman, M. et M. Dillinger (2011). "Real-time Multi-media translation for healthcare: a usability study." Proceedings of the 13th Machine Translation Summit: 19-23.

Sheng, V. S., F. Provost et P. G. Ipeirotis (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Sherief, N., N. Jiang, M. Hosseini, K. Phalp et R. Ali (2014). Crowdsourcing software evaluation. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, ACM.

Snover, M., B. J. Dorr, R. Schwartz, L. Micciulla et J. Makhoul (2006). A Study of Translation Edit Rate with Targeted Human Annotation. 7th Conference of the Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA.

Soller, W. R., P. Chan et A. Higa (2012). "Performance of a New Speech Translation Device in Translating Verbal Recommendations of Medication Action Plans for Patients with Diabetes." Journal of Diabetes Science and Technology **6**(4).

Somers, H. (2003). Sublanguage. Computers and Translation: A translator's guide. H. Somers. Amsterdam; Philadelphia, PA, John Benjamins: 283-295.

Somers, H. et Y. Sugita (2003). Evaluating Commercial Spoken Language Translation Software. MT Summit IX, New Orleans.

Somers, H. L. et H. J. Lovel (2006). Can AAC technology facilitate communication for patients with limited English?

Specia, L. (2011). Exploiting Objective Annotations for Measuring Translation Post-editing Effort. 15th Conference of the European Association for Machine Translation (EAMT), Leuven, Belgium.

Specia, L., N. Hajlaoui, C. Hallett et W. Aziz (2011). "Predicting machine translation adequacy." Machine Translation Summit XIII, Xiamen, China.

Specia, L., D. Raj et M. Turchi (2010). "Machine translation evaluation versus quality estimation." Machine Translation 24(1): 39-50.

Starlander, M. (2007). Using a Wizard of Oz as a baseline to determine which system architecture is the best for a spoken language translation system. 16th Nordic Conference of Computational Linguistics (NODALIDA), Tartu, Estonia.

Starlander, M. (2014). TURKOISE: a Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain. Workshop on Automatic and Manual Metrics for Operational Translation Evaluation - within LREC 2014, Reykjavik, Islande, Keith J. Miller, Lucia Specia, Kim Harris, Stacey Bailey.

Starlander, M. (2015). Let the EAGLES Fly into New Standards: Adapting our CAT Tool Evaluation Methodology to the ISO 25000 Series, Londres.

Starlander, M., P. Bouillon, G. Flores, M. Rayner et N. Tsourakis (2008). Comparing two different bidirectional versions of the limited domain medical spoken language translator MedSLT. 12th annual conference of the European Association for Machine Translation, Hamburg.

Starlander, M. et L. Morado Vazquez (2013). Training translation students to evaluate CAT tools using Eagles: a case study. Aslib: Translating and the Computer 35. Londres, Aslib.

Starlander, M. et A. Popescu-Belis (2002). Corpus-based Evaluation of a French Spelling and Grammar Checker, Starlander M. and Popescu-Belis A., in: Proceedings of LREC 2002, p.268-274, Las Palmas de Gran Canaria, Spain, 2002. LREC, Las Palmas de Gran Canaria, Spain.

Surowiecki, J. (2005). The wisdom of crowds, Anchor.

Tate, C. R. (2008). A Statistical Analysis of Automated MT Evaluation Metrics for Assessments in Task-Based MT Evaluation. Association for Machine Translation in the Americas (AMTA), Waikiki, Hawai.

Tillmann, C., S. Vogel, H. Ney et H. Sawaf (2000). "Statistical Translation of Text and Speech: First Results with the RWTH System." Machine Translation **15**: 43-74.

Tsourakis, N. (2013). Development & evaluation of multilingual multimodal dialogue systems on mobile devices, University of Geneva.

Tsourakis, N., P. Bouillon et M. Rayner (2009). Design Issues for a Bidirectional Mobile Medical Speech Translator. Proceedings of the SiMPE Workshop at MobileHCI: 7.

Turian, J. P., L. Shen et D. Melamed (2003). Evaluation of Machine Translation and its evaluation. Machine Translation Summit IX, New Orleans.

Turner, A. M., M. K. Brownstein, K. Cole, H. Karasz et K. Kirchhoff (2015). "Modeling workflow to design machine translation applications for public health practice." Journal of Biomedical Informatics **53**(0): 136-146.

Van Slype, G. (1979). Critical Methods for Evaluating the Quality of Machine Translation. Bruxelles, European Commission Directorate General Scientific and Technical Information and Information Management. Bureau Marcel van Dijk: 188.

Wahlster, W. (2000). Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbomil System. Verbomobil: Foundations of speech-to-speech translation. W. Wahlster. Berlin, Springer: 3-21.

Waibel , A., A. Badran, A. W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. M. Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna et J. Zhang (2003). Speechalator: Two-way Speech-to-Speech Translation in your Hand. HLT-NAACL, Edmonton.

Waibel , A., A. Badran, A. W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. M. Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna et J. Zhang (2003). Speechalator: two-way speech-to-speech translation on a consumer PDA. EUROSPEECH, Geneva, Switzerland.

Waibel , A. et C. Fügen (2008). "Spoken Language Translation : Enabling cross-lingual human-human communication." IEEE Signal Processing Magazine **25**(3): 70-79.

Weiss, B. et C. Schlenoff (2011). "Performance Assessments of Two-Way, Free-Form, Speech-to-Speech Translation Systems for Tactical Use." International Test and Evaluation Association Journal (ITEA Journal) **32**(1): 69-75.

Xiaodong, H. et D. Li (2011). "Speech Recognition, Machine Translation, and Speech Translation; A Unified Discriminative Learning Paradigm [Lecture Notes]." Signal Processing Magazine, IEEE **28**(5): 126-133.

Zhao, Y. et Q. Zhu (2014). "Evaluation on crowdsourcing research: Current status and future direction." Information Systems Frontiers **16**(3): 417-434.

Zhou, B., L. Besacier et Y. Gao (2007). On efficient coupling of ASR and SMT For Speech Translation. International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawaii.

Zhou, B., X. Cui, S. Huang, M. Cmejrek, W. Zhang, J. Xue, J. Cui, B. Xiang, G. Daggett, U. Chaudhari, S. Maskey et E. Marcheret (2013). "The IBM speech-to-speech translation system for smartphone: Improvements for resource-constrained tasks." Comput. Speech Lang. **27**(2): 592-618.

Zhou, B., X. Cui, S. Huang, M. Cmejrek, W. Zhang, J. Xue, J. Cui, B. Xiang, G. Daggett, U. Chaudhari, S. Maskey et E. Marcheret (2013). "The IBM speech-to-speech translation system for smartphone: Improvements for resource-constrained tasks." Computer Speech & Language **27**(2): 592-618.

Zong, C. et M. Seligman (2005). "Toward Practical Spoken Language Translation." Machine Translation **19**(2): 113-137.

Sites web par ordre alphabétique

Agreestat, http://www.agreestat.com/agreestat_excel.html, consulté le 24.03.2016.

Amazon Mechanical Turk - AMT, <https://www.mturk.com/mturk/welcome>, consulté le 01.07.2015.

Amazon Web Services, AMT, <https://aws.amazon.com/fr/mturk/details/>, consulté le 08.05.2014.

BabelDr, <http://babeldr.unige.ch/project-description/>, consulté le 01.03.2016.

BOLT par DARPA, <http://www.darpa.mil/program/broad-operational-language-translation>, consulté le 18.02.2016.

BOLT par l'Université de Columbia, Colorado, <http://www.cs.columbia.edu/speech/projects.cgi>, consulté le 24.04.2015.

CCLINC, sur le site de recherche militaire, <http://www.dtic.mil/docs/citations/ADA460574>, consultée le 27.06.2014.

Converser, <http://spokentranslation.com/technology/>, consulté le 16.02.2016

Crowdfower, <http://www.crowdfower.com/>, consulté le 01.07.2015.

Dallas Children's Hospital, <https://www.childrens.com/>, consulté le 18.02.2016.

DARPA, <http://www.darpa.mil/default.aspx>, consulté le 26.06.2014.

Digital Trends, <http://www.digitaltrends.com/mobile/best-language-translation-apps>, consulté le 05.08.2014.

EAGLES Final Report - Seven Steps, <http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html>, consulté le 02.08.2015.

EAGLES Project, <http://www.issco.unige.ch/en/research/projects/eagles/>, consulté le 30.06.2015.

Euromatrix, <http://www.euromatrix.net/publications/>, consulté le 23.06.2015.

Evalda-Cesta, <http://www.technolangue.net/article199.html?pagewanted=all>, consulté le 02.08.2011.

Fluential, <http://www.fluentialinc.com/>, consulté le 16.02.2016.

FranceTerme sur culture.fr, <http://www.culture.fr/franceterme/>, consulté le 28 juillet 2014.

GALE, [http://www.darpa.mil/Our_Work/I2O/Programs/Global_Autonomous_Language_Exploitation_\(GALE\).aspx](http://www.darpa.mil/Our_Work/I2O/Programs/Global_Autonomous_Language_Exploitation_(GALE).aspx), consulté le 02.08.2011

Google Blog, <http://googlesystem.blogspot.ch/2007/10/google-translate-switches-to-google.html>, consulté le 16.02.2016.

Google Translate App pour Android,
<https://play.google.com/store/apps/details?id=com.google.android.apps.translate&hl=fr>,
consulté le 23.06.2015

Google Translate, <https://support.google.com/translate/answer/6142474>, consulté le
24.04.2015.

Grand dictionnaire, <http://www.granddictionnaire.com/>, consulté le 28 juillet 2014.

HTK d'Entropics, <http://htk.eng.cam.ac.uk>, consulté le 16.02.2016.

Jibbigo, <http://www.cmu.edu/silicon-valley/news-events/news/2009/jibbigo.html>,
consulté le 05.08.2014.

Language and translation - online magazine,
[http://ec.europa.eu/dgs/translation/publications/magazines/languagestranslation/documents/iss
ue_06_en.pdf](http://ec.europa.eu/dgs/translation/publications/magazines/languagestranslation/documents/issue_06_en.pdf), consulté le 23.06.2015.

Lingenio, <http://www.lingenio.de/>, consulté le 16.02.2016.

Mastor d'IBM,
[http://domino.research.ibm.com/comm/research.nsf/pages/r.uit.innovation.html?Open&printa
ble](http://domino.research.ibm.com/comm/research.nsf/pages/r.uit.innovation.html?Open&printable), consulté le 16.02.2016.

Matecat, <https://www.matecat.com/features/>, consulté le 16.02.2016.

Microworkers, <https://microworkers.com/>, accédé le 01.07.2015.

Minifreelance, <http://minifreelance.com/>, accédé le 01.07.2015.

MITRE, <https://www.mitre.org/>, consulté le 15.04.2014.

Nespole, sur le site de l'Institut Fondazione Bruno Kessler,
<http://hlt.fbk.eu/projects/nespole-speech-speech-translation-e-commerceservice-scenarios>,
consultée le 27.06.2014.

Nuance, <http://www.nuance.com/naturallyspeaking/>, consulté le 16.01.2015.

Phraselator, <http://www.voxtec.com/phraselator/features/>, consulté le 27.06.2014.

SayHi Translate, <http://www.sayhitranslate.com/>, consulté le 17.04.2015.

SICS Swedish ICT, <http://www.sics.se/isl/sicstuswww/site/index.html>, consulté le
18.02.2016.

Speechlinks, Université South California, USC, <http://sail.usc.edu/transonics/s2s.php>, consulté le 17.04.2015.

SRI EduSpeak, <http://www.speechatsri.com/products/eduspeak.shtml>, consulté le 16.02.2016.

SRI International, <http://www.speech.sri.com/>, consultée le 27.06.2014.

TC-STAR Consortium, <http://tcstar.org/pages/info.htm>, consulté le 19.01.2015

Ted Talks, <https://www.ted.com/talks>, consulté le 17.04.2015.

Termdat, <https://www.termdat.bk.admin.ch/Search/Search>, consulté le 28.07.2014.

The Tartan, journal universitaire de Carnagie Mellon, <http://www.thetartan.org/2005/9/26/scitech/interact>, consulté le 27.06.2014.

TIDES, <http://projects ldc.upenn.edu/TIDES/Translation/TransAssess02.pdf>, consulté le 18.08.2011.

Tradessan hébergé par Migranthealth.eu, <http://migranthealth.eu/index.php/en/news/40-tradassan-an-app-to-facilitate-communication-with-foreign-patients>, consulté le 16.02.2016.

Traducmed, <http://traducmed.fr/>, consulté le 24.04.2015.

Universaldocor, <http://www.universaldocor.com/>, consulté le 16.02.2015

University Southwest, <http://www.utsouthwestern.edu/>, consulté le 18.02.2016.

U-star Consortium, <http://www.ustar-consortium.com/app/app.html>, consulté le 17.04.2015.

ViaVoice, d'IBM, http://www-306.ibm.com/software/pervasive/embedded_viavoice/, consulté le 16.02.2016.

Voice Response Translator, <http://www.miltrans.com/>, consulté le 16.02.2016.

Wordmagicsoft, <http://www.wordmagicsoft.com/>, consulté le 16.02.2016.

Yahoo Blog, <http://ca.news.yahoo.com/blogs/right-click/google-instant-translation-devices-could-turn-science-fiction-172350808.html>, consulté le 30.07.2013.

9 Liste des tableaux

Tableau 2.1 : Récapitulatif des campagnes et consortiums en TAP (Agency 2008 : p.98)	39
Tableau 2.2 : Exemple de dialogue avec la version unidirectionnelle.....	42
Tableau 2.3 : Exemple de dialogue avec la version bidirectionnelle.....	43
Tableau 2.4 : Exemple d'ellipse	50
Tableau 2.5 : Résolution d'ellipse par substitution.....	51
Tableau 2.6 : Classes d'ellipses pour le lieu	51
Tableau 2.7 : Généralisation des patrons à partir des exemples du Tableau 2.6	52
Tableau 2.8 : Exemple d'interaction avec le bidirectionnel	52
Tableau 2.9 : Résolution de l'ellipse de l'interaction de Fig. 3.10 par concaténation....	53
Tableau 2.10 : Récapitulatif des principaux systèmes de TAP du domaine médical	58
Tableau 3.1 : Principales étapes de l'évaluation selon EAGLES	62
Tableau 3.2 : Comparatif des caractéristiques et sous-caractéristiques de la qualité à l'usage selon ISO/IEC 9126 :1 et SQuaRE ISO/IEC 25010	66
Tableau 3.3 : Échelle de Likert et sa traduction en français	70
Tableau 3.4 : Omission d'un article lors de la reconnaissance par MedSLT	75
Tableau 3.5 : Substitution d'un article lors de la reconnaissance par MedSLT.	76
Tableau 3.6 : Échelle NIST pour la fluidité et l'adéquation	80
Tableau 3.7 : Tableau récapitulatif traduit et extrait de Hallgren, 2012	84
Tableau 3.8 : Échelle d'interprétation de (Landis et Koch, 1976) traduite en français. .	84
Tableau 3.9 : Correspondance entre deux phrases pour WER.....	86
Tableau 3.10 : Correspondance entre deux phrases pour BLEU	88
Tableau 3.11 : Correspondance entre deux phrases pour BLEU	88
Tableau 3.12 : Alignements TER de l'hypothèse, de la référence et les éditions effectuées pour deux exemples	91
Tableau 3.13 : Alignement METEOR de l'hypothèse, de la référence	92
Tableau 4.1 : Récapitulatif des questions de recherche de Q1_principale	103
Tableau 4.2 : Récapitulatif des questions de recherche de Q2_principale	107
Tableau 4.3 : Traduction du modèle de qualité retenu en sous-questions de recherche, variables dépendantes et indépendante	113
Tableau 4.4 : Extrait de protocole pour les sessions 1 et 2 (doc. = docteur, pat. = patient)	115

Tableau 4.5 : Instructions de LC pour V1	116
Tableau 4.6 : Instructions de LC pour V2	117
Tableau 4.7 : Extrait brut d'un dialogue retranscrit, scénario 1, patient 1 et médecin 1	118
Tableau 4.8: Variables dépendantes assorties des mesures choisies	120
Tableau 4.9 : Variables dépendantes et interprétation des mesures choisies.....	123
Tableau 4.10 : Répartition des 229 phrases (anglais-espagnol) entre les 14 évaluateurs de l'Hôpital de Dallas	125
Tableau 4.11 : Exemple d'application de l'échelle HMS	125
Tableau 4.12 : Total du nombre de phrases évaluées pour EXPQ2_S1	126
Tableau 4.13: Tableau récapitulatif de l'expérience EXPQ2_S1	127
Tableau 4.14: Tableau récapitulatif de l'expérience EXPQ2_S2	128
Tableau 4.15 : Données de l'expérience EXPQ2_S3	130
Tableau 4.16: Tableau récapitulatif de l'expérience EXPQ2_S3	130
Tableau 4.17 : Récapitulatif des groupes d'évaluateurs et du nombre d'évaluations...	133
Tableau 4.18: Tableau récapitulatif de l'expérience EXPQ2_S4-S6	134
Tableau 4.19: Tableau récapitulatif de l'expérience EXPQ2_S7	135
Tableau 4.20 : Tableau récapitulatif des expériences effectuées pour Q2_principale..	136
Tableau 5.1 : Reprise du tableau 4.3 su modèle de qualité et des variables d'EXPQ1	142
Tableau 5.2 : Reprise du tableau d'interprétation de la section 4.1.2.5, Tableau 4.9 ...	143
Tableau 5.3 : Accomplissement de la tâche en pourcentage de diagnostics posés.....	144
Tableau 5.4 : Diagnostic par médecin et par système – vert = V1, bleu = V2, 1=correct, 0=incorrect, ps =pas sûr, strep = angine à streptocoque, virale = angine virale	144
Tableau 5.5 : Extrait traduit du tableau de collecte d'information (0= info non-recueillie, 1= info recueillie).....	146
Tableau 5.6 : Temps et nombre d'interactions par scénario et par système	150
Tableau 5.7 : Taux de préférence par système pour chaque type d'utilisateur.....	153
Tableau 5.8 : Huit catégories de classement des interactions	154
Tableau 5.9 : Analyse des interactions des utilisateurs avec V1 et V2, en pourcentage	155
Tableau 5.10 : Exemple de formes négatives mal reconnues, produisant potentiellement des erreurs d'interaction.....	156
Tableau 5.11: Principaux résultats pour la qualité à l'usage et interprétation	158
Tableau 5.12 : Évaluation de la capacité fonctionnelle (extrait du Tableau 4.9).....	160

Tableau 5.13 : Taux d'erreur au mot (WER) et à la phrase (SER) pour tout le corpus	161
Tableau 5.14 : Taux d'erreur au mot (WER) et à la phrase (SER) pour les phrases couvertes par les systèmes V1 et V2	161
Tableau 5.15 : Pourcentage de phrases compris dans la couverture par système.....	162
Tableau 5.16 : Types de phrases employées par les patients dans V1 et V2.....	162
Tableau 5.17 : Types d'erreur de reconnaissance vocale par langue et par système....	163
Tableau 5.18 : Pourcentage de phrases rejetées à cause d'une mauvaise reconnaissance	164
Tableau 5.19 : Récapitulatif pour la precision de la RV	165
Tableau 5.20 : Extrait de l'évaluation binaire de la rétro-traduction.....	167
Tableau 5.21 : Exemples de désaccord entre utilisateurs en EXPQ1 et hors EXPQ1, en bleu V1, en blanc V2.	167
Tableau 5.22 : Exemples de phrases AC_IE lors de l'utilisation réelle, rejetées par les évaluateurs hispanophones hors expérience.	168
Tableau 5.23 : Exemple de formulation impersonnelle, rejetés pour les utilisateurs et évaluateurs	169
Tableau 5.24 : Exemple de décalage de temps pour V2.....	169
Tableau 5.25 : Pourcentage d'accord entre évaluateurs et kappa pour V1 et V2 pour l'évaluation de la rétro-traduction.....	170
Tableau 5.26 : Récapitulatif de la précision de la rétro-traduction.....	171
Tableau 5.27 : Fluidité et adéquation pour V1 et V2.....	172
Tableau 5.28 : Récapitulatif de la précision de la TA – évaluation humaine	173
Tableau 5.29 : Résultats de l'évaluation automatique	173
Tableau 5.30 : Récapitulation des scores de précision de la RV, la RétroTra et la TA	175
Tableau 5.31 : Taux de phrases hors couverture parmi les phrases mal reconnues.....	177
Tableau 5.32 : Exemples de phrases hors couverture	177
Tableau 5.33 : Extrait du questionnaire pour V1 concernant la facilité d'utilisation...	179
Tableau 5.34 : Extrait du Tableau 5.9.....	180
Tableau 5.35 : Evolution entre première et quatrième utilisation du médecin	182
Tableau 5.36 : Evolution entre première et deuxième utilisation du patient	183
Tableau 5.37: Récapitulation des scores obtenus pour la facilité d'utilisation.....	184
Tableau 5.38 : Score final pour V1 et V2	185
Tableau 6.1 : Moyenne des scores obtenus sur une échelle de 1-4 (1=DANG, 4=CCOR), sur toutes les phrases du corpus (avec répétition).....	193

Tableau 6.2 : Evaluation de la qualité de la traduction par catégorie	194
Tableau 6.3 : Extrait de l'évaluation par les évaluateurs de Dallas	195
Tableau 6.4 : Estimation du kappa de Fleiss avec Agreestat.....	195
Tableau 6.5 : Pourcentage d'accord entre les trois évaluations	196
Tableau 6.6 : Traduction unique pour multiples phrases LS	198
Tableau 6.7 : Corrélation entre les différentes mesures automatiques et HMS.....	200
Tableau 6.8 : Evaluation humaine, comparaison entre MedSLT et GT	202
Tableau 6.9 : Evaluation humaine, comparaison entre MedSLT et GT par catégories	202
Tableau 6.10 : Kappa de Fleiss pour MedSLT et GT pour différents groupes d'évaluateurs	203
Tableau 6.11 : Pourcentage d'accord entre évaluateurs pour MedSLT et GT	203
Tableau 6.12 : Résultats des mesures automatiques	205
Tableau 6.13 : Extrait de l'évaluation phrase par phrase.....	205
Tableau 6.14 : Corrélation Pearson entre les différentes mesures	207
Tableau 6.15 : Fluidité pour expert- maison comparés aux travailleurs AMT	214
Tableau 6.16 : Adéquation pour experts maison comparés aux travailleurs AMT	215
Tableau 6.17 : Pourcentage d'accord unanime entre évaluateurs.....	215
Tableau 6.18 : TURKoise	216
Tableau 6.19 : Accord par catégorie pour TURKoise.	217
Tableau 6.20 : Tableau récapitulatif du temps et des coûts	218
Tableau 6.21 : Évolution du kappa selon le nombre d'évaluations	219

10 Liste des figures

Figure 2.a : Architecture d'un système de reconnaissance vocale.....	25
Figure 2.b : Évolution de la technologie de la TAP, tiré de l'article (Nakamura 2009).	40
Figure 2.c : Fonctionnement de MedSLT	46
Figure 2.d : Flux de traitement avec MedSLT	48
Figure 2.e : Règles de traduction pour le bidirectionnel anglais-espagnol	49
Figure 2.f : Capture d'écran du bidirectionnel anglais-espagnol, onglet médecin	54
Figure 2.g : Capture d'écran du bidirectionnel anglais-espagnol, onglet patient	55
Figure 3.a : Relations entre qualité interne, externe et qualité à l'usage	63
Figure 3.b : Modèle de qualité externe de produit selon ISO/IEC 9126 :1 (2001).....	71
Figure 4.a : Modèle de qualité retenu pour notre contexte d'utilisation.....	108
Figure 4.b : Caractéristiques de qualité externe de notre modèle de qualité	110
Figure 4.c: Schéma récapitulatif de notre méthode de recherche	138
Figure 5.a : Accomplissement de la tâche en éliminant les deux scénarios difficiles. .	145
Figure 5.b : Récapitulation des mesures pour chaque version.....	148
Figure 5.c : Rappel et précision par version et par scénario	149
Figure 5.d : Comparaison entre temps et nombre d'interactions moyens pour chaque scénario avec V1 et V2	151
Figure 5.e : Première caractéristique externe retenu - capacité fonctionnelle	159
Figure 5.f : Types de phrases employées par les patients dans V1 et V2	162
Figure 5.g : Type de phrase hors couverture par système pour les patients	178
Figure 5.h : Comparaison de l'opinion des patients sur V1 et V2.....	179

11 Liste des annexes

Annexe 1 : Corpus d'origine pour les maux de gorge et scénarios

Annexe 2 : Protocole de l'EXPQ1

Annexe 3 : ExpQ2_S1 : fichiers Excel envoyés aux évaluateurs de Dallas

Annexe 4 : Instructions données aux utilisateurs pour MedSLT V1 et V2 (EXPQ1)

Annexe 5 : Scénarios de l'EXPQ1

Annexe 6 : Corpus récoltés et retranscrits (EXPQ1)

Annexe 7 : Tableau des informations récoltées par les médecins lors de l'EXPQ1

Annexe 8 : Evaluations HMS des experts de l'hôpital de Dallas

Annexe 9 : Questionnaires distribués aux utilisateurs lors de l'EXPQ1

Annexe 10 : Résultats du questionnaire de satisfaction concernant MedSLT en général

Annexe 11 : Exemple de fichier d'évaluation Excel pour EXPQ2_S4-S6

Annexe 1 : Corpus d'origine pour les maux de gorge et scénarios

Strep Pharyngitis History Text for English/Spanish Voice Recognition Project

Glenn Flores, MD
Medical College of Wisconsin

May 2, 2005

Please note: Each “do you” question should also have an option of “does your child?” (the question may be asked of the parent when the child is younger).

Do you have a sore throat? Yes/No

For how long have you had your sore throat? __ hours/ __ days

Do you have a fever? Yes/No

If yes, how high? __ degrees

Do you have a headache? Yes/No

Do you have a stomachache? Yes/No

Do you have a rash? Yes/No

Do you have swollen glands? Yes/No

Do you have congestion? Yes/No

Do you have a cough? Yes/No

Have you been in contact with anyone in the last week who has strep throat? Yes/No

Have you seen a doctor or nurse for your sore throat? Yes/No

If yes, when? __ days ago.

Was the diagnosis a strep throat? Yes/No

Was a throat culture or rapid strep test done? Yes/No

If Yes, what were the results? Strep throat diagnosis/not strep throat

Are you allergic to any antibiotic? Yes/No

If Yes, is it Penicillin, Amoxicillin, Augmentin, or a cephalosporin? Yes/No

If Yes, which one? _____

Annexe 2 : Protocole de l'EXPQ1

1.	Install machine and headphones		
2.	Launch MedSLT restricted		
3.	Check mikes and headphones		
4.	Check system with 2-3 sentences		
5.	Prepare instructions for Physician 1		
	Timing	Temps	
1.	ID_docteur:	-	
2.	Accueil du médecin 1	8h10	
Intro_speech :			
	Dear Dr. X, during this session, we are going to test two different versions of the spoken language translation system MedSLT. The aim of the system, as you might know, is to facilitate the dialogue between a physician and a patient that doesn't speak your language.		
	In order to test this system, you are going to simulate a patient interview. You will see four different patients and each patient twice. These patients suffer from a sore throat speak only Spanish; you will thus use the system to find out if they suffer from a strep throat or merely from a viral sore throat.		
	To interact with the system, we need to talk into a mike and a headphone. Before we welcome the patient, I would like you to get familiar with the system. Please, read these instructions carefully, and then, we'll test the mike position by running the system briefly.		
	<i>Give instructions Doc ENG</i>		
	But before we start, CONSENT SCRIPT		
Phase 1 : Doc training			
1.	Let doc read carefully		
2.	Test a couple of sentences to check mike and system understood		
3.	Remind task : make diagnosis and fill in info form		
4.	<i>Hand out info form page 1</i>		
Phase 2 : diagnosis			
Patient 1			
5.	Patient 1 enters Room 1	8h15	
6.	Give instructions to read (restricted)		
7.	Start visit 1	8h20	
8.	Physician fills in diagnosis 1, and info		
9.	End of visit 1	8h30	

10.	Patient 1 goes to Room 2	→13h35	
11.	RELAUNCH MEDSLT		
Patient 2			
12.	Patient 2 enters Room 1		
13.	Give instructions to read (restricted)		
14.	Start visit 2		
15.	Physician fills in diagnosis 2, and info		
16.	End of visit 2		
17.	Patient 2 fills in form		
18.	Gets 50 \$, sings receipt and leaves		
Physician 1 takes a 10 min break			
Patient 3			
19.	Patient 3 enters Room 1		
20.	Welcome and consent		
21.	Give instructions to read (unrestricted)		
22.	Start visit 3		
23.	Physician fills in diagnosis 3, and info		
24.	End of visit 3		
25.	Patient 3 goes to Room 2		
Patient 4			
26.	Patient 4 enters Room 1		
27.	Give instructions to read (unrestricted)		
28.	Start visit 4		
29.	Physician fills in diagnosis 4, and info		
30.	End of visit 4		
31.	Patient 4 fills in form		
32.	Physician 1 fills in questionnaire		
33.	Patient 4 gets 50 \$, signs receipt and leaves		
34.	Phys 1 gets 50 \$, signs receipt and leaves		
END SESSION 1			

Annexe 3 : ExpQ2_S1 : fichiers Excel envoyés aux évaluateurs de Dallas

<u>Instructions:</u> Please rate the following sentences according to the scale given below, by picking the corresponding choice from the list. Please note, that the first 5 lines in grey are examples of how to apply the scale. Please DO give us a comment in the "Comments" column or a better translation suggestion, whenever you have a doubt or if you really don't like the translation. The context (1st column) should be taken into consideration, since sometimes the Source is an incomplete sentence (cf. example 11). Thank you.				
Scale	Description	Short		
Completely correct meaning	The translation is completely correct. All the meaning from the source is present in the target sentence.	CCOR		
Not completely correct meaning but not dangerous	The translation is not completely correct. The meaning is slightly different by it represents no danger of miscommunication between doctor and patient	MEAN		
Dangerous (False Sense)	This translation is incorrect. The meaning in the target and in the source are very different. It is a false sense, dangerous for communication between doctor and patient.	DANG		
Nonsense	This translation doesn't make any sense. It is giberish. This translation is not a correct in the target language.	NONS		
EXAMPLE EVALUATION				
CONTEXT	SOURCE	TARGET	RATING	Comments
do you vomit	are you better	¿el dolor disminuye?	MEAN	give comment here
(context irrelevant)	are you congested	¿tiene la nariz tapada?	CCOR	give comment here
is the pain getting better	are you having fever	¿el dolor está aliviado cuando tiene fiebre?	DANG	give comment here
did the doctor do a throat culture test	are you taking any medicine	¿toma medicina?	CCOR	give comment here
have you taken any medicine for the sore throat	did the doctor run a strep test	¿le han realizado a menudo la prueba rápida por estreptococo?	MEAN	give comment here
CONTEXT	SOURCE	TARGET	EVAL	COMMENTS
have you had a headache	any stomache ache	¿el estómago le ha dolido?	?	
is anybody around you suffering from a sore throat	are you alergic to any antibiotics	¿tiene alergia a algunos antibióticos?	?	

have you been in contact with anyone with a strep throat	are you allergic to antibiotics	¿los antibióticos le dan alergia?	?	
(context irrelevant)	are you allergic to any antibiotic	¿tiene alergia a algunos antibióticos?	?	
(context irrelevant)	are you allergic to any antibiotics	¿ha tenido alergia a algunos antibióticos?	?	
is anybody around you suffering from a sore throat	are you allergic to any antibiotics	¿tiene alergia a algunos antibióticos?	?	
have you taken any medicine for the pain	are you allergic to any medicine	¿tiene alergia a alguna medicina?	?	
did the doctor do a rapid strep test	are you allergic to any medicines	¿tiene alergia a alguna medicina?	?	
have you been in contact with anyone who has a sore throat	are you allergic to any medication	¿ha tenido alergia a alguna medicina?	?	
are you taking penicillin	are you allergic to any medication	¿tiene alergia a alguna medicina?	?	
(context irrelevant)	are you allergic to any medications	¿ha tenido alergia a alguna medicina?	?	
have you recently been in contact with anyone with a strep throat	are you allergic to any medications	¿tiene alergia a alguna medicina?	?	
have you taken any medicine for the sore throat	are you allergic to any medicine	¿ha tenido alergia a alguna medicina?	?	
(context irrelevant)	are you allergic to any medicine	¿tiene alergia a alguna medicina?	?	
(context irrelevant)	are you allergic to any medicines	¿tiene alergia a alguna medicina?	?	
(context irrelevant)	are you allergic to anything	¿medicina le da alergia?	?	
do you vomit	are you better	¿el dolor disminuye?	?	
(context irrelevant)	are you congested	¿tiene la nariz tapada?	?	
have you seen the doctor this week	are you coughing	¿tiene tos?	?	
do you vomiting	are you getting better	¿el dolor disminuye?	?	
are you congested	are you glands swollen	¿tiene los ganglios inflamados?	?	
is the pain getting better	are you having fever	¿el dolor está aliviado cuando tiene fiebre?	?	
did the doctor do a throat culture test	are you taking any medicine	¿toma medicina?	?	
is anybody around you suffering from a sore throat	are you taking any medicine for the pain	¿ha tomado medicina para el dolor?	?	
what is the character of the pain	are you taking anything for the pain	¿toma medicina para el dolor?	?	
(context irrelevant)	are you taking medicines	¿toma medicina?	?	

have you taken any medicine for the sore throat	are you taking penicillin	¿ha tomado la penicilina?	?	
do you have vomiting	are you vomiting	¿vomita?	?	
(context irrelevant)	did the doctor do a rapid strep test	¿le han realizado la prueba rápida por estreptococo?	?	
have you seen a doctor	did the doctor do a strep test	¿le han realizado la prueba rápida por estreptococo?	?	
(context irrelevant)	did the doctor do a throat culture test	¿le han realizado un cultivo de la garganta?	?	
have you taken any medicine for the sore throat	did the doctor run a strep test	¿le han realizado a menudo la prueba rápida por estreptococo?	?	
have you seen a doctor	did they do a strep test	¿le han realizado la prueba rápida por estreptococo?	?	
did you see a doctor this week	did you have a throat culture	¿le han realizado un cultivo de la garganta?	?	
(context irrelevant)	did you have any test	¿le han realizado un examen?	?	
have you been to another doctor	did you have any tests	¿le han realizado un examen?	?	
have you seen a doctor	did you have any tests done	¿le han realizado un examen?	?	
have you seen a doctor	did you have any tests done	¿le realizan un examen?	?	
have you been in contact with anyone recently who has a sore throat	did you see a doctor this week	¿ha consultado un médico esta semana?	?	
(context irrelevant)	did you see the doctor this week	¿ha consultado un médico esta	?	
are you allergic to any medicine	did you see the doctor today	¿ha consultado hoy un médico?	?	
have you been to the doctor	did you see the doctor two days ago	¿ha consultado un médico hace dos días?	?	
sí me tome unas pastillas para el dolor de garganta	did you take an antibiotic	¿ha tomado antibióticos?	?	
(context irrelevant)	did you take antibiotics	¿ha tomado antibióticos?	?	
do you have any fever	do you have congestion	¿tiene la nariz tapada?	?	
do you have a rash	do you feel sick	¿tiene náuseas?	?	
do you have a headache	do you have	¿tiene el dolor?	?	
how high is your fever	do you have a cough	¿ha tenido tos?	?	
(context irrelevant)	do you have a cough	¿tiene tos?	?	
(context irrelevant)	do you have a fever	¿tiene fiebre?	?	
(context irrelevant)	do you have a headache	¿le duele la cabeza?	?	

do you have a cough	do you have a headache	¿le ha dolido la cabeza?	?	
do you have a cough	do you have a headache	¿tiene tos ayer?	?	
do you have a headache	do you have a rash	¿ha tenido una erupción cutánea?	?	
(context irrelevant)	do you have a rash	¿tiene una erupción cutánea?	?	
(context irrelevant)	do you have a runny nose	¿ha mucho tenido moco?	?	
(context irrelevant)	do you have a runny nose	¿tiene mucho moco?	?	
(context irrelevant)	do you have a sore throat	¿le duele la garganta?	?	
no_context	do you have a sore throat today	¿le duele la garganta hoy?	?	
(context irrelevant)	do you have a stomachache	¿el estómago le duele?	?	
(context irrelevant)	do you have a stomach ache	¿el estómago le duele?	?	
(context irrelevant)	do you have any congestion	¿tiene la nariz tapada?	?	
me duele mucho y me arde por dentro	do you have any drainage in your throat	¿tiene mucho una infección de la garganta por estreptococo?	?	
how long has your throat been hurting you for	do you have any fever	¿ha tenido fiebre?	?	
dos días	do you have any fever	¿tiene fiebre?	?	
for how long have you had your sore throat	do you have any fever with it	¿tiene fiebre?	?	
out of context	do you have any pain	¿tiene el dolor?	?	
do you have a cough	do you have any rashes	¿tiene una erupción cutánea?	?	
(context irrelevant)	do you have congestion	¿tiene la nariz tapada?	?	
(context irrelevant)	do you have cough	¿tiene tos?	?	
do you have swollen glands	do you have fever	¿ha tenido fiebre?	?	
(context irrelevant)	do you have fever	¿tiene fiebre?	?	
(context irrelevant)	do you have headache	¿le duele la cabeza?	?	
(context irrelevant)	do you have runny nose	¿tiene mucho moco?	?	
no_context	do you have sore throat	¿le duele la garganta?	?	
do you have	do you have stomach ache	¿el estómago le duele?	?	
have you taken any medicine	do you have strep throat	¿tiene una infección de la garganta por estreptococo?	?	
(context irrelevant)	do you have swollen glands	¿tiene los ganglios inflamados?	?	
do you feel sick	do you have vomiting	¿el estómago le duele?	?	
what were the results of this test	do you take any medicines	¿toma medicina?	?	
are you allergic to any medicine	do you think you are getting better	¿el dolor disminuye?	?	
do you have a headache	do you vomit	¿vomita?	?	
do you have congestion	do you vomiting	¿vomita?	?	

are you congested	does anyone have a sore throat	¿alguien tiene el dolor de garganta?	?	
no_context	does your throat hurt	¿le duele la garganta?	?	
no_context	does your throat hurt	¿le ha dolido la garganta?	?	
have you had runny nose	fever	¿ha tenido fiebre?	?	
(context irrelevant)	for how long have you had a sore throat	¿desde cuándo le duele la garganta?	?	
where is your pain	for how long have you had this pain	¿desde cuándo tiene el dolor?	?	
(context irrelevant)	for how long have you had your sore throat	¿desde cuándo le duele la garganta?	?	
do you have a sore throat today	for how many days	¿desde cuándo tiene el dolor?	?	
(context irrelevant)	good morning	¿hola?	?	
have you felt tired	have you been around anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	?	
is the pain getting better	have you been around anyone with a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	?	
do you have congestion	have you been around anyone with sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	?	
is the pain getting better	have you been around someone with a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	?	
(context irrelevant)	have you been in contact with a sick person	¿ha estado en contacto con una persona enferma?	?	
do yo have congestion	have you been in contact with an ill person	¿ha estado en contacto con esa persona?	?	
have you been in contact with anyone who has a sore throat	have you been in contact with anyone in the last week who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta en la semana pasada?	?	
for how long have you had your sore throat	have you been in contact with anyone in the last week who has strep throat	¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo en la semana pasada?	?	
(context irrelevant)	have you been in contact with anyone recently who has a sore throat	¿ha estado recientemente en contacto con alguien que tiene el dolor de garganta?	?	
(context irrelevant)	have you been in contact with anyone recently who has strep throat	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	?	

(context irrelevant)	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	?	
do you have a fever	have you been in contact with anyone who has had a strep throat recently	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	?	
(context irrelevant)	have you been in contact with anyone with a strep throat	¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo?	?	
(context irrelevant)	have you been in contact with anyone with a strep throat in the last week	¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo en la semana pasada?	?	
(context irrelevant)	have you been in contact with anyone with a strep throat recently	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	?	
(context irrelevant)	have you been in contact with someone with a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	?	
(context irrelevant)	have you been to a doctor	¿ha consultado un médico?	?	
(context irrelevant)	have you been to another doctor	¿ha consultado un otro médico?	?	
do you have a sore throat	have you been to see a doctor	¿ha consultado un médico?	?	
(context irrelevant)	have you been to the doctor	¿ha consultado un médico?	?	
is anybody around you suffering from a sore throat	have you been to visit the doctor	¿ha consultado un médico?	?	
how long have you had a headache	have you felt tired	¿ha estado cansado?	?	
(context irrelevant)	have you had a cough	¿ha tenido tos?	?	
(context irrelevant)	have you had a fever	¿ha tenido fiebre?	?	
(context irrelevant)	have you had a headache	¿le ha dolido la cabeza?	?	
(context irrelevant)	have you had a rash	¿ha tenido una erupción cutánea?	?	
(context irrelevant)	have you had a runny nose	¿ha mucho tenido moco?	?	
how long have you had a sore throat	have you had a sore throat for several days	¿le duele la garganta desde hace algunos días?	?	
(context irrelevant)	have you had a stomach ache	¿el estómago le ha dolido?	?	
do you have a stomachache	have you had a strep test done	¿le han realizado la prueba rápida por estreptococo la noche?	?	
what medicine did you take	have you had any congestion	¿le han realizado un examen?	?	
have you had a rash	have you had any tests done	¿le han realizado un examen?	?	
have you had a stomach ache	have you had any tests run	¿le han realizado un examen?	?	

(context irrelevant)	have you had congestion	¿ha tenido la nariz tapada?	?	
have you had a headache	have you had cough	¿ha tenido tos?	?	
(context irrelevant)	have you had fever	¿ha tenido fiebre?	?	
have you had congestion	have you had runny nose	¿ha mucho tenido moco?	?	
(context irrelevant)	have you recently been in contact with anyone with a strep throat	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	?	
(context irrelevant)	have you seen a doctor	¿ha consultado un médico?	?	
(context irrelevant)	have you seen a doctor for this sore throat	¿ha consultado un médico para el dolor de garganta?	?	
(context irrelevant)	have you seen a doctor for your sore throat	¿ha consultado un médico para el dolor de garganta?	?	
have you taken any medication	have you seen a doctor lately	¿ha consultado recientemente un médico?	?	
(context irrelevant)	have you seen a doctor recently	¿ha consultado recientemente un médico?	?	
have you recently been in contact with anyone with a strep throat	have you seen another doctor	¿ha consultado un otro médico?	?	
how long have you had the pain	have you seen any other doctors	¿ha consultado un otro médico?	?	
are you allergic to any antibiotics	have you seen the doctor this week	¿consulta un médico desde hace esta semana?	?	
(context irrelevant)	have you taken any medication	¿ha tomado medicina?	?	
(context irrelevant)	have you taken any medicine	¿ha tomado medicina?	?	
have you been in contact with anyone who has a sore throat	have you taken any medicine for the pain	¿ha tomado medicina para el dolor?	?	
(context irrelevant)	have you taken any medicine for the sore throat	¿ha tomado medicina para el dolor de garganta?	?	
(context irrelevant)	have you taken any medicine for your pain	¿ha tomado medicina para el dolor?	?	
sí	have you taken any medicines	¿ha tomado medicina?	?	
no todavia no	have you taking any medicines	¿ha tomado medicina?	?	
(context irrelevant)	have you visited a doctor	¿ha consultado un médico?	?	
no_source	have you visited the doctor	¿ha consultado un médico?	?	
are you allergic to any antibiotic	have you vomited	¿ha vomitado?	?	
do you have a cough	headache	¿le duele la cabeza?	?	
do you have a fever	how high is the fever	¿cuántos grados tiene?	?	

do you have a fever	how high is the temperature	¿cuántos grados tiene?	?	
(context irrelevant)	how high is your fever	¿cuántos grados tiene?	?	
sí	how high was the fever	¿cuántos grados ha tenido?	?	
(context irrelevant)	how high was your fever	¿cuántos grados ha tenido?	?	
fever	how high was your temperature	¿cuántos grados ha tenido?	?	
do you have a headache	how long did you have a headache	¿desde cuándo le duele la cabeza?	?	
is your throat hurting	how long has your throat been hurting you for	¿desde cuándo le duele la garganta?	?	
(context irrelevant)	how long has your throat hurt	¿desde cuándo le duele la garganta?	?	
do you have a fever	how long have you had a fever	¿desde cuándo tiene fiebre?	?	
(context irrelevant)	how long have you had a headache	¿desde cuándo le duele la cabeza?	?	
do you have a runny nose	how long have you had a runny nose	¿desde cuándo tiene mucho moco?	?	
(context irrelevant)	how long have you had a sore throat	¿desde cuándo le duele la garganta?	?	
where is the pain	how long have you had a sore throat	¿desde cuándo le ha dolido la garganta?	?	
(context irrelevant)	how long have you had a stomachache	¿desde cuándo le duele el estómago?	?	
(context irrelevant)	how long have you had congestion	¿desde cuándo tiene la nariz tapada?	?	
where is your pain	how long have you had strep throat	¿desde cuándo le duele la garganta?	?	
do you have congestion	how long have you had the cough	¿desde cuándo tiene tos?	?	
have you had a headache	how long have you had the headache	¿desde cuándo le duele la cabeza?	?	
(context irrelevant)	how long have you had the pain	¿desde cuándo tiene el dolor?	?	
(context irrelevant)	how long have you had your sore throat	¿desde cuándo le duele la garganta?	?	
do you have a fever	how many days of sore throat	¿cuál es la intensidad del dolor en la garganta?	?	
for how long have you had your sore throat	how severe is it	¿cuál es la intensidad del dolor?	?	
(context irrelevant)	how severe is the pain	¿cuál es la intensidad del dolor?	?	
how long have you had a sore throat	how severe is your pain	¿cuál es la intensidad del dolor?	?	

do you have a headache	how severe is your sore throat	¿cuál es la intensidad del dolor en la garganta?	?	
(context irrelevant)	is anybody around you suffering from a sore throat	¿alguien en su entorno tiene el dolor de garganta?	?	
do you have a rash	is anyone around you suffering from a sore throat	¿alguien en su entorno tiene el dolor de garganta?	?	
when did your sore throat start	is the pain accompanied by fever	¿tiene fiebre cuando tiene el dolor?	?	
is the pain getting better	is the pain becoming more frequent	¿le duele más a menudo?	?	
(context irrelevant)	is the pain better	¿el dolor disminuye?	?	
(context irrelevant)	is the pain getting better	¿el dolor disminuye?	?	
did the doctor do a rapid strep test	is the pain less frequent	¿le duele menos a menudo?	?	
(context irrelevant)	is the pain less severe	¿el dolor disminuye?	?	
what is the type of pain	is the pain severe	¿el dolor es severo?	?	
what kind of pain is it	is there a cough	¿tiene tos?	?	
where is your pain	is your throat hurting	¿le duele la garganta?	?	
no_context	out of context	¿desde cuándo tiene fiebre?	?	
la cabeza se siente muy caliente	out of context	¿ha tenido la nariz tapada?	?	
headache	runny nose	¿tiene mucho moco?	?	
have you had congestion	sí	¿ha tenido la nariz tapada?	?	
are you taking anything for the pain	was a rapid strep test done	¿le han realizado la prueba rápida por estreptococo?	?	
(context irrelevant)	was a strep test done	¿le han realizado la prueba rápida por estreptococo?	?	
have you visited a doctor	was a throat culture test carried out	¿le han realizado un cultivo de la garganta?	?	
did you see the doctor this week	was a throat culture test done	¿le han realizado un cultivo de la garganta?	?	
did the doctor do a rapid strep test	was it positive for streptococcus	¿el resultado ha sido positivo?	?	
(context irrelevant)	was the diagnosis a strep throat	¿le han diagnosticado una infección de la garganta por estreptococo?	?	
did the doctor run a strep test	were the results of the test positive	¿el resultado del examen ha sido positivo?	?	
did the doctor do a rapid strep test	were the results positive	¿el resultado ha sido positivo?	?	
are you allergic to anything	what are you allergic to	¿qué le da alergia?	?	
(context irrelevant)	what is the character of the pain	¿cómo es el dolor?	?	

(context irrelevant)	what is the character of the pain	¿cómo es el dolor?	?	
how long have you had a sore throat	what is the character of your pain	¿cómo es el dolor?	?	
(context irrelevant)	what is the type of pain	¿cómo es el dolor?	?	
do you have a fever	what is your temperature	¿cuántos grados tiene?	?	
(context irrelevant)	what kind of pain is it	¿cómo es el dolor?	?	
have you taken any medicine for the sore throat	what medicine are you taking	¿qué medicina toma?	?	
have you taken any medicines	what medicine did you take	¿qué medicina ha tomado?	?	
(context irrelevant)	what was the result	¿cuál ha sido el resultado?	?	
(context irrelevant)	what was the result of the test	¿cuál ha sido el resultado del examen?	?	
was a throat culture test carried out	what was their diagnosis	¿qué le han diagnosticado?	?	
have you had fever	what was your temperature	¿cuántos grados ha tenido?	?	
(context irrelevant)	what were the results	¿cuál ha sido el resultado?	?	
(context irrelevant)	what were the results of the test	¿cuál ha sido el resultado del examen?	?	
did the doctor do a throat culture test	what were the results of the throat culture test	¿cuál ha sido el resultado del cultivo de la garganta?	?	
(context irrelevant)	what were the results of this test	¿cuál ha sido el resultado del examen?	?	
was a rapid strep test done	what were the results of this testing	¿cuál ha sido el resultado del examen?	?	
where is the pain	when did it start	¿cuándo ha empezado el dolor?	?	
(context irrelevant)	when did the pain start	¿cuándo ha empezado el dolor?	?	
(context irrelevant)	when did you see the doctor	¿cuándo ha consultado un médico?	?	
have you seen a doctor	when did you see the nurse	¿cuándo ha consultado una enfermera?	?	
did you take an antibiotic	when did you take the antibiotics	¿cuándo ha tomado antibióticos?	?	
(context irrelevant)	when did you visit the doctor	¿cuándo ha consultado un médico?	?	
(context irrelevant)	when did your fever start	¿desde cuándo tiene fiebre?	?	
when did your fever start	when did your sore throat start	¿cuándo empieza el dolor de garganta?	?	
do you have a sore throat	when did your sore throat start	¿cuándo ha empezado el dolor de garganta?	?	
(context irrelevant)	when have you been in contact with a person who has strep throat	¿cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo?	?	
(context irrelevant)	where is the pain	¿dónde le duele?	?	
(context irrelevant)	where is your pain	¿dónde le duele?	?	
how high is your fever	which antibiotics are you allergic to	¿qué antibiótico le da alergia?	?	

Annexe 5 : Instructions données aux utilisateurs pour MedSLT V1 et V2 (EXPQ1)

MedSLT 1 : Restricted patient

- Traduce lo que dice el médico al idioma que habla el paciente (en este caso del inglés al español)
- Traduce lo que contesta el paciente al médico

Su tarea

- USTED es el paciente, no habla inglés
- El médico a su lado no habla español
- Tiene dolores de garganta
- Conteste a las preguntas del médico (traducidas por el sistema), usando el sistema MedSLT 2 según los síntomas descritos en la página que recibirá justo antes de empezar.

Desarrollo del test

- Lea atentamente estas instrucciones
- No muestre estas instrucciones al médico
- No muestre la lista de síntomas
- Conteste a las preguntas que el médico le hace por medio del sistema MedSLT 2
- La tarea está completada cuando el médico emita un diagnóstico

¿Cómo hablar con el sistema?

Puede usar solo frases incompletas o contestar por *sí / no*

	DIGA:	NO DIGA:
<u>Médico:</u> ¿desde cuándo le duele la garganta?	• desde hace más de una semana	• me duele desde hace más de una semana
<u>Médico:</u> ¿Tiene fiebre?	• sí, un poco • sí, más de 38 grados • no • si • no sé	• sí, tengo un poco de fiebre • sí, tengo más de 38 grados • no no tengo fiebre • sí, tengo fiebre

Introducción a la interfaz

The screenshot shows the MedSLT application window. It has a menu bar (File, Settings, Help) and a toolbar with flags for Doctor and Patient. The main area is divided into sections: History, Paciente, and Respuesta. The History section on the left shows a list of previous interactions. The Paciente section has buttons for 'Pulsa y habla' and 'Pulsa y escucha'. The Respuesta section displays the system's understanding of the patient's input and a list of possible responses. A translation section at the bottom shows the English translation of the patient's input. Callouts are numbered 1 through 6, providing instructions on how to use the interface.

(1) Habla cuando la pestaña "Patient" está activa

(2) Haga clic en este botón justo antes de hablar

(3) Verifique que el sistema ha entendido correctamente

(4) Si lo necesita, verifique lo que el sistema ha reconocido

(5) Si el sistema no ha entendido correctamente, puede seleccionar una frase de la lista de respuestas de la ventana de Ayuda

(6) Haga clic en este botón para hacer escuchar la traducción de su respuesta en inglés

Historial del diálogo

Traducción de su respuesta en inglés

Cómo usar el sistema

- El médico y el paciente comparten el ratón
- Antes de hablar, haga clic sobre el botón "Pulsa y habla" (2)
- Después, espera que aparezca algo en los campos: "Respuesta" (3), "El sistema ha entendido" (4) y "Traducción" (5)
- Si el sistema le ha entendido correctamente, haga clic sobre el botón "Pulsa y escucha" (6) para que el médico oiga su respuesta en inglés
- Si el sistema NO le ha entendido correctamente, haga clic de nuevo sobre el botón « Pulsa y habla » (1) y reformule.
- También puede seleccionar directamente en la ventana "Ayuda" la respuesta que quería dar si está propuesta en la lista.
- Hable claramente

MedSLT V2 : Unrestricted patient

- Traduce lo que dice el médico al idioma que habla el paciente (en este caso del inglés al español)
- Traduce lo que contesta el paciente al médico

Su tarea

- USTED es el paciente, no habla inglés
- El médico a su lado no habla español
- Tiene dolores de garganta
- Conteste a las preguntas del médico (traducidas por el sistema), usando el sistema MedSLT 1 según los síntomas descritos en la página que recibirá justo antes de empezar.

Desarrollo del test

- Lea atentamente estas instrucciones
- No muestre estas instrucciones al médico
- No muestre la lista de síntomas
- Conteste a las preguntas que el médico le hace por medio del sistema MedSLT1
- La tarea está completada cuando el médico emita un diagnóstico

¿Cómo hablar con el sistema?

1. Expresa una idea por respuesta

<u>Médico:</u>	Diga:	No diga
¿Dónde le duele?	• me duele la garganta • me duele la cabeza	• me duele la garganta y la cabeza

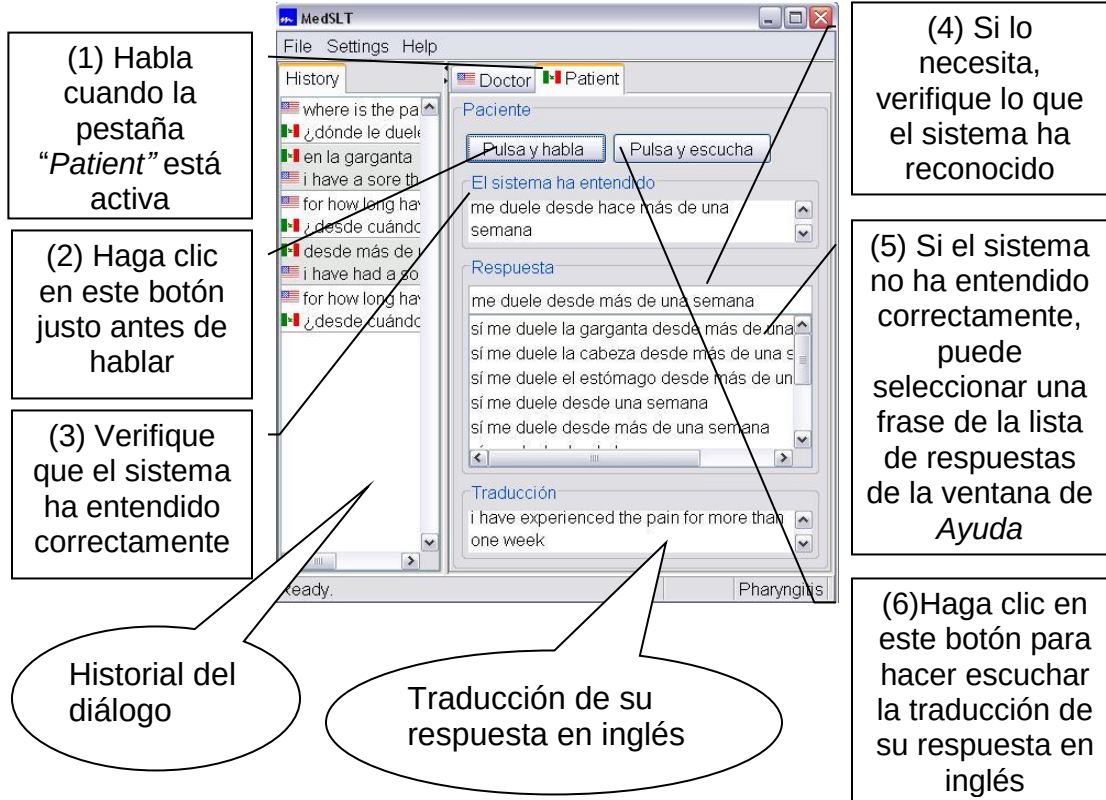
2. Puede usar

	frases enteras	frases incompletas
<u>Médico:</u> ¿Desde cuándo le duele la garganta?	• me duele desde hace más de una semana	• desde hace más de una semana
<u>Médico:</u> ¿Tiene fiebre?	• sí, tengo mucha fiebre • no, tengo nauseas	• sí, tengo mucha fiebre • no, tengo nauseas • si • no sé

3. No puede usar frases negativas:

<u>Médico:</u> ¿Tiene fiebre?-	Diga:	No diga
	• no	• no tengo fiebre

Introducción a la interfaz



Cómo usar el sistema

- El médico y el paciente comparten el ratón
- Antes de hablar, haga clic sobre el botón "Pulsa y habla" (2)
- Después, espera que aparezca algo en los campos: "Respuesta" (3), "El sistema ha entendido" (4) y "Traducción" (5)
- Si el sistema le ha entendido correctamente, haga clic sobre el botón "Pulsa y escucha" (6) para que el médico oiga su respuesta en inglés
- Si el sistema NO le ha entendido correctamente, haga clic de nuevo sobre el botón « Pulsa y habla » (1) y reformule.
- También puede seleccionar directamente en la ventana "Ayuda" la respuesta que quería dar si está propuesta en la lista.
- Hable claramente

MedSLT V1 and V2 : Physician

- Translates what the physician says (in English) into Spanish
- Translates what the patient answers (in Spanish) into English

Your task:

- YOU are the doctor, you ONLY speak English
- Your patient, sitting next to you, ONLY speaks Spanish
- You have to ask him/her questions to determine if he/she has a strep throat or a viral sore throat

Steps to follow:

1. Read these instructions carefully
2. Ask the patient diagnosis questions using the MedSLT system
3. The task is completed once you have made your diagnosis, i.e. when you say either: « you have a strep throat » or « you have a viral sore throat ».

How to formulate questions:

4. simple sentences, one idea per question

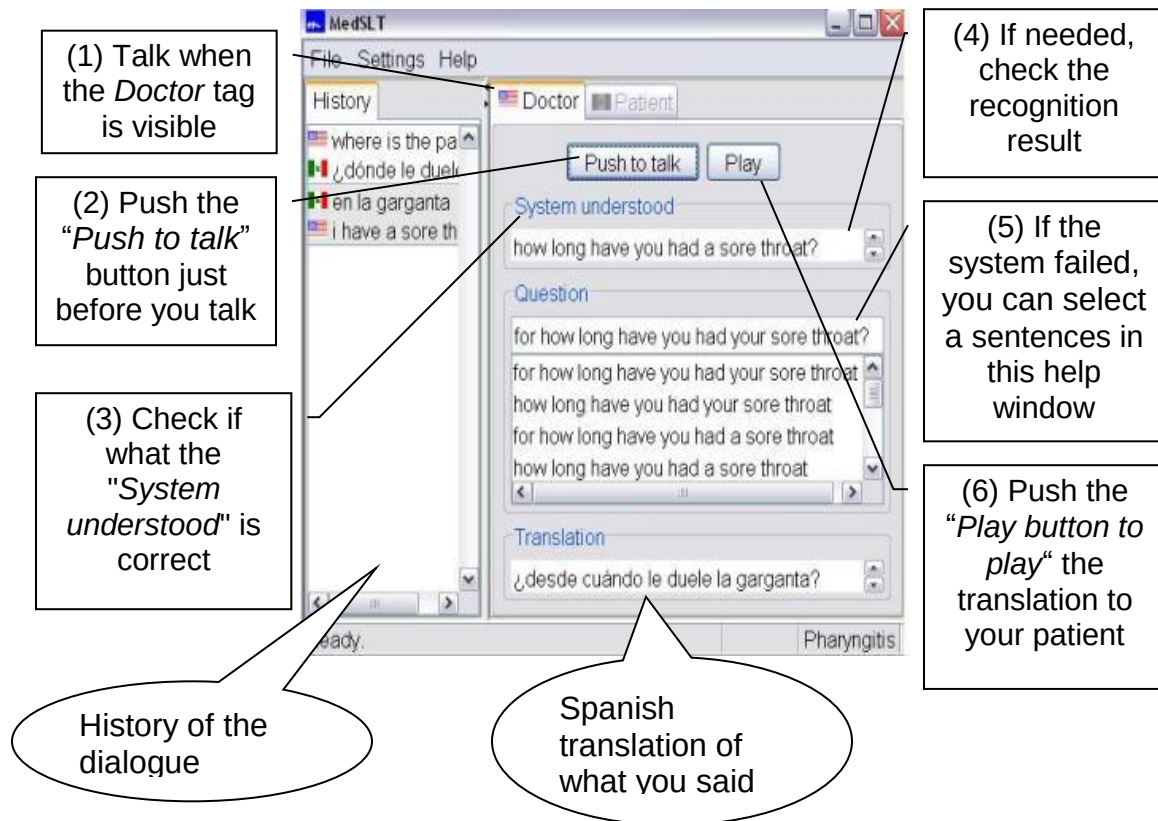
- **SAY:**
 - For how long have you had your sore throat? [Patient answers]
 - For more than a week? [Patient answers]
- **DO NOT SAY:**
 - For how long have you had your sore throat, for more than a week?
- **DO NOT SAY:**
 - Do you have a sore throat **or** a headache?

5. you don't have to repeat yourself

If the question you want to ask is very similar to the previous one, try just saying the part that is different. For example

- **YOUR PREVIOUS QUESTION WAS**
 - Do you have a sore throat?
- **YOU CAN SAY EITHER**
 - Do you have a headache?
 - A headache?
- **YOUR PREVIOUS QUESTION WAS**
 - When did you visit the doctor?
- **YOU CAN SAY EITHER**
 - Did you visit the doctor this week?
 - This week?

Introduction to the interface



How to use MedSLT

- Before speaking, click the “*Push to talk*” button (2).
- Speak
- Check if the system understood you correctly in the “*System understood*” (3)
- If what the system has understood seems correct, push the “*Play*” button (4) so that the patient can hear the translation into Spanish
- If what the system has understood is incorrect, or if the recognition failed, then either
 - start again by clicking the «*Push to talk*» (1) button
 - or select a sentence directly from the help window (5) and play it by pushing the “*Play*” button.
- Wait until the patient has answered using the system before asking the next question.

Test the following sentences with the person supervising your session (Don't forget to push the “*Play*” button)

- Where is your pain?
- For how long have you had your sore throat?
- Do you have a fever?

Annexe 5 : Scénarios de l'EXPQ1

Scenario 1

The strength of the symptom is represented by the number of + in the brackets.

	You have
Pain where	throat
Intensity of pain	(++)
Duration	more than a week
Other symptoms	fever (39,5 C°) headache (++) swollen glands (++) rash (+)
Contact with ill person	yes, 2 weeks ago
Previous medical visit for pain	yes doctor last Monday no strep diagnosed
Test carried out	rapid strep test result negative
Allergies to antibiotics	augmentin

Scenario 2

The strength of the symptom is represented by the number of + in the brackets.

	Scenario 2
Pain where	throat
Intensity of pain	(+)
Duration	less than a week
Other symptoms	fever (40 C°) 2 days runny nose stomachache
Contact with ill person	no
Previous medical visit for pain	no
Test carried out	no
Allergies to antibiotics	penicillin

Scenario 3

The strength of the symptom is represented by the number of + in the brackets.

	Scenario 3
Pain where	throat
Intensity of pain	(+)
Duration	4 days
Other symptoms	fever (39 C°) headache (+) swollen glands (+) stomachache
Contact with ill person	no
Previous medical visit	yes, nurse a week ago
Test carried out	throat culture test no result yet
Allergies to antibiotics	no

Scenario 4

The strength of the symptom is represented by the number of + in the brackets.

	Scenario 4
Pain where	throat
Intensity of pain	(++)
Duration	more than 3 days
Other symptoms	fever (++) swollen glands congestion headache
Contact with ill person	yes, yesterday
Previous medical visit	yes
Test carried out	rapid strep test
Allergies to antibiotics	amoxicillin

Scenario 5

The strength of the symptom is represented by the number of + in the brackets.

	Scenario 5
Pain where	throat
Intensity of pain	++
Duration	3-4 days
Other symptoms	fever (38 C°) headache congestion (++) cough (++)
Contact with ill person	Yes, a month ago
Previous medical visit	no
Test carried out	no
Allergies to antibiotics	no

Scenario 6

The strength of the symptom is represented by the number of + in the brackets.

	Scenario 6
Pain where	throat
Intensity of pain	+
Duration	more than a week
Other symptoms	headache congestion (++) (start) runny nose (now) cough (+)
Contact with ill person	don't remember
Previous medical visit	yes, a nurse
Test carried out	don't know
Allergies to antibiotics	no

Scenario 7

The strength of the symptom is represented by the number of + in the brackets.

	Scenario 7
Pain where	throat
Intensity of pain	+
Duration	10 days
Other symptoms	fever (38,5 C°) headache (+) runny nose (++) cough (+)
Contact with ill person	no
Previous medical visit	yes, a week ago nurse no strep-diagnosed
Test carried out	no
Allergies to antibiotics	yes, amoxicillin

Scenario 8

The strength of the symptom is represented by the number of + in the brackets.

	Scenario 8
Pain where	throat
Intensity of pain	++
Duration	2 days
Other symptoms	fever (low) headache (++) congestion (++) cough (++)
Contact with ill person	don't remember
Previous medical visit	yes
Test carried out	rapid strep test result negative
Allergies to antibiotics	no

Annexe 6 : Corpus récoltés et retranscrits (EXPQ1)

MedSLT V1								
Actor	Source- Transcript	Recognition	Backtranslation	Translation	Accepted	System NB	Session/ DOC	pat
doctor	what brings you into the office today	was pain in any often today			NA_BR	1	1	1
doctor	what brings you to the clinic today	was congestion accompanied today			NA_BR	1	1	1
doctor	what are your symptoms	was a testing done	was the test carried out?	¿le han realizado un examen?	NA_BR	1	1	1
doctor	do you have any pain	are you having pain	do you experience the pain?	¿tiene el dolor?	AC	1	1	1
patient	sí en la garganta	sí la garganta	me duele la garganta	i have a sore throat	AC	1	1	1
doctor	for how long					1	1	1
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	1	1
patient	desde hace dos días	desde hace los días	me duele la garganta desde hace días	i have had a sore throat for days	NA_IE	1	1	1
patient	sí desde hace dos días	sí desde hace dos días	me duele la garganta desde hace dos días	i have had a sore throat for two days	AC	1	1	1
doctor	have you had a fever	have you had a fever	have you had fever?	¿ha tenido fiebre?	AC	1	1	1
patient	sí de treinta y nueve punto cinco grados	sí treinta y ocho grados	he tenido treinta y ocho grados de fiebre	my temperature was one hundred one degrees	NA_BR	1	1	1
patient	sí treinta y nueve grados y medio	sí treinta y nueve horas mes día			NA_BR	1	1	1
patient	más de treinta y nueve grados	más de treinta y nueve grados	he tenido más de treinta y nueve grados de fiebre	my temperature was more than one hundred three degrees	AC_help	1	1	1
doctor	have you had a cough	have you had cough	have you had a cough?	¿ha tenido tos?	AC	1	1	1
patient	no no	no una			NA_BR	1	1	1
patient	no	no	no	no	AC	1	1	1
doctor	headache	headache	have you had the headaches?	¿le ha dolido la cabeza?	NA_BR	1	1	1

doctor	how long have you had a headache	how long have you had a headache	how long have you had the headaches?	¿desde cuándo le duele la cabeza?	AC_help	1	1	1
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC_help	1	1	1
patient	sí mucho	sí mucho	me duele mucho la cabeza	i experience a severe pain in the head	AC	1	1	1
doctor	have you had a runny nose	've you had runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	1	1	1
doctor			have you had a runny nose?	¿ha mucho tenido moco?	REP	1	1	1
patient	no	no	no	no	AC	1	1	1
doctor	has anyone else been sick	is anyone often sick	does anyone often feel nauseous?	¿alguien tiene a menudo náuseas?	NA_BR	1	1	1
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	1	1	1
patient	sí hace días	sí hace días	he estado en contacto con alguien que tiene el dolor de garganta hace días	i was in contact with someone who had a sore throat days ago	AC	1	1	1
doctor	did they need to see a doctor	is pain intense near doctor			NA_BR	1	1	1
doctor	was it strep throat	what is strep throat			NA_BR	1	1	1
doctor	have you been in contact with anyone with a strep throat	have you been in contact with anyone with a strep throat	have you been in contact with anyone with a strep throat	¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	1	1	1
patient	no sé	no sé	no sé	i don't know	AC	1	1	1
doctor	have you had these symptoms before	have you had the intense cough for day	have you had a severe cough for several days?	¿tiene mucho tos desde hace algunos días?	NA_BR	1	1	1
doctor	do you have any allergies to medicines	have any allergies been intense	have allergies been severe?		NA_BR	1	1	1
doctor	are you allergic to antibiotics	are you allergic to antibiotic	are you allergic to antibiotics?	¿los antibióticos le dan alergia?	AC	1	1	1
patient	no	no	no	no	AC	1	1	1

doctor	have you had strep throat before	have you had strep throat for week	have you had strep throat for several weeks?	¿tiene una infección de la garganta por estreptococo desde hace algunas semanas?	NA_BR	1	1	1
doctor	did you see a doctor before	did you see a doctor for it	did you visit the doctor for the pain?	¿ha consultado un médico para el dolor?	NA_BR	1	1	1
doctor	did you see the doctor this week	did you see the doctor this week	did you visit the doctor this week?	¿ha consultado un médico esta semana?	AC_help	1	1	1
patient	no	no	no	no	AC	1	1	1
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	1	1	1
patient						1	1	1
patient	medio fuerte	adiós cuarenta			NA_BR	1	1	1
patient	más o menos fuerte	no muy fuerte	el dolor es muy fuerte	the pain is very severe	NA_BR	1	1	1
patient	me duele	la cabeza			NA_BR	1	1	1
patient	fuerte	fuerte	el dolor es severo	the pain is severe	AC	1	1	1
doctor	where is your pain		where is the pain?	¿dónde le duele?	AC	1	1	2
patient	me duele la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	1	1	2
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	1	2
patient	me duele desde hace cuatro días	no me acuerdo desde hace cuatro días			NA_BR	1	1	2
patient	me duele desde hace cuatro días	no me acuerdo desde hace cuatro días			NA_BR	1	1	2
patient	desde hace cuatro días	desde hace cuatro días	me duele la garganta desde hace cuatro días	i have had a sore throat for four days	AC	1	1	2
doctor	how severe is your pain	how severe is your pain		¿cuál es la intensidad del dolor?	AC	1	1	2
patient	me duele poco	pienso un poco			NA_BR	1	1	2
patient	un poco	un poco			NA_BR	1	1	2
patient	un poco	un poco			NA_BR	1	1	2
patient	sí un poquito	sí un poquito	sí un poquito		AC_help	1	1	2
patient	sí un poquito	sí un poquito			NA_BR	1	1	2
patient	sí un poquito	sí un poquito			NA_BR	1	1	2
patient	sí un poquito	sí un poquito			NA_BR	1	1	2
patient	sí un poco	sí un poco			NA_BR	1	1	2

patient	un poco	un poco			NA_BR	1	1	2
doctor				¿cuál es la intensidad del dolor?	REP	1	1	2
patient	un poco	un poco			NA_BR	1	1	2
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	1	2
patient	sí yo tengo fiebre	sí desde los meses	tengo fiebre desde hace meses	i have had fever for months	NA_BR	1	1	2
patient	treinta y ocho grados de				NA_BR	1	1	2
patient	treinta y ocho grados de fiebre	treinta y ocho grados de fiebre	tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	AC	1	1	2
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	1	2
patient	no	no	no	no	AC	1	1	2
doctor	do you have a runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	1	2
patient	no	no	no	no	AC	1	1	2
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	1	1	2
patient	sí tengo dolor de cabeza	sí el cultivo de cabeza			C	1	1	2
patient	tengo dolor de cabeza	sí cultivo de cabeza			NA_BR	1	1	2
patient	sí	sí	sí	yes	AC	1	1	2
doctor	how severe is your headache	how severe is your headache			NA_NT	1	1	2
doctor	how bad is your headache	how bad is your headache			NA_NT	1	1	2
doctor	has anyone else had a sore throat	had anyone often had a sore throat	had anyone often had a sore throat?		NA_BR	1	1	2
doctor	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who have a sore throat?	¿ha estado recientemente en contacto con alguien que tiene el dolor de garganta?	AC_help	1	1	2
patient	no	no	no	no	AC	1	1	2
doctor	have you seen a doctor recently	have you seen that doctor recently	have you recently visited the doctor?	¿ha consultado recientemente un médico?	AC	1	1	2
patient	sí	sí	sí	yes	NA	1	1	2
patient	sí viernes pasado	sí viernes pasado	he consultado un médico el viernes pasado	i visited the doctor last friday	NA	1	1	2

patient	sí viernes pasado	sí viernes pasado	he consultado un médico el viernes pasado	i visited the doctor last friday	AC	1	1	2
doctor	did they do any test	did you do the test	did you carry out the test?		NA	1	1	2
doctor	did the doctor do any test	did doctor do the test	did the doctor carry out the test?	¿le han realizado un examen?	NA	1	1	2
doctor	did the doctor do a throat culture test	did the doctor do a throat culture test	did the doctor carry out a throat culture test?	¿le han realizado un cultivo de la garganta?	AC_help	1	1	2
patient	no	no	no	no	AC	1	1	2
doctor	are you taking any medicine	are you taking any medicine	do you take medicine?	¿toma medicina?	AC	1	1	2
patient	no	no	no	no	AC	1	1	2
doctor	are you allergic to medecine	are you allergic to medicine	are you allergic to medicine?	¿medicina le da alergia?	NA_BR	1	1	2
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC_help	1	1	2
patient	sí a la penicilina	sí a la penicilina	la penicilina me da alergias	i am allergic to penicillin	AC	1	1	2
doctor	good morning	good morning	good morning?	¿hola?	AC	1	2	2
patient	buenos días	buenos días	hola	good morning	NA	1	2	2
patient	buenos días	buenos días	hola	good morning	NA	1	2	2
patient	dos días				NA	1	2	2
patient	hola	hola	hola	hello	AC	1	2	2
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	2	2
patient	la garganta	la garganta	me duele la garganta	i have a sore throat	AC	1	2	2
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	2	2
patient	más de tres días	más de tres días	me duele la garganta desde hace más de tres días	i have had a sore throat for more than three days	AC	1	2	2
doctor	what is the character of your pain	what is the character of your pain	what is the character of the pain?	¿cómo es el dolor?	AC	1	2	2
patient	fuerte	fuerte	el dolor es severo	the pain is severe	AC	1	2	2
patient	el dolor es severo	el lunes desde hoy			NA_BR	1	2	2
patient	fuerte	fuerte	el dolor es severo	the pain is severe	AC	1	2	2
doctor	have you had a cough	have you had cough	have you had a cough?	¿ha tenido tos?	AC	1	2	2

patient	no	no	no	no	AC	1	2	2
doctor	have you had a fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	1	2	2
patient	no	no	no	no	AC	1	2	2
doctor	have you had a runny nose	have you had a runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	1	2	2
patient	no	no	no	no	AC	1	2	2
doctor	have you had a headache	have you had headache	have you had the headaches?	¿le ha dolido la cabeza?	AC	1	2	2
patient	sí me ha dolido la cabeza	viernes en la cabeza			NA_BR	1	2	2
patient	sí	sí	sí	yes	AC	1	2	2
patient	sí un poco	sí un poco	la cabeza me ha un poco dolido	i experienced a bit of pain in the head	AC	1	2	2
doctor	have you had a stomachache	have you had any today			NA_BR	1	2	2
doctor	any stomachache	any stomachache	have you had a stomachache?	¿el estómago le ha dolido?	AC	1	2	2
patient	sí	sí	sí	yes	AC	1	2	2
doctor	is your pain improving	is your pain intense pain	is the pain severe?	¿el dolor es severo?	AC	1	2	2
doctor	is the pain better	is the pain intense	is the pain severe?	¿el dolor es severo?	AC	1	2	2
doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC_help	1	2	2
patient	no	no	no	no	AC	1	2	2
doctor	have you been around anyone who is sick	have you been around anyone tuesday			NA_BR	1	2	2
doctor	have you been around someone with a sore throat	have you been around someone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	1	2	2
patient	no	no	no	no	AC	1	2	2
doctor	have you had a rash	have you had a rash	have you had a rash?	¿ha tenido una erupción cutánea?	AC	1	2	2
patient	no	no	no	no	AC	1	2	2
doctor	have you had congestion	have you had congestion	have you felt congested?	¿ha tenido la nariz tapada?	AC	1	2	2
patient	no	no	no	no	AC	1	2	2

doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	2	2
patient	sí una enfermera	sí una enfermera	he consultado una enfermera	i visited a nurse	AC	1	2	2
doctor	how long ago	someone now			NA_BR	1	2	2
doctor	when did you see the nurse	when did you see the nurse	when did you visit a nurse?	¿cuándo ha consultado una enfermera?	AC	1	2	2
patient	hace tres días	hace tres días	he consultado una enfermera hace tres días	i visited a nurse three days ago	AC	1	2	2
doctor	are you on any medication	anyone to any medication			NA_BR	1	2	2
doctor	do you take any medication	do you take any medication	do you take medication?	¿toma medicina?	AC	1	2	2
doctor	no	now			NA_BR	1	2	2
doctor	no	now			NA_BR	1	2	2
doctor	good morning	good morning	good morning?	¿hola?	AC	1	2	1
patient	buenos días	los días			NA_BR	1	2	1
patient	buenos días	buenos días	hola	good morning	AC	1	2	1
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	2	1
patient	en la garganta	la garganta	me duele la garganta	i have a sore throat	AC	1	2	1
doctor	how long have you had a sore throat	how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	2	1
patient	no sé	no sé	no sé	i don't know	AC	1	2	1
doctor	is your sore throat getting better	is your sore throat any other			NA_BR	1	2	1
doctor	is your pain better	is your pain another			NA_BR	1	2	1
doctor	are you still having pain	are you having pain	do you experience the pain?	¿tiene el dolor?	NA_BR	1	2	1
doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC_help	1	2	1
patient	no	no	no	no	AC	1	2	1
doctor	when did the pain start	when did the pain start	when did the pain start?	¿cuándo ha empezado el dolor?	AC	1	2	1
patient	no sé	no sé	no sé	i don't know	AC	1	2	1
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	2	1
patient	sí mucho	sí mucho	tengo mucha fiebre	i have a severe fever	AC	1	2	1

doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	1	2	1
patient	no sé	no sé	no sé	i don't know	AC	1	2	1
doctor	what is the type of pain	what is this kind of pain	what is the character of the pain?	¿cómo es el dolor?	AC	1	2	1
patient	fuerte	cuarenta			NA_BR	1	2	1
patient	fuerte	fuerte	el dolor es severo	the pain is severe	AC	1	2	1
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	2	1
patient	no	no	no	no	AC	1	2	1
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	2	1
patient	sí congestion	sí al infección	tengo mucho el dolor de garganta		NA_BR	1	2	1
patient	sí congestion	sí infección	tengo mucho el dolor de garganta		NA_BR	1	2	1
patient	sí tapada	día pasado		i have a runny nose during the last day	NA_BR	1	2	1
patient	sí con la nariz tapada	sí tarde pasada			NA_BR	1	2	1
patient	sí	sí	sí	yes	AC	1	2	1
doctor	do you have a rash	do you have rash	do you have a rash?	¿tiene una erupción cutánea?	AC	1	2	1
patient	no	no	no	no	AC	1	2	1
doctor	do you have a headache	yesterday	do you have a rash since yesterday?	¿tiene una erupción cutánea ayer?	NA_BR	1	2	1
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	1	2	1
patient	sí	sí	sí	yes	AC	1	2	1
doctor	how long did you have a headache	how long have you had headache	how long have you had the headaches?	¿desde cuándo le duele la cabeza?	AC	1	2	1
patient	no sé	no sé	no sé	i don't know	AC	1	2	1
doctor	do you have a stomachache	do you have a stomachache	do you have a stomachache?	¿el estómago le duele?	AC	1	2	1
patient	no	no	no	no	NA_IE	1	2	1
patient	no	no	no	no	AC	1	2	1

doctor	have you been in contact with someone with a sore throat	have you been in contact with someone with a sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	1	2	1
patient	sí hace dos días	sí hace dos días	he estado en contacto con alguien que tiene el dolor de garganta hace dos días	i was in contact with someone two days ago who had a sore throat	AC	1	2	1
doctor	have you seen a doctor	hadn\'t you seen doctor	hadn't you visited the doctor?		NA_BR	1	2	1
doctor	have you been to a doctor	have you been to a doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	2	1
patient	sí	sí	sí	yes	AC	1	2	1
doctor	when did you visit the doctor	when did you visit the doctor	when did you visit the doctor?	¿cuándo ha consultado un médico?	AC	1	2	1
patient	no sé	no sé	no sé	i don't know	AC	1	2	1
doctor	are you taking any medicine	have you taken any medicine	have you taken medicine?	¿ha tomado medicina?	NA_BR	1	2	1
doctor	have you taken any medicine for the sore throat	have you taken any medicine for the sore throat	have you taken medicine for your sore throat?	¿ha tomado medicina para el dolor de garganta?	AC	1	2	1
patient	no sé	no sé	no sé	i don't know	AC	1	2	1
doctor	did the doctor run a strep test	did the doctor often do strep test	did the doctor often carry out a rapid strep test?	¿le han realizado a menudo la prueba rápida por estreptococo?	AC	1	2	1
patient	sí era positiva	sí amoxicilina			NA_BR	1	2	1
doctor	was the test positive	was the test positive	was the test positive?		NA	1	2	1
doctor	were the results of the test positive	were the results of the test positive	were the results of the test positive?	¿el resultado del examen ha sido positivo?	AC_help	1	2	1
patient	sí	sí	sí	yes	AC	1	2	1
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	1	2	1
patient	sí a la moxicil amoxicilina	en amoxicilina mucho día			NA_BR	1	2	1
patient	sí a la amoxicilina	sí a la amoxicilina	la amoxicilina me da alergias	i am allergic to amoxicillin	AC_help	1	2	1
doctor	where is your pain	[rien]			NA_BR	1	1	3

doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC_help	1	1	3
patient	en la garganta	garganta	me duele la garganta	i have a sore throat	AC	1	1	3
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	1	3
patient	desde hace más de una semana	desde hace más de una semana	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	AC	1	1	3
doctor	how severe is the pain	how severe is pain		¿cuál es la intensidad del dolor?	NA_NB	1	1	3
doctor	how bad is the pain	how bad is the pain			NA_BR	1	1	3
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	1	1	3
patient	muy fuerte	muy fuerte	el dolor es muy fuerte	the pain is very severe	AC	1	1	3
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	1	3
patient	sí un poco	sí un poco	tengo un poco de fiebre	i have a bit of fever	AC	1	1	3
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	1	3
patient	sí	sí	sí	yes	AC	1	1	3
doctor	do you have a headache	are you having any			NA_BR	1	1	3
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	1	1	3
patient	sí mucho	sí mucho	me duele mucho la cabeza	i experience a severe pain in the head	AC	1	1	3
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	1	3
patient	sí	sí	sí	yes	AC	1	1	3
doctor	have you been near someone who has had a sore throat	have you been near someone who has that sore throat			NA_BR	1	1	3
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	1	1	3
patient	sí hace un mes	sí hace mes	he estado en contacto con alguien que tiene el dolor de garganta hace meses	i was in contact with someone months ago who had a sore throat	AC_IE	1	1	3

doctor	have you been in contact with someone within the last week	have you been in contact with someone in that the last week			NA	1	1	3
doctor	have you been in contact with anyone in the last week who has a sore throat	have you been in contact with anyone in the last week who has a sore throat	have you been in contact with anyone in the last week who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta en la semana pasada?	AC_help	1	1	3
patient	no	no	no	no	AC	1	1	3
doctor	did you go to the doctor within the last week	did you visit the doctor when that			NA_BR	1	1	3
doctor	when did you visit the doctor	when did you visit the doctor	when did you visit the doctor?	¿cuándo ha consultado un médico?	AC_help	1	1	3
patient	no he ido	no días	consulto un médico desde hace días		NA_BR	1	1	3
patient	no fui	no muy			NA_BR	1	1	3
patient	no un médico	no un médico	no un médico		NA_help	1	1	3
patient	no fui	no sé	no sé	i don't know	NA_BR	1	1	3
doctor			when did you visit the doctor?	¿cuándo ha consultado un médico?	REP	1	1	3
patient	no	no	no	no	AC	1	1	3
doctor	are you taking any medicine	are you taking any medicine	do you take medicine?	¿toma medicina?	NA	1	1	3
doctor	have you taken any medicine for your pain	have you taken any medicine for your	have you taken medicine for the pain?	¿ha tomado medicina para el dolor?	AC_help	1	1	3
patient	no	no	no	no	AC	1	1	3
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	1	1	3
patient	no	no	no	no	AC	1	1	3
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	1	4
patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	1	1	4
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	1	4
patient	algunas semanas	algunas semanas	me duele la garganta desde hace algunas semanas	i have had a sore throat for several weeks	AC	1	1	4

doctor	how bad is your sore throat	how bad is your sore throat			NA_NT	1	1	4
doctor	how bad is your pain	how bad is your pain			NA_NT	1	1	4
doctor	how severe is your pain	how severe is your pain		¿cuál es la intensidad del dolor?	NA	1	1	4
doctor	what is the pain like	what is the pain lately	what is a recent pain?	¿qué es recientemente el dolor?	NA	1	1	4
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	1	1	4
patient	muy fuerte	muy fuerte	el dolor es muy fuerte	the pain is very severe	AC	1	1	4
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	1	4
patient	sí	sí	sí	yes	AC	1	1	4
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	1	1	4
patient	treinta y nueve grados	treinta y nueve grados	tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	AC	1	1	4
doctor	do you have a cough	\ 've you had a cough	have you had a cough?	¿ha tenido tos?	AC	1	1	4
patient	desde hace más de un día	desde hace más de un día			NA_NT	1	1	4
patient	desde hace más de un día	desde hace más de un día			NA_NT	1	1	4
patient	sí más de un día	sí más de un día	tengo tos desde hace más de un día	i have coughed for more than one day	AC	1	1	4
doctor	do you have a headache	\ 've you had headache	have you had the headaches?	¿le ha dolido la cabeza?	AC	1	1	4
patient			me duele la cabeza desde hace más de dos días algunas horas		AC	1	1	4
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	1	4
patient	sí	sí	sí	yes	AC	1	1	4
doctor	have you seen a doctor recently	have you seen the doctor recently	have you recently visited the doctor?	¿ha consultado recientemente un médico?	NA	1	1	4
doctor	have you seen a doctor for your sore throat	have you seen a doctor for your sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC	1	1	4

patient	no	no	no	no	AC	1	1	4
doctor	have you been around someone who has had a sore throat	have you been around someone who had a sore throat	have you been in contact with anyone who had a sore throat?	¿ha estado en contacto con alguien que ha tenido el dolor de garganta?	NA	1	1	4
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	1	1	4
patient	no pienso	no pienso	no pienso	i don't think so	AC	1	1	4
doctor	have you taken any medicine for the pain	have you taken any medicine for the pain	have you taken medicine for the pain?	¿ha tomado medicina para el dolor?	AC	1	1	4
patient	sí	sí	sí	yes	AC	1	1	4
doctor	what medecine did you take	what medicine is the pain			NA_BR	1	1	4
doctor	what medicine	what medicine			NA_BR	1	1	4
doctor	are you allergic to any medecine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	1	1	4
patient	sí a la penicilina	sí a la penicilina	la penicilina me da alergias	i am allergic to penicillin	AC	1	1	4
doctor	good morning	good morning	good morning?	¿hola?	AC	1	2	4
patient	buenos días	buenos días	hola	good morning	AC	1	2	4
doctor	where is your pain	which are pain	what causes the pain?	¿qué le da el dolor?	NA_BR	1	2	4
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC_help	1	2	4
patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	1	2	4
doctor	how long have you had a sore throat	how long did you have a sore throat	how long did you have a sore throat?	¿desde cuándo le ha dolido la garganta?	NA_BR	1	2	4
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	2	4
patient	más de una semana	más de una semana	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	AC	1	2	4
doctor	what is the charachter of the pain	this character of the pain			NA_BR	1	2	4
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	1	2	4
patient	muy fuerte	muy fuerte	el dolor es muy fuerte	the pain is very severe	AC	1	2	4
doctor	is the pain improving	is the pain intense	is the pain severe?	¿el dolor es severo?	NA_BR	1	2	4

doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC_help	1	2	4
patient	no	no	no	no	AC	1	2	4
doctor	do you have a fever	the bad fever	is a severe pain better when you have fever?	¿el dolor está mucho aliviado cuando tiene fiebre?	NA_BR	1	2	4
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	2	4
patient	sí	sí	sí	yes	AC	1	2	4
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	1	2	4
patient	treinta y nueve grados de fiebre	[rien]			NA_BR	1	2	4
patient	treinta y nueve grados de fiebre	treinta y nueve grados de fiebre	tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	AC	1	2	4
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	2	4
patient	sí	sí	sí	yes	AC	1	2	4
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	2	4
patient	sí	sí	sí	yes	AC	1	2	4
doctor	have you had a rash	have you had a rash	have you had a rash?	¿ha tenido una erupción cutánea?	AC	1	2	4
patient	desde hace un mes	desde hace mes			NA_BR	1	2	4
patient	desde hace un mes	desde hace mes			NA_BR	1	2	4
patient	desde hace un mes	desde hace mes			NA_BR	1	2	4
patient	desde hace un mes	desde hace un mes	desde hace un mes		AC_help	1	2	4
patient	desde hace un mes	desde hace mes			NA_BR	1	2	4
patient	sí	sí	sí	yes	AC	1	2	4
doctor	where is the rash	where is a rash	where is a rash?		NA_NT	1	2	4
doctor	have you had a headache	have you had headache	have you had the headaches?	¿le ha dolido la cabeza?	AC	1	2	4
patient	más de una semana	más de una semana	me duele la cabeza desde hace más de una semana	i have experienced the pain in the head for more than one week	AC	1	2	4
doctor	have you taken any medicine	have you taken any medicine	have you taken medicine?	¿ha tomado medicina?	AC	1	2	4
patient	sí	sí	sí	yes	AC	1	2	4

doctor	what medicine have you	was medicine taken	was medicine taken?		NA_BR	1	2	4
doctor	have you had fatigue	have you had the pain	have you experienced the pain?	¿ha tenido el dolor?	NA_BR	1	2	4
doctor	have you had fatigue	have you had it this week	have you experienced the pain this week?	¿le duele desde hace esta semana?	NA_BR	1	2	4
doctor	are you tired	the character			NA_BR	1	2	4
doctor	has anyone else had a sore throat	did anyone have this sore throat	did anyone have a sore throat?	¿alguien ha tenido el dolor de garganta?	AC	1	2	4
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	1	2	4
patient	[rien]	[rien]	he estado en contacto con alguien que tiene el dolor de garganta hace dos días semanas		AC	1	2	4
doctor	have you seen a doctor	have you seen doctor	have you visited the doctor?	¿ha consultado un médico?	NA_IE	1	2	4
doctor	no	now			NA_IE	1	2	4
doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	2	4
patient	no	no	no	no	AC	1	2	4
doctor	are you allergic to any medicines	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	1	2	4
patient	sí a la penicilina	sí a la penicilina	la penicilina me da alergias	i am allergic to penicillin	AC	1	2	4
doctor	have you had a stomachache	have you had stomachache	have you had a stomachache?	¿el estómago le ha dolido?	AC	1	2	4
patient	sí	sí	sí	yes	AC	1	2	4
doctor	have you had any tests run	have you had a testing	have you had the test?	¿le han realizado un examen?	AC	1	2	4
patient	no	no	no	no	AC	1	2	4
doctor	good morning	good morning	good morning?	¿hola?	AC	1	2	3
patient	buenos días	buenos días	hola	good morning	AC	1	2	3
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	2	3
patient	en la garganta	la garganta	me duele la garganta	i have a sore throat	AC	1	2	3

doctor	how long have you had the pain	how long have you had the pain	how long have you experienced the pain?	¿desde cuándo tiene el dolor?	AC	1	2	3
patient	dos días	dos días	me duele desde hace dos días	i have experienced the pain for two days	AC	1	2	3
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC	1	2	3
patient	muy fuerte	muy fuerte	el dolor es muy fuerte	the pain is very severe	AC	1	2	3
doctor	is the pain getting worse	is the pain really	is the pain very worse?		NA_BR	1	2	3
doctor	is the pain getting better	is the pain intense	is the pain severe?	¿el dolor es severo?	NA_BR	1	2	3
doctor	is the pain less severe	is the pain less severe	is the pain better?	¿el dolor disminuye?	AC_help	1	2	3
patient	no	no	no	no	AC	1	2	3
doctor	have you had fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	1	2	3
patient	sí poca	sí horas	he tenido fiebre horas	i had fever hours	NA_BR	1	2	3
patient	sí un poquito	sí un poquito	he tenido un poco de fiebre	i had a bit of fever	AC_help	1	2	3
doctor	what was your temperature	what was your temperature	what was your temperature?	¿cuántos grados ha tenido?	AC	1	2	3
patient	treinta y siete	treinta y siete	he tenido treinta y siete grados de fiebre	my temperature was ninety nine degrees	AC	1	2	3
doctor	have you had a runny nose	have you had a runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	1	2	3
patient	sí mucho	sí mucho	he tenido mucho moco	i had a runny nose	AC	1	2	3
doctor	have you had a cough	have you had a cough	have you had a cough?	¿ha tenido tos?	AC	1	2	3
patient	sí	sí	sí	yes	AC	1	2	3
doctor	have you had a rash	have you had rash	have you had a rash?	¿ha tenido una erupción cutánea?	AC	1	2	3
patient	no	no	no	no	AC	1	2	3
doctor	have you had a headache	have you had headache	have you had the headaches?	¿le ha dolido la cabeza?	AC	1	2	3
patient	sí bastante	sí muy punzante		i experienced a stabbing severe pain in the head	AC_IE	1	2	3
doctor	how long have you had the headache	how long do you have headache	how long do you have the headaches?	¿desde cuándo le duele la cabeza?	AC	1	2	3
patient	dos días	dos días	me duele la cabeza desde hace dos días	i have experienced the pain in the head for two days	AC	1	2	3

doctor	is the headache improving	is the headache intense	are the headaches severe?		NA_BR	1	2	3
doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC_help	1	2	3
patient	un poco	sí un poco	el dolor mejora un poco	a bit of pain is decreasing	AC	1	2	3
doctor	have you had a stomachache	have you had a stomachache	have you had a stomachache?	¿el estómago le ha dolido?	AC	1	2	3
patient	no	no	no	no	AC	1	2	3
doctor	have you been in contact with someone with a sore throat	have you been in contact with someone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	1	2	3
patient	no sé	no sé	no sé	i don't know	AC	1	2	3
doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	2	3
patient	sí	sí	sí	yes	AC	1	2	3
doctor	when did you see the doctor	when did you see the doctor	when did you visit the doctor?	¿cuándo ha consultado un médico?	AC	1	2	3
patient	ayer	ayer	he consultado un médico ayer	i visited the doctor yesterday	AC	1	2	3
doctor	did the doctor run any tests	is the doctor having a test	does the doctor have the test?		NA_BR	1	2	3
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC_help	1	2	3
patient	sí	sí	sí	yes	AC	1	2	3
doctor	do you know the results of the tests	is another results of the test			NA_BR	1	2	3
doctor	what were the results of the test	what were the results of the test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC	1	2	3
patient	fue negativo	muy negativo			NA_BR	1	2	3
patient	negativo	negativo	el resultado del examen ha sido negativo	the results of the test were negative	AC	1	2	3
doctor	are you allergic to any medications	were you allergic to any medication	were you allergic to any medication?	¿ha tenido alergia a alguna medicina?	AC	1	2	3
patient	no	no	no	no	AC	1	2	3
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	3	5

doctor	for how long have you had your sore throat	for how long have you had your test	how long have you had the test?		NA_BR	1	3	5
doctor	how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	3	5
patient	desde cuando tiene desde cuando duele la garganta	desde horas vienes desde hoy pasado en la garganta	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	NA_IE	1	3	5
patient	desde hace más de una semana	desde hace más de una semana	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	AC	1	3	5
doctor	for how long have you had a sore throat	for how long have you had this throat	how long have you had your throat?	¿desde cuándo tiene la garganta?	NA_BR	1	3	5
doctor	for how long have you had a sore throat	for how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	3	5
patient	dos días	dos días	me duele la garganta desde hace dos días	i have had a sore throat for two days	AC	1	3	5
doctor	do you have runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	3	5
patient	no	no	no	no	AC	1	3	5
patient	out of context		tengo mucho moco hace dies semanas horas		NA_IE	1	3	5
patient	out of context		tengo mucho moco desde hace meses	i have had a runny nose for months	NA_IE	1	3	5
doctor	is there anyone sick in your family	is anyone visiting person	does anyone visit the person?	¿esa persona está en contacto con alguien?		1	3	5
doctor	have you been in contact with anyone with a strep throat in the last week	have you been in contact with anyone with a strep throat in the last week	have you been in contact with anyone with a strep throat in the last week	¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo en la semana pasada?	AC_help	1	3	5
patient	sí hace días	sí hace días	he estado en contacto con alguien que tiene una infección de la garganta por estreptococo hace días	i was in contact with someone days ago who had strep throat	AC	1	3	5
doctor	do you have a headache	\`ve you had any			NA_BR	1	3	5

doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC_help	1	3	5
patient	sí	sí	sí	yes	AC	1	3	5
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	1	3	5
doctor	for how long have you had a sore throat	do you have this swollen throat who has a sore throat			NA_BR	1	3	6
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	3	6
patient	cuatro días	cuatro días	me duele la garganta desde hace cuatro días	i have had a sore throat for four days	AC	1	3	6
doctor	how bad is the pain	how bad is this pain			NA_BR	1	3	6
doctor	how bad is the pain	how bad is this pain			NA_BR	1	3	6
doctor	how bad is this pain	how bad is this pain			NA_BR	1	3	6
doctor	how severe is the pain	how severe is the pain		¿cuál es la intensidad del dolor?	AC	1	3	6
patient	mucho	[rien]			NA_BR	1	3	6
doctor				¿cuál es la intensidad del dolor?	REP	1	3	6
patient	mucho	[rien]			NA_BR	1	3	6
patient	sí mucho	sí mucho	sí mucho		NA_NT	1	3	6
patient	mucho	mucho	mucho		AC_help	1	3	6
doctor				¿cuál es la intensidad del dolor?	REP	1	3	6
patient	bastante	punzante	el dolor es punzante	the pain is stabbing	NA_BR	1	3	6
patient	duele mucho	no mucho			NA_BR	1	3	6
patient	muy fuerte	muy fuerte	el dolor es muy fuerte	the pain is very severe	AC_help	1	3	6
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	1	3	6
patient	sí	sí	sí	yes	AC	1	3	6
doctor	do you have a runny nose	\ve you had runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	1	3	6
patient	no entiendo	no entiendo	no entiendo	i don't understand	AC	1	3	6
doctor	have you had a runny nose	have you had runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	1	3	6
patient	no	no	no	no	AC	1	3	6

doctor	have you had a fever	have you had a fever	have you had fever?	¿ha tenido fiebre?	AC	1	3	6
patient	sí	sí	sí	yes	AC	1	3	6
doctor	when have you been in contact with a person who has strep throat	when have you been in contact with a person who has strep throat	when have you been in contact with the person who has strep throat?	¿cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo?	NA_IE	1	3	6
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	1	3	6
patient	no	no	no	no	AC	1	3	6
doctor	have you had any medical tests	have you had difficulty swallowing	have you had difficulty swallowing?	¿ha tenido dificultades para tragar?	NA_BR	1	3	6
doctor	have you had a test for strep throat	have you had the test for several hours	have you had the test for several hours?	¿le realizan un examen desde hace algunas horas?	NA_BR	1	3	6
doctor	have you had any tests	have you had the test	have you had the test?	¿le han realizado un examen?	NA_BR	1	3	6
doctor	have you had	have you had it	have you experienced the pain?	¿ha tenido el dolor?	NA	1	3	6
doctor	are you allergic to antibiotic	are you allergic to medicine	are you allergic to medicine?	¿medicina le da alergia?	NA	1	3	6
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC_help	1	3	6
patient	sí a la penicilina	sí penicilina	la penicilina me da alergias	i am allergic to penicillin	NA	1	3	6
patient	sí a la penicilina	sí a la penicilina	la penicilina me da alergias	i am allergic to penicillin	AC	1	3	6
doctor	have you had a culture done for strep	did you have this who sees the person			NA_BR	1	3	6
doctor	where do you hurt	were you worse	was the pain worse?		NA_BR	1	4	6
doctor	does your throat hurt	did your throat hurt	did you have a sore throat?	¿le ha dolido la garganta?	AC	1	4	6
patient	sí	sí	sí	yes	AC	1	4	6
doctor	for how long	[rien]				1	4	6

doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	4	6
patient	hace cuatro días	hace cuatro días	me duele la garganta hace cuatro días	i have had a sore throat for four days	AC	1	4	6
doctor	how severe is it	how severe is it		¿cuál es la intensidad del dolor?	AC	1	4	6
patient	mucho	mucho			NA_NT	1	4	6
patient	sí mucho	sí mucho	sí mucho		NA_NT	1	4	6
patient	mucho	mucho	mucho		NA_NT	1	4	6
patient	bastante	punzante	el dolor es punzante	the pain is stabbing	NA_BR	1	4	6
patient	severo	severo	el dolor es severo	the pain is severe	AC	1	4	6
doctor	do you have other symptoms	the positive testing			NA_BR	1	4	6
doctor	do you have cough	your cough			NA_BR	1	4	6
doctor	do you have cough	do you have cough	do you have a cough?	¿tiene tos?	AC	1	4	6
patient	no	no	no	no	AC	1	4	6
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	1	4	6
patient	no	no	no	no	AC	1	4	6
doctor	who else is sick	\ 've you felt sick	have you felt nauseous?	¿ha tenido náuseas?	NA	1	4	6
doctor	anyone else with sore throat	\ 's anyone now with sore throat			NA_BR	1	4	6
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	1	4	6
patient	no	una			NA_BR	1	4	6
patient	no	no	no	no	AC	1	4	6
doctor	do you have fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	1	4	6
patient	nada	grados			NA_BR	1	4	6
patient	no	no	no	no	AC	1	4	6
doctor	do you vomit	do you swallow	do you swallow?	¿traga?	NA_BR	1	4	6
doctor	do you vomit	your augmentin			NA_BR	1	4	6
doctor	do you vomit	do you swallow	do you swallow?	¿traga?	NA_BR	1	4	6
doctor	is there vomiting	is it accompanied			NA_BR	1	4	6
doctor	do you have diarrhea	your saturday			NA_BR	1	4	6
doctor	do you have allergies to medicine	have your allergies lasted	have allergies lasted?		NA_BR	1	4	6

doctor	have you taken any medicine for your pain	have you taken any medicine for your	have you taken medicine for the pain?	¿ha tomado medicina para el dolor?	AC_help	1	4	6
patient	no	no	no	no	AC	1	4	6
doctor	do you have headache	your day		¿toma medicina para el dolor desde hace algunos días?	NA_BR	1	4	6
doctor	is there a headache	severe headache			NA_BR	1	4	6
doctor	headache	any	have you taken medicine for the pain?	¿ha tomado medicina para el dolor?	NA_BR	1	4	6
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC_help	1	4	6
patient	sí mucho	sí mucho	me duele mucho la cabeza	i experience a severe pain in the head	AC	1	4	6
doctor	do you have a sore throat	your sore throat			NA_BR	1	4	5
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC_help	1	4	5
patient	sí un poco	sí un poco	me duele un poco la garganta	i have a bit of a sore throat	NA	1	4	5
patient	sí un poco	sí un poco	me duele un poco la garganta	i have a bit of a sore throat	AC	1	4	5
patient			me duele un poco la garganta	i have a bit of a sore throat	yes	1	4	5
doctor	how long	hello	hello?	¿hola?	NA_BR	1	4	5
doctor	how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	4	5
patient	desde hace más de una semana	desde hoy desde hace más de una semana			NA_BR	1	4	5
patient	desde hace más de una semana	desde hace más de una semana	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	AC	1	4	5
doctor	how severe	[rien]				1	4	5
doctor	does it hurt a lot	does it hurt last week	do you experience the pain last week?		NA_BR	1	4	5
doctor	does it hurt in your throat	does it hurt in your throat	do you have a sore throat?	¿le duele la garganta?	AC_help	1	4	5
doctor	are you getting better	are you doing another			NA_BR	1	4	5

doctor	do you have fever	the fever			NA_BR	1	4	5
doctor	a fever	a fever	a fever		NA_help	1	4	5
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC_help	1	4	5
patient	sí un poco	sí un poco	tengo un poco de fiebre	i have a bit of fever	AC	1	4	5
doctor	how high	hello	hello?	¿hola?	NA_BR	1	4	5
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC_help	1	4	5
patient	no sé	no sé	no sé	i don't know	AC	1	4	5
doctor	do you have cough	your cough			NA_BR	1	4	5
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC_help	1	4	5
patient	no	no	no	no	AC	1	4	5
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	1	4	5
patient	sí	sí	sí	yes	AC	1	4	5
doctor	who else is sick	were you often sick	did you often feel nauseous?	¿ha tenido a menudo náuseas?	NA_BR	1	4	5
doctor	who else has a sore throat	was this a sore throat	did you have a sore throat?	¿le ha dolido la garganta?	NA_BR	1	4	5
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	1	4	5
patient	sí hace unos cuantos días	sí hace cuatro días	he estado en contacto con alguien que tiene el dolor de garganta hace cuatro días	i was in contact with someone four days ago who had a sore throat	NA_BR	1	4	5
patient	sí hace unos cuantos días	sí hace cuatro días	he estado en contacto con alguien que tiene el dolor de garganta hace cuatro días	i was in contact with someone four days ago who had a sore throat	AC	1	4	5
doctor	do you have a headache	your medicine			NA_BR	1	4	5
doctor	does your head hurt	the character			NA_BR	1	4	5
doctor	how severe is your pain	how severe is your pain		¿cuál es la intensidad del dolor?	NA	1	4	5
doctor	have you taken any medicine for your pain	have you taken any medicine for your	have you taken medicine for the pain?	¿ha tomado medicina para el dolor?	AC_help	1	4	5
doctor	how severe is the pain	how severe \s the pain		¿cuál es la intensidad del dolor?	AC	1	4	5

patient	tengo fiebre	pienso fiebre			NA_BR	1	4	5
patient	tengo fiebre	al fiebre			NA_BR	1	4	5
patient	tengo mucha fiebre	en mucho viernes			NA_BR	1	4	5
patient	tengo mucha fiebre	en mucho viernes			NA_BR	1	4	5
patient	treinta y nueve grados de fiebre	treinta y nueve grados de fiebre	treinta y nueve grados de fiebre		AC_help	1	4	5
doctor	are you getting better	are you in another			NA_BR	1	4	5
doctor	are you taking medecines	are you taking medicine	do you take medicine?	¿toma medicina?	NA_BR	1	4	5
doctor	have you taken any medicine for the sore throat	have you taken any medicine for the sore throat	have you taken medicine for your sore throat?	¿ha tomado medicina para el dolor de garganta?	AC	1	4	5
patient	no	no	no	no	AC	1	4	5
doctor	have you had any tests for your throat	have you had intense sore throat	have you had a severe sore throat?	¿la garganta le ha mucho dolido?	NA_BR	1	4	5
doctor	have you had any tests	have you had the test	have you had the test?	¿le han realizado un examen?	NA	1	4	5
doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	4	5
patient	sí	sí	sí	yes	AC	1	4	5
doctor	do you have a sore throat	do you have this testing	do you have the test?	¿le realizan un examen?	NA_BR	1	3	7
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC	1	3	7
patient	sí un poco	sí un poco	me duele un poco la garganta	i have a bit of a sore throat	AC	1	3	7
doctor	have you been in contact with anyone with strep	are you in contact with anyone with this			NA_BR	1	3	7
doctor	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	1	3	7

patient	sí hace un mes	sí hace un mes	he estado en contacto con alguien que tiene una infección de la garganta por estreptococo hace un mes	i was in contact with someone one month ago who had strep throat	AC	1	3	7
doctor	do you have a runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	3	7
patient	no	no	no	no	AC	1	3	7
doctor	do you have swollen glands	do you have swollen glands	do you have swollen glands?	¿tiene los ganglios inflamados?	AC	1	3	7
patient	no	no	no	no	AC	1	3	7
doctor	do you have fever	your fever			NA_BR	1	3	7
doctor	do you have fever	\ve you had fever	have you had fever?	¿ha tenido fiebre?	AC	1	3	7
patient	sí de treinta y ocho	sí treinta y ocho	he tenido treinta y ocho grados de fiebre	my temperature was one hundred one degrees	AC	1	3	7
doctor	do you have a rash	do you have a rash	do you have a rash?	¿tiene una erupción cutánea?	AC	1	3	7
patient	no	no	no	no	AC	1	3	7
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	NA_IE	1	3	7
doctor	do you have headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	1	3	7
patient	sí	sí	sí	yes	AC	1	3	7
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC	1	3	8
patient	sí un poco	sí un poco	me duele un poco la garganta	i have a bit of a sore throat	AC	1	3	8
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	3	8
patient	no sé	no sé	no sé	i don't know	AC	1	3	8
doctor	have you been in contact with anyone who has had strep recently	have you been in contact with anyone who has test recently	have you been in contact with anyone recently who have the test?		NA_BR	1	3	8

doctor	have you been in contact with anyone recently who has strep throat	have you been in contact with anyone recently who has strep throat	have you been in contact with anyone recently who have strep throat?	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	1	3	8
patient	no lo recuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	1	3	8
doctor	do you have swollen glands	do you have swollen glands	do you have swollen glands?	¿tiene los ganglios inflamados?	AC	1	3	8
patient	sí los tengo	sí dos grados			NA_BR	1	3	8
patient	sí los tengo	sí dos semanas	tengo los ganglios inflamados dos semanas	i have swollen glands two weeks	NA_BR	1	3	8
patient	sí un poco	sí un poco	tengo un poco los ganglios inflamados	my glands are slightly swollen	AC	1	3	8
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	1	3	8
patient	sí	sí	sí	yes	AC	1	3	8
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	3	8
patient	sí sí tengo	sí muy intenso	tengo mucho moco	i have a severe runny	AC	1	3	8
doctor	do you have a rash	sí muy intenso			NA_BR	1	3	8
doctor	do you have a rash	do you have rash	do you have a rash?	¿tiene una erupción cutánea?	AC	1	3	8
patient	sí tengo una erupción cutánea	cultivo de mucho pecho			NA_BR	1	3	8
patient	sí desde hace tres días	sí desde hace tres días	tengo una erupción cutánea desde hace tres días	i have had a rash for three days	AC	1	3	8
doctor	does your throat hurt	did your throat hurt	did you have a sore throat?	¿le ha dolido la garganta?	AC	1	4	8
patient	sí un poco	sí un poco	la garganta me ha un poco dolido	i had a bit of a sore throat	AC	1	4	8
doctor	how severe	\ 's cough severe	is a cough severe?		NA_BR	1	4	8
doctor	does it hurt a lot	do several last			NA_BR	1	4	8
doctor	do you have fever	the fever			NA_BR	1	4	8
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC_help	1	4	8

patient	sí más de treinta y ocho grados	sí más de treinta y ocho grados	tengo más de treinta y ocho grados de fiebre	my temperature is more than one hundred one degrees	AC	1	4	8
doctor	how many days of sore throat	how high is a sore throat		¿cuál es la intensidad del dolor en la garganta?	AC_IE	1	4	8
patient	me duele poco	médico			NA_BR	1	4	8
patient	un poco	un poco	un poco		NA_help	1	4	8
patient	muy poco	un poco			NA_BR	1	4	8
patient	un poco	un poco			NA_BR	1	4	8
patient	me duele poco	médico			NA_BR	1	4	8
patient	no me duele mucho	no me acuerdo mucho			NA_BR	1	4	8
patient	sí mucho	sí mucho	sí mucho		AC_help	1	4	8
patient	me duele poco	desde horas				1	4	8
patient	sí un poco	sí un poco	sí un poco		AC_help	1	4	8
patient	la garganta me duele poco	en garganta lunes			NA_BR	1	4	8
doctor	do you have cough	the cough	do you have a cough?	¿tiene tos?	NA_IE	1	4	8
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC_help	1	4	8
patient	sí un poco	sí un poco	tengo un poco de tos	i have a bit of a cough	AC	1	4	8
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	1	4	8
patient	no	no	no	no	AC	1	4	8
doctor	how long has your throat hurt	how long has your throat hurt	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	4	8
patient	desde hace tres días	desde hace tres días	me duele la garganta desde hace tres días	i have had a sore throat for three days	AC	1	4	8
doctor	anyone with a sore throat	anyone with sore throat			NA_BR	1	4	8
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	1	4	8
patient	no lo sé	no sé	no sé	i don't know	AC	1	4	8
doctor	do you have a headache	your medication			NA_BR	1	4	8
doctor	does your head hurt	did your nose hurt			NA_BR	1	4	8
doctor	headaches	headache			NA_BR	1	4	8
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	1	4	8

patient	me duele un poco	no creo un poco	no pienso la cabeza me duele un poco		NA_BR	1	4	8
patient	sí un poco	sí un poco	me duele un poco la cabeza	i experience a bit of pain in the head	AC_help	1	4	8
doctor	are you getting better	are you seeing doctor	do you visit the doctor?	¿consulta un médico?	NA_BR	1	4	8
doctor	how severe is your sore throat	how severe is your sore throat		¿cuál es la intensidad del dolor en la garganta?	AC	1	4	8
patient	me duele un poco	hola un poco			NA_BR	1	4	8
patient	un poco	un poco			NA_BR	1	4	8
patient	un poco	un poco			NA_BR	1	4	8
patient	sí un poco	sí un poco	sí un poco		AC_help	1	4	8
patient	sí me duele	sí lunes			NA_BR	1	4	8
doctor	are you taking medecines	are you taking medicine	do you take medicine?	¿toma medicina?	AC	1	4	8
patient	sí desde ayer	sí desde días	tomo medicina desde hace días	i have taken medication for days	NA_BR	1	4	8
patient	sí desde ayer	sí desde ayer	tomo medicina desde ayer	i have taken medication since yesterday	AC_help	1	4	8
doctor	does your throat hurt	does your throat hurt	do you have a sore throat?	¿le duele la garganta?	AC	1	4	7
patient	sí me duele	sí viernes			NA_BR	1	4	7
patient	sí	sí	sí	yes	AC	1	4	7
doctor	for how long	[rien]			NA_BR	1	4	7
doctor	for how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	4	7
patient	desde hace dos días	desde hace dos días	me duele la garganta desde hace dos días	i have had a sore throat for two days	AC	1	4	7
doctor	how severe is your sore throat	how severe \s your sore throat		¿cuál es la intensidad del dolor en la garganta?	NA_IE	1	4	7
doctor	does it hurt a lot	does another last			NA_BR	1	4	7
doctor	have you been in contact with anyone in the last week who has strep throat	have you been in contact with anyone in the last week who has strep throat	have you been in contact with anyone in the last week who have strep throat?	¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo en la semana pasada?	AC_help	1	4	7

patient	no lo sé	no noche	he estado en contacto con alguien que tiene una infección de la garganta por estreptococo la noche	i was in contact with someone the evening who had strep throat	NA_BR	1	4	7
patient	no lo sé	no noche	he estado en contacto con alguien que tiene una infección de la garganta por estreptococo la noche	i was in contact with someone the evening who had strep throat	NA_BR	1	4	7
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	1	4	7
doctor	do you have cough	your cough		¿ha estado en contacto con alguien en tos la semana pasada?	NA_BR	1	4	7
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	4	7
patient	sí	sí	sí	yes	AC	1	4	7
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	1	4	7
patient	sí mucho	sí mucho	tengo la nariz muy tapada	i feel severely congested	AC	1	4	7
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	4	7
patient	sí poca	sí hora	tengo fiebre horas	i have fever hours	NA_BR	1	4	7
patient	sí un poco	sí un poco	tengo un poco de fiebre	i have a bit of fever	AC_help	1	4	7
doctor	does your throat hurt a lot	did your throat hurt	did you have a sore throat?	¿le ha dolido la garganta?	NA_BR	1	4	7
doctor	is your throat getting better	is strep throat in another			NA_BR	1	4	7
doctor	are you getting worse	are you really worse	is the pain very worse?		NA_BR	1	4	7
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC_help	1	4	7
doctor	are you taking medecines	are you taking medicine	do you take medicine?	¿toma medicina?	AC	1	4	7
patient	no ahora no	no grados			NA_BR	1	4	7
patient	no	no	no	no	AC	1	4	7
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	5	9
patient	en la cabeza	en la cabeza	me duele la cabeza	i experience a pain in the head	AC	1	5	9

doctor	is your throat hurting	is your throat hurting	do you have a sore throat?	¿le duele la garganta?	AC	1	5	9
patient	sí	sí	sí	yes	AC	1	5	9
doctor	how long has your throat been hurting you for	how long is your throat hurting	how long do you have a sore throat?	¿desde cuándo le duele la garganta?	AC	1	5	9
patient	hace dos días	hace dos días	me duele la garganta hace dos días	i have had a sore throat for two days	AC	1	5	9
doctor	do you have any fever	did you have any	did you have fever?	¿ha tenido fiebre?	AC	1	5	9
patient	sí de treinta y nueve punto cinco	sí desde hace algunos días			NA_BR	1	5	9
patient	sí desde hace algunos días	sí desde hace algunos días			NA_BR	1	5	9
doctor			did you have fever?	¿ha tenido fiebre?	REP	1	5	9
patient	sí	sí	sí	yes	AC	1	5	9
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	1	5	9
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	1	5	9
patient	treinta y nueve con cinco	el examen positivo			NA_BR	1	5	9
patient	treinta y nueve y cinco	treinta y nueve	tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	NA_BR	1	5	9
patient	treinta y nueve punto cinco	sí examen positivo			NA_BR	1	5	9
patient	sí treinta y nueve	sí treinta y nueve	tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	AC_help	1	5	9
doctor	do you have any runny nose	are you having any now			NA_BR	1	5	9
doctor	do you have rhinorrhea	your medication			NA_BR	1	5	9
doctor	do you have runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	5	9
patient	no	no	no	no	AC	1	5	9
doctor	has anybody been sick at home	has anybody been sick in several			NA_BR	1	5	9

doctor	when have you been in contact with a person who has strep throat	when have you been in contact with a person who has strep throat	when have you been in contact with the person who has strep throat?	¿cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo?	AC_help	1	5	9
patient	hace días	hace días	he estado en contacto con una persona que tiene una infección de la garganta por estreptococo hace días	i was in contact with the person who had strep throat days ago	AC	1	5	9
doctor	how bad is your pain	how bad is your pain			NA_BR	1	5	9
doctor	how bad is your pain	how bad is your pain			NA_BR	1	5	9
doctor	out of ten how much is your pain	how intense was severe pain		¿cuál ha mucho sido la intensidad del dolor?	NA_BR	1	5	9
doctor	is it hurting a lot	is it hurting now			NA_BR	1	5	9
doctor	what kind of pain is it	what kind of pain is it	what is the character of the pain?	¿cómo es el dolor?	AC_help	1	5	9
patient	muy fuerte	muy fuerte	el dolor es muy fuerte	the pain is very severe	AC	1	5	9
doctor	when was your last visit to the doctor	when was congested nurse to the doctor			NA_BR	1	5	9
doctor	when did you visit the doctor	when did you visit the doctor	when did you visit the doctor?	¿cuándo ha consultado un médico?	AC_help	1	5	9
patient	hace mucho	hace mucho mes	he mucho consultado un médico hace meses		NA_BR	1	5	9
patient	hace muchos meses	hace mucho mes	he mucho consultado un médico hace meses		NA_BR	1	5	9
patient	hace mucho tiempo	hace mucho día pasado			NA_BR	1	5	9
patient	hace mucho tiempo	hace mucho día	he mucho consultado un médico hace días		NA_BR	1	5	9
patient	no	no	no	no	NA_BR	1	5	9
patient	hace	hace	he consultado un médico hace meses	i visited the doctor months ago	AC	1	5	9
doctor	how old are you	how long were you			NA_BR	1	5	9
doctor	I think you have strep throat	this intensity of this			NA_BR	1	5	9
doctor		is the pain steady	is the pain steady?	¿el dolor es estable?	NA_BR	1	5	9

doctor		is the pain better	is the pain better?	¿el dolor disminuye?	NA_BR	1	5	9
doctor		what was the result of this	what were the results?	¿cuál ha sido el resultado?	NA_BR	1	5	9
doctor	what is hurting you today	which are your today			NA_BR	1	5	10
doctor	why are you here today	when are you today			NA_BR	1	5	10
doctor	do you have a sore throat today	do you have a sore throat today	do you have a sore throat since today?	¿le duele la garganta hoy?	AC	1	5	10
patient	sí la garganta me duele	en la garganta tarde	me duele la garganta hoy en la noche		NA_BR	1	5	10
patient	sí la garganta	sí la garganta	me duele la garganta hoy	i have a sore throat today	AC	1	5	10
doctor	how long have you had the sore throat for	how long had this throat hurt	how long had you had a sore throat?		NA_BR	1	5	10
doctor	for how long have you had a sore throat	for how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	5	10
patient	cuatro días	cuatro días	me duele la garganta desde hace cuatro días	i have had a sore throat for four days	AC	1	5	10
doctor	do you have any fever with it	had any fever lasted	had fever lasted?		NA_BR	1	5	10
doctor	do you have fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	1	5	10
patient	sí un poco	sí un poco	tengo un poco de fiebre	i have a bit of fever	AC	1	5	10
doctor	do you have any headache	any sunday			NA_BR	1	5	10
doctor	do you have any congestion	is any congestion	do you feel congested?	¿tiene la nariz tapada?	AC	1	5	10
patient	no realmente no	no augmentin			NA_BR	1	5	10
patient	no	no	no	no	AC	1	5	10
doctor	have you been in contact with anybody with strep throat	have you been in contact with anybody with strep throat		¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo?	NA_BR	1	5	10
doctor	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	1	5	10
patient	no	no	no	no	AC	1	5	10
doctor	has anybody been sick around you	has anybody been sick around you			NA_BR	1	5	10

doctor	is anybody around you suffering from a sore throat	is anybody around you suffering from a sore throat	does anyone around you have your sore throat?	¿alguien en su entorno tiene el dolor de garganta?	AC	1	5	10
patient	no sé	no sé	no sé	i don't know	AC	1	5	10
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	1	5	10
patient	sí un poco	sí un poco	me duele un poco la cabeza	i experience a bit of pain in the head	AC	1	5	10
doctor	have you seen a doctor recently	have you seen a doctor recently	have you recently visited the doctor?	¿ha consultado recientemente un médico?	AC	1	5	10
patient	sí hace dos días y me diagnostico con infección viral	sí hace dos días mañana hace dos horas un día	he consultado mañana un médico recientemente hace dos días una hora		NA_BR	1	5	10
patient	desde hace dos días	desde hace dos días	desde hace dos días		AC_help	1	5	10
patient	sí desde hace dos días	sí hace dos días	sí desde hace dos días		NA_BR	1	5	10
patient	sí hace dos días	sí desde hace dos días	he consultado un médico hace dos días	i visited the doctor two days ago	AC_help	1	5	10
doctor	did he do any lab studies on you	does anyone to medicine visit you			NA_BR	1	5	10
doctor	did he do any tests	did you do a test	did you carry out the test?		NA_BR	1	5	10
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC	1	5	10
patient	no	no	no	no	AC	1	5	10
doctor	is the pain improving	is the pain starting	does the pain start?	¿el dolor empieza?	NA_BR	1	5	10
doctor	is the pain less frequent	is the pain less frequent	is the frequency of the pain decreasing?	¿le duele menos a menudo?	AC	1	5	10
patient	sí un poco	sí un poco			NA_BR	1	5	10
patient	sí a los antibióticos	sí a los antibióticos	sí a los antibióticos		NA_help	1	5	10
patient	sí un poco	sí un poco			AC	1	5	10
doctor	how long have you had a sore throat	how long is your sore throat	how long do you have a sore throat?	¿desde cuándo le duele la garganta?	NA_IE	1	6	10
doctor	how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	6	10

patient	desde hace cuatro días	desde hace cuatro días	me duele la garganta desde hace cuatro días	i have had a sore throat for four days	AC	1	6	10
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	6	10
patient	no	no	no	no	AC	1	6	10
doctor			do you have fever?	¿tiene fiebre?	REP	1	6	10
doctor	do you have a cough	your cough	do you have a cough?	¿tiene tos?	NA_BR	1	6	10
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC_help	1	6	10
patient	sí un poco	sí un poco	tengo un poco de tos	i have a bit of a cough	AC	1	6	10
doctor	do you have a runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	1	6	10
patient	no	no	no	no	AC	1	6	10
doctor			do you have a runny nose?	¿tiene mucho moco?	REP	1	6	10
doctor	do you have a rash	your drugs			NA_BR	1	6	10
doctor	do you have a rash	do you have rash	do you have a rash?	¿tiene una erupción cutánea?	AC	1	6	10
patient	no tan poco	no un poco	tengo una erupción cutánea leve	i have a bit of a rash	NA_BR	1	6	10
patient	no	no	no	no	AC	1	6	10
doctor	does anything else hurt	did any feel worse			NA_BR	1	6	10
doctor	do you feel weak	do you feel weak	do you feel weak		NA_help	1	6	10
doctor	has anyone else been sick		does anyone often feel nauseous?	¿alguien tiene a menudo náuseas?	NA_BR	1	6	10
doctor	is anyone around you suffering from a sore throat	is anyone around you suffering from sore throat	does anyone around you have your sore throat?	¿alguien en su entorno tiene el dolor de garganta?	AC	1	6	10
patient	no tan poco	profundo		someone around you has a deep sore throat	NA_BR	1	6	10
patient	en verdad no	grados			NA_BR	1	6	10
patient	no	no	no	no	AC	1	6	10
doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	6	10
patient	una enfermera especialista tres días atras	hace semanas desde hace día pasado			NA_BR	1	6	10

patient	hace tres días con una enfermera	hace tres días una semana	he consultado un médico hace tres días una semana		NA_BR	1	6	10
patient	hace tres días	hace tres días	he consultado un médico hace tres días	i visited the doctor three days ago	AC_help	1	6	10
doctor	did they do a strep test	is a cough your strep throat			NA_BR	1	6	10
doctor	was the diagnosis a strep throat	is a cough your strep throat	was the diagnosis strep throat?	¿le han diagnosticado una infección de la garganta por estreptococo?	AC_help	1	6	10
patient	sí hicieron una	sí grados			NA_BR	1	6	10
patient	hicieron una	desde lunes			NA_BR	1	6	10
patient	sí	sí	sí	yes	AC	1	6	10
doctor	are you allergic to any antibiotics	were you allergic to any antibiotic	were you allergic to any antibiotic?	¿ha tenido alergia a algunos antibióticos?	NA_BR	1	6	10
doctor	are you allergic to any antibiotic	are you allergic to any antibiotic	are you allergic to any antibiotic?	¿tiene alergia a algunos antibióticos?	AC_help	1	6	10
patient	no	no	no	no	AC	1	6	10
doctor	how long have you been suffering from the sore throat	how long have you been with sore throat			NA_NT	1	6	9
doctor	how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	6	9
patient	no sé	no sé	no sé	i don't know	AC	1	6	9
doctor	do you have a fever	your fever			NA_BR	1	6	9
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	6	9
patient	sí	sí	sí	yes	AC	1	6	9
doctor	how long have you had a fever	how long have you had a fever	how long have you had fever?	¿desde cuándo tiene fiebre?	AC	1	6	9
patient	no sé	no sé	no sé	i don't know	AC	1	6	9
doctor	how high was your fever	how high was your fever	what was your temperature?	¿cuántos grados ha tenido?	AC	1	6	9
patient	muy alta	muy rápida			NA_BR	1	6	9
patient	alta	pasada			NA_BR	1	6	9
patient	mucho	mucho			NA_NT	1	6	9
patient	mucho	mucho			NA_NT	1	6	9
patient	sí mucho		sí mucho		AC_help	1	6	9

patient	sí mucho	[rien]			NA_BR	1	6	9
patient	no sé	no sé arto			NA_BR	1	6	9
patient	no sé	no sé	no sé	i don't know	AC_help	1	6	9
doctor	what is your temperature	what is your temperature	what is your temperature?	¿cuántos grados tiene?	NA_help	1	6	9
doctor	do you have a runny nose	did you have runny nose	did you have a runny nose?	¿ha mucho tenido moco?	AC	1	6	9
patient	sí	sí	sí	yes	AC	1	6	9
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	6	9
patient	no	no	no	no	AC	1	6	9
doctor	do you have a sore throat	do you have sore throat	do you have a sore throat?	¿le duele la garganta?	AC	1	6	9
patient	sí	sí	sí	yes	AC	1	6	9
doctor	has anyone around you been sick	has anyone around you been sick			NA_NT	1	6	9
doctor	is anyone around you ill	does anyone around you feel sick	does anyone around you feel nauseous?	¿alguien en su entorno tiene náuseas?	NA_BR	1	6	9
doctor	is anyone in your family sick	is anyone near anyone sick			NA_BR	1	6	9
doctor	have you been to see a doctor	have you visited doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	6	9
patient	sí	sí	sí	yes	AC	1	6	9
doctor	when did you visit the doctor	when do you visit the doctor	when do you visit the doctor?	¿cuándo consulta un médico?	NA_BR	1	6	9
doctor	when did you visit the doctor	when did you visit the doctor	when did you visit the doctor?	¿cuándo ha consultado un médico?	AC_help	1	6	9
patient	no sé	no sé	no sé	i don't know	AC	1	6	9
doctor	did they run a throat test	the rapid strep test			NA_BR	1	6	9
doctor	did the doctor do a throat culture test	did the doctor do a throat culture test	did the doctor carry out a throat culture test?	¿le han realizado un cultivo de la garganta?	AC-help	1	6	9
patient	sí	sí	sí	yes	AC	1	6	9
doctor	what was the result	what was the results	what were the results?	¿cuál ha sido el resultado?	AC	1	6	9
patient	positivo	positivo	el resultado ha sido positivo	the results were positive	AC	1	6	9
doctor	positive of strep throat	positive results			NA_BR	1	6	9

doctor	where the results positive for strep throat	were the results often this worse strep			NA_BR	1	6	9
doctor	was the diagnosis a strep throat	was the diagnosis a strep throat	was the diagnosis strep throat?	¿le han diagnosticado una infección de la garganta por estreptococo?	AC_help	1	6	9
patient	sí	sí	sí	yes	AC	1	6	9
doctor	when have you been in contact with a person who has strep throat	when have you been in contact with a person who has strep throat	when have you been in contact with the person who has strep throat?	¿cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo?	AC_help	1	6	9
patient	hace unos días	hace algunos días	he estado en contacto con una persona que tiene una infección de la garganta por estreptococo hace algunos días	i was in contact with the person who had strep throat several days ago	AC	1	6	9
doctor	are you allergic to antibiotics	are you allergic to antibiotic	are you allergic to antibiotics?	¿los antibióticos le dan alergia?	NA	1	6	9
doctor	are you allergic to any antibiotic	are you allergic to any antibiotic	are you allergic to any antibiotic?	¿tiene alergia a algunos antibióticos?	AC_help	1	6	9
patient	sí	sí	sí	yes	AC	1	6	9
doctor	do you have a sore throat today	do you have a sore throat today	do you have a sore throat since today?	¿le duele la garganta hoy?	AC	1	5	11
patient	sí	sí	sí	yes	AC	1	5	11
doctor	for how many days	for how long is pain	how long do you experience the pain?	¿desde cuándo tiene el dolor?	AC	1	5	11
patient	hace una semana	adiós semanas			NA_BR	1	5	11
patient	más de una semana	más de una semana	me duele desde hace más de una semana	i have experienced the pain for more than one week	AC	1	5	11
doctor	is it better today	is it bad your today	is the pain severe since today?	¿el dolor es hoy severo?	NA_BR	1	5	11
doctor	is the pain worse now	is the pain worse now			NA_NT	1	5	11
doctor	is the pain less severe	is the pain less severe	is the pain better?	¿el dolor disminuye?	AC_help	1	5	11

patient	el dolor disminuye	desde los meses en hoy	el dolor mejora en hoy desde hace meses		NA_BR	1	5	11
patient	el dolor disminuye	pienso los días en lunes	sí pienso el dolor mejora en lunes		NA_BR	1	5	11
patient	el dolor es menos	[rien]			NA_BR	1	5	11
patient	el dolor es menos	[rien]			NA_BR	1	5	11
patient	el dolor es menos	pienso lunes mes			NA_BR	1	5	11
patient	sí	sí	sí	yes	AC	1	5	11
doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor?	¿ha consultado un médico?	NA	1	5	11
doctor	have you seen a doctor for your sore throat	have you seen a doctor for your sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC_help	1	5	11
patient	no	no	no	no	NA	1	5	11
patient	no	no	no	no	AC	1	5	11
doctor	do you have any congestion	are you having congestion	do you feel congested?	¿tiene la nariz tapada?	AC	1	5	11
patient	sí	sí	sí	yes	AC	1	5	11
doctor	do you have any cough	do you have any	do you have a cough?	¿tiene tos?	NA	1	5	11
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	5	11
patient	sí	sí	sí	yes	AC	1	5	11
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	6	12
patient	desde esta mañana	desde la mañana			NA_NT	1	6	12
patient	desde la mañana	desde la mañana			NA_NT	1	6	12
patient	sí desde hoy	sí desde hoy			NA_NT	1	6	12
patient	sí desde ayer	sí desde ayer	sí desde ayer		NA_help	1	6	12
patient	sí desde hoy	sí desde hoy	sí desde hoy		NA_help	1	6	12
patient	sí desde ayer	sí desde ayer	sí desde ayer		NA_help	1	6	12
patient	sí desde hace algunos días	sí desde hace algunos días	me duele la garganta desde hace algunos días	i have had a sore throat for several days	NA_help	1	6	12
patient	sí desde hoy	sí desde hoy	sí desde hoy		NA_help	1	6	12
patient	empezo hoy	infección			NA_NT	1	6	12
patient	sí desde hoy	sí desde hoy			NA_NT	1	6	12
patient	sí desde hace algunos días	sí desde hace algunos días	me duele la garganta desde hace algunos días	i have had a sore throat for several days	AC	1	6	12

doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	6	12
patient	sí	sí	sí	yes	AC	1	6	12
doctor	what is your temperature	what is your temperature	what is your temperature?	¿cuántos grados tiene?	AC	1	6	12
patient	treinta y ocho punto cinco	treinta y ocho meses en pecho			NA_BR	1	6	12
patient	treinta y ocho grados y medio	[rien]			NA_BR	1	6	12
patient	treinta y ocho grados	treinta y ocho grados	tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	AC	1	6	12
doctor	do you have a runny nose	your runny nose			NA_BR	1	6	12
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC_help	1	6	12
patient	sí un poco	sí un poco	tengo un poco de moco	i have a bit of a runny	AC	1	6	12
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	6	12
patient	sí un poco	sí un poco	tengo un poco de tos	i have a bit of a cough	AC	1	6	12
doctor	do you have a headache	yesterday	do you have a cough since yesterday?	¿tiene tos ayer?	AC_IE	1	6	12
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	NA_IE	1	6	12
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC_help	1	6	12
patient	un poco	un poco	me duele un poco la cabeza	i experience a bit of pain in the head	AC	1	6	12
doctor	do you have a rash	your rash	do you have a rash?	¿tiene una erupción cutánea?	AC	1	6	12
patient	no	no	no	no	AC	1	6	12
doctor	have you visited a doctor	have you visited the doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	6	12
patient	sí hace una semana	días una semana	consulto un médico desde hace una semana días		NA_BR	1	6	12
patient	sí hace una semana	sí hace una semana	he consultado un médico hace una semana	i visited the doctor one week ago	AC_help	1	6	12
doctor	did they perform a throat culture test	did person have throat culture test	did the person have a throat culture test?		NA_BR	1	6	12

doctor	was a throat culture test carried out	was a throat culture test carried out	was a throat culture test carried out?	¿le han realizado un cultivo de la garganta?	AC_help	1	6	12
patient	no	no	no	no	AC	1	6	12
doctor	what was their diagnosis	what was the diagnosis	what was the diagnosis?	¿qué le han diagnosticado?	AC	1	6	12
patient	no sé	no sé	no sé	i don't know	AC	1	6	12
doctor	has anyone around you had a sore throat	did anyone around you have a sore	did anyone around you have a sore throat?		NA_BR	1	6	12
doctor	is anybody around you suffering from a sore throat	is anybody around you suffering from a sore throat	does anyone around you have your sore throat?	¿alguien en su entorno tiene el dolor de garganta?	AC_help	1	6	12
patient	no	no	no	no	AC	1	6	12
doctor	are you allergic to anti	were you allergic to any			NA_IE	1	6	12
doctor	are you allergic to any antibiotics	are you allergic to any antibiotics	are you allergic to any antibiotic?	¿tiene alergia a algunos antibióticos?	AC_help	1	6	12
patient	sí a la amoxicilina	sí a la amoxicilina	la amoxicilina me da alergias	i am allergic to amoxicillin	AC	1	6	12
doctor	how long have you had a sore throat	how long \ 've you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	6	11
patient	algunas semanas	[rien]			NA_BR	1	6	11
patient	hace algunas semanas	hace algunas semanas	me duele la garganta hace algunas semanas	i have had a sore throat for several weeks	AC	1	6	11
doctor	have you had a fever	have you had a fever	have you had fever?	¿ha tenido fiebre?	AC	1	6	11
patient	sí	sí	sí	yes	AC	1	6	11
doctor			have you had fever?	¿ha tenido fiebre?	REP	1	6	11
doctor	how high was the fever	how high was this fever	what was your temperature?	¿cuántos grados ha tenido?	AC	1	6	11
patient	treinta y nueve grados	treinta y nueve grados	he tenido treinta y nueve grados de fiebre	my temperature was one hundred three degrees	AC	1	6	11
doctor	have you had a runny nose	have you had runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	1	6	11
patient	sí	sí	sí	yes	AC	1	6	11
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	1	6	11

patient	sí mucho	sí mucho	me duele mucho la cabeza	i experience a severe pain in the head	AC	1	6	11
doctor			do you have the headaches?	¿le duele la cabeza?	REP	1	6	11
doctor	do you have a rash	do you have rash	do you have a rash?	¿tiene una erupción cutánea?	AC	1	6	11
patient	sí un poco	sí un poco	tengo una erupción cutánea leve	i have a bit of a rash	AC	1	6	11
doctor	where is the rash	where is a rash	where is a rash?		NA_NT	1	6	11
doctor	what does the rash look like	what was a rash lately	what was a recent rash?	¿qué ha sido recientemente una erupción cutánea?	NA_BR	1	6	11
doctor	has anyone around you had a sore throat	\s anyone around you had a sore throat	has anyone around you had a sore throat?		NA_BR	1	6	11
doctor	is anybody around you suffering from a sore throat	is anybody around you suffering from a sore throat	does anyone around you have your sore throat?	¿alguien en su entorno tiene el dolor de garganta?	AC_help	1	6	11
patient	sí	sí	sí	yes	AC	1	6	11
doctor	have you been to visit the doctor	have you visited the doctor	have you visited the doctor?	¿ha consultado un médico?	AC	1	6	11
patient	sí hace una semana	sí hace una semana	he consultado un médico hace una semana	i visited the doctor one week ago	AC	1	6	11
doctor	did they do a throat culture test	did you do that throat culture test	did you carry out a throat culture test?		NA_BR	1	6	11
doctor	did the doctor do a throat culture test	did the doctor do a throat culture test	did the doctor carry out a throat culture test?	¿le han realizado un cultivo de la garganta?	AC_help	1	6	11
patient	sí	[rien]			NA_BR	1	6	11
patient	sí	sí	sí	yes	AC	1	6	11
doctor	what were the results of the throat culture test	what were the results of the throat culture test	what were the results of a throat culture test?	¿cuál ha sido el resultado del cultivo de la garganta?	AC_help	1	6	11
patient	infección por estreptococo	infección por estreptococo	el resultado de un cultivo de la garganta ha sido una infección de la garganta por estreptococo	the results of a throat culture test were strep throat	AC	1	6	11

doctor	are you allergic to antibiotics	were you allergic to antibiotics	were you allergic to antibiotics?	¿los antibióticos le han dado alergia?	NA_BR	1	6	11
doctor	are you allergic to any antibiotic	are you allergic to any antibiotic	are you allergic to any antibiotic?	¿tiene alergia a algunos antibióticos?	AC_help	1	6	11
patient	sí a la penicilina	sí a la penicilina	la penicilina me da alergias	i am allergic to penicillin	AC	1	6	11
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	7	13
patient	un poco aqui en este lado	pienso cultivo de estómago			NA_BR	1	7	13
patient	sí más de treinta y nueve grados	sí más de treinta y nueve grados	sí más de treinta y nueve grados		NA_help	1	7	13
patient	sí cuarenta grados de fiebre	sí cuarenta grados de fiebre	sí cuarenta grados de fiebre		NA_help	1	7	13
doctor			where is the pain?	¿dónde le duele?	REP	1	7	13
patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC_help	1	7	13
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	7	13
patient	como desde hace tres días	hoy desde hace tres días			NA_BR	1	7	13
patient	desde hace tres días	desde hace tres días	me duele la garganta desde hace tres días	i have had a sore throat for three days	AC_help	1	7	13
doctor	why have you come here today	was headache severe today	were the headaches severe since today?		NA_BR	1	7	13
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	7	13
patient	aqui en este lado de la garganta	ayer pasado en la garganta			NA_BR	1	7	13
patient	aqui en este lado de la garganta	ayer pasado en la garganta			NA_BR	1	7	13
patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC_help	1	7	13
doctor	for how long have you had this pain	for how long have you had this pain	how long have you experienced the pain?	¿desde cuándo tiene el dolor?	AC	1	7	13
patient	hace dos días	hace dos días	tengo el dolor hace dos días	i have experienced the pain for two days	AC	1	7	13
doctor	do you have any other symtoms	do you have any other intense cough			NA_BR	1	7	13

doctor	do you have a runny nose	did you have a runny nose	did you have a runny nose?	¿ha mucho tenido moco?	AC	1	7	13
patient	no	no	no	no	AC	1	7	13
doctor	do you have a headache	are you having headache	do you have the headaches?	¿le duele la cabeza?	AC	1	7	13
patient	no	no	no	no	AC	1	7	13
doctor	do you have pain anywhere else	do you have pain in several			NA_BR	1	7	13
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	1	7	13
patient	no	no	no	no	AC	1	7	13
doctor	is anyone	anyone			NA_BR	1	7	13
doctor	has anyone around you been sick	has anyone around you been sick			NA_NT	1	7	13
doctor	have you been around anyone with sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	1	7	13
patient	sí creo que hace algunos días alguien en el trabajo le tenía dolor de garganta	si pienso hace los días mañana pasada hoy desde horas en garganta			NA_BR	1	7	13
patient	sí lunes pasado	sí lunes pasado	he estado en contacto con alguien el lunes pasado que tiene el dolor de garganta	i was in contact with someone last monday who had a sore throat	AC_help	1	7	13
doctor	have you had fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	1	7	13
patient	creo que sí	creo que sí	sí pienso	i think so	AC	1	7	13
doctor	what was your temperature	what was your temperature	what was your temperature?	¿cuántos grados ha tenido?	AC	1	7	13
patient	me llegó a treinta y nueve punto cinco	pienso treinta y nueve grados			NA_BR	1	7	13
patient	más de treinta y nueve grados	más de treinta y nueve grados	he tenido más de treinta y nueve grados de fiebre	my temperature was more than one hundred three degrees	AC_help	1	7	13
doctor	has anything helped the pain	has any been with the pain			NA_BR	1	7	13

doctor	how bad is the pain	how bad is the pain			NA_BR	1	7	13
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	1	7	13
patient	me duele mucho y me arde por dentro	desde mucho mes tarde tarde			NA_BR	1	7	13
doctor			what is the character of the pain?	¿cómo es el dolor?	REP	1	7	13
patient	quemante	quemante	el dolor es quemante	the pain is burning	AC_help	1	7	13
doctor	do you have any drainage in your throat	do you have any intense strep throat		¿tiene mucho una infección de la garganta por estreptococo?	AC_IE	1	7	13
patient	no no entendí qué me dijo	no noche dies meses			NA_BR	1	7	13
patient	puede repetir	fuerte	tengo mucho una infección de la garganta por estreptococo		NA_BR	1	7	13
doctor	have you seen a doctor for this problem	have you seen a doctor for this strong cough	have you visited the doctor for a severe cough?	¿ha mucho consultado un médico para tos?	NA_BR	1	7	13
doctor	have you seen a doctor for your sore throat	have you seen a doctor for your sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC_help	1	7	13
patient	no todavía no	no entiendo	no entiendo	i don't understand	AC_IE	1	7	13
doctor			have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	REP	1	7	13
patient	no	no	no	no	AC	1	7	13
doctor	have you taking any medicines	have you taken any medicine	have you taken medicine?	¿ha tomado medicina?	AC	1	7	13
patient	sí me tome unas pastillas para el dolor de garganta	sí mes semanas hace día pasado en la garganta			NA_BR	1	7	13
doctor			have you taken medicine?	¿ha tomado medicina?	REP	1	7	13
patient	un poco	un poco	he un poco tomado medicina	i took a bit of medication	AC	1	7	13
doctor	what kind of medicine	what kind of medicine			NA_NT	1	7	13

doctor	did you take an antibiotic	did you take any antibiotic	did you take antibiotics?	¿ha tomado antibióticos?	AC	1	7	13
patient	tomé un poco de amoxicilina	no creo de			NA_BR	1	7	13
patient	amoxicilina	amoxicilina	he tomado la amoxicilina	i took amoxicillin	AC_help	1	7	13
doctor	when did you take the antibiotics	when did you take any antibiotic	when did you take antibiotics?	¿cuándo ha tomado antibióticos?	AC	1	7	13
patient	desde hace una semana	desde hace una semana			NA_NT	1	7	13
patient	desde hace una semana	desde hace una semana			NA_NT	1	7	13
doctor	[rien]	when is the pain	when do you experience the pain?	¿cuándo tiene el dolor?	AC	1	7	13
patient	[rien]	la semana pasada	me duele la semana pasada	i experience a pain during the last week	AC	1	7	13
doctor	[rien]	is the pain less frequent	is the frequency of the pain decreasing?	¿le duele menos a menudo?	AC	1	7	13
patient	[rien]	disminuye	este síntoma disminuye	the symptoms are decreasing	NA	1	7	13
doctor	where is the pain	[rien]			NA_BR	1	7	14
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC_help	1	7	14
patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	1	7	14
patient			me duele la garganta	i have a sore throat	REP	1	7	14
doctor	for how long	[rien]			NA_BR	1	7	14
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	1	7	14
patient	desde hace menos de	desde hace mes	me duele la garganta desde hace meses	i have had a sore throat for months	AC_IE	1	7	14
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	1	7	14
patient	sí	sí	sí	yes	AC	1	7	14
doctor	do you have a stomachache	do you have a stomachache	do you have a stomachache?	¿el estómago le duele?	AC	1	7	14
patient	un poco	un poco	el estómago me duele un poco	i experience a bit of pain in the stomach	AC	1	7	14
doctor	have you been vomiting	how feverish is augmentin			NA_BR	1	7	14

doctor	do you have nausea	do you have your	do you have fever?	¿tiene fiebre?	NA_BR	1	7	14
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	1	7	14
patient	sí un poco	sí un poco	tengo la nariz un poco tapada	i feel a bit congested	AC	1	7	14
doctor	do you have cough	do you have cough	do you have a cough?	¿tiene tos?	AC	1	7	14
patient	mucho	mucho	tengo mucha tos	i have a severe cough	AC	1	7	14
doctor	have you been around anyone that has been sick	have you been around anyone who is sick	have you been in contact with anyone who feel nauseous?		NA_NT	1	7	14
doctor	have you been around anyone with sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA_IE	1	7	14
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	1	7	14
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	1	7	14
doctor	have you seen a doctor for this	have you seen a doctor for this			NA_NT	1	7	14
doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor?	¿ha consultado un médico?	AC_help	1	7	14
patient	sí	sí	sí	yes	NA_IE	1	7	14
patient	sí	sí	sí	yes	AC	1	7	14
doctor	did you have any tests done	did you have the testing	did you have the test?	¿le han realizado un examen?	AC	1	7	14
patient	sí	sí	sí	yes	AC	1	7	14
doctor	what test	what rash			NA_BR	1	7	14
doctor	what were the results of this test	what were the results of this test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC_help	1	7	14
patient	positivo	positivo	el resultado del examen ha sido positivo	the results of the test were positive	AC	1	7	14
doctor	for what	what			NA_BR	1	7	14
doctor	what test did you have	what rash is your cough			NA_BR	1	7	14
doctor	have you taken any medicine	have you taken any medicine	have you taken medicine?	¿ha tomado medicina?	AC	1	7	14
patient	sí	sí	sí	yes	AC	1	7	14

doctor	what kind of medicine	what kind of medicine			NA_NT	1	7	14
doctor	did you take antibiotics	did you take	did you take antibiotics?	¿ha tomado antibióticos?	AC	1	7	14
patient	sí	sí	sí	yes	AC	1	7	14
doctor	are you allergic to anything	are you allergic to			NA_BR	1	7	14
doctor	are you allergic to any medicines	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	1	7	14
patient	no	no	no	no	AC	1	7	14
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	1	8	14
patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	1	8	14
doctor	how long have you had your sore throat	how long have you had this sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	8	14
patient	más de tres días	más de tres días	me duele la garganta desde hace más de tres días	i have had a sore throat for more than three days	AC	1	8	14
doctor	have you had any congestion	have you had intense congestion	have you felt severely congested?		NA_BR	1	8	14
doctor	how long have you had congestion	how long have you had congestion	how long have you had congestion	¿desde cuándo tiene la nariz tapada?	AC_help	1	8	14
patient	más de un día	más de un día	tengo la nariz tapada desde hace más de un día	i have felt congested for more than one day	AC	1	8	14
doctor	have you had fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	1	8	14
patient	no	no	no	no	AC	1	8	14
doctor	have you had a rash	have you had rash	have you had a rash?	¿ha tenido una erupción cutánea?	AC	1	8	14
patient	no	no	no	no	AC	1	8	14
doctor	have you had any tests done	have you had a testing	have you had the test?	¿le han realizado un examen?	AC	1	8	14
patient	sí	sí	sí	yes	AC	1	8	14
doctor	what test have you had done	when has anyone done this			NA_BR	1	8	14
doctor	what were the results of the test	what were the results of the test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC	1	8	14
patient	no sé no sé	no sé noche			NA_BR	1	8	14
patient	no sé	no sé	no sé	i don't know	AC	1	8	14

doctor	has anyone else been sick at home	is anyone often sick in several			NA_BR	1	8	14
doctor	have you recently been in contact with anyone with a strep throat	have you recently been in contact with anyone with a strep throat	have you recently been in contact with anyone with a strep throat	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	1	8	14
patient	no	no	no	no	AC	1	8	14
doctor	are you allergic to any medications	are you allergic to any medication	are you allergic to any medication?	¿tiene alergia a alguna medicina?	AC	1	8	14
patient	no	no	no	no	AC	1	8	14
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	1	8	14
patient	sí	sí	sí	yes	AC	1	8	14
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	1	8	14
patient	sí	sí	sí	yes	AC	1	8	14
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	1	8	13
patient	en la garganta	la garganta	me duele la garganta	i have a sore throat	AC	1	8	13
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	1	8	13
patient	sí un poco	sí un poco	tengo la nariz un poco tapada	i feel a bit congested	AC	1	8	13
doctor	do you have fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	1	8	13
patient	sí un poco creo	sí un poco severo			NA_BR	1	8	13
patient	un poco	un poco	tengo un poco de fiebre	i have a bit of fever	AC_help	1	8	13
doctor	how high did the fever go	how high is your throat	how severe is your throat?		NA_BR	1	8	13
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	1	8	13
patient	no me tomé la temperatura	no me acuerdo una			NA_BR	1	8	13
doctor			what is your temperature?	¿cuántos grados tiene?	REP	1	8	13
patient	no sé	no sé	no sé	i don't know	AC	1	8	13
doctor	do you have a rash	do you have rash	do you have a rash?	¿tiene una erupción cutánea?	AC	1	8	13

patient	no	no	no	no	AC	1	8	13
doctor		has anyone now been			NA	1	8	13
doctor	have you recently been in contact with anyone with a strep throat	have you recently been in contact with anyone with a strep throat	have you recently been in contact with anyone with a strep throat	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	1	8	13
patient	hace unos días	algunos días	he estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo algunos días	i was in contact with someone several days recently who had strep throat	AC	1	8	13
doctor	have you seen another doctor	have you seen another doctor	have you visited another doctor?	¿ha consultado un otro médico?	AC	1	8	13
patient	sí	sí	sí	yes	AC	1	8	13
doctor	where any tests done	were any that lately			NA_BR	1	8	13
doctor	was a strep test done	was a strep test done	was a rapid strep test carried out?	¿le han realizado la prueba rápida por estreptococo?	AC_help	1	8	13
patient	sí me hicieron una prueba rápida	sí pienso la prueba rápida	sí pienso me han realizado una prueba rápida por estreptococo		NA_BR	1	8	13
patient	sí me hicieron una prueba rápida	sí pienso una prueba rápida	sí pienso me han realizado una prueba rápida por estreptococo		NA_BR	1	8	13
patient	sí el día lunes	sí el día lunes	me han realizado una prueba rápida por estreptococo lunes		NA_help	1	8	13
patient	me hicieron una prueba rápida	sí pienso una prueba rápida	sí pienso me han realizado una prueba rápida por estreptococo		NA_BR	1	8	13
patient	sí una prueba rápida	sí una prueba rápida	me han realizado una prueba rápida por estreptococo	a rapid strep test was carried out	AC	1	8	13

doctor	what were the results	what were the results	what were the results?	¿cuál ha sido el resultado?	AC	1	8	13
patient	fue positivo	positivo	el resultado ha sido positivo	the results were positive	AC	1	8	13
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	1	7	15
patient	en la garganta	adiós la garganta			NA_BR	1	7	15
patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC_help	1	7	15
doctor	for how long have you had a sore throat	for how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	1	7	15
patient	más de un mes	más de un mes	me duele la garganta desde hace más de un mes	i have had a sore throat for more than one month	AC	1	7	15
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	1	7	15
patient	sí	sí	sí	yes	AC	1	7	15
doctor	do you have a stomachache	yesterday	do you have the headaches since yesterday?	¿le duele la cabeza ayer?	NA_BR	1	7	15
doctor	do you have a stomachache	do you have a stomachache	do you have a stomachache?	¿el estómago le duele?	AC_help	1	7	15
patient	no	no	no	no	AC	1	7	15
doctor	do you have swollen glands	do you have swollen glands	do you have swollen glands?	¿tiene los ganglios inflamados?	AC	1	7	15
patient	sí	sí	sí	yes	NA_IE	1	7	15
patient	sí	sí	sí	yes	AC	1	7	15
doctor	have you been around anyone with sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA_NT	1	7	15
doctor	have you	any	do you have swollen glands?	¿tiene los ganglios inflamados?	NA_BR	1	7	15
doctor	have you been around anyone with sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA_NT	1	7	15
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	1	7	15

patient	sí	sí	sí	yes	AC	1	7	15
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	1	7	15
patient	no	no	no	no	AC	1	7	15
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	1	7	15
patient	sí	sí	sí	yes	AC	1	7	15
doctor	have you seen a doctor for this problem	have you seen a doctor for this cough	have you visited the doctor for a cough?	¿ha consultado un médico para tos?	NA_BR	1	7	15
doctor	have you seen a doctor for your sore throat	have you seen a doctor for your sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC_help	1	7	15
patient	sí	sí	sí	yes	AC	1	7	15
doctor	did the doctor perform any tests	did the doctor have stomachache	did the doctor have a stomachache?		NA_BR	1	7	15
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	NA_IE	1	7	15
doctor	did the doctor perform any tests	is the doctor seen for any rash	is the doctor visited for a rash?		NA_BR	1	7	15
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC_help	1	7	15
patient	sí	sí	sí	yes	NA_IE	1	7	15
patient	sí	sí	sí	yes	AC	1	7	15
doctor	what was the result	what was the results	what were the results?	¿cuál ha sido el resultado?	AC	1	7	15
patient	positivo	positivo	el resultado ha sido positivo	the results were positive	AC	1	7	15
doctor	are you allergic to anything	are you allergic to medicine	are you allergic to medicine?	¿medicina le da alergia?	AC	1	7	15
patient	sí	sí	sí	yes	AC	1	7	15
doctor	what are you allergic to	what are you allergic to	what are you allergic to?	¿qué le da alergia?	AC	1	7	15
patient	a la penicilina	a la penicilina			NA_BR	1	7	15
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	1	7	16
patient	me duele la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	1	7	16

doctor	how long have you had the pain	how long have you had the pain	how long have you experienced the pain?	¿desde cuándo tiene el dolor?	AC	1	7	16
patient	desde hace más de una semana	sí desde hace más de una semana	me duele desde hace más de una semana	i have experienced the pain for more than one week	AC	1	7	16
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	1	7	16
patient	sí	sí	sí	yes	AC	1	7	16
doctor	do you have a stomachache	do you have stomachache	do you have a stomachache?	¿el estómago le duele?	AC	1	7	16
patient	no	no	no	no	AC	1	7	16
doctor	do you have a cough	are you having cough	do you have a cough?	¿tiene tos?	AC	1	7	16
patient	sí un poco	sí un poco	tengo un poco de tos	i have a bit of a cough	AC	1	7	16
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	1	7	16
patient	sí	sí	sí	yes	AC	1	7	16
doctor	are your glands swollen	are your glands swollen	do you have swollen glands?	¿tiene los ganglios inflamados?	AC	1	7	16
patient	no	no	no	no	AC	1	7	16
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	1	7	16
patient	no	no	no	no	AC	1	7	16
doctor	have you been around anyone with sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA_NT	1	7	16
doctor	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who have a sore throat?	¿ha estado recientemente en contacto con alguien que tiene el dolor de garganta?	NA_help	1	7	16
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	1	7	16
patient	no	no	no	no	AC	1	7	16
doctor	have you seen a doctor for your sore throat	have you seen a doctor for your sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC	1	7	16

patient	consulté a una enfermera	buenos días una enfermera			NA_BR	1	7	16
patient	consulté a una enfermera	desde lunes desde día			NA_BR	1	7	16
patient	una enfermera	una enfermera	he consultado una enfermera para el dolor de garganta	i visited a nurse for my sore throat	AC_help	1	7	16
doctor	did she test	is the test done	is the test carried out?	¿le realizan un examen?	NA_BR	1	7	16
doctor	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC	1	7	16
patient	no sé	no sé	no sé	i don't know	AC	1	7	16
doctor	are you allergic to any medecines	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	1	7	16
patient	muy alérgico a las penicilinas	hola médico			NA_BR	1	7	16
patient	la penicilina	la penicilina	la penicilina me da alergias	i am allergic to penicillin	AC_help	1	7	16
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	1	8	16
patient	me duele la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	1	8	16
doctor	how long have you had the pain	how long have you had the pain	how long have you experienced the pain?	¿desde cuándo tiene el dolor?	AC	1	8	16
patient	desde hace más de una semana	desde hace más de una semana	me duele desde hace más de una semana	i have experienced the pain for more than one week	AC	1	8	16
doctor	have you had congestion	have you had congestion	have you felt congested?	¿ha tenido la nariz tapada?	AC	1	8	16
patient	no	no	no	no	AC	1	8	16
doctor	have you had fever	have you had fever	have you had fever?	¿ha tenido fiebre?	NA_IE	1	8	16
doctor	sí treinta y ocho con cinco	the thursday			NA_IE	1	8	16
doctor	sí treinta y ocho con cinco	any medicine			NA_IE	1	8	16
doctor	sí	any	have you felt congested?	¿ha tenido la nariz tapada?	AC_IE	1	8	16
patient	no	no	no	no	AC	1	8	16
doctor	have you been to another physician	have you been to another feverish person			NA_BR	1	8	16

doctor	have you been to a doctor	have you been to a doctor	have you visited the doctor?	¿ha consultado un médico?	AC_help	1	8	16
patient	una enfermera hace una semana	una enfermera	he consultado una enfermera	i visited a nurse	AC	1	8	16
doctor	did you have any test	did you have the test	did you have the test?	¿le han realizado un examen?	AC	1	8	16
patient	no	sí dos			NA_BR	1	8	16
patient	no	no	no	no	AC	1	8	16
doctor	have you been	have you been in it			NA_BR	1	8	16
doctor	have you been in contact with anyone with a strep throat	have you been in contact with anyone with a strep throat	have you been in contact with anyone with a strep throat	¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	1	8	16
doctor	are you allergic to any medications	were you allergic to any medication	were you allergic to any medication?	¿ha tenido alergia a alguna medicina?	AC	1	8	16
patient	sí a la amoxicilina	sí la amoxicilina	la amoxicilina me ha dado alergias	i was allergic to amoxicillin	AC	1	8	16
doctor	how severe is the pain	how severe is the pain		¿cuál es la intensidad del dolor?	AC	1	8	16
doctor	me duele un poco	when do you swallow	when do you swallow?	¿cuándo traga?	NA_IE	1	8	16
doctor	me duele un poco	when do you swallow	when do you swallow?	¿cuándo traga?	NA_IE	1	8	16
doctor	sí un poco	the people			NA_IE	1	8	16
doctor	sí un poco	the people			NA_IE	1	8	16
doctor	when have you been in contact with a person who has strep throat	when have you been in contact with a person who has strep throat	when have you been in contact with the person who has strep throat?	¿cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo?	AC_help	1	8	16
patient	hace dos días	hace dos días	he estado en contacto con una persona que tiene una infección de la garganta por estreptococo hace dos días	i was in contact with the person who had strep throat two days ago	AC	1	8	16
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	NA_NT	1	8	15
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	1	8	15

patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	1	8	15
doctor	how long have you had a sore throat	how long did you have a sore throat	how long did you have a sore throat?	¿desde cuándo le ha dolido la garganta?	AC	1	8	15
patient	dos días	dos días			NA_BR	1	8	15
patient	dos días	dos días			NA_BR	1	8	15
patient	sí desde algunos días	sí desde algunos días	sí desde algunos días		NA_BR	1	8	15
doctor	do you have any fever	are you having fever	do you have fever?	¿tiene fiebre?	AC	1	8	15
patient	no	grados			NA_BR	1	8	15
patient	sí	sí	sí	yes	AC	1	8	15
doctor	do yo have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	1	8	15
patient	out of context	días	tengo la nariz tapada días	i feel congested days	AC	1	8	15
doctor	have you been in contact with an ill person	have you been in contact with your person	have you been in contact with the person?	¿ha estado en contacto con esa persona?	AC	1	8	15
patient	no	no	no	no	AC	1	8	15
doctor	have you been to another doctor	have you been to another doctor	have you visited another doctor?	¿ha consultado un otro médico?	AC	1	8	15
patient	sí la semana pasada	sí la semana pasada		i visited another doctor during the last week	AC	1	8	15
doctor	did you have any tests	did you have the test	did you have the test?	¿le han realizado un examen?	AC	1	8	15
patient	sí	sí	sí	yes	AC	1	8	15
doctor	what were the results	what were the results	what were the results?	¿cuál ha sido el resultado?	AC	1	8	15
patient	negativo	negativo	el resultado ha sido negativo	the results were negative	AC	1	8	15
doctor	are you allergic to any antibiotics	were you allergic to any antibiotics	were you allergic to any antibiotic?	¿ha tenido alergia a algunos antibióticos?	AC	1	8	15
patient	no	no	no	no	AC	1	8	15

MedSLT V2								
Actor	Source- Transcript	Recognition	Backtranslation	Translation	Accepted	System NB	Session/ DOC	pat
doctor	good morning my name is doctor calandro	this doctor early in now			NA_BR	2	2	2
doctor	good morning	good morning	good morning?	¿hola?	AC	2	2	2
patient	buenos días	buenos días	hola	good morning	AC	2	2	2
doctor	how are you today	how are you today			NA_NT	2	2	2
doctor	how are you feeling	how are any			NA_NT	2	2	2
doctor	how do you feel	amoxicillin			NA_BR	2	2	2
doctor	what is bothering you	[rien]			NA_BR	2	2	2
doctor	how are you	how are you			NA_BR	2	2	2
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	2	2	2
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	2	2
doctor	how long have you had a sore throat	how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	NA_IE	2	2	2
doctor	how long have you had a sore throat	how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	2	2
patient	me duele desde hace dos días	me duele desde hace dos días	me duele desde hace dos días	i have experienced the pain for two days	AC	2	2	2
doctor	how bad is the pain	how bad is the pain			NA_NT	2	2	2
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	2	2
patient	fuerte	fuerte	el dolor es severo	the pain is severe	AC	2	2	2
doctor	do you have a fever	the sudden fever			NA_BR	2	2	2
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC_help	2	2	2
patient	sí yo he tenido fiebre	sí fiebre ha venido siempre	fiebre ha empezado siempre	yes fever always started	AC_IE	2	2	2
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	2	2	2
patient	treinta y nueve punto cinco grados centigrados	treinta y nueve horas si tengo grados			NA_BR	2	2	2
patient	tengo treinta y nueve grados de fiebre	tengo treinta y nueve grados de fiebre	tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	AC_help	2	2	2

doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	2	2
patient	no no tengo tos	no horas tengo tos	tengo tos horas	no i cough hours	NA_BR	2	2	2
patient	no no tengo tos	no noche tengo tos	tengo tos la noche	no i cough the evening	NA_BR	2	2	2
patient	no	no	no	no	AC	2	2	2
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	2	2
patient	no	no	no	no	AC	2	2	2
doctor	have you had any vomiting	have you had pain if vomiting	have you experienced the pain when you vomit?	¿le ha dolido cuando vomita?	NA_BR	2	2	2
doctor	any vomiting	any augmentin			NA_BR	2	2	2
doctor	have you had a headache	have you had headache	have you had the headaches?	¿le ha dolido la cabeza?	AC	2	2	2
patient	sí sí he tenido dolor de cabeza	sí siete días lo ví a veces	consulto a veces un médico desde hace siete días		NA_BR	2	2	2
patient	sí me duele la cabeza desde hace algunos días	sí me duele la cabeza desde hace algunos días	me duele la cabeza desde hace algunos días	yes i have experienced the pain in the head for several days	AC_help	2	2	2
doctor	have you seen a doctor already	have you seen an doctor to augmentin			NA_BR	2	2	2
doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor	ha consultado un médico	AC_help	2	2	2
patient	no	no	no	no	AC	2	2	2
doctor	have you been around anyone who is sick	have you been around anyone tuesday			NA_BR	2	2	2
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	2	2	2
patient	sí hace dos días	sí hace dos días	he estado en contacto con alguien que tiene el dolor de garganta hace dos días	i was in contact with someone two days ago who had a sore throat	AC	2	2	2
doctor	are you allergic to any medication	were you allergic to any medication	were you allergic to any medication?	¿ha tenido alergia a alguna medicina?	AC	2	2	2
patient	no	no	no	no	AC	2	2	2

doctor	have you had a stomachache	have you had a stabbing pain	have you experienced a stabbing pain?		NA_BR	2	2	2
doctor	have you had a stomachache	have you had a stabbing pain	have you experienced a stabbing pain?		NA_BR	2	2	2
doctor	how long have you had a stomachache	how long have you had a stomachache	how long have you had a stomachache?	¿desde cuándo le duele el estómago?	AC_help	2	2	2
patient	desde hace dos días	desde hace dos días	el estómago me duele desde hace dos días	i have experienced the pain in the stomach for two days	AC	2	2	2
doctor	have you had any vomiting	have you had any if coming			NA_BR	2	2	2
doctor	have you had any diarrhea	have you had pain in that area			NA_BR	2	2	2
doctor	have you had any congestion	have you had intense congestion	have you felt severely congested?		NA_BR	2	2	2
doctor	how long have you had congestion	how long have you had congestion	how long have you had congestion	¿desde cuándo tiene la nariz tapada?	AC_help	2	2	2
patient	no		no	no	AC	2	2	2
doctor	i think you have a strep throat	did you have strep throat	did you have strep throat?	¿ha tenido una infección de la garganta por estreptococo?	NA_BR	2	2	2
doctor	good morning	good morning	good morning?	¿hola?	AC	2	2	1
patient	buenos días	buenos días	hola	good morning	AC	2	2	1
doctor	how are you this morning	how high is this morning	how severe is this morning?	¿cuál es la intensidad la mañana?	NA_BR	2	2	1
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	2	2	1
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	2	1
doctor	how long have you had a sore throat	how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	2	1
patient	me duele desde hace cuatro días	me duele desde hace cuatro días	me duele desde hace cuatro días	i have experienced the pain for four days	AC	2	2	1
doctor	what is the type of pain	what is this kind of pain	what is the character of the pain?	¿cómo es el dolor?	AC	2	2	1
patient	no es tan fuerte	no esta hora			NA_BR	2	2	1
patient	me duele poco	me da moco			NA_BR	2	2	1
patient	duele poco	hola moco			NA_BR	2	2	1

patient	no entendí	no augmentin			NA_BR	2	2	1
patient	no entiendo	no entiendo	no entiendo	i don't understand	AC	2	2	1
doctor	is the pain severe	is the pain severe	is the pain severe?	¿el dolor es severo?	AC	2	2	1
patient	no	no	no	no	AC	2	2	1
doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC	2	2	1
patient	no sé	no sé	no sé	i don't know	AC	2	2	1
doctor	are you having fever	by having fever	is the pain better when you have fever?	¿el dolor está aliviado cuando tiene fiebre?	AC_IE	2	2	1
doctor		[rien]	is the pain better when you have fever?	¿el dolor está aliviado cuando tiene fiebre?	REP	2	2	1
patient	no	no	no	no	AC	2	2	1
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	2	2	1
patient	no sé	no sé arto			NA_BR	2	2	1
patient	tengo fiebre cuando me toco	tengo fiebre cuando la tengo	tengo fiebre cuando presento este síntoma	i have fever when i experience the symptoms	NA_BR	2	2	1
patient	la siento a tocar	la temperatura			NA_BR	2	2	1
patient	no está medida	no esta medicina			NA_BR	2	2	1
patient	no sé	no sé	no sé	i don't know	AC	2	2	1
doctor	have you had a cough	have you had cough	have you had a cough?	¿ha tenido tos?	AC	2	2	1
patient	no	no	no	no	AC	2	2	1
doctor	have you had a runny nose	have you had a runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	2	2	1
patient	tan poco	un poco	he tenido un poco de moco	i had a bit of a runny nose	NA_BR	2	2	1
patient	no	no	no	no	AC	2	2	1
doctor	have you had a rash	have you had rash	have you had a rash?	¿ha tenido una erupción cutánea?	AC	2	2	1
patient	no	no	no	no	AC	2	2	1
doctor	have you had a stomachache	have you had stabbing pains	have you experienced a stabbing pain?		NA_BR	2	2	1
doctor	how long have you had a stomachache	how long have you had a stomachache	how long have you had a stomachache?	¿desde cuándo le duele el estómago?	AC_help	2	2	1
patient	no me duele	no me duele	me duele	no i experience a pain	AC_IE	2	2	1

doctor	have you had a headache	have you had headache	have you had the headaches?	¿le ha dolido la cabeza?	AC	2	2	1
patient	sí	sí	sí	yes	AC	2	2	1
doctor	how long have you had a headache	how long have you had headache	how long have you had the headaches?	¿desde cuándo le duele la cabeza?	AC	2	2	1
patient	desde hace cuatro días	desde hace cuatro días	me duele la cabeza desde hace cuatro días	i have experienced the pain in the head for four days	AC	2	2	1
doctor	is the headache improving	is the headache increasing	are the headaches increasing?		NA_BR	2	2	1
doctor	is the headache getting better	is the headache getting better			NA_NT	2	2	1
doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC_help	2	2	1
patient	no sé	once			NA_BR	2	2	1
patient	no sé	no tres			NA_BR	2	2	1
patient	no sé	no sé	no sé	i don't know	AC_help	2	2	1
doctor	have you been around anyone who is sick	have you been around anyone who is sick	have you been in contact with anyone who feel nauseous?		NA_NT	2	2	1
doctor	have you been around anyone with a sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	2	2	1
patient	no	no	no	no	AC	2	2	1
doctor	have you seen a doctor	have you seen an doctor	have you visited the doctor?	¿ha consultado un médico?	NA_IE	2	2	1
doctor	have you seen a doctor for your sore throat	have you seen a doctor for your sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC	2	2	1
patient	sí hace dos días fui	sí hace los días hoy	he consultado hoy un médico para el dolor de garganta hace días		NA_BR	2	2	1
patient	sí hace dos días	sí hace dos días	he consultado un médico para el dolor de garganta hace dos días	i visited the doctor for my sore throat two days ago	AC	2	2	1
doctor	did they run any tests	is a bad pain intense	is the pain severe?	¿el dolor es severo?	NA_BR	2	2	1

doctor	have you had any medicine for the pain	have you had pain when medicine is helping			NA_BR	2	2	1
doctor	have you taken any medicine	have you taken any medicine	have you taken medicine?	¿ha tomado medicina?	AC	2	2	1
patient	no	no	no	no	AC	2	2	1
doctor	do you have strep throat	do you have strep throat	do you have strep throat?	¿tiene una infección de la garganta por estreptococo?	AC	2	2	1
patient	no	no	no	no	AC	2	2	1
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	2	2	1
patient	sí soy alérgica a la penicilina	estoy alérgica al penicilina	la penicilina me da alergias	i am allergic to penicillin	NA	2	2	1
patient	sí tengo alergia a la penicilina	sí tengo alergia a la penicilina	la penicilina me da alergias	yes i am allergic to penicillin	AC_help	2	2	1
doctor	do you think you are getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC	2	2	1
patient	no sé	doce			NA_BR	2	2	1
patient	no sé	no sé	no sé	i don't know	AC	2	2	1
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	2	1	1
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	1	1
doctor	how long have you had a sore throat	how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	1	1
patient	por cuatro días	cuatro días	me duele la garganta desde hace cuatro días	i have had a sore throat for four days	AC	2	1	1
doctor	how severe is the pain	how severe is the pain		¿cuál es la intensidad del dolor?	NA_NB	2	1	1
doctor	what is the character of the pain	what is the character of the pain?	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	1	1
patient	no muy fuerte	no muy fuerte	el dolor es muy fuerte	the pain is very severe	AC_IE	2	1	1
doctor	do you have a fever	've you had a fever	have you had fever?	¿ha tenido fiebre?	AC_BR	2	1	1
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC_help	2	1	1
patient	creo que no	profundo		i have a deep fever	AC_BR	2	1	1
patient	no	no	no	no	AC	2	1	1

doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	2	1	1
patient	sí un poco	sí un poco	me duele un poco la cabeza	i experience a bit of pain in the head	AC	2	1	1
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	1	1
patient	no	no	no	no	AC	2	1	1
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	1	1
patient	no	no	no	no	AC_help	2	1	1
doctor	did you see a doctor earlier	did you see a doctor the early year			AC_BR	2	1	1
doctor	did you see the doctor this week	did you see the doctor this week	did you visit the doctor this week?	¿ha consultado un médico esta semana?	AC_help	2	1	1
patient	sí ví a una enfermera de practica medica hace tres días	sí ví hoy enfermera últimamente hace tres días	he consultado hoy una enfermera recientemente hace tres días		NA_BR	2	1	1
patient	sí una enfermera especializada hace tres días	sí una enfermera ha estado sabado hace tres días			NA_BR	2	1	1
patient	sí he consultado una enfermera hace tres días	sí he consultado una enfermera hace tres días	he consultado una enfermera hace tres días	yes i visited a nurse three days ago	AC_help	2	1	1
doctor	did he do a throat culture	did you do a throat culture	did you carry out a throat culture test?		NA_BR	2	1	1
doctor	was a throat culture test done	was a throat culture test done	was a throat culture test carried out?	¿le han realizado un cultivo de la garganta?	AC	2	1	1
patient	sí	sí	sí	yes	AC	2	1	1
doctor	what were the results	what were the results	what were the results?	¿cuál ha sido el resultado?	NA_IE	2	1	1
doctor	what were the results of this test	what were the results of this test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC_help	2	1	1
patient	todavia no sé	lo ha diagnosticado	me han diagnosticado este síntoma	the diagnosis was the symptoms	NA_BR	2	1	1
patient	no sé	no sé	no sé	i don't know	AC_help	2	1	1
doctor	do you take any medicines	are you taking medicine	do you take medicine?	¿toma medicina?	AC	2	1	1

patient	no	no	no	no	AC	2	1	1
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	2	1	1
patient	no	no	no	no	AC	2	1	1
doctor	where is your pain	where is the pain	where is the pain?	¿dónde le duele?	AC	2	1	2
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	1	2
doctor	how long has your throat hurt	how long has your throat hurt	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	1	2
patient	no sé	no sé	no sé	i don't know	AC	2	1	2
doctor	how bad is the pain	how bad is the pain			NA_NB	2	1	2
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	1	2
patient	muy fuerte	muy fuerte	el dolor es muy fuerte	the pain is very severe	AC	2	1	2
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	1	2
patient	tengo fiebre	[rien]			NA_BR	2	1	2
patient	sí	sí	sí	yes	AC	2	1	2
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	1	2
patient	no	no	no	no	AC	2	1	2
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	2	1	2
patient	sí mucho	sí mucho	me duele mucho la cabeza	i experience a severe pain in the head	AC	2	1	2
doctor	has anyone else had a sore throat	has anyone now had a sore throat			NA_NB	2	1	2
doctor	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who have a sore throat?	¿ha estado recientemente en contacto con alguien que tiene el dolor de garganta?	AC_help	2	1	2
patient	sí he estado con alguien que tiene dolor de garganta	sí he estado una noche que tiene dolor de garganta			NA_BR	2	1	2
patient	sí he estado en contacto con alguien que tiene una infección de la garganta	si he estado en contacto con alguien que tiene una infección de la	he estado en contacto con alguien que tiene el dolor de garganta	yes i was in contact with someone who had a sore throat	AC_help	2	1	2

doctor	did you see a doctor this week	did you see a doctor this week	did you visit the doctor this week?	¿ha consultado un médico esta semana?	AC	2	1	2
patient	sí	sí	sí	yes	AC	2	1	2
doctor	did you have a throat culture	did you have a throat culture	did you have a throat culture test?	¿le han realizado un cultivo de la garganta?	AC	2	1	2
patient	sí	sí	sí	yes	AC	2	1	2
doctor	what were the results	what were the results	what were the results?	¿cuál ha sido el resultado?	AC	2	1	2
patient	el resultado fue positivo	el resultado fue positivo	el resultado ha sido positivo	the results were positive	AC	2	1	2
doctor	are you taking any medicine	are you eating any medicine	do you eat medicine?	¿come medicina?	NA_BR	2	1	2
doctor	have you taken any medicine for the sore throat	have you taken any medicine for the sore throat	have you taken medicine for your sore throat?	¿ha tomado medicina para el dolor de garganta?	AC_help	2	1	2
patient	no	no	no	no	AC	2	1	2
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	2	1	2
patient	sí a la moxilina amoxicilina	día a la amoxicilina amoxicilina			NA_BR	2	1	2
patient	sí tengo alergia a la amoxicilina	sí tengo alergia a la amoxicilina	la amoxicilina me da alergias	yes i am allergic to amoxicillin	AC_help	2	1	2
doctor	good morning	good morning	good morning?	¿hola?	AC	2	2	4
patient	buenos días	buenas tardes	buenas tardes	good afternoon	NA_BR	2	2	4
patient	buenos días	buenos días	hola	good morning	AC	2	2	4
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	2	2	4
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	2	4
doctor	how long have you had a sore throat	how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	2	4
patient	más de una semana	más de una semana	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	AC	2	2	4
doctor	what is the type of pain	what is that kind of pain	what is the character of the pain?	¿cómo es el dolor?	AC	2	2	4
patient	el dolor es muy fuerte	el dolor es muy fuerte	el dolor es muy fuerte	the pain is very severe	NA	2	2	4
patient	el dolor es muy fuerte	el dolor es muy fuerte	el dolor es muy fuerte	the pain is very severe	AC	2	2	4

doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC	2	2	4
patient	sí	sí	sí	yes	NA	2	2	4
patient	sí	sí	sí	yes	AC	2	2	4
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	2	2	4
patient	sí	sí	sí	yes	AC	2	2	4
doctor	what is your temperature	what is your temperature	what is your temperature?	¿cuántos grados tiene?	AC	2	2	4
patient	tengo treinta y nueve grados de fiebre	tengo treinta y nueve grados de fiebre	tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	AC	2	2	4
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	2	4
patient	sí	sí	sí	yes	AC	2	2	4
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	2	2	4
patient	sí	sí	sí	yes	AC	2	2	4
doctor	how long have you had the cough	how long have you had the cough	how long have you had a cough?	¿desde cuándo tiene tos?	AC	2	2	4
patient	hace algunas semanas	hace algunas semanas	tengo tos hace algunas semanas	i have coughed for several weeks	AC	2	2	4
doctor	have you had a rash	have you had rash	have you had a rash?	¿ha tenido una erupción cutánea?	AC	2	2	4
patient	sí	sí	sí	yes	AC	2	2	4
doctor	what type of rash	what kind of rash			NA_NT	2	2	4
doctor	what does the rash look like	what piercing rash was night	what was a piercing rash night?	¿qué ha sido una erupción cutánea penetrante la noche?	NA_BR	2	2	4
doctor	where is the rash	where is rash	where is a rash?		NA_NT	2	2	4
doctor	is the rash better	is a rash better	is a rash better?		NA_NT	2	2	4
doctor	do you have a headache	the bad headache			NA_BR	2	2	4
doctor	have you had a headache	have you had headache	have you had the headaches?	¿le ha dolido la cabeza?	AC	2	2	4
patient	sí	sí	sí	yes	AC	2	2	4
doctor	how long have you had the headache	how long have you had headache	how long have you had the headaches?	¿desde cuándo le duele la cabeza?	AC	2	2	4

patient	más de una semana	más de una semana	me duele la cabeza desde hace más de una semana	i have experienced the pain in the head for more than one week	AC	2	2	4
doctor	have you had a stomachache	have you had a stabbing pain	have you experienced a stabbing pain?		NA_BR	2	2	4
doctor	how long have you had a stomachache	how long have you had a stomachache	how long have you had a stomachache?	¿desde cuándo le duele el estómago?	AC_help	2	2	4
patient	hoy	hoy	el estómago me duele desde hoy	i have experienced the pain in the stomach since today	AC	2	2	4
doctor	have you been around anyone with a sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA	2	2	4
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	2	2	4
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	2	4
doctor	have you seen a doctor	any doctor			NA_BR	2	2	4
doctor	have you seen a doctor	have you seen a doctor	have you seen a doctor	¿ha consultado un médico?	AC_help	2	2	4
patient	sí hace una semana	sí hace una semana	he consultado un médico hace una semana	i visited the doctor one week ago	AC	2	2	4
doctor	did the doctor do a strep test	did the doctor do strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC	2	2	4
patient	sí la semana pasada	sí la semana pasada	me han realizado una prueba rápida por estreptococo la semana pasada	a rapid strep test was carried out during the last week	AC	2	2	4
doctor	what was the test result	while fatigue has helped			NA_BR	2	2	4
doctor	what was the result of the test	what was the result of the test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC	2	2	4
patient	el resultado fue positivo	el resultado fue positivo	el resultado ha sido positivo	the results were positive	AC	2	2	4

doctor	are you on any medication	anyone to any medication			NA_BR	2	2	4
doctor	have you taken any medicine for the sore throat	have you taken any medicine for the sore throat	have you taken medicine for your sore throat?	¿ha tomado medicina para el dolor de garganta?	AC_help	2	2	4
patient	sí	sí	sí	yes	AC	2	2	4
doctor	what medication are you taking	what medication are they taking			NA_BR	2	2	4
doctor	what medication	what medication			NA_BR	2	2	4
doctor	are you taking penicillin	have you taken penicillin	have you taken penicillin?	¿ha tomado la penicilina?	AC	2	2	4
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	2	4
doctor	are you allergic to any medication	are you allergic to any medication	are you allergic to any medication?	¿tiene alergia a alguna medicina?	AC	2	2	4
patient	no sé	no sé	no sé	i don't know	AC	2	2	4
doctor	good morning	good morning	good morning?	¿hola?	AC	2	2	3
patient	buenos días	buenos días	hola	good morning	AC	2	2	3
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	2	2	3
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	2	3
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	2	3
patient	más de una semana	no entiendo semana			NA_BR	2	2	3
patient	más de una semana	más de una semana	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	AC	2	2	3
doctor	what is the type of pain	is the pain this pain	do you experience the pain?	¿tiene el dolor?	AC	2	2	3
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC	2	2	3
patient	el dolor es profundo	el dolor es profundo	el dolor es profundo	the pain is deep	AC	2	2	3
doctor	is the pain improving	is the pain increasing	is the pain worse?	¿el dolor está	NA	2	2	3
doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC	2	2	3
patient	no	no	no	no	AC	2	2	3
doctor	is the pain becoming more frequent	is the pain becoming more frequent	is the frequency of the pain increasing?	¿le duele más a menudo?	AC	2	2	3

patient	sí	sí	sí	yes	AC	2	2	3
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	2	2	3
patient	no	no	no	no	AC	2	2	3
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	2	2	3
patient	sí tengo congestion	sí tengo operación			NA_BR	2	2	3
patient	sí sí tengo	sí cinco			NA_BR	2	2	3
patient	sí	sí	sí	yes	AC	2	2	3
doctor	have you had a rash	another rash			NA_BR	2	2	3
doctor	do you have a rash	do you have rash	do you have a rash?	¿tiene una erupción cutánea?	AC	2	2	3
patient	no	no	no	no	AC	2	2	3
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	2	2	3
patient	sí	sí	sí	yes	AC	2	2	3
doctor	how long have you had a headache	how long have you had headache	how long have you had the headaches?	¿desde cuándo le duele la cabeza?	AC	2	2	3
patient	más de un día	la tiene días	el médico presenta este síntoma días	the doctor experiences the symptoms days	AC	2	2	3
doctor	do you have a stomach ache	do you have stabbing pains	do you experience a stabbing pain?		NA	2	2	3
doctor	have you had a stomach ache	have you had a stabbing pain	have you experienced a stabbing pain?		NA	2	2	3
doctor	do you have a stomach ache	do you have a stabbing pain	do you experience a stabbing pain?		NA	2	2	3
doctor	have you felt tired	have you felt tired	have you felt tired?	¿ha estado cansado?	AC	2	2	3
patient	mucho	mucho			NA_NT	2	2	3
patient	sí	sí	sí	yes	AC	2	2	3
doctor	have you been around anyone who has a sore throat	have you been around anyone who has sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	2	2	3
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	2	3
doctor	have you already seen a doctor	have you lately seen doctor	have you recently visited the doctor?	¿ha consultado recientemente un médico?	AC	2	2	3
doctor	have you seen a doctor	have you seen a doctor	have you visited the doctor?	¿ha consultado un médico?	AC	2	2	3

patient	sí una enfermera	sí una enfermera	he consultado una enfermera	i visited a nurse	AC	2	2	3
doctor	did they run any tests	did pain around any last			NA_BR	2	2	3
doctor	did they do a strep test	did they do a strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC	2	2	3
patient	no sé	no sé	no sé	i don't know	AC	2	2	3
doctor	when did you see the doctor	when did you see the doctor	when did you visit the doctor?	¿cuándo ha consultado un médico?	AC	2	2	3
patient	hace dos días	hace dos días	he consultado un médico hace dos días	i visited the doctor two days ago	AC	2	2	3
doctor	are you taking any medicine	by taking any			NA_BR	2	2	3
doctor	have you taken any medicine for the sore throat	have you taken any medicine for the sore throat	have you taken medicine for your sore throat?	¿ha tomado medicina para el dolor de garganta?	AC	2	2	3
patient	no	no	no	no	AC	2	2	3
doctor	are you allergic to any medicine	were you allergic to any medicine	were you allergic to any medicine?	¿ha tenido alergia a alguna medicina?	AC	2	2	3
patient	no	no	no	no	AC	2	2	3
doctor	where is your pain	where are your pains	where is the pain?	¿dónde le duele?	AC	2	1	3
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	1	3
doctor	how bad is the pain	how bad is the pain			NA_NT	2	1	3
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	1	3
patient	es fuerte	es fuerte	el dolor es severo	the pain is severe	AC	2	1	3
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	1	3
patient	me duele desde la mañana	me duele desde la mañana	me duele desde la mañana	i have experienced the pain this morning	AC	2	1	3
doctor	do you have a headache	me duele desde la mañana	do you have the headaches?	¿le duele la cabeza?	AC	2	1	3
patient	sí tengo dolor en la cabeza	sí tengo dolor en la cabeza	me duele la cabeza	yes i experience a pain in the head	AC	2	1	3
doctor	do you have a fever	your fever	do you have fever?	¿tiene fiebre?	AC	2	1	3
patient	sí tengo mucha fiebre	sí tengo mucha fiebre	tengo mucha fiebre	yes i have a severe fever	AC	2	1	3

doctor	what is your temperature	what is your temperature	what is your temperature?	¿cuántos grados tiene?	AC	2	1	3
patient	treinta y ocho punto cinco	treinta y ocho grados	tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	NA_BR	2	1	3
patient	treinta y ocho grados de fiebre	treinta y ocho grados de fiebre	tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	AC_help	2	1	3
doctor	do you have a cough	do you have cough	do you have a cough?	¿tiene tos?	AC	2	1	3
patient	sí tengo poca tos	sí tengo una tos	tengo tos	yes i cough	AC	2	1	3
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	1	3
patient	sí tengo mucho moco	sí tengo mucho moco	tengo mucho moco	yes i have a runny nose	AC	2	1	3
doctor	has anyone around you had a sore throat	had anyone around you had sore throat	had anyone around you had a sore throat?		NA_BR	2	1	3
doctor	is anybody around you suffering from a sore throat	is anybody around you suffering from a sore throat	does anyone around you have your sore throat?	¿alguien en su entorno tiene el dolor de garganta?	AC_help	2	1	3
patient	no nadie tiene	no nariz			NA_BR	2	1	3
patient	no nadie	no nariz			NA_BR	2	1	3
patient	no	no	no	no	AC	2	1	3
doctor	are you taking any medicine for the pain	have you taken any medicine for the pain	have you taken medicine for the pain?	¿ha tomado medicina para el dolor?	AC	2	1	3
patient	no he tomado medicina	no he tomado medicinas	he tomado medicina	no i took medication	AC_IE	2	1	3
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	2	1	3
patient	sí la amoxicilina	sí la amoxicilina	la amoxicilina me da alergias	i am allergic to amoxicillin	AC	2	1	3
doctor	did you see another doctor today	another saturday			NA_BR	2	1	3
doctor	did you see the doctor today	did you see the doctor today	did you visit the doctor since today?	¿ha consultado hoy un médico?	AC_help	2	1	3
patient	sí hace una semana ví una enfermera	sí hace una semana ví enfermera	he consultado una enfermera hace una semana	yes i visited a nurse one week ago	AC	2	1	3
doctor	did she do any tests	did you do a test	did you carry out the test?		NA_BR	2	1	3

doctor	was a strep test done	was a strep test done	was a rapid strep test carried out?	¿le han realizado la prueba rápida por estreptococo?	AC_help	2	1	3
patient	no	no	no	no		2	1	3
doctor	where is your pain	where are pains	where is the pain?	¿dónde le duele?	AC	2	1	4
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	1	4
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	1	4
patient	me duele desde hace más de una semana	me duele desde hace más de una semana	me duele desde hace más de una semana	i have experienced the pain for more than one week	AC	2	1	4
doctor	how bad is the pain	how bad is the pain			NA_BR	2	1	4
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	1	4
patient	muy fuerte	muy fuerte	el dolor es muy fuerte	the pain is very severe	AC	2	1	4
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	2	1	4
patient	sí	sí	sí	yes	AC	2	1	4
doctor	what is your temperature	what is your temperature	what is your temperature?	¿cuántos grados tiene?	AC	2	1	4
patient	treinta y nueve grados	sí nueve grados	tengo nueve grados de fiebre	my temperature is nine degrees	AC_IE	2	1	4
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	1	4
patient	sí	sí	sí	yes	AC	2	1	4
doctor	headache	headache	do you have the headaches?	¿le duele la cabeza?	AC	2	1	4
patient	sí	sí	sí	yes	AC	2	1	4
doctor	runny nose	runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	1	4
patient	sí	sí	sí	yes	AC	2	1	4
doctor	has anyone around you been sick	had anyone around you been sick			NA_NT	2	1	4
doctor	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC	2	1	4

patient	sí	sí	sí	yes	AC	2	1	4
doctor	have you been to the doctor	have you been to the doctor	have you visited the doctor?	¿ha consultado un médico?	AC	2	1	4
patient	sí	sí	sí	yes	AC	2	1	4
doctor	did they do a strep throat test	did a few make strep throat intense			NA_BR	2	1	4
doctor	did the doctor do a throat culture test	did the doctor do a throat culture test	did the doctor carry out a throat culture test?	¿le han realizado un cultivo de la garganta?	AC	2	1	4
patient	sí	sí	sí	yes	AC	2	1	4
doctor	what was the result	what was the results	what were the results?	¿cuál ha sido el resultado?	AC	2	1	4
patient	negativo	negativo	el resultado ha sido negativo	the results were negative	AC	2	1	4
doctor	are you taking any medicine for the pain	have you taken any medicine for the pain	have you taken medicine for the pain?	¿ha tomado medicina para el dolor?	NA_BR	2	1	4
doctor	have you taken any medicine for the sore throat	have you taken any medicine for the sore throat	have you taken medicine for your sore throat?	¿ha tomado medicina para el dolor de garganta?	AC	2	1	4
patient	sí	sí	sí	yes	AC	2	1	4
doctor	what medecine are you taking	what medicine are you taking	what medicine do you take?	¿qué medicina toma?	AC	2	1	4
patient	tilenol	ayer hoy			NA_BR	2	1	4
patient	tilenol	ayer hoy			NA_BR	2	1	4
patient	acetaminofein	hace los meses			NA_BR	2	1	4
patient	tilenol	ayer hoy			NA_BR	2	1	4
patient	tilenol	ayer hoy			NA_BR	2	1	4
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	2	1	4
patient	no	no	no	no	AC	2	1	4
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	2	4	6
doctor	when did your problem start	when does your cough start	when does a cough start?	¿cuándo empieza tos?	NA_BR	2	4	6
doctor	when did it start	when did it start	when did the pain start?	¿cuándo ha empezado el dolor?	AC	2	4	6
doctor	when did your fever start	when did your fever start	when did fever start?	¿desde cuándo tiene fiebre?	AC	2	4	6

patient	empezo desde ayer	ha empezado trece días	este síntoma empieza desde hace trece días	the symptoms have started for thirteen days	NA_BR	2	4	6
patient	la fiebre empezo ayer	la fiebre ha empezado ayer	fiebre ha empezado ayer	fever started yesterday	AC	2	4	6
doctor	when did your sore throat start	when does your sore throat start	when does your sore throat start?	¿cuándo empieza el dolor de garganta?	AC	2	4	6
patient	hace dos días	hace dos días	el dolor de garganta empieza hace dos días	my sore throat has started for two days	AC	2	4	6
doctor	do you have fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	4	6
patient	sí	sí	sí	yes	AC	2	4	6
doctor	do you have cough	your cough	do you have a cough?	¿tiene tos?	AC	2	4	6
patient	no	no	no	no	AC	2	4	6
doctor	do you have congestion	your congestion	do you feel congested?	¿tiene la nariz tapada?	AC	2	4	6
patient	no	no	no	no	AC	2	4	6
doctor	do you vomiting	are you vomiting	do you vomit?	¿vomita?	AC	2	4	6
patient	no	no	no	no	AC	2	4	6
doctor	do you have diarrhea	your area			NA_BR	2	4	6
doctor	do you have diarrhea	your area			NA_BR	2	4	6
doctor	are you getting better	are you getting better	is the pain better?	¿el dolor disminuye?	AC	2	4	6
patient	no sé	no sé	no sé	i don't know	AC	2	4	6
doctor	do you have runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	4	6
patient	no	no	no	no	AC	2	4	6
doctor	are your eyes red	for your times day			NA_BR	2	4	6
doctor	do your eyes hurt	your times your day			NA_BR	2	4	6
doctor	are your eyes red	[rien]			NA_BR	2	4	6
doctor	how high was your fever	how high was your fever	what was your temperature?	¿cuántos grados ha tenido?	AC	2	4	6
patient	treinta y nueve punto cinco	treinta y nueve grados celsius	he tenido treinta y nueve grados de fiebre	my temperature was one hundred three degrees	NA_BR	2	4	6
patient	treinta y nueve punto cinco	sesenta meses en tos cinco días			NA_BR	2	4	6
patient	treinta y nueve cinco	sesenta meses cinco horas			NA_BR	2	4	6
patient	he tenido treinta y nueve punto cinco	sí tengo treinta y nueve grados celsius	tengo treinta y nueve grados de fiebre	yes my temperature is one hundred three degrees	AC	2	4	6

doctor	is anyone else sick	is anyone now sick			NA_BR	2	4	6
doctor	is anyone else sick	is anyone hour sick	does anyone feel nauseous for hours?		NA_BR	2	4	6
doctor	when have you been in contact with a person who has strep throat	when have you been in contact with a person who has strep throat	when have you been in contact with the person who has strep throat?	¿cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo?	AC_help	2	4	6
patient	no sé	no sé	no sé	i don't know	AC	2	4	6
doctor	sorry what is the problem	are they with several			NA_BR	2	4	5
doctor	what is the problem	what is several			NA_BR	2	4	5
doctor	where is your pain	where are your pains	where is the pain?	¿dónde le duele?	AC	2	4	5
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	4	5
doctor	for how long	[rien]			NA_BR	2	4	5
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	4	5
patient	me duele hace cuatro días	me duele hace cuatro días	me duele hace cuatro días	i have experienced the pain for four days	AC	2	4	5
doctor	is anyone else sick	is anyone hour sick	does anyone feel nauseous for hours?		NA_BR	2	4	5
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	2	4	5
patient	tengo mucha fiebre	tengo mucho fiebre	tengo mucha fiebre	i have a severe fever	AC	2	4	5
doctor	do you know the number	do you have another			NA_BR	2	4	5
doctor	how high is the fever	how high is the fever	what is your temperature?	¿cuántos grados tiene?	NA_IE	2	4	5
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC_help	2	4	5
patient	fiebre al tocar	fiebre dura	fiebre dura	fever lasts	AC	2	4	5
doctor	do you have cough	your cough			NA_BR	2	4	5
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	4	5
patient	no no tengo tos	no tengo tos	tengo tos	no i cough	AC_IE	2	4	5
doctor			do you have a cough?	¿tiene tos?	REP	2	4	5
patient	no no tengo tos	no noche tengo tos	tengo tos la noche	no i cough the evening	AC_IE	2	4	5
doctor	do you have congestion	your congestion	do you feel congested?	¿tiene la nariz tapada?	AC	2	4	5

patient	no tengo la nariz tapada	no tengo la nariz tapada	tengo la nariz tapada	no i feel congested	AC	2	4	5
doctor	do you have a headache	your medication				2	4	5
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC_help	2	4	5
patient	sí me duele la cabeza	sí si tengo más a menudo la cabeza			NA_BR	2	4	5
patient	sí tengo dolor de cabeza	sí si tengo dolor de cabeza			NA_BR	2	4	5
patient	sí	sí	sí	yes	AC	2	4	5
doctor	do you vomit	do you vomit	do you vomit?	¿vomita?	AC	2	4	5
patient	sí	sí	sí	yes	AC	2	4	5
doctor	do you have diarrhea	your area			NA_BR	2	4	5
doctor	do you have diarrhea	your area			NA_BR	2	4	5
doctor	do you have a rash	your rash			NA_BR	2	4	5
doctor	is there a rash	[rien]			NA_BR	2	4	5
doctor	are you better	are you better	is the pain better?	¿el dolor disminuye?	AC	2	4	5
patient	no ha he disminuido	no tengo a veces			NA_BR	2	4	5
patient	no ha desminuido	no dolor ha venido	el dolor ha empezado	no the pain started	AC	2	4	5
doctor	is the pain better	is the pain better	is the pain better?	¿el dolor disminuye?	AC	2	4	5
patient	no	no	no	no	AC	2	4	5
doctor	who else is sick	were you often sick	did you often feel nauseous?	¿ha tenido a menudo náuseas?	NA_BR	2	4	5
doctor	have you been in contact with anyone with a strep throat in the last week	have you been in contact with anyone with a strep throat in the last week	have you been in contact with anyone with a strep throat in the last week	¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo en la semana pasada?	AC	2	4	5
patient	no no he tenido contacto	no lo tengo hoy			NA_BR	2	4	5
doctor	for how many days have you had a sore throat	how many days have you had to food			NA_BR	2	3	5
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	3	5
patient	me duele desde hace cuatro días	me duele desde hace cuatro días	me duele desde hace cuatro días	i have experienced the pain for four days	AC	2	3	5

doctor	how severe is the pain	how severe is the pain		¿cuál es la intensidad del dolor?	NA	2	3	5
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	3	5
patient	me duele la cabeza	me duele la cabeza	me duele la cabeza	i experience a pain in the head	AC	2	3	5
doctor	do you have a runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	3	5
patient	no no tengo moco	no tengo moco			NA_BR	2	3	5
patient	no no tengo moco	no tengo moco			NA_BR	2	3	5
patient	no no tengo	no lo tengo	presento este síntoma	no i experience the symptoms	AC_IE	2	3	5
doctor	have you had a fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	2	3	5
patient	no he tenido fiebre	no sé tengo fiebre	no sé tengo fiebre		NA_BR	2	3	5
patient	no no he tenido fiebre	no noche tengo fiebre	tengo fiebre la noche	no i have fever the	NA_BR	2	3	5
patient	no no tengo fiebre		presento siempre este síntoma	no i always experience the symptoms	AC_IE	2	3	5
doctor	have you been in contact with anyone who had strep	have you been in contact with anyone who has test	have you been in contact with anyone who have the test?		NA_BR	2	3	5
doctor	have you been in contact with anyone recently who has strep throat	have you been in contact with anyone recently who has strep throat	have you been in contact with anyone recently who have strep throat?	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	2	3	5
patient	no no he estado en contacto con alguien	no noche en contacto con alguien	he estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo la noche	i was in contact with someone the evening recently who had strep throat	AC_IE	2	3	5
doctor	do you have a sore throat	do you have sore throat	do you have a sore throat?	¿le duele la garganta?	AC	2	3	5
patient	no me duele la garganta	no me duele la garganta	me duele la garganta	no i have a sore throat	AC_IE	2	3	5
doctor	do you have cough	do you have cough	do you have a cough?	¿tiene tos?	AC	2	3	5
patient	no no tengo tos	no tengo tos	tengo tos	no i cough	NA_BR	2	3	5

doctor	what symptoms do you have	less than an year	do you have a sore throat for less than one year?	¿le duele la garganta menos de un año?	NA_BR	2	3	5
doctor	do you have a rash	do you have a rash	do you have a rash?	¿tiene una erupción cutánea?	AC_help	2	3	5
patient	no no tengo	no lo tengo	presento este síntoma	no i experience the symptoms	AC_IE	2	3	5
doctor	how do you feel	have you seen you			NA_BR	2	3	5
doctor	do you feel sick	do you feel sick	do you feel nauseous?	¿tiene náuseas?	NA_IE	2	3	5
doctor	do you feel sick	do you feel sick	do you feel nauseous?	¿tiene náuseas?	AC	2	3	5
patient	no no tengo nausea	no algunos			NA_BR	2	3	5
patient	no no tengo nausea	no tengo náuseas	tengo náuseas	no i feel nauseous	AC_IE	2	3	5
doctor	do you have diarrhea	do you have the pain year	do you experience the pain for several years?	¿le duele algunos años?	NA_BR	2	3	5
doctor	do you have diarrhea	do you have the pain early			NA_BR	2	3	5
doctor	do you have vomiting	do you have stomachache	do you have a stomachache?	¿el estómago le duele?	AC	2	3	5
patient	sí me duele	sí me duele	me duele	yes i experience a pain	AC	2	3	5
doctor	are you vomiting	are you vomiting	do you vomit?	¿vomita?	AC	2	3	5
patient	no vomito	no mucho	vomito mucho		NA_BR	2	3	5
doctor	do you have diarrhea	do you have nausea	do you feel nauseous?	¿tiene náuseas?	NA_BR	2	3	5
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC	2	3	6
patient	sí	sí	sí	yes	AC	2	3	6
doctor	for how many days have you had a sore throat	how many recently cause sore throat			NA_BR	2	3	6
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	3	6
patient	no sé	náuseas			NA_BR	2	3	6
patient	no sé	no once			NA_BR	2	3	6
patient	no sé	no sé	no sé	i don't know	AC_help	2	3	6
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	2	3	6
patient	sí	sí	sí	yes	AC	2	3	6
doctor	do you have a runny nose	another runny nose			NA_BR	2	3	6

doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC_help	2	3	6
patient	no	sólo			NA_BR	2	3	6
patient	no	no	no	no	AC	2	3	6
doctor	have you had a fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	2	3	6
patient	sí mucho fiebre	sí mucha fiebre	he tenido mucha fiebre	i had a severe fever	AC	2	3	6
doctor	have you been in contact with anyone who has had a strep throat	have you been in contact with anyone with a strep throat		¿ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo?	NA_BR	2	3	6
doctor	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC_help	2	3	6
patient	sí	medicina			NA_BR	2	3	6
patient	sí	sí	sí	yes	AC	2	3	6
doctor	do you have a rash	do you have rash	do you have a rash?	¿tiene una erupción cutánea?	AC	2	3	6
patient	no	no	no	no	AC	2	3	6
doctor	do you have a stomachache	constricting	do you have a constricting rash?	¿tiene una erupción cutánea constrictiva?	NA_BR	2	3	6
doctor	do you have a stomachache	another augmentin			NA_BR	2	3	6
doctor	do you have a stomachache	constricting	do you have a constricting rash?	¿tiene una erupción cutánea constrictiva?	NA_BR	2	3	6
doctor	do you have a stomachache	do you have a stomachache	do you have a stomachache?	¿el estómago le duele?	AC_help	2	3	6
patient	no	no	no	no	AC	2	3	6
doctor	do you have fever	your fever			NA_BR	2	4	8
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC_help	2	4	8
patient	no tengo fiebre	no tengo fiebre	tengo fiebre	no i have fever	NA_BR	2	4	8
patient	sí tengo fiebre desde hace algunos días	sí tengo fiebre desde algunos días	tengo fiebre desde hace algunos días	yes i have had fever for several days	NA_help	2	4	8
patient	sí tengo fiebre	sí tengo fiebre	tengo fiebre	yes i have fever	NA_help	2	4	8

patient	sí tengo una erupción cutánea muy fuerte	sí tengo una erupción cutánea muy fuerte	tengo muy fuerte una erupción cutánea	yes i have a very severe rash	NA_help	2	4	8
patient	sí tengo fiebre desde algunos días	sí tengo fiebre desde algunos días	tengo fiebre desde hace algunos días	yes i have had fever for several days	NA_help	2	4	8
patient	sí tengo fiebre desde hace algunos días	sí tengo fiebre desde hace algunos días	tengo fiebre desde hace algunos días	yes i have had fever for several days	NA_help	2	4	8
patient	sí tengo fiebre	sí tengo fiebre	tengo fiebre	yes i have fever	AC	2	4	8
doctor	how are you	how are you			NA_NT	2	4	8
doctor	which antibiotics are you allergic to	which antibiotics are you allergic to	what antibiotics are you allergic to?	¿qué antibiótico le da alergia?	AC_help	2	4	8
patient	tengo tos desde hace menos de dos semanas	tengo tos desde hace menos de tres semanas	tengo tos desde hace menos de dos semanas	i have coughed for less than two weeks	NA_help	2	4	8
patient	tengo tos desde hace menos de tres semanas		tengo tos desde hace menos de tres semanas	i have coughed for less than three weeks	NA_help	2	4	8
patient	la penecilina me da alergia	la penicilina me da alergias	la penicilina me da alergias	i am allergic to penicillin	AC	2	4	8
doctor	how high is the fever	high fatigue			NA_BR	2	4	8
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC_help	2	4	8
patient	tengo treinta y ocho grados	tengo treinta y ocho grados	tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	AC	2	4	8
patient			tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	yes	2	4	8
doctor	do you have a sore throat	your sore throat			NA_BR	2	4	8
doctor	is your throat sore	is your throat swollen			NA_BR	2	4	8
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	2	4	8
patient	me duele en la garganta	me duele en la garganta	me duele la garganta	i have a sore throat	AC	2	4	8
doctor	how long	hello	hello?	¿hola?	NA_BR	2	4	8
doctor	how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	4	8
patient	me duele desde hace dos días	me duele desde hace dos días	me duele desde hace dos días	i have experienced the pain for two days	AC	2	4	8
doctor	how severe	[rien]			NA_BR	2	4	8
doctor	does it hurt a lot	does it ever last	does the pain ever last?	¿el dolor ha ya durado?	NA_BR	2	4	8

doctor	how severe	[rien]			NA_BR	2	4	8
doctor	is the pain less severe	is the pain less severe	is the pain better?	¿el dolor disminuye?	AC_help	2	4	8
patient	no	no	no	no	NA_BR	2	4	8
patient	no me duele más cada día	no me duele martres cada día			NA_BR	2	4	8
patient	me duele la garganta cada día	me duele la garganta cada día	me duele la garganta cada día	i have a sore throat every day	AC	2	4	8
doctor	do you have fever	your fever	is the pain better when you have fever?	¿el dolor está aliviado cuando tiene fiebre?	NA_BR	2	4	8
doctor	when did your fever start	when did your fever start	when did fever start?	¿desde cuándo tiene fiebre?	AC	2	4	8
patient	la fiebre empezo anoche	la fiebre ha empezado noche	fiebre ha empezado la noche	fever started the evening	NA_BR	2	4	8
patient	la fiebre empezo en la noche	la fiebre ha empezado una noche			NA_BR	2	4	8
patient	la fiebre empezo anoche	la fiebre ha empezado la noche	fiebre ha empezado la noche	fever started the evening	NA_BR	2	4	8
patient	desde ayer en al noche	desde ayer la noche			NA_BR	2	4	8
patient	desde ayer en la noche	trece días la noche			NA_BR	2	4	8
patient	desde ayer en la noche	no sé fiebre dura mucho	no sé mucha fiebre dura		NA_BR	2	4	8
patient	sí en la noche	sí en la noche	sí en la noche		NA_help	2	4	8
patient	sí tengo náuseas en la noche	sí tengo náuseas en la noche	tengo náuseas en la noche	yes i feel nauseous in the evening	NA_help	2	4	8
patient	sí en la noche	sí en la noche	sí en la noche		NA_help	2	4	8
patient	sí tengo mucho moco en la tarde	sí tengo mucho moco en la tarde	tengo mucho moco en la noche	yes i have a runny nose in the evening	NA_help	2	4	8
patient	sí en la noche	sí en la noche	sí en la noche		NA_help	2	4	8
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	4	8
patient	sí tengo mucha tos	sí tengo mucho tos	tengo mucha tos	yes i have a severe cough	AC	2	4	8
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	2	4	8
patient	sí tengo la nariz tapada	sí tengo la nariz	tengo la nariz tapada	yes i feel congested	AC	2	4	8
doctor	does anyone have a sore throat	does anyone have a sore throat	does anyone have a sore throat?	¿alguien tiene el dolor de garganta?	AC	2	4	8
patient	mis dos hijos	sí doce grados			NA_BR	2	4	8
patient	mis dos hijos	muy positivos			NA_BR	2	4	8

patient	mi familia tiene dolor de garganta	mucho día que tiene dolor de garganta	alguien que tiene el dolor de garganta tiene mucho el dolor de garganta días	someone days has a severe sore throat who have a sore throat	NA_BR	2	4	8
patient	mi hermana	normalmente	alguien tiene normalmente el dolor de garganta	someone usually has a sore throat	NA_BR	2	4	8
patient	mi hermana tiene dolor de garganta	na que tiene dolor de garganta	alguien que tiene el dolor de garganta tiene el dolor de garganta semanas	someone weeks has a sore throat who have a sore throat	NA_BR	2	4	8
patient	headache	tres			NA_IE	2	4	8
patient	do you have a headache	últimamente	alguien tiene recientemente el dolor de garganta	someone has a recent sore throat	NA_IE	2	4	8
doctor	do you have a headache	do you have any			NA_BR	2	4	8
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC	2	4	8
patient	sí me duele mucho la cabeza	sí me duele mucho la cabeza	me duele mucho la cabeza	yes i experience a severe pain in the head	AC	2	4	8
doctor	are you taking medecines	when taking medicine	do you have the headaches when you take medicine?	¿le duele la cabeza cuando toma medicina?	NA_BR	2	4	8
doctor	have you taken any medicine for your pain	have you taken any medicine for your	have you taken medicine for the pain?	¿ha tomado medicina para el dolor?	AC_help	2	4	8
patient	sí tomé medicina anoche	sí tomo medicinas la noche	tomo medicina la noche	yes i take medication the evening	AC	2	4	8
doctor	does your throat hurt	does your throat hurt	do you have a sore throat?	¿le duele la garganta?	AC	2	4	7
patient	sí me duele la garganta	sí me duele la	me duele la garganta	yes i have a sore throat	AC	2	4	7
doctor	how long	hello	hello?	¿hola?	NA_BR	2	4	7
doctor	how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	4	7
patient	más de una semana	más de una semana	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	AC	2	4	7

doctor	how severe is it	how severe is it		¿cuál es la intensidad del dolor?	NA	2	4	7
doctor	what kind of pain is it	what kind of pain is it	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	4	7
patient	me duele más o menos	me duele más a menudo	me duele más a menudo		NA_BR	2	4	7
patient	me duele un poco	me duele más a menudo	me duele un poco	i experience a bit of pain	AC_help	2	4	7
doctor	do you have cough	your cough			NA_BR	2	4	7
doctor	is there a cough	is there cough	do you have a cough?	¿tiene tos?	AC	2	4	7
patient	sí tengo tos	sí tengo tos	tengo tos	yes i cough	AC	2	4	7
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	2	4	7
patient	sí estoy empenzando estar congestionada	sí desde hace semanas algunas semanas	tengo la nariz tapada desde hace algunas semanas	i have felt congested for several weeks	NA_BR	2	4	7
patient	sí tengo congestion	sí tengo operación			NA_BR	2	4	7
patient	sí	sí	sí	yes	AC	2	4	7
doctor	do you have fever	your fever			NA_BR	2	4	7
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC_help	2	4	7
patient	no	no	no	no	AC	2	4	7
doctor	do you have headache	your headache	do you have the headaches?	¿le duele la cabeza?	NA_BR	2	4	7
patient	sí me duele la cabeza	sí me duele la cabeza	me duele la cabeza	yes i experience a pain in the head	AC	2	4	7
doctor	do you have a sore throat	do you have them	do you experience the pain?	¿tiene el dolor?	NA_BR	2	3	7
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC_help	2	3	7
patient	sí me duele la garganta	sí me duele la	me duele la garganta	yes i have a sore throat	AC	2	3	7
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	3	7
patient	sí de treinta y ocho con cinco	sí más de treinta y ocho horas tiempo			NA_BR	2	3	7
patient	sí tengo treinta y ocho con cinco grados	sí tengo treinta hecho con cinco grados			NA_BR	2	3	7
patient	sí tengo treinta y ocho grados de fiebre	sí tengo treinta y ocho grados de fiebre	tengo treinta y ocho grados de fiebre	yes my temperature is one hundred one degrees	AC	2	3	7

doctor	do you have swollen glands	do you have swollen glands	do you have swollen glands?	¿tiene los ganglios inflamados?	AC	2	3	7
patient	no no he notado	no lunes pasado	tengo los ganglios inflamados el lunes pasado		NA_BR	2	3	7
patient	no	no	no	no	AC	2	3	7
doctor	have you had contact with anyone with strep recently	how negative is anyone with this recently			NA_BR	2	3	7
doctor	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC	2	3	7
patient	no	no	no	no	AC	2	3	7
doctor	do you have runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	3	7
patient	sí bastante	sí la tarde	tengo mucho moco la noche	i have a runny nose the evening	NA_BR	2	3	7
patient	sí mucho	sí mucho	tengo mucho moco	i have a runny nose	AC	2	3	7
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC	2	3	8
patient	sí me duele desde hace más de una semana	sí me duele desde hace más de una semana	me duele desde hace más de una semana	yes i have experienced the pain for more than one week	AC	2	3	8
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	3	8
patient	sí tengo fiebre	sí tengo fiebre	tengo fiebre	yes i have fever	AC	2	3	8
doctor	have you been in contact with anyone who has had a strep throat recently	have you been in contact with anyone who has strep throat recently	have you been in contact with anyone recently who have strep throat?	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC	2	3	8
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	3	8
doctor	do you have swollen glands	do you have swollen glands	do you have swollen glands?	¿tiene los ganglios inflamados?	AC	2	3	8
patient	no lo sé	no doce			NA_BR	2	3	8

patient	no sé	doce			NA_BR	2	3	8
patient	no sé	no sé	no sé	i don't know	AC	2	3	8
doctor	have you had a runny nose	have you had runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	2	3	8
patient	no tengo congestion	no tengo infección	tengo el dolor de garganta	no i have my sore throat	NA_BR	2	3	8
patient	tengo congestion	tengo infección	tengo el dolor de garganta	i have my sore throat	NA_BR	2	3	8
patient	no	no	no	no	AC	2	3	8
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	2	3	8
patient	sí me duele	sí me duele	me duele	yes i experience a pain	AC	2	3	8
doctor	do you have a rash	\ve you had a rash	have you had a rash?	¿ha tenido una erupción cutánea?	AC	2	3	8
patient	si tengo una erupción	sí tengo el pecho	tengo el pecho	yes i have the chest	NA_BR	2	3	8
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	2	6	10
patient	me duele la cabeza	me duele la cabeza	me duele la cabeza	i experience a pain in the head	AC	2	6	10
doctor	how long have you had a sore throat	how long \ve you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	6	10
patient	hace tres días	hace tres días	me duele la garganta hace tres días	i have had a sore throat for three days	AC	2	6	10
doctor	have you had a fever	ve you had fever	have you had fever?	¿ha tenido fiebre?	AC	2	6	10
patient	sí	sí	sí	yes	AC	2	6	10
doctor	do you have a runny nose	your runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	2	6	10
patient	no	no	no	no	AC	2	6	10
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	6	10
doctor	was the diagnosis a strep throat	was the diagnosis a strep throat	was the diagnosis strep throat?	¿le han diagnosticado una infección de la garganta por estreptococo?	AC_help	2	6	10
doctor	is the pain less severe	is the pain less severe	is the pain better?	¿el dolor disminuye?	AC_help	2	6	10
doctor	is the pain less frequent	is the pain less frequent	is the frequency of the pain decreasing?	¿le duele menos a menudo?	AC_help	2	6	10

doctor	was the diagnosis a strep throat	was the diagnosis a strep throat	was the diagnosis strep throat?	¿le han diagnosticado una infección de la garganta por estreptococo?	AC_help	2	6	10
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	6	10
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC_help	2	6	10
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	6	10
doctor	is the pain getting better	is the pain getting better	is the pain better?	¿el dolor disminuye?	AC_help	2	6	10
doctor	has anyone else been sick around you	is anyone often sick around you			NA_BR	2	6	10
doctor	has anyone else been sick around you	is anyone often sick around you			NA_BR	2	6	10
doctor	is anyone around you sick	is anyone around you sick			NA_BR	2	6	10
doctor	is anybody around you suffering from a sore throat	is anybody around you suffering from a sore throat	does anyone around you have your sore throat?	¿alguien en su entorno tiene el dolor de garganta?	AC	2	6	10
patient	sí	sí	sí	yes	AC	2	6	10
doctor	are you allergic to any antibiotics	are you allergic to any antibiotics	are you allergic to any antibiotic?	¿tiene alergia a algunos antibióticos?	AC	2	6	10
patient	ninguno	ninguna			NA_NT	2	6	10
patient	ninguno	ninguna			NA_NT	2	6	10
patient	no	no	no	no	AC	2	6	10
doctor	have you seen the doctor this week	have you seen the doctor this week	have you visited the doctor this week?	¿consulta un médico desde hace esta semana?	AC	2	6	10
patient	no	no	no	no	AC	2	6	10
doctor	how intense is the pain	how intense is the pain		¿cuál es la intensidad del dolor?	NA_BR	2	6	10
doctor	how bad is the pain	how bad is the pain			NA_BR	2	6	10
doctor	have you been coughing	does eating cause pains	do you experience the pain when you eat?	¿le duele cuando come?	NA_BR	2	6	10
doctor	are you coughing	are you coughing	do you cough?	¿tiene tos?	AC	2	6	10
patient	no	no	no	no	AC	2	6	10
doctor	do you have any rashes	ny medicine			NA_BR	2	6	10

doctor	do you have a rash	your rash			NA_BR	2	6	10
doctor	do you have a rash	do you have a rash	do you have a rash?	¿tiene una erupción cutánea?	AC	2	6	10
patient	sí tengo una erupción	sí tengo una erupción cutánea	tengo una erupción cutánea	yes i have a rash	AC	2	6	10
doctor	how long have you been sick	alleviating it			NA_BR	2	6	9
doctor	how long have you been sick	how long is the pain	how long do you experience the pain?	¿desde cuándo tiene el dolor?	NA_BR	2	6	9
doctor	how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	6	9
patient	hace cuatro días	hace cuatro días	me duele la garganta hace cuatro días	i have had a sore throat for four days	AC	2	6	9
doctor	have you had a fever as well	any other feverish people			NA_BR	2	6	9
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	6	9
patient	sí	sí	sí	yes	AC	2	6	9
doctor	do you have runny nose	your runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	6	9
patient	no	no	no	no	AC	2	6	9
doctor			do you have a runny nose?	¿tiene mucho moco?	REP	2	6	9
doctor	do you have a cough	how can you have a cough			NA_BR	2	6	9
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC_help	2	6	9
patient	no	no	no	no	AC	2	6	9
doctor	do you have any rashes	is any rash this	do you have a rash?	¿tiene una erupción cutánea?	AC	2	6	9
patient	no	no	no	no	AC	2	6	9
doctor	has anyone been sick around you	when is this around you			NA_BR	2	6	9
doctor	is anyone around you sick	is anyone around you sick			NA_NT	2	6	9
doctor	is anybody around you suffering from a sore throat	is anybody around you suffering from a sore throat	does anyone around you have your sore throat?	¿alguien en su entorno tiene el dolor de garganta?	AC_help	2	6	9
patient	no	no	no	no	AC	2	6	9

doctor	how high was your fever	how high was your fever	what was your temperature?	¿cuántos grados ha tenido?	AC	2	6	9
patient	caliente pero no lo sé	alguien tiene el dolor los meses	alguien tiene el dolor meses	someone experiences the pain months	NA_BR	2	6	9
patient	no lo sé	no doce	he tenido doce grados de fiebre	my temperature was twelve degrees	NA_BR	2	6	9
patient	no sé	no sé	no sé	i don't know	AC	2	6	9
doctor	have you been to the doctor	have you been to the doctor	have you visited the doctor?	¿ha consultado un médico?	AC	2	6	9
patient	sí hace dos días	sí hace los días	he consultado un médico hace días	i visited the doctor days ago	AC	2	6	9
doctor	did you see the doctor two days ago	did you see the doctor two days ago	did you visit the doctor two days ago?	¿ha consultado un médico hace dos días?	AC_help	2	6	9
doctor			did you visit the doctor two days ago?	¿ha consultado un médico hace dos días?	REP	2	6	9
patient	sí	sí	sí	yes	AC	2	6	9
doctor	did they test you then	if they have eaten			NA_BR	2	6	9
doctor	did the doctor do a throat culture test	did the doctor do a throat culture test	did the doctor carry out a throat culture test?	¿le han realizado un cultivo de la garganta?	AC_help	2	6	9
patient	qué quiere decir	de el remedio			NA_BR	2	6	9
patient	qué quiere decir	tengo fiebre desde días	tengo fiebre desde hace días	i have had fever for days	NA_BR	2	6	9
patient	sí	sí	sí	yes	AC	2	6	9
doctor	what were the results of the test	what were the results of the test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC_help	2	6	9
patient	un virus	un viernes			NA_BR	2	6	9
patient	un virus	una nariz un poco tapada			NA_BR	2	6	9
patient	un virus	hola ningún			NA_BR	2	6	9
patient	infección viral	infección viral	el resultado del examen ha sido una infección viral	the results of the test were a viral sore throat	AC	2	6	9
doctor	are you allergic to any antibiotics	were you allergic to any antibiotic	were you allergic to any antibiotic?	¿ha tenido alergia a algunos antibióticos?	AC	2	6	9
patient	sí a la penicilina	sí a la penicilina	la penicilina me ha dado alergias	i was allergic to penicillin	AC	2	6	9

doctor	how are you doing today	how weak is day			NA_BR	2	5	9
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC	2	5	9
patient	sí	sí	sí	yes	AC	2	5	9
doctor	since when	this month	do you have a sore throat this month?	¿le duele la garganta este mes?	NA_BR	2	5	9
doctor	when did your sore throat start	when did your sore throat start	when did your sore throat start?	¿cuándo ha empezado el dolor de garganta?	AC_help	2	5	9
patient	hace cuatro días	hace cuatro días	el dolor de garganta ha empezado hace cuatro días	my sore throat started four days ago	AC	2	5	9
doctor	is the pain accompanied by fever	is the pain accompanied by fever	do you have fever when you experience the pain?	¿tiene fiebre cuando tiene el dolor?	AC_help	2	5	9
patient	no	no	no	no	AC	2	5	9
doctor	are you able do eat	anybody			NA_BR	2	5	9
doctor	do you have any congestion	getting congestion			NA_BR	2	5	9
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC_help	2	5	9
patient	no	no	no	no	AC	2	5	9
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	2	5	9
patient	sí	sí	sí	yes	AC	2	5	9
doctor	for how many days	how many days			NA_BR	2	5	9
doctor	how many days of headaches	how many lately made them steady			NA_BR	2	5	9
doctor	how long have you had a headache	how long have you had a headache	how long have you had the headaches?	¿desde cuándo le duele la cabeza?	AC_help	2	5	9
patient	hace cuatro días	hace cuatro días	me duele la cabeza hace cuatro días	i have experienced the pain in the head for four days	AC	2	5	9
doctor	have you been in contact with an ill person	have you been in contact with any person	have you been in contact with the person?	¿ha estado en contacto con esa persona?	NA_BR	2	5	9
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	2	5	9

patient	no	no	no	no	AC	2	5	9
doctor	how bad is the pain	how many is the pain			NA_BR	2	5	9
doctor	what is the character of the pain	what is the character of the pain	what is the character of the pain?	¿cómo es el dolor?	AC_help	2	5	9
patient	fuerte	fuerte	el dolor es severo	the pain is severe	AC	2	5	9
doctor	are you taking anything for the pain	are you taking medicine for the pain	do you take medicine for the pain?	¿toma medicina para el dolor?	AC	2	5	9
patient	no	no	no	no	AC	2	5	9
doctor	have you had a strep test done	have you had this strep test months	have you had a rapid strep test for several months?	¿le realizan la prueba rápida por estreptococo desde hace algunos meses?	NA_BR	2	5	9
doctor	was a rapid strep test done	was a rapid strep test done	was a rapid strep test carried out?	¿le han realizado la prueba rápida por estreptococo?	AC_help	2	5	9
patient	sí	sí	sí	yes	AC	2	5	9
doctor	was it positive	was this positive			NA_BR	2	5	9
doctor	what were the results of this testing	what were the results of this testing	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC_help	2	5	9
patient	todavía no hay resultados	hola día jueves			NA_BR	2	5	9
patient	no hay resultados	no algunas grados			NA_BR	2	5	9
patient	todavía no	hola día			NA_BR	2	5	9
patient	no	no	no	no	AC	2	5	9
doctor	do you have any rashes	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	NA_BR	2	5	9
doctor	do you have a rash	do you have a rash	do you have a rash?	¿tiene una erupción cutánea?	AC_help	2	5	9
patient	no	no	no	no	AC	2	5	9
doctor	have you taken any medication	have you taken any medication	have you taken medication?	¿ha tomado medicina?	AC	2	5	9
patient	no	no	no	no	AC	2	5	9
doctor	is the pain better	is the pain better	is the pain better?	¿el dolor disminuye?	AC	2	5	9
patient	no	no	no	no	AC	2	5	9
doctor	when was the strep test done	was this strep test done	was a rapid strep test carried out?	¿le han realizado la prueba rápida por estreptococo?	NA_BR	2	5	9

doctor	when did you do the test	when did they do the test	when did the doctor carry out the test?		NA_BR	2	5	9
doctor	did the doctor do a throat culture test	did the doctor do a throat culture test	did the doctor carry out a throat culture test?	¿le han realizado un cultivo de la garganta?	NA_help	2	5	9
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC_help	2	5	9
patient	sí	sí	sí	yes	AC	2	5	9
doctor	when	sweating	did the doctor carry out a rapid strep test when you sweat?	¿le han realizado la prueba rápida por estreptococo cuando suda?	NA_BR	2	5	9
doctor	when did you do the test	when did you do the test	when did you carry out the test?		NA_NT	2	5	9
doctor	have you ever had strep throat	have they ever had strep throat			NA_NT	2	5	9
doctor	do you have any congestion	having congestion	did the doctor carry out a rapid strep test when you feel congested?	¿le han realizado la prueba rápida por estreptococo cuando tiene la nariz tapada?	NA_BR	2	5	9
doctor	how long have you had congestion	how long have you had congestion	how long have you had congestion	¿desde cuándo tiene la nariz tapada?	AC_help	2	5	9
doctor	do you have a are you congested	do you have high congestion			NA_BR	2	5	9
doctor	are you congested	are you congested	do you feel congested?	¿tiene la nariz tapada?	AC	2	5	9
patient	no	no	no	no	AC	2	5	9
doctor	are you in pain today	are you moving today	do you move since today?	¿mueve hoy?	NA_BR	2	5	10
doctor	do you have pain	the testing			NA_BR	2	5	10
doctor	is it hurting	constricting			NA_BR	2	5	10
doctor	are you in pain	are your pains pain	do you experience the pain?	¿tiene el dolor?	NA_BR	2	5	10
doctor	do you have a sore throat	\ve you had a sore throat	have you had a sore throat?	¿le ha dolido la garganta?	NA_BR	2	5	10
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	5	10

patient	no estoy seguro	positivo			NA_BR	2	5	10
patient	no sé	doce			NA_BR	2	5	10
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	5	10
doctor	days	days	have you had a sore throat for several days?	¿le duele la garganta desde hace algunos días?	NA_BR	2	5	10
doctor	have you had a sore throat for several days	have you had a sore throat for several	have you had a sore throat for several days?	¿le duele la garganta desde hace algunos días?	AC_help	2	5	10
patient	no estoy seguro	no seis grados			NA_BR	2	5	10
patient	dos días	dos días	me duele la garganta desde hace dos días	i have had a sore throat for two days	AC	2	5	10
doctor	do you have fever	can you have fever	do you have fever?	¿tiene fiebre?	NA_BR	2	5	10
doctor	do you have fever	the mild fever			NA_BR	2	5	10
doctor	do you have fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	5	10
doctor			do you have fever?	¿tiene fiebre?	REP	2	5	10
patient	sí he tenido mucha fiebre	siete días	tengo fiebre siete días	i have fever seven days	NA_BR	2	5	10
patient	sí he tenido fiebre	siete grados a fiebre	tengo siete grados de fiebre	my temperature is seven degrees	NA_BR	2	5	10
patient	sí tengo fiebre	sí tengo fiebre	tengo fiebre	yes i have fever	AC_help	2	5	10
doctor	do you have any cough	getting the cough			NA_BR	2	5	10
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC_help	2	5	10
patient	sí un poco	sí un poco	tengo un poco de tos	i have a bit of a cough	AC	2	5	10
doctor	runny nose	runny nose	do you have a runny nose?	¿tiene mucho moco?	NA_BR	2	5	10
doctor	is your nose running	is your nose running	do you have a runny nose?	¿tiene mucho moco?	NA_help	2	5	10
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC_help	2	5	10
patient	también	alguien			NA_BR	2	5	10
patient	también	alguien			NA_BR	2	5	10
patient	sí mucho moco	sí mucho moco	tengo mucho moco	i have a runny nose	AC	2	5	10
doctor	how many days of runny nose	how many gave them runny nose			NA_BR	2	5	10
doctor	how long have you had a runny nose	how long have you had runny nose	how long have you had a runny nose?	¿desde cuándo tiene mucho moco?	AC	2	5	10
patient	no estoy seguro	no sé dolor			NA_BR	2	5	10
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	5	10

doctor	have you been in contact with a sick person	have you been in contact with sick person	have you been in contact with sick people?	¿ha estado en contacto con una persona enferma?	AC	2	5	10
patient	sí unos días atrás	[rien]			NA_BR	2	5	10
patient	sí hace dos semanas he estado en contacto con alguien que tiene una infección de la garganta	si hace dos semanas he estado en contacto con alguien que tiene una infección de la garganta	he estado en contacto con alguien que tiene el dolor de garganta hace dos semanas	yes i was in contact with someone two weeks ago who had a sore throat	AC_help	2	5	10
doctor	did you have strep throat	did you have strep throat	did you have strep throat?	¿ha tenido una infección de la garganta por estreptococo?	NA	2	5	10
doctor	when have you been in contact with a person who has strep throat	when have you been in contact with a person who has strep throat	when have you been in contact with the person who has strep throat?	¿cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo?	AC_help	2	5	10
doctor			when have you been in contact with the person who has strep throat?	¿cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo?	REP	2	5	10
patient	sí he estado en contacto	[rien]			NA_BR	2	5	10
patient	sí tenía infección	sí tiene infección	el médico tiene el dolor de garganta	yes the doctor has my sore throat	NA_BR	2	5	10
patient	sí tengo	severo			NA_BR	2	5	10
patient	sí	sí	sí	yes	AC	2	5	10
doctor	have you taken any medication	have you taken any medication	have you taken medication?	¿ha tomado medicina?	AC	2	5	10
patient	no	no	no	no	AC	2	5	10
doctor	have you seen a doctor lately	have you seen a doctor lately	have you recently visited the doctor?	¿ha consultado recientemente un médico?	AC	2	5	10
patient	sí	sí	sí	yes	AC	2	5	10
doctor	did he do a strep test on you	did seeing you help at this morning			NA_BR	2	5	10

doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC_help	2	5	10
patient	sí también	examen	me han realizado el examen	a test was carried out	AC	2	5	10
doctor	was it positive for streptococcus	was positive result positive	were the results positive?	¿el resultado ha sido positivo?	AC	2	5	10
patient	sí fue positivo	sí fue positivo nueve días	el dolor es positivo desde hace nueve días		NA_BR	2	5	10
patient	sí el resultado ha sido positivo	sí el resultado ha sido positivo	el resultado ha sido positivo	yes the results were positive	AC_help	2	5	10
doctor	when was the test done	was the test done	was the test carried out?	¿le han realizado un examen?	NA_BR	2	5	10
doctor	when did you do the strep test	when did you do the strep test	when did you carry out a rapid strep test?		NA_NT	2	5	10
doctor	when did you do the strep test	when did you do this strep test	when did you carry out a rapid strep test?		NA_NT	2	5	10
doctor	how long have you had a sore throat	how long did you have this			NA_BR	2	6	12
doctor	how long have you had your sore throat	how long have you had your sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	6	12
patient	desde hace más de tres días	desde hace más de tres días sabado			NA_BR	2	6	12
patient	tengo tos desde hace más de tres días	tengo tos desde hace más de tres días	tengo tos desde hace más de tres días	i have coughed for more than three days	AC_help	2	6	12
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	6	12
patient	sí	sí	sí	yes	AC	2	6	12
doctor	how high is the fever	how high is this fever	what is your temperature?	¿cuántos grados tiene?	AC	2	6	12
patient	treinta y ocho grados	treinta y ocho grados	tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	AC	2	6	12
doctor	do you have a runny nose	your runny nose			NA_BR	2	6	12
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC_help	2	6	12
patient	no	no	no	no	AC	2	6	12

doctor	do you have a rash	your rash	do you have a rash?	¿tiene una erupción cutánea?	AC	2	6	12
patient	no	no	no	no	AC	2	6	12
doctor	has anyone around you been sick	\s anyone around you been sick			NA_NT	2	6	12
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	2	6	12
patient			he estado en contacto con alguien que tiene el dolor de garganta un día meses		AC	2	6	12
patient			he estado en contacto con alguien que tiene el dolor de garganta un día meses		AC	2	6	12
doctor			have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	REP	2	6	12
patient	sí he estado en contacto con una persona que esta enferma de la garganta	[rien]			NA_BR	2	6	12
patient	sí he estado en contacto con alguien que ha tenido dolor de garganta	si he estado en contacto con alguien que tiene dolor de garganta	he estado en contacto con alguien que tiene el dolor de garganta	yes i was in contact with someone who had a sore throat	AC	2	6	12
patient	sí he estado en contacto con alguien que ha tenido dolor de garganta	si he estado en contacto con alguien que tiene dolor agudo de garganta		yes i was in contact with someone who had a sharp sore throat	NA_BR	2	6	12
patient	sí he estado en contacto con alguien que tiene una infección de la garganta	si he estado en contacto con alguien que tiene una infección de la	he estado en contacto con alguien que tiene el dolor de garganta	yes i was in contact with someone who had a sore throat	AC_help	2	6	12
doctor	have you been to visit a doctor	seeing this dizzy doctor			NA_BR	2	6	12

doctor	have you visited the doctor	have you visited the doctor	have you visited the doctor?	¿ha consultado un médico?	AC	2	6	12
patient	no	no	no	no	AC	2	6	12
doctor	are you allergic to antibiotics	were you allergic to antibiotic	were you allergic to antibiotics?	¿los antibióticos le han dado alergia?	NA_BR	2	6	12
doctor	are you allergic to any antibiotic	are you allergic to any antibiotic	are you allergic to any antibiotic?	¿tiene alergia a algunos antibióticos?	AC_help	2	6	12
patient	no	no	no	no	AC	2	6	12
doctor	have you had any vomiting or diarrhea	is anybody tired	does anyone feel tired?	¿alguien está cansado?	NA_BR	2	6	12
doctor	have you vomited	have you vomited	have you vomited?	¿ha vomitado?	AC	2	6	12
patient	no	no	no	no	AC	2	6	12
doctor	diarrhea	area			NA_BR	2	6	12
doctor	do you have diarrhea	your area			NA_BR	2	6	12
doctor	have you had diarrhea	have you had nausea early			NA_BR	2	6	12
doctor	do you have sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC	2	6	12
patient	sí	sí	sí	yes	AC	2	6	12
doctor	for how many days	four hundred days	do you have a sore throat for four hundred days?		NA_BR	2	6	12
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	6	12
patient	más de una semana	más de una semana	me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	AC	2	6	12
doctor	do you have any fever with it	do you have any fever	do you have fever?	¿tiene fiebre?	AC	2	6	12
patient	no	no	no	no	AC	2	6	12
doctor	do you have a runny nose	do you have a runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	6	12
patient	sí	sí	sí	yes	AC	2	6	12
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	6	12
patient	sí un poco	[rien]			NA_BR	2	6	12
patient	sí un poco	sí un poco	tengo un poco de tos	i have a bit of a cough	AC	2	6	12

doctor	have you been in contact with someone with a sore throat	have you been in contact with someone with a sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA	2	6	12
doctor	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who have a sore throat?	¿ha estado recientemente en contacto con alguien que tiene el dolor de garganta?	AC_help	2	6	12
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	6	12
doctor	have you seen a doctor recently	have you seen a doctor recently	have you recently visited the doctor?	¿ha consultado recientemente un médico?	AC	2	6	12
patient	sí	sí	sí	yes	AC	2	6	12
doctor	did he do any lab tests	did you do that test	did you carry out the test?		NA_BR	2	6	12
doctor	did he do a test	did you do a test	did you carry out the test?		NA_BR	2	6	12
doctor	did he do a strep test	did you do a strep test	did you carry out a rapid strep test?		NA_BR	2	6	12
doctor	did he do a strep test	did you do a strep test	did you carry out a rapid strep test?		NA_BR	2	6	12
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC_help	2	6	12
patient	no sé	no sé	no sé	i don't know	AC	2	6	12
doctor	are you feeling better today	are you seeing doctor today	do you visit the doctor since today?	¿consulta hoy un médico?	NA_BR	2	6	12
doctor	is the pain better today	is the pain bad your today	is the pain severe since today?	¿el dolor es hoy severo?	NA_BR	2	6	12
doctor	is the pain less severe	is the pain less severe	is the pain better?	¿el dolor disminuye?	AC	2	6	12
patient	sí un poco	sí un poco	el dolor mejora un poco	a bit of pain is decreasing	AC	2	6	12
doctor	where is the pain	when is the pain	where is the pain?	¿dónde le duele?	NA_BR	2	6	12
doctor		where is the pain	where is the pain?	¿dónde le duele?	AC_help	2	6	12
patient	tengo tos	tengo tos	tengo tos	i cough	AC	2	6	11
doctor	how long have you had a sore throat	how long have you had a sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	6	11
patient	mas de un mes	[rien]			NA_BR	2	6	11

patient	tengo tos desde hace un mes	tengo tos desde hace un mes	tengo tos desde hace un mes	i have coughed for one month	AC	2	6	11
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	6	11
patient	sí	sí	sí	yes	AC	2	6	11
doctor	how high is the fever	how high is the fever	what is your temperature?	¿cuántos grados tiene?	AC	2	6	11
doctor	how long have you had a sore throat	how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	6	11
patient	más de un mes	más de un mes	me duele la garganta desde hace más de un mes	i have had a sore throat for more than one month	AC	2	6	11
doctor	do you have a fever	having fever			NA_BR	2	6	11
doctor	do you have a fever	do you have a fever	do you have fever?	¿tiene fiebre?	AC	2	6	11
patient	sí	sí	sí	yes	AC	2	6	11
doctor	how high is the temperature	how high is a temperature	what is your temperature?	¿cuántos grados tiene?	AC	2	6	11
patient	treinta y ocho grados	treinta y ocho grados	tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	AC	2	6	11
doctor	have you been coughing	relieving cough			NA_BR	2	6	11
doctor	do you have a cough	are you ever	do you ever cough?	¿ha ya tenido tos?	NA_BR	2	6	11
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC_help	2	6	11
patient	sí	sí	sí	yes	AC	2	6	11
doctor	do you have a runny nose	your runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	6	11
patient	sí	sí	sí	yes	AC	2	6	11
doctor	do you have a headache	yesterday	do you have a runny nose since yesterday?	¿tiene mucho moco ayer?	NA_BR	2	6	11
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	NA_BR	2	6	11
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC_help	2	6	11
patient	sí	sí	sí	yes	AC	2	6	11
doctor	have you visited a doctor	have you visited the doctor	have you visited the doctor?	¿ha consultado un médico?	AC	2	6	11
patient	sí	[rien]			NA_BR	2	6	11

patient	sí hace una semana	hace una semana	he consultado un médico hace una semana	i visited the doctor one week ago	AC	2	6	11
doctor	did they do a throat test throat culture test	did they do a strep test for cold seconds			NA_BR	2	6	11
doctor	did the doctor do a throat culture test	did the doctor do a throat culture test	did the doctor carry out a throat culture test?	¿le han realizado un cultivo de la garganta?	AC_help	2	6	11
patient	no	no	no	no	AC	2	6	11
doctor	are you allergic to any antibiotics	were you allergic to any antibiotic	were you allergic to any antibiotic?	¿ha tenido alergia a algunos antibióticos?	AC	2	6	11
patient	no	no	no	no	AC	2	6	11
doctor	do you have a sore throat today	having this testing			NA_BR	2	5	11
doctor	do you have a sore throat	this sore throat			NA_BR	2	5	11
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC_help	2	5	11
patient	sí mucho	sí mucho	me duele mucho la garganta	i have a severe sore throat	AC	2	5	11
doctor	since when	when sweating	do you have a sore throat when you sweat?	¿le duele la garganta cuando suda?	NA_BR	2	5	11
doctor	when did it start	when this starts			NA_BR	2	5	11
doctor	when did your sore throat start	when did your sore throat start	when did your sore throat start?	¿cuándo ha empezado el dolor de garganta?	AC_help	2	5	11
patient	la semana pasada	la semana pasada	el dolor de garganta ha empezado la semana pasada	my sore throat started during the last week	AC	2	5	11
doctor	do you have fever with it	is fever intense	is fever severe?	¿la fiebre es severa?	NA_BR	2	5	11
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	5	11
patient	sí	[rien]			NA_BR	2	5	11
patient	sí	sí	sí	yes	AC	2	5	11
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC	2	5	11
patient	treinta y ocho	treinta y ocho	tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	AC	2	5	11
doctor	have you been in contact with a sick person	have you been in contact with the sick person	have you been in contact with sick people?	¿ha estado en contacto con una persona enferma?	AC	2	5	11

patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	5	11
doctor	have you seen a doctor for this sore throat	have you seen a doctor for this sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC	2	5	11
patient	hace tres días	hace tres días	he consultado un médico para el dolor de garganta hace tres días	i visited the doctor for my sore throat three days ago	AC	2	5	11
doctor	did he do a strep test	did you do a strep test	did you carry out a rapid strep test?		NA_BR	2	5	11
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC_help	2	5	11
patient	sí hace dos días	sí hace dos días	me han realizado una prueba rápida por estreptococo hace dos días	a rapid strep test was carried out two days ago	AC_help	2	5	11
doctor	was it positive	was that positive			NA_BR	2	5	11
doctor	what were the results of the test	what were the results of the test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC_help	2	5	11
patient	el resultado fue positivo	el resultado fue positivo	el resultado ha sido positivo	the results were positive	AC	2	5	11
doctor	did you take antibiotics	did you take	did you take antibiotics?	¿ha tomado antibióticos?	AC	2	5	11
patient	ayer	ayer	he tomado ayer los antibióticos	i took antibiotics yesterday	AC	2	5	11
doctor	is your throat hurting you today	is it constricting today	is the pain constricting since today?	¿el dolor es hoy constrictivo?	NA_BR	2	5	12
doctor	do you have a sore throat	do you have a sore throat	do you have a sore throat?	¿le duele la garganta?	AC	2	5	12
patient	sí un poco	sí un poco	me duele un poco la garganta	i have a bit of a sore throat	AC	2	5	12
doctor	for how many days	[rien]			NA_BR	2	5	12
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	5	12
patient	desde hace dos días	desde hace los días	me duele la garganta desde hace días	i have had a sore throat for days	AC	2	5	12
doctor	do you have any fever	getting fever			NA_BR	2	5	12

doctor	is the pain accompanied by fever	is the pain accompanied by fever	do you have fever when you experience the pain?	¿tiene fiebre cuando tiene el dolor?	NA_help	2	5	12
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	5	12
patient	un poco	un poco	tengo un poco de fiebre	i have a bit of fever	AC	2	5	12
doctor	how high was it	how high was this		¿cuál ha sido la intensidad?	NA_BR	2	5	12
doctor	how high is your fever	how high is your fever	what is your temperature?	¿cuántos grados tiene?	AC_help	2	5	12
patient	no sé	no sé	no sé	i don't know	AC	2	5	12
doctor	do you have a runny nose	do you have runny nose	do you have a runny nose?	¿tiene mucho moco?	AC	2	5	12
patient	no no tengo moco	no tengo moco			NA_BR	2	5	12
patient	no no tengo moco	no horas tengo moco			NA_BR	2	5	12
patient	no tengo moco	no tengo moco			NA_BR	2	5	12
patient	no	no	no	no	AC	2	5	12
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	5	12
patient	un poco	un poco	tengo un poco de tos	i have a bit of a cough	AC	2	5	12
doctor	have you been in contact with a person with a sore throat	have you been in contact with a person this onset			NA_BR	2	5	12
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	2	5	12
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC_help	2	5	12
doctor	do you have any chills	your medications			NA_BR	2	5	12
doctor	have you seen a doctor	have you seen the doctor	have you visited the doctor?	¿ha consultado un médico?	AC	2	5	12
patient	sí	sí	sí	yes	AC	2	5	12
doctor	did he do any tests	do you do a test	do you carry out the test?		NA_BR	2	5	12
doctor	did he do a test	did you do a test	did you carry out the test?		NA_BR	2	5	12
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC_help	2	5	12
patient	sí	sí	sí	yes	AC	2	5	12

doctor	was it positive	was that positive			NA_BR	2	5	12
doctor	were the results positive	were the results positive	were the results positive?	¿el resultado ha sido positivo?	AC_help	2	5	12
patient	no	no	no	no	AC	2	5	12
doctor	how are you feeling today	how were a few in today			NA_BR	2	8	14
doctor	where is your pain	where is sharp pain	where is a sharp pain?		NA_BR	2	8	14
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC_help	2	8	14
patient	me duele la garganta me duele la cabeza	me duele la garganta a menudo en la			NA_BR	2	8	14
patient	sí me duele la garganta	sí me duele la	me duele la garganta	yes i have a sore throat	AC_help	2	8	14
doctor	when did the pain start	when did the pain start	when did the pain start?	¿cuándo ha empezado el dolor?	AC	2	8	14
patient	desde hace un mes	desde hace un mes	desde hace un mes		NA_help	2	8	14
patient	desde hace una semana	desde hace una semana			NA_NT	2	8	14
doctor	when did the pain start	when did the pain start	when did the pain start?	¿cuándo ha empezado el dolor?	AC	2	8	14
patient	desde hace un mes	desde hace un mes			NA_NT	2	8	14
patient	me duele la garganta desde hace un mes	me duele la garganta desde hace un mes	me duele la garganta desde hace un mes	i have had a sore throat for one month	AC_help	2	8	14
doctor	when did the pain start	at last year	when did the pain start?	¿cuándo ha empezado el dolor?	NA_BR	2	8	14
doctor	when did the pain start	when did the pain start	when did the pain start?	¿cuándo ha empezado el dolor?	AC	2	8	14
patient	la semana pasada	la semana pasada	el dolor ha empezado la semana pasada	the pain started during the last week	AC	2	8	14
doctor	how severe is the pain	how severe is the pain		¿cuál es la intensidad del dolor?	AC	2	8	14
patient	más o menos me duele	más a menudo			NA_BR	2	8	14
patient	me duele muy a menudo	me duele muy a menudo	me duele muy a menudo	i very often experience a pain	AC_help	2	8	14
doctor	do you have other symptoms	do you have other kind of glands			NA_BR	2	8	14
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	2	8	14

patient	mucho	no mucho	tengo la nariz muy tapada	i feel severely congested	AC	2	8	14
doctor	do you have a fever	do you have fever	do you have fever?	¿tiene fiebre?	AC	2	8	14
patient	sí	dies	tengo diez grados de fiebre	my temperature is ten degrees	NA_BR	2	8	14
patient	tengo diez grados de fiebre	tengo diez grados de fiebre	tengo diez grados de fiebre	my temperature is ten degrees	NA_help	2	8	14
patient	tengo treinta y nueve grados de fiebre	tengo treinta y nueve grados de fiebre	tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	AC_help	2	8	14
doctor	do you have a rash	do you have rash	do you have a rash?	¿tiene una erupción cutánea?	AC	2	8	14
patient	sí	sí	sí	yes	AC	2	8	14
doctor	is the rash	[rien]			NA_BR	2	8	14
doctor	do you have any abdominal pain	\ve you had any other dull muscle			NA_BR	2	8	14
doctor	do you have a stomachache	do you have a stomachache	do you have a stomachache?	¿el estómago le duele?	AC	2	8	14
patient	sí	sí	sí	yes	AC	2	8	14
doctor	have you had a strep test done	have you had a strep test night	have you had a rapid strep test night?	¿le han realizado la prueba rápida por estreptococo la noche?	AC_IE	2	8	14
patient	sí	sí	sí	yes	AC	2	8	14
doctor	what were the results	what were the results	what were the results?	¿cuál ha sido el resultado?	AC	2	8	14
patient	positivo	es positivo once días	el dolor es positivo once días		NA_BR	2	8	14
patient	positivo	positivo	el resultado ha sido positivo	the results were positive	AC	2	8	14
doctor	[rien]	does it occur	do you experience the pain?	¿tiene el dolor?	NA_BR	2	8	13
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC_help	2	8	13
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	8	13
doctor	have you had a fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	2	8	13
patient	la cabeza se siente muy caliente	en la cabeza desde mucho ayer			NA_BR	2	8	13
doctor			have you had fever?	¿ha tenido fiebre?	REP	2	8	13

patient	mi cabeza se siente caliente	ví la persona recientemente	he estado recientemente en contacto con la persona	i recently visited the person	AC_IE	2	8	13
doctor	out of context	did you feel congested	did you feel congested?	¿ha tenido la nariz tapada?	AC	2	8	13
doctor	have you taken any medicine for the sore throat	have you taken any medicine for the sore throat	have you taken medicine for your sore throat?	¿ha tomado medicina para el dolor de garganta?	AC_help	2	8	13
doctor	how long have you had the pain	how long have you had the pains	how long have you experienced the pain?	¿desde cuándo tiene el dolor?	AC	2	8	13
patient	desde hace cuatros días	desde hace cuatro días	me duele desde hace cuatro días	i have experienced the pain for four days	AC	2	8	13
doctor	have you seen any other doctors	have you seen any other doctor	have you visited another doctor?	¿ha consultado un otro médico?	AC	2	8	13
patient	sí	sí pienso dos			NA_BR	2	8	13
doctor			have you visited another doctor?	¿ha consultado un otro médico?	REP	2	8	13
patient	sí	sí	sí	yes	AC	2	8	13
doctor	have you taken any medicines	have you taken any medicine	have you taken medicine?	¿ha tomado medicina?	AC	2	8	13
patient	sí he tomado medicina	he bien tomado medicina			NA_BR	2	8	13
patient	sí he tomado medicina	sí he tomado	he tomado medicina	yes i took medication	AC_help	2	8	13
doctor	what medicine did you take	what medicine did you take	what medicine did you take?	¿qué medicina ha tomado?	AC	2	8	13
patient	tome un jarabe para el dolor	domingo pasado para el dolor			NA_BR	2	8	13
patient	sí tomo medicina para el dolor de garganta	sí tomo medicina para el dolor de garganta	tomo medicina para el dolor de garganta	yes i take medication for my sore throat	AC_help	2	8	13
doctor	have you had any congestion	have you had the testing	have you had the test?	¿le han realizado un examen?	AC_IE	2	8	13
patient	sí me diagnosticaron una infeccción viral	tiene ya dos grados a una infección viral			NA_BR	2	8	13
patient	tengo una infección viral	tengo una infección viral	tengo una infección viral	i have a viral sore throat	AC_help	2	8	13
doctor	where is your pain	where is your pain	where is the pain?	¿dónde le duele?	AC	2	7	13

patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	7	13
doctor	how long have you had strep throat	how long have you had sore throat	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC	2	7	13
patient	me duele desde hace cuatro días	me duele desde hace cuatro días	me duele desde hace cuatro días	i have experienced the pain for four days	AC	2	7	13
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	2	7	13
patient	sí me duele mucho	sí me duele mucho	me duele mucho	yes i experience a severe pain	AC	2	7	13
doctor	do you have pain anywhere else	have pains been mild	has the pain been mild?	¿el dolor ha sido leve?	NA_BR	2	7	13
doctor	do you have	do you have it	do you experience the pain?	¿tiene el dolor?	AC	2	7	13
patient	sí me duele mucho la cabeza	sí me duele mucho la cabeza	me duele mucho la cabeza	yes i experience a severe pain in the head	AC	2	7	13
doctor	do you have	your times			NA_BR	2	7	13
doctor	do you have stomachache	do you have stomachache	do you have a stomachache?	¿el estómago le duele?	AC	2	7	13
patient	un poco	un poco	el estómago me duele un poco	i experience a bit of pain in the stomach	AC	2	7	13
doctor	do you have swollen glands	do you have swollen glands	do you have swollen glands?	¿tiene los ganglios inflamados?	AC	2	7	13
patient	sí mucho	sí mucho	tengo los ganglios muy inflamados	i have severely swollen glands	AC	2	7	13
doctor	have you been around anyone that has had	have you been around anyone to that	have you been in contact with anyone?	¿ha estado en contacto con alguien?	NA_BR	2	7	13
doctor	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who has a sore throat	have you been in contact with anyone recently who have a sore throat?	¿ha estado recientemente en contacto con alguien que tiene el dolor de garganta?	AC_help	2	7	13
patient	creo que no	lo tengo	presento este síntoma	i experience the symptoms	AC_IE	2	7	13
doctor	have you seen a doctor for your sore throat	have you seen the doctor for your sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC	2	7	13

patient	sí fui con una enfermera de práctica médica hace tres días	si vi sólo la enfermera en contacto con médico hace tres días			NA_BR	2	7	13
patient	sí he visto un médico hace tres días	sí he visto un médico hace tres días	he consultado un médico hace tres días	yes i visited the doctor three days ago	AC_help	2	7	13
doctor	did you have any tests	did you have a test	did you have the test?	¿le han realizado un examen?	NA_IE	2	7	13
doctor	did they test you for sore strep throat	did they have the worse sore throat for your period			NA_BR	2	7	13
doctor	was the diagnosis a strep throat	was the diagnosis a strep throat	was the diagnosis strep throat?	¿le han diagnosticado una infección de la garganta por estreptococo?	AC_help	2	7	13
patient	me hicieron un cultivo	es recientemente positivo	el dolor es recientemente positivo		NA_BR	2	7	13
doctor			was the diagnosis strep throat?	¿le han diagnosticado una infección de la garganta por estreptococo?	REP	2	7	13
patient	me hicieron una prueba de cultivo	mes que tiene una prueba de una tos	me diagnostican una infección de la garganta por estreptococo que tiene el examen de tos desde hace meses		NA_BR	2	7	13
patient	sí ha hecho una prueba rápida la semana pasada	sí ha hecho una prueba rápida la semana pasada	me han realizado una prueba rápida por estreptococo la semana pasada	yes the doctor carried out a rapid strep test during the last week	AC_help	2	7	13
doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	2	7	13
patient	no sé	no sé	no sé	i don't know	AC	2	7	13
doctor	do you have a sore throat	do you have sore throat	do you have a sore throat?	¿le duele la garganta?	AC	2	7	14
patient	sí	sí	sí	yes	AC	2	7	14
doctor	for how long	[rien]				2	7	14
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	7	14

patient	no estoy segura	no desde hace hora	me duele la garganta desde hace horas	i have had a sore throat for hours	AC_IE	2	7	14
doctor	do you have a headache	yesterday			NA_BR	2	7	14
doctor	do you have a headache	do you have a headache	do you have the headaches?	¿le duele la cabeza?	AC_help	2	7	14
patient	sí	sí	sí	yes	AC	2	7	14
doctor	do you have a cough	do you have a cough	do you have a cough?	¿tiene tos?	AC	2	7	14
patient	no	no	no	no	AC	2	7	14
doctor	do you have congestion	is yesterday sudden			NA_BR	2	7	14
doctor	do you have congestion	do you have congestion	do you feel congested?	¿tiene la nariz tapada?	AC	2	7	14
patient	sí	sí	sí	yes	AC	2	7	14
doctor	swollen glands	swollen glands			NA_BR	2	7	14
doctor	do you have swollen glands	do you have swollen glands	do you have swollen glands?	¿tiene los ganglios inflamados?	AC_help	2	7	14
patient	sí	sí	sí	yes	AC	2	7	14
doctor	have you been around anyone with sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA_NT	2	7	14
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who have a sore throat?	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC_help	2	7	14
doctor	have you seen a doctor	have you seen the doctor	have you visited the doctor?	¿ha consultado un médico?	AC	2	7	14
patient	sí	sí	sí	yes	AC	2	7	14
doctor	have you been tested for strep throat	has the congested person had strep throat			NA_BR	2	7	14
doctor	did you have a test for strep throat	did you have a test for seconds	did you have the test for several seconds?	¿le han realizado un examen algunos segundos?	NA_BR	2	7	14
doctor	did you have any tests done	do you have the testing	do you have the test?	¿le realizan un examen?	AC	2	7	14
patient	sí	sí	sí	yes	AC	2	7	14
doctor	what did the test show	was there the test			NA_BR	2	7	14

doctor	were the results of the test positive	were the results of the test positive	were the results of the test positive?	¿el resultado del examen ha sido positivo?	NA	2	7	14
doctor	what were the results of this test	what were the results of this test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC_help	2	7	14
patient	positivo	positivo	el resultado del examen ha sido positivo	the results of the test were positive	AC	2	7	14
doctor	are you allergic to anything	are you allergic to medicine	are you allergic to medicine?	¿medicina le da alergia?	AC	2	7	14
patient	sí tengo alergia a la amoxicilina	sí tengo alergia al amoxicilina	la amoxicilina me da alergias	yes i am allergic to amoxicillin	AC	2	7	14
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	2	8	16
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	8	16
doctor	how long have you had the pain	how long have you had pains	how long have you experienced the pain?	¿desde cuándo tiene el dolor?	AC	2	8	16
patient	desde hace tres o cuatro días	desde hace treinta y ocho días	me duele desde hace treinta y ocho días	i have experienced the pain for thirty eight days	AC_IE	2	8	16
doctor	have you had fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	2	8	16
patient	sí he tenido fiebre	sí es severo vienes			NA_BR	2	8	16
patient	sí la fiebre ha venido gradualmente	sí la fiebre ha venido gradualmente	fiebre ha empezado gradualmente	yes fever started gradually	AC	2	8	16
doctor	have you had congestion	have you had congestion	have you felt congested?	¿ha tenido la nariz tapada?	AC	2	8	16
patient	sí he tenido congestion	siete días en cuello	he tenido la nariz tapada en la nuca siete días		NA_BR	2	8	16
patient	sí he tenido la nariz tapada	sí siete días en nariz tapada	he tenido la nariz tapada en siete días		NA_BR	2	8	16
patient	sí he tenido la nariz tapada	sí siete días en nariz tapada	he tenido la nariz tapada en siete días		NA_BR	2	8	16
patient	sí tengo la nariz tapada	sí tengo la nariz	tengo la nariz tapada	yes i feel congested	AC	2	8	16
doctor	has anyone else been sick	has anyone now been sick			NA_BR	2	8	16
doctor	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	have you been in contact with anyone who has a sore throat	¿ha estado en contacto con alguien que tiene el dolor de garganta?	AC	2	8	16
patient	no	no	no	no	AC	2	8	16

doctor	have you been to another doctor	have you been to another doctor	have you visited another doctor?	¿ha consultado un otro médico?	AC	2	8	16
patient	no	no	no	no	NA_IE	2	8	16
patient	have any tests been done	ha venido hace setenta días	este síntoma ha empezado hace setenta días	the symptoms started seventy days ago	NA_IE	2	8	16
doctor	have any tests been done	have any helped in the month	has a better pain been in this month?		NA_BR	2	8	16
doctor	was a strep test done	was a strep test done	was a rapid strep test carried out?	¿le han realizado la prueba rápida por estreptococo?	AC	2	8	16
patient	no	no	no	no	AC	2	8	16
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	2	8	15
patient	en la garganta	en la garganta	me duele la garganta	i have a sore throat	AC	2	8	15
doctor	how long have you had the pain	how long have you had the pain	how long have you experienced the pain?	¿desde cuándo tiene el dolor?	AC	2	8	15
patient	más de una semana	más de una semana	me duele desde hace más de una semana	i have experienced the pain for more than one week	AC	2	8	15
doctor	have you had fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	2	8	15
patient	have you had congestion	adiós a muchas alergias			NA_BR	2	8	15
doctor	have you had congestion	have you had congestion	have you felt congested?	¿ha tenido la nariz tapada?	AC	2	8	15
patient	sí	sí	sí	yes	AC	2	8	15
doctor	has anyone else been ill	has anyone now been ill			NA_BR	2	8	15
doctor	have you recently been in contact with anyone with a strep throat	have you recently been in contact with anyone with a strep throat	have you recently been in contact with anyone with a strep throat	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC	2	8	15
doctor	have you been to another doctor	have you been to another doctor	have you visited another doctor?	¿ha consultado un otro médico?	AC	2	8	15
doctor	did you have any test	did you have the test	did you have the test?	¿le han realizado un examen?	AC	2	8	15

doctor	what were the results	what were the results	what were the results?	¿cuál ha sido el resultado?	AC	2	8	15
patient	el resultado fue positivo	el resultado fue positivo	el resultado ha sido positivo	the results were positive	AC	2	8	15
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC	2	7	15
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	7	15
doctor	for how long how long have you had the pain	for how long has heat often alleviated the pain			NA_BR	2	7	15
doctor	how long have you had the pain	how long have you had the pain	how long have you experienced the pain?	¿desde cuándo tiene el dolor?	AC	2	7	15
patient	una semana	una semana	me duele desde hace una semana	i have experienced the pain for one week	AC	2	7	15
doctor	do you hurt anywhere else	are you tired in your now			NA_BR	2	7	15
doctor	do you have a headache	do you have headache	do you have the headaches?	¿le duele la cabeza?	AC	2	7	15
patient	no	no	no	no	AC_help	2	7	15
doctor	do you have a stomachache	getting stomachache			NA_BR	2	7	15
doctor	do you have a stomachache	do you have a stomachache	do you have a stomachache?	¿el estómago le duele?	AC	2	7	15
patient	no	no	no	no	AC	2	7	15
doctor	do you have a cough	are you having cough	do you have a cough?	¿tiene tos?	AC	2	7	15
patient	un poco	un poco	tengo un poco de tos	i have a bit of a cough	AC_help	2	7	15
doctor	do you have swollen glands	can swollen glands hurt	do you experience the pain in swollen glands?	¿le duele en los ganglios inflamados?	NA_BR	2	7	15
doctor	do you have swollen glands	do you have swollen glands	do you have swollen glands?	¿tiene los ganglios inflamados?	AC_help	2	7	15
patient	sí un poco	sí un poco	tengo un poco los ganglios inflamados	my glands are slightly swollen	AC_help	2	7	15
doctor	has anyone around you been sick with sore throat	has anyone around you been sick when sore throat came			NA_BR	2	7	15
doctor	is anybody around you suffering from a sore throat	is anybody around you suffering from a sore throat	does anyone around you have your sore throat?	¿alguien en su entorno tiene el dolor de garganta?	AC_help	2	7	15

patient	sí	sí	sí	yes	AC	2	7	15
doctor	have you seen a doctor for your sore throat	have you seen a doctor for your sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC	2	7	15
patient	sí hace una semana	sí hace una semana	he consultado un médico para el dolor de garganta hace una semana	i visited the doctor for my sore throat one week ago	AC_help	2	7	15
doctor	did they test you for strep throat	did they have the worse strep throat			NA_BR	2	7	15
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC	2	7	15
patient	sí	sí	sí	yes	AC	2	7	15
doctor	what was the result of the test	what was the results of the test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC	2	7	15
patient	positivo	positivo	el resultado del examen ha sido positivo	the results of the test were positive	AC_help	2	7	15
doctor	are you allergic to any medicines	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	2	7	15
patient	sí tengo alergia a la penicilina	sí tengo alergia a la penicilina	la penicilina me da alergias	yes i am allergic to penicillin	AC_help	2	7	15
doctor	where is the pain	where is the pain	where is the pain?	¿dónde le duele?	AC_help	2	7	16
patient	me duele la garganta	me duele la garganta	me duele la garganta	i have a sore throat	AC	2	7	16
doctor	for how long have you had the sore throat	for how long \ 's any other occurred			NA_BR	2	7	16
doctor	for how long have you had your sore throat	for how long have you had your sore	how long have you had a sore throat?	¿desde cuándo le duele la garganta?	AC_help	2	7	16
patient	desde hace dos días	desde hace dos días	me duele la garganta desde hace dos días	i have had a sore throat for two days	AC	2	7	16
doctor	have you had a fever	have you had fever	have you had fever?	¿ha tenido fiebre?	AC	2	7	16
patient	sí he tenido un poco	sí disminuye un poco	este síntoma disminuye un poco	yes a bit of the symptoms is decreasing	AC_IE	2	7	16
doctor	have you had a fever	have you had a fever	have you had fever?	¿ha tenido fiebre?	AC	2	7	16
patient	sí me duele un poco	sí me duele un poco	me duele un poco	yes i experience a bit of pain	AC	2	7	16

doctor	have you had a headache	have you had a headache	have you had the headaches?	¿le ha dolido la cabeza?	AC	2	7	16
patient	sí mucho	sí mucho	la cabeza me ha mucho dolido	i experienced a severe pain in the head	AC	2	7	16
doctor	have you had cough	have you had cough	have you had a cough?	¿ha tenido tos?	AC	2	7	16
patient	sí bastante	sí bruscamente			NA_BR	2	7	16
patient	sí bastante	sí catorce			NA_BR	2	7	16
patient	mucha tos	mucha tos	he tenido mucha tos	i had a severe cough	AC	2	7	16
doctor	have you had congestion	have you had congestion	have you felt congested?	¿ha tenido la nariz tapada?	AC	2	7	16
patient	no	no	no	no	AC	2	7	16
doctor	have you had runny nose	have you had runny nose	have you had a runny nose?	¿ha mucho tenido moco?	AC	2	7	16
patient	sí mucho escurrimiento nasal	sí ha mucho sufrido náuseas			NA_BR	2	7	16
patient	sí mucho	sí mucho	he tenido mucho moco	i had a runny nose	AC	2	7	16
doctor	fever	fever	have you had fever?	¿ha tenido fiebre?	AC	2	7	16
patient	sí poca	sí nuca			NA_BR	2	7	16
patient	sí poca	sí nuca			NA_BR	2	7	16
patient	sí	sí	sí	yes	AC	2	7	16
doctor	how high was your temperature	how high was your temperature	what was your temperature?	¿cuántos grados ha tenido?	AC	2	7	16
patient	treinta y ocho punto cinco	treinta y ocho grados de moco			NA_BR	2	7	16
patient	treinta y ocho grados	treinta y ocho grados	he tenido treinta y ocho grados de fiebre	my temperature was one hundred one degrees	AC	2	7	16
doctor	have you been around anyone with sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA_BR	2	7	16
doctor	have you been around anyone with sore throat	have you been around anyone with sore throat		¿ha estado en contacto con alguien que tiene el dolor de garganta?	NA_BR	2	7	16

doctor	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	have you been in contact with anyone with a strep throat recently	¿ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo?	AC	2	7	16
patient	no me acuerdo	no me acuerdo	no me acuerdo	i don't remember	AC	2	7	16
doctor	have you seen a doctor for this sore throat	have you seen a doctor for this sore throat	have you visited the doctor for your sore throat?	¿ha consultado un médico para el dolor de garganta?	AC	2	7	16
patient	sí	sí	sí	yes	AC	2	7	16
doctor	when	glands			NA_BR	2	7	16
doctor	when did you see the doctor	when did you see the doctor	when did you visit the doctor?	¿cuándo ha consultado un médico?	NA	2	7	16
doctor	when did you visit the doctor	when you visit the doctor			NA_BR	2	7	16
doctor	when did you visit the doctor	when did you visit the doctor	when did you visit the doctor?	¿cuándo ha consultado un médico?	AC_help	2	7	16
patient	la semana pasada	la semana pasada	he consultado un médico la semana pasada	i visited the doctor during the last week	NA_help	2	7	16
patient	ayer	ayer	he consultado un médico ayer	i visited the doctor yesterday	AC	2	7	16
doctor	did the doctor test you for strep throat	did the doctor have your strep throat	did the doctor have strep throat?	¿el médico ha tenido una infección de la garganta por estreptococo?	NA	2	7	16
doctor	did the doctor do a throat culture test	did the doctor do a throat culture test	did the doctor carry out a throat culture test?	¿le han realizado un cultivo de la garganta?	NA_help	2	7	16
doctor	did the doctor do a rapid strep test	did the doctor do a rapid strep test	did the doctor carry out a rapid strep test?	¿le han realizado la prueba rápida por estreptococo?	AC_help	2	7	16
patient	sí	sí	sí	yes	AC	2	7	16
doctor	what was the result of the test	what was the results of the test	what were the results of the test?	¿cuál ha sido el resultado del examen?	AC	2	7	16
patient	negativo	es negativo	el dolor es negativo		NA_NT	2	7	16
patient	negativo	negativo	el resultado del examen ha sido negativo	the results of the test were negative	AC	2	7	16

doctor	are you allergic to any medicine	are you allergic to any medicine	are you allergic to any medicine?	¿tiene alergia a alguna medicina?	AC	2	7	16
patient	no	no	no	no	AC	2	7	16

Annexe 7 : Tableau des informations récoltées par les médecins lors de l'EXPQ1

Relevant Clinical Data	Case 1	Phys 1	Phys2	Phys3	Phys4
Severity of pain	++	severe	severe	severe	?
Symptom duration	two days	2 days	2 days		yes
Symptoms	fever (39.5 C°)	fever	fever - 103		yes
	headache (++)	headache	headache +	headache	
	swollen glands (++)	no congestion	no rhin.	no runny nose	
	red, sandpaper-like rash (+)	no cough	no cough		
	stomachache		stomachache		
Contact with ill person	yes, days ago	yes	yes	yes	not sure
Recent medical visit for condition	no	no			did not ask
Recent diagnostic test performed	no				did not ask
Allergy to antibiotics	none	no	no		did not ask
	Comments	(erased strep)			easy to do
	Diagnosis	not sure	strep	strep	strep
System		1	2	1	2
ID_SESSION		15.10.33.281	15.05.55.453	17.40.36.343	17.45.30.140
inter_num	utterances_NT_OCC_COUN	52 à 95	1 à 51	723 à 740	742 à 811
Number of interactions		43	51	17	69

<i>Relevant Clinical Data</i>	Case 2	Phys 1	Phys2	Phys3	Phys4
Severity of pain	+	little	not severe	severe	yes
Symptom duration	four days	4 days	4 days	4 days	yes
Symptoms	fever to touch	fever -101°	no cough	fever	yes and no
	nausea	no cough	no rhin.	no runny nose	
	vomiting	no congestion	no stomacheache		
	headache	headache	headache 4 days	headache	
			no rash		
Contact with ill person	no	no	no	no runny nose	yes
Recent medical visit for condition	yes	last Friday	2 days ago		did not ask
	doctor				
	two days ago				
	diagnosed with viral infection				
Recent diagnostic test performed	no	no	no test, no meds		did not ask
Allergy to antibiotics	penicillin	penicillin	penicillin	penicillin	did not ask
	<i>Comments</i>		hard to get all of symptoms especially temperature, tests done, and prior docotr visit		difficult interview, I had to repeat and rephrase questions
	<i>Diagnosis</i>	viral	viral	strep	not sure
System		1	2	1	2
ID_SESSION		15.41.43.921	15.42.49.828	17.55.22.062	17.58.32.203
inter_num	<i>utterances_NT_OCC_COUN</i>	96 à 143	144 à 208	776 à 811	812 à 887
Number of interactions		47	64	35	75

<i>Relevant Clinical Data</i>	Case 3	Phys 1	Phys2	Phys3	Phys4
Severity of pain	+	severe	severe	no sore throat	yes
Symptom duration	4 days	4 days	3 days	4 days	yes
Symptoms	none	no cough	no cough, no rhino, no congestion	no runny nose	yes
	headache (+)	headache	headache 4 days	headache	
	swollen glands (+)	no fever	no fever, no rash	no fever, no rash	
	stomachache		stomacheache	no nausea, stomacheache, no vomiting	
Contact with ill person	no	no	no	no	yes
Recent medical visit for condition	yes, nurse practitioner	yes	yes, nurse		did not ask
	three days ago		3 days ago		
Recent diagnostic test performed	throat culture test	throat culture			did not ask
	no result yet	unknown result	no meds		
Allergies to antibiotics	no	no	no meds		tried to ask but could not get proper translation
	Comments				easiest interview so far. System still does not recognise all my comments and questions
	Diagnosis	viral	viral	not sure	viral
System		2	1	2	1
ID_SESSION		15.57.51.140	16.04.49.359	18.08.58.031	18.12.40.218
inter_num	utterances_NT_OCC_COUN	209 à 241	242 à 287	852 à 887	888 à 926
Number of interactions		32	45	35	38

<i>Relevant Clinical Data</i>	Case 4	Phys 1	Phys2	Phys3	Phys4
Severity of pain	++	very severe	severe	severe	not sure
Symptom duration	not sure	?	?	?	yes
Symptoms	fever (++)	fever	fever		yes
	swollen glands	no cough	no cough, no rash, no stomacheache	no runny nose, no rash, no stomacheache	
	congestion		congestion	fever	
	headache	headache	headache	headache	
Contact with ill person	yes, a couple of days ago	yes	yes, 2 days ago	yes	yes
Recent medical visit for condition	yes	yes	yes, when?		yes
Recent diagnostic test performed	rapid strep test positive	troat test positive	strep test		no
Allergy to antibiotics	amoxicillin	amoxicillin	amoxicillin		did not ask
	Comments				difficult to ask if pat had headache also difficult to access severity
	Diagnosis	strep	strep	strep	viral
System		2	1	2	1
ID_SESSION		16.15.29.062	16.19.57.031	18.22.26.375	18.23.18.109
inter_num	<i>utterances_NT_OCC_COUN</i>	288 à 317	318 à 374	927 à 952	953 à 1000
Number of interactions		29	56	25	47

<i>Relevant Clinical Data</i>	Case 5	Phys 1	Phys2	Phys3	Phys4
Severity of pain	++	very severe	>improved	severe	yes
Symptom duration	3-4 days	> 1 week	1 week		yes
Symptoms	fever (38 C°)	fever	fever (103°)	fever	yes
	headache	headache	headache, >1 week	headache	
	congestion (++)	runny nose	congestion, rash, stomach ache > 1 day	no runny nose, no swollen glands	
	cough (++)	cough	weeks	no rash	
Contact with ill person	yes, a month ago	not recently	?	yes one month ago	no
Recent medical visit for condition	no	no	yes, doctor, 1 wk ago		did not ask
Recent diagnostic test performed	no	—	strep test, +		did not ask
Allergy to antibiotics	no	no	?		did not ask
	Comments		unable to get rash characteristics		difficult interview
	Diagnosis	viral	strep	strep	viral
System		1	2	1	2
ID_SESSION		16.29.03.265	16.32.35.343	18.33.59.796	18.41.37.000
inter_num	<i>utterances_NT_OCC_COUN</i>	375 à 411	412 à 468	1001 à 1018	1019 à 1081
Number of interactions		36	56	17	62

<i>Relevant Clinical Data</i>	Case 6				
		Phys 1	Phys2	Phys3	Phys4
Severity of pain	+	very severe	pain	severe	yes
Symptom duration	more than a week	several weeks	> 1 week		yes
Symptoms	headache	headache	headache	swollen glands, headache	yes
	congestion (++) (start)	cough	congestion	runny nose	
	runny nose (now)	fever	no fever, no rash	rash	
	cough (+)		fatigue	? Fever	
Contact with ill person	don't remember	no	?	?	did not ask
Recent medical visit for condition	yes, a nurse practitioner	no	yes, doctor 2 days ago		did not ask
Recent diagnostic test performed	don't know	—	?		did not ask
Allergy to antibiotics	no	penicillin	no		did not ask
	Comments	erased viral			yes/no answers worked best
	Diagnosis	not sure	viral	strep	viral
System		1	2	1	2
ID_SESSION		16.42.41.218	16.46.31.796	19.07.40.515	19.08.17.328
inter_num	<i>utterances_NT_OCC_COUN</i>	469 à 502	503 à 551	1082 à 1100	1101 à 1121
Number of interactions		33	48	18	20

<i>Relevant Clinical Data</i>	Case 7	Phys 1	Phys2	Phys3	Phys4
Severity of pain	+	severe	very severe pain, no improvement	severe	no
Symptom duration	since this morning	1 day	> 1 week		yes
Symptoms	fever (38,5 C°)	fever -101	fever (103°)	fever	yes
	headache (+)	headache	headache 1 wk		
	runny nose (++)	runny nose	rhinorrhea, stomacheache	runny nose	
	cough (+)	cough	cough, rash	swollen glands?	
Contact with ill person	no	no	yes	no	not clear
Recent medical visit for condition	yes, a week ago	1 week ago	no		did not ask
	nurse	nurse			
	no strep-diagnosed				
Recent diagnostic test performed	no	no	no		did not ask
Allergies to antibiotics	yes, amoxicillin	amoxicillin	penicilin		did not ask
	Comments		again, no rash chracteristics could not ask what meds were taken		hard to get severity
	Diagnosis	viral	not sure	viral	viral
System		2	1	2	1
ID_SESSION		16.56.47.921	17.01.20.156	19.19.40.421	19.24.29.453
inter_num	<i>utterances_NT_OCC_COUN</i>	552 à 584	585 à 637	1122 à 1137	1138 à 1181
Number of interactions		32	52	15	43

<i>Relevant Clinical Data</i>	Case 8	Phys 1	Phys2	Phys3	Phys4
Severity of pain	++	severe	very severe	severe	no
Symptom duration	2 days	> 1 week	2 day		yes
Symptoms	fever (low)	fever - 39°	temp 99°	fever	yes
	headache (++)	headahce	headache, 2 days	headache	
	congestion (++)	runny nose	rhinorrhea, cough	runny nose	
	cough (++)	cough	no rash, no stomachache	rash, swollen glands ?	
Contact with ill person	don't remember	yes	?	?	not clear
Recent medical visit for condition	yes	yes	yesterday		did not ask
Recent diagnostic test performed	rapid strep test	culture	strep test		did not ask
	result negative	negative	negative		
Allergies to antibiotics	no	no	no		did not ask
	Comments				yes/no answers worked best
	Diagnosis	viral	viral	strep	viral
System		2	1	2	1
ID_SESSION		17.06.44.734	17.14.50.968	19.30.58.562	19.35.48.359
inter_num	utterances_NT_OCC_COUN τ	638 à 675	676 à 722	1182 à 1199	1200 à 1226
Number of interactions		37	46	17	26

<i>Relevant Clinical Data</i>	Case 1	Phys5	Phys6	Phys7	Phys8
Severity of pain	++	yes	unable	severe	
Symptom duration	two days	2 days	3 days	a few days	last week
Symptoms	fever (39.5 C°)	yes	fever		fever 103°
	headache (++)			headache	
	swollen glands (++)		no runny nose, no cough		congestion
	red, sandpaper-like rash (+)		rash		rash
	stomachache				abdominal pain
Contact with ill person	yes, days ago	yes	yes	yes	
Recent medical visit for condition	no		yes	no	
Recent diagnostic test performed	no		no	no	strep test positive
Allergy to antibiotics	none		yes		
	<i>Comments</i>				
	<i>Diagnosis</i>	strep	strep	not sure	strep
System		1	2	1	2
ID_SESSION		21.10.44.515	21.12.31.828	16.00.56.531	16.09.23.093
inter_num	<i>utterances_NT_OCC_COUNT</i>	1227 à 1269	1270 à 1306	1827 à 1889	1890 à 1924
Number of interactions		42	36	62	34

<i>Relevant Clinical Data</i>	Case 2	Phys5	Phys6	Phys7	Phys8
Severity of pain	+		cross		
Symptom duration	four days	tick	tick	months	4 days
Symptoms	fever to touch	tick	tick		
	nausea			cough	
	vomiting			congestion	
	headache			headache	
Contact with ill person	no	tick	tick	don't remember	
Recent medical visit for condition	yes	tick	tick	yes	yes
	doctor				
	two days ago				
	diagnosed with viral infection				
Recent diagnostic test performed	no	tick	tick	yes positive, unknown test	
Allergy to antibiotics	penicillin		tick	penicillin	
	<i>Comments</i>				took medicine for sore throat
	<i>Diagnosis</i>	viral	viral	viral	viral
System		1	2	1	2
ID_SESSION		21.33.34.140	21.35.44.312	16.37.34.500	16.43.49.640
inter_num	<i>utterances_NT_OCC_COUNT</i>	1307 à 1343	1344 à 1383	1925 à 1964	1965 à 1988
Number of interactions		36	39	39	23

<i>Relevant Clinical Data</i>	Case 3	Phys5	Phys6	Phys7	Phys8
Severity of pain	+		cross	severe	
Symptom duration	4 days	tick	tick	4 days	3 days, >1 day of congetion
Symptoms	none	tick	tick		no fever, no rash
	headache (+)			headache	headache
	swollen glands (+)			swollen glands	congestion
	stomachache				
Contact with ill person	no	tick	tick	no	no
Recent medical visit for condition	yes, nurse practitioner		tick	3 days ago	
	three days ago				
Recent diagnostic test performed	throat culture test		tick	strep throat CX	yes, don't know results
	no result yet				
Allergies to antibiotics	no		tick	unknown	no
	Comments				
	Diagnosis	not sure	not sure	strep	viral
System		2	1	2	1
ID_SESSION		21.46.15.046	21.51.48.468	16.48.14.921	16.54.39.625
inter_num	<i>utterances_NT_OCC_COUNT</i>	1384 à 1438	1439 à 1473	1989 à 2017	2018 à 2043
Number of interactions		54	34	28	25

<i>Relevant Clinical Data</i>	Case 4	Phys5	Phys6	Phys7	Phys8
Severity of pain	++		cross		
Symptom duration	not sure	tick	cross	4 hours	
Symptoms	fever (++)	tick	tick		low fever
	swollen glands			swollen glands	no rash
	congestion			congestion	congestion a bit
	headache				yes
Contact with ill person	yes, a couple of days ago	tick	cross	don't know	
Recent medical visit for condition	yes		tick	yes	yes
Recent diagnostic test performed	rapid strep test positive		tick	yes	rapid strep
Allergy to antibiotics	amoxicillin		tick	amoxicillin	
	Comments				
	Diagnosis	strep	viral	strep	strep
System		2	1	2	1
ID_SESSION		21.59.51.390	22.01.13.062	16.59.17.265	17.02.24.218
inter_num	<i>utterances_NT_OCC_COUNT</i>	1474 à 1529	1530 à 1575	2044 à 2073	2074 à 2101
Number of interactions		55	45	29	27

<i>Relevant Clinical Data</i>	Case 5	Phys5	Phys6	Phys7	Phys8
Severity of pain	++		cross		
Symptom duration	3-4 days	tick	tick	> 1 month	3-4 days
Symptoms	fever (38 C°)	tick	tick	no congestion	fever
	headache			headache	
	congestion (++)			stomacheache	congestion
	cough (++)			cough	
Contact with ill person	yes, a month ago		tick	yes	no
Recent medical visit for condition	no	tick	tick	yes	no
Recent diagnostic test performed	no		cross	rapid strep +	no
Allergy to antibiotics	no		tick	penicillin	
	Comments				
	Diagnosis	viral	viral	strep	not sure
System		1	2	1	2
ID_SESSION		22.11.55.171	22.22.28.359	17.07.24.812	17.15.03.062
inter_num	<i>utterances_NT_OCC_COUNT</i>	1576 à 1598	1599 à 1661	2102 à 2138	2139 à 2159
Number of interactions		22	62	36	20

<i>Relevant Clinical Data</i>	Case 6				
		Phys5	Phys6	Phys7	Phys8
Severity of pain	+	tick	cross		
Symptom duration	more than a week	tick	tick	> 1 wk	1 week
Symptoms	headache	tick	tick	headache	fever
	congestion (++) (start)			congestion	congestion
	runny nose (now)			no stomachache, no fever	
	cough (+)			cough	
Contact with ill person	don't remember	tick	tick	no	no
Recent medical visit for condition	yes, a nurse practitioner	tick	tick	yes, nurse	yes
Recent diagnostic test performed	don't know	tick	tick	don't know	yes, positive
Allergy to antibiotics	no		tick	penicillin	
	Comments				
	Diagnosis	viral	viral	viral	strep
System		1	2	1	2
ID_SESSION		22.22.41.359	22.33.29.531	17.19.33.640	17.24.57.453
inter_num	<i>utterances_NT_OCC_COUNT</i>	1633 à 1661	1662 à 1693	2160 à 2189	2190 à 2203
Number of interactions		28	31	29	13

<i>Relevant Clinical Data</i>	Case 7	Phys5	Phys6	Phys7	Phys8
Severity of pain	+		cross		a little
Symptom duration	since this morning	tick	tick	1 week	1 week
Symptoms	fever (38,5 C°)	tick	tick		fever +
	headache (+)			no headache	no congestion
	runny nose (++)				
	cough (+)				
Contact with ill person	no	tick	tick	don't know	
Recent medical visit for condition	yes, a week ago	tick	tick	doctor one week ago	yes a nurse
	nurse				
	no strep-diagnosed				
Recent diagnostic test performed	no	tick	tick	rapid strep test, negative	no test
Allergies to antibiotics	yes, amoxicillin		tick	penicillin	yes, amoxil
	Comments				
	Diagnosis	strep	viral	strep	not sure
ID_SESSION		2	1	2	1
inter_num	<i>utterances_NT_OCC_COUNT</i>	22.34.46.578	22.46.09.984	17.27.41.046	17.33.26.562
Number of interactions		1694 à 1719	1720 à 1762	2204 à 2231	2232 à 2259
		25	42	27	27

<i>Relevant Clinical Data</i>	Case 8	Phys5	Phys6	Phys7	Phys8
Severity of pain	++	tick	cross		
Symptom duration	2 days	tick	cross and tick	2 days	2 days
Symptoms	fever (low)	tick	tick	fever 101°	fever
	headache (++)			headache	
	congestion (++)			rhinorrehea, no congestion	congestion
	cough (++)			cough	
Contact with ill person	don't remember	tick	?	don't know	no
Recent medical visit for condition	yes	tick	tick	yes, doctor yesterday	yes, last week
Recent diagnostic test performed	rapid strep test	tick	tick	rapid strep negative	yes, neg
	result negative				
Allergies to antibiotics	no		tick	no	no
	Comments		erased not sure		
	Diagnosis	viral	strep	viral	not sure
System		2	1	2	1
ID_SESSION		22.48.57.203	23.02.10.093	17.37.57.812	17.51.48.171
inter_num	<i>utterances_NT_OCC_COUNT</i>	1763 à 1795	1796 à 1826	2260 à 2307	2308 à 2329
Number of interactions		32	30	47	21

Annexe 8 : Evaluations HMS des experts de l'hôpital de Dallas

MedSLT V1 : Résultats de l'évaluation HMS

Backtranslation	Translation	Eval1	Eval2	Eval3	ReferenceTrans 1	ReferenceTrans 2	ReferenceTrans 3
me duele la semana pasada	i experience a pain during the last week						
me duele la garganta desde hace algunas semanas	i have had a sore throat for several weeks	3	4	4	I have had a sore throat for a few weeks	I have had a sore throat for the past few weeks	I have had a sore throat for a few weeks
he tomado la amoxicilina	i took amoxicillin	4	4	4	yes i took amoxicillin	I took the antibiotic amoxicillin	I took amoxicillin
he consultado un médico ayer	i visited the doctor yesterday	4	3	4	i saw the doctor yesterday	I saw the doctor yesterday	i saw the doctor yesterday
hola	good morning	4	4	4	good day	Good morning	good morning
hola	good morning	4	4	4	good day	Good morning	good morning
hola	good morning	4	4	4	good day	Good morning	good morning
sí pienso	i think so	4	4	4	yes i think that i have had fever	I think I have a fever	i think so
me duele la garganta desde hace cuatro días	i have had a sore throat for four days	4	4	4	my throat has been sore for four days	I have had my sore throat for four days	i have had a sore throat for four days
me duele la garganta desde hace cuatro días	i have had a sore throat for four days	4	4	4	i have had a sore throat for four days	I have had a sore throat for four days	i have had a sore throat for four days
me duele la garganta desde hace cuatro días	i have had a sore throat for four days	4	4	4	i have had a sore throat for the last four days	I have had a sore throat for the past four days	i have had a sore throat since four days ago
me duele la garganta desde hace cuatro días	i have had a sore throat for four days	4	3	4	my throat has been sore for the last four days	I have had my sore throat for the past four	i have had a sore throat since four days ago
me duele la garganta desde hace dos días	i have had a sore throat for two days	4	4	4	my throat has been sore for the last two days	I have had my sore throat for the past two	i have had a sore throat since two days ago
me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	4	4	4	i have had a sore throat for more than a week	I have had a sore throat for over a week	I have had a sore throat since over a week ago

[illegible]

me duele la garganta	i have a sore throat	3	4	3	the pain is in my throat	My pain is in the throat	the pain is in my throat
me duele la garganta	i have a sore throat	3	4	3	the pain is in my throat	My pain is in the throat	the pain is in my throat
me duele la garganta	i have a sore throat	3	4	3	the pain is in my throat	My pain is in the throat	the pain is in my throat
me duele la garganta	i have a sore throat	3	4	3	the pain is in my throat	My pain is in the throat	the pain is in my throat
me duele la garganta	i have a sore throat	3	4	3	the pain is in my throat	My pain is in the throat	the pain is in my throat
el resultado ha sido positivo	the results were positive	4	4	4	the results were positive	the results were positive	the results were positive
el dolor es severo	the pain is severe	3	4	4	the pain i feel is strong	The character of my pain is strong	the pain is severe
el dolor es severo	the pain is severe	3	3	3	it is a strong type of pain	The type of pain is strong	the pain is severe
el dolor es severo	the pain is severe	3	4	4	the pain is strong	The character of the pain is strong	the pain is severe
el dolor es severo	the pain is severe	3	4	4	the pain is strong	The character of the pain is strong	the pain is severe
he consultado un médico hace meses	i visited the doctor months ago	1	2	1	i visited the doctor ago	I visited the doctor yesterday	I went to the doctor ago
me duele la garganta hace algunas semanas	i have had a sore throat for several weeks	3	3	3	i have had a sore throat for a few weeks	I have had a sore throat for the past few weeks	I have had a sore throat for a few weeks
me duele la garganta hace cuatro días	i have had a sore throat for four days	4	3	4	i have had a sore throat for four days	I have had my sore throat for the past four	i have had a sore throat for four days
he estado en contacto con una persona que tiene una	i was in contact with the person who had strep	4	2	4	i was in contact with a person who has strep	I was in contact with a person who has strep	i have not been in contact with a person who has strep
me duele la garganta hace dos días	i have had a sore throat for two days	4	3	4	my throat has been hurting for the last two	My throat has been hurting me for the past	my throat has been hurting since two days ago
tengo el dolor hace dos días	i have experienced the pain for two days	4	3	4	i have had this pain for the last two days	I have had this pain for the past two days	i have had this pain since two days ago
he estado en contacto con una persona que tiene una	i was in contact with the person who had strep	4	2	4	two days ago	Two days ago	two days ago
he consultado una enfermera hace tres días	i visited a nurse three days ago	4	4	4	i saw the nurse three days ago	I saw the nurse three days ago	i saw the nurse three days ago

he consultado un médico hace tres días	i visited the doctor three days ago	4	2	4	yes i saw a doctor three days ago	I saw a doctor three days ago	yes i saw a doctor three days ago
he estado en contacto con una persona que tiene una	i was in contact with the person who had strep	3	3	3	yes i was in contact with a person with strep	I was in contact with a person who has strep	i was in contact with someone with strep throat a
he estado recientemente en contacto con alguien que	i was in contact with someone several days	1	3	1	yes i was in contact with someone with strep	Yes I was in contact with a person who has	yes i was in contact with someone with strep throat a
hola	hello	4	4	4	hello	hello	good morning
el resultado de un cultivo de la garganta ha sido una	the results of a throat culture test were strep	3	3	4	the result of the throat culture test showed an	The result of the throat culture test was	the results of the throat culture test indicated a
me duele la garganta	i have a sore throat	3	4	3	the pain is in my throat	My pain is in the throat	the pain is in my throat
la penicilina me da alergias	i am allergic to penicillin	4	4	4	yes i am allergic to penicillin	I am allergic to penicillin	i'm allergic to penicillin
he tenido más de treinta y nueve grados de fiebre	my temperature was more than one hundred	3	4	3	yes i had a fever of over thirty nine degrees	Yes I have had a fever of over thirty-nine	i have had a fever of over hundred and two degrees
he tenido más de treinta y nueve grados de fiebre	my temperature was more than one hundred	3	4	3	my temperature was more than thirty nine	My temperature was over thirty-nine degrees	my temperature was higher than hundred and two
me duele la garganta desde hace más de tres días	i have had a sore throat for more than three days	4	4	4	i have had a sore throat for more than three days	I have had my sore throat for over three	i have had a sore throat for longer than three days
me duele la garganta desde hace más de tres días	i have had a sore throat for more than three days	4	4	4	i have had a sore throat for more than three days	I have had a sore throat for over three days	i have had a sore throat for longer than three days
tengo la nariz tapada desde hace más de un día	i have felt congested for more than one day	1	1	2	i have had a headache for more than a day	I have had a headache for over a day	i have had a headache for more than a day
me duele la garganta desde hace más de un mes	i have had a sore throat for more than one month	4	4	4	i have had a sore throat for more than a month	I have had a sore throat for over a month	i have had a sore throat for more than a month
me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	4	4	4	i have had a sore throat for more than a month	I have had a sore throat for over a month	i have had a sore throat for more than a month
me duele la cabeza desde hace más de una semana	i have experienced the pain in the head for more	3	3	4	yes i have had a headache for more than	Yes I have had a headache for over a	i have had a headache for more than a week
me duele desde hace más de una semana	i have experienced the pain for more than one	4	4	4	i have had the pain for more than a week	I have had the pain for over a week	i have had the pain for more than a week
me duele la garganta	i have a sore throat	3	4	3	my throat hurts	My pain is in the throat	the pain is in my throat
me duele la garganta	i have a sore throat	3	4	3	my throat hurts	My pain is in the throat	the pain is in my throat

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no sé	i don't know	4	4	4	i don't know	i don't know	i don't know
no entiendo	i don't understand	2	2	1	no i haven't seen a doctor for my sore throat	No I have not yet seen a doctor for my sore	no i still haven't seen a doctor for my sore throat
el resultado ha sido positivo	the results were positive	4	4	4	the result was positive	the result was positive	the result was positive
el resultado del examen ha sido positivo	the results of the test were positive	4	4	4	the result of the test was positive	the result of the test was positive	the result was positive
el resultado ha sido positivo	the results were positive	4	4	4	the results were positive	the results were positive	the result were positive
el dolor es quemante	the pain is burning	4	4	2	i feel a burning pain	It really hurts and I have a burning sensation	my throat is burning
el dolor es severo	the pain is severe	4	4	3	it is severe	it is severe	it's severe
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes

[illegible]

[illegible]

[illegible]

sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
la amoxicilina me da alergias	i am allergic to amoxicillin	4	4	4	yes i am allergic to amoxicillin	yes i am allergic to amoxicillin	yes i'm allergic to amoxicillin
la amoxicilina me ha dado alergias	i was allergic to amoxicillin	3	1	1	yes i am allergic to amoxicillin	yes i am allergic to amoxicillin	yes i'm allergic to amoxicillin
la amoxicilina me da alergias	i am allergic to amoxicillin	4	4	4	yes i am allergic to amoxicillin	yes i am allergic to amoxicillin	yes i'm allergic to amoxicillin
la penicilina me da alergias	i am allergic to penicillin	4	4	4	yes i am allergic to penicillin	yes i am allergic to penicillin	yes i'm allergic to penicillin
la penicilina me da alergias	i am allergic to penicillin	4	4	4	yes i am allergic to penicillin	yes i am allergic to penicillin	yes i'm allergic to penicillin

la penicilina me da alergias	i am allergic to penicillin	4	4	4	yes i am allergic to penicillin	yes i am allergic to penicillin	yes i'm allergic to penicillin
la penicilina me da alergias	i am allergic to penicillin	4	4	4	yes i am allergic to penicillin	yes i am allergic to penicillin	yes i'm allergic to penicillin
la penicilina me da alergias	i am allergic to penicillin	4	4	4	yes i am allergic to penicillin	yes i am allergic to penicillin	yes i'm allergic to penicillin
	i experienced a stabbing severe pain in the head	3	3	1	yes i have had a bit of a headache	yes i have had quite a lot of headaches	yes i have had a moderate headache
he tenido treinta y ocho grados de fiebre	my temperature was one hundred one degrees	1	4	1	yes i have a temperature of thirty eight	yes i have a fever of thirty eight	yes i have a fever of thirty eight degrees
tomo medicina desde ayer	i have taken medication since yesterday	4	4	4	yes i have been taking medicines since	yes i have been taking medicines since	yes i have been taking medicine since yesterday
me duele la garganta desde hace algunos días	i have had a sore throat for several days	4	4	4	i have had a sore throat for a few days	i have had a sore throat for the last few days	yes i have had a sore throat since a few days
me duele la garganta desde hace dos días	i have had a sore throat for two days	4	4	4	i have had my sore throat for two days	i have had a sore throat for the last few days	yes i have had a sore throat since two days
tengo una erupción cutánea desde hace tres días	i have had a rash for three days	4	4	4	yes i have had a rash for three days	yes i have had a rash for the last three days	rashsore throat since three days ago
me duele la garganta	i have a sore throat	4	4	4	yes i have pain in my throat	yes my throat hurts	yes my throat hurts
he estado en contacto con alguien que tiene el dolor de	i was in contact with someone who had a sore	4	4	4	yes i was in contact with someone who had a	yes i was in contact with someone with a sore	yes days ago i was in contact with someone who
he estado en contacto con alguien que tiene una infección de la garganta por estreptococo hace días	i was in contact with someone days ago who had strep throat	4	4	4	yes i was in contact with someone who had a strep throat two days ago	yes i was in contact with someone with a strep thoat a few days ago	yes days ago i was in contact with someone who has a strep throat
he estado en contacto con alguien que tiene el dolor de	i was in contact with someone two days ago	4	4	4	yes i was in contact with someone with a sore	yes i was in contact with someone with a sore	yes i was in contact with someone who has a sore
he consultado un médico hace dos días	i visited the doctor two days ago	4	4	4	yes i saw a doctor two days ago	yes i went to the doctor two days ago	yes i saw a doctor two days ago
he estado en contacto con alguien que tiene una	i was in contact with someone one month ago	4	4	4	yes i was in contact with someone with a strep	yes i was in contact with someone with strep	yes a month ago i was in contact with someone who
he estado en contacto con alguien que tiene el dolor de	i was in contact with someone months ago	3	3	3	yes i was in contact with someone who had a	yes i was in contact with someone with a sore	yes a month ago i was in contact with someone who
he consultado un médico hace una semana	i visited the doctor one week ago	4	4	4	yes i went to visit the doctor one week ago	yes i visited the doctor a week ago	yes i went to the doctor a week ago

he consultado un médico hace una semana	i visited the doctor one week ago	4	4	4	yes i went to visit the doctor one week ago	yes i visited the doctor a week ago	yes i went to the doctor a week ago
he estado en contacto con alguien que tiene el dolor de	i was in contact with someone four days ago	3	4	3	yes i was in contact with someone who had a	yes i was in contact with someone with a sore	yes i was in contact a few days ago with someone with
me duele la garganta hoy	i have a sore throat today	3	3	3	yes i have a sore throat today	yes i have had a sore throat today	yes i have a sore throat today
	i visited another doctor during the last week	4	4	4	yes i went to another doctor last week	yes i went to another doctor last week	yes i went to another doctor last week
he estado en contacto con alguien el lunes pasado que	i was in contact with someone last monday	4	3	3	yes i was around someone with (a) sore	yes i was around someone with a sore	yes last Monday I was around someone with a sore
tengo más de treinta y ocho grados de fiebre	my temperature is more than one hundred one	3	3	3	yes i have a fever above 38 degrees	yes i have a fever of more than thirty eight	yes i have a fever of over thirty eight degrees
tengo tos desde hace más de un día	i have coughed for more than one day	4	4	4	yes i have had a cough for more than one day	yes i have had a cough for more than a day	yes i have had a cough for over a day
me duele mucho la cabeza	i experience a severe pain in the head	4	3	3	yes i have a big headache	yes i have a very bad headache	yes i have a bad headache
tengo mucha fiebre	i have a severe fever	3	4	4	yes i have a high fever	yes i have a very high fever	yes i have a high fever
me duele mucho la cabeza	i experience a severe pain in the head	4	3	3	yes i have had a big headache	yes i have had a very bad headache	yes i have had a bad headache
he tenido mucho moco	i had a runny nose	3	3	3	yes i have had a very runny nose	yes i have had a very runny nose	yes i have had a very runny nose
me duele mucho la cabeza	i experience a severe pain in the head	4	3	3	yes i have had a big headache	yes i have had a very bad headache	yes i have had a bad headache
tengo la nariz muy tapada	i feel severely congested	4	3	3	yes i am very congested	yes i am very congested	yes i am very congested
me duele mucho la cabeza	i experience a severe pain in the head	4	3	3	yes i have a big headache	yes i have a very bad headache	yes i have a bad headache
tengo mucho moco	i have a severe runny nose	1	1	1	yes i do have a runny nose	yes i have a runny nose	yes yes i have a runny nose
tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	1	1	3	i have a temperature of 39	my fever is at thirty nine	i have a fever of thirty nine degrees
la cabeza me ha un poco dolido	i experienced a bit of pain in the head	3	4	3	yes i have a slight headache	yes i have a bit of a headache	yes i have a bit of a headache
tengo un poco de fiebre	i have a bit of fever	4	4	3	yes i have (a) slight fever	yes i have a bit of fever	yes i have a slight fever

me duele un poco la garganta	i have a bit of a sore throat	4	4	3	yes my throat is a little sore	yes i have a bit of a sore throat	yes my throat is a little sore
tengo un poco de fiebre	i have a bit of fever	4	4	3	yes i have (a) slight fever	yes i have a bit of fever	yes i have a slight fever
me duele un poco la garganta	i have a bit of a sore throat	4	4	3	yes my throat is a little sore	yes i have a bit of a sore throat	yes my throat is a little sore
me duele un poco la garganta	i have a bit of a sore throat	4	4	3	yes my throat is a little sore	yes i have a bit of a sore throat	yes my throat is a little sore
tengo un poco los ganglios inflamados	my glands are slightly swollen	4	4	3	yes my glands are a little swollen	yes my glands are slightly swollen	yes my glands are a little swollen
la garganta me ha un poco dolido	i had a bit of a sore throat	4	4	3	yes my throat hurts a bit	yes my throat hurts a bit	yes my thorat hurts a little
tengo un poco de tos	i have a bit of a cough	4	4	3	yes i have a slight cough	yes i have a bit of a cough	yes i have a bit of a cough
tengo un poco de fiebre	i have a bit of fever	4	4	3	yes i have a slight fever	yes i have a bit of fever	yes i have a slight fever
me duele un poco la cabeza	i experience a bit of pain in the head	3	4	3	yes i have had a slight headache	yes i have had a bit of a headache	yes i have had a slight headache
tengo un poco de tos	i have a bit of a cough	4	4	3	yes i have a slight cough	yes i have a bit of a cough	yes i have a bit of a cough
tengo un poco de moco	i have a bit of a runny nose	4	4	3	yes my nose is a little runny/i have a slightly	yes i have a bit of a runny nose	yes my nose is a little runny
tengo un poco de tos	i have a bit of a cough	4	4	3	yes i have a slight cough	yes i have a bit of a cough	yes i have a bit of a cough
tengo una erupción cutánea leve	i have a bit of a rash	4	4	3	yes i have a slight rash	yes i have a bit of a rash	yes i have a slight rash
tengo la nariz un poco tapada	i feel a bit congested	3	4	3	yes i have a bit of congestion	yes i have a bit of congestion	yes i am a little congested
tengo la nariz un poco tapada	i feel a bit congested	3	4	3	yes i have a bit of congestion	yes i have a bit of congestion	yes i am a little congested
tengo un poco de tos	i have a bit of a cough	4	4	3	yes i have a slight cough	yes i have a bit of a cough	yes i have a bit of a cough
me duele un poco la cabeza	i experience a bit of pain in the head	3	4	3	yes i have had a slight headache	yes i have had a bit of a headache	yes i have had a slight headache
tengo un poco de fiebre	i have a bit of fever	4	4	3	yes i have a slight fever	yes i have a bit of fever	yes i have a slight fever

he tenido un poco de fiebre	i had a bit of fever	4	4	4	yes i have had a very slight fever	yes i have had a bit of fever	yes i have a very slight fever
he consultado una enfermera	i visited a nurse	4	4	4	yes i have seen a nurse	yes i saw a nurse	yes i say a nurse
me han realizado una prueba rápida por estreptococo	a rapid strep test was carried out	3	4	4	yes a rapid strep test was done	yes a rapid strep test was done	yes a rapid strep test was done
he consultado un médico el viernes pasado	i visited the doctor last friday	3	4	3	yes i saw a doctor last Friday	yes i saw the doctor last Friday	yes i saw a doctor last Friday
tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	1	1	4	i have a temperature of 39 degrees	i have a fever of thirty nine degrees	i have a fever of thirty nine degrees
he tenido treinta y nueve grados de fiebre	my temperature was one hundred three degrees	1	1	4	i had a temperature of 39 degrees	i have a fever of thirty nine degrees	i have a fever of thirty nine degrees
tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	1	1	4	i have a temperature of 39 degrees	i have a fever of thirty nine degrees	i have a fever of thirty nine degrees
tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	3	3	4	my temperature was 38 degrees	my temperature was thirty eight degrees	i have had a fever of thirty eight degrees
tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	3	3	3	my temperature is 38 degrees	my temperature is thirty eight degrees	i have a temperature of thirty eight degrees
he tenido treinta y siete grados de fiebre	my temperature was ninety nine degrees	3	3	3	my temperature was 37	my temperature is thirty seven degrees	i had a temperature of thirty seven degrees
el dolor mejora un poco	a bit of pain is decreasing	1	4	3	the pain is getting a little better	yes the pain is getting slightly better	the pain is getting a little better
me duele un poco la cabeza	i experience a bit of pain in the head	3	4	3	i have a slight headache	yes i have a bit of a headache	i have a slight headache
he un poco tomado medicina	i took a bit of medication	1	1	1	yes i tok some medicicine for my sore	yes i took some pills for my sore throat	yes i took some medicine for my sore throat
el estómago me duele un poco	i experience a bit of pain in the stomach	3	4	3	i have (a) slight stomache ache	yes i have a bit of a stomachache	i have a bit of a stomachache
tengo un poco de fiebre	i have a bit of fever	4	4	4	i have a slight fever	yes i have a bit of a temperature	i have a low-grade fever
he consultado una enfermera para el dolor de garganta	i visited a nurse for my sore throat	3	4	3	i have seen a nurse for my sore throat	i saw a nurse for my sore throat	I saw a nurse a week ago
he consultado una enfermera	i visited a nurse	3	1	1	i saw a nurse a week ago	i went to a nurse a week ago	i went to a nurse a week ago

MedSLT V2 : Résultats de l'évaluation HMS							
Backtranslation	Translation	Eval 1	Eval 2	Eval 3	Ref 1	Ref 2	Ref 3
he consultado un médico ayer	i visited the doctor yesterday	4	4	4	i visited the doctor yesterday	I visited the doctor yesterday	i went to the doctor yestreday
he tomado ayer los antibióticos	i took antibiotics yesterday	4	4	4	yes i took antibiotics yesterday	I took antibiotics yesterday	i took antibiotics yesterday
hola	good morning	4	4	4	good day	Good morning	good morning
hola	good morning	4	4	4	good day	Good morning	good morning
hola	good morning	4	4	4	good day	Good morning	good morning
hola	good morning	4	4	4	good day	Good morning	good morning
presento este síntoma	i experience the symptoms	1	1	2	i don't think that I have been in contact with	I don't think I have been in contact with anyone	i don't think i have been in contact with anyone who has
me duele la cabeza desde hace cuatro días	i have experienced the pain in the head for four	4	3	4	i have had the pain for the last four days	I have had the pain for the past four days	i have had the pain since four days ago
me duele desde hace cuatro días	i have experienced the pain for four days	4	3	4	i have had the pain for the last four days	I have had the pain for the past four days	i have had the pain since four days ago
el estómago me duele desde hace dos días	i have experienced the pain in the stomach for	3	3	3	i have had a stomachache for the last	I have had a stomachache for the	i have had a stomachache since two days ago
me duele la garganta desde hace dos días	i have had a sore throat for two days	4	4	4	my throat has been sore for the last two days	I have had my sore throat for the past two	i have had a sore throat since two days ago
me duele la garganta desde hace días	i have had a sore throat for days	1	1	3	my throat has been sore for the last two days	I have had my sore throat for the past two	i have had a sore throat since two days ago
me duele desde hace treinta y ocho días	i have experienced the pain for thirty eight days	1	1	1	i have had the pain for the last three or four	I have had the pain for the past three or four	i have had the pain since three or four days ago
me duele la garganta desde hace dos días	i have had a sore throat for two days	4	4	4	i have had a sore throat for two days	I have had a sore throat for two days	I have had a sore throat for two days
el dolor es muy fuerte	the pain is very severe	3	4	4	the pain is very strong	The type of pain is very strong	the pain is very severe
el dolor es profundo	the pain is deep	4	4	4	the pain is very intense	The character of the pain is deep	the pain is deep

el resultado ha sido positivo	the results were positive	4	3	4	the results were positive	the results were positive	the results were positive
el resultado ha sido positivo	the results were positive	4	4	4	the rests of the tests were positive	the result of the test was positive	the test result was positive
el resultado ha sido positivo	the results were positive	4	3	4	the results were positive	the results were positive	the results were positive
el resultado ha sido positivo	the results were positive	4	4	4	the rests of the tests were positive	the result of the test was positive	the test result was positive
me duele la garganta	i have a sore throat	3	3	3	the pain is in the throat	The pain is in the throat	the pain is in my throat
el dolor es severo	the pain is severe	3	4	4	the pain is very strong	The character of the pain is strong	the pain is severe
fiebre dura	fever lasts	1	2	2	you can tell by touching	My fever is feverish to the touch	i am hot to the touch
el dolor es severo	the pain is severe	3	4	4	the pain i feel is strong	The character of my pain is strong	the pain is severe
el dolor es severo	the pain is severe	3	4	4	the pain i feel is strong	The character of my pain is strong	the pain is severe
tengo tos hace algunas semanas	i have coughed for several weeks	3	4	4	i have had the cough for a few weeks	I have had the cough for the past few weeks	i have had the cough for a few weeks
me duele la cabeza hace cuatro días	i have experienced the pain in the head for four	4	3	4	i have had a headache for four days	I have had a headache for the past four days	i have had a headache for four days
el dolor de garganta ha empezado hace cuatro días	my sore throat started four days ago	4	3	4	my throat started to be sore four days ago	My sore throat started four days ago	my sore throat started four days ago
me duele la garganta hace cuatro días	i have had a sore throat for four days	4	3	4	i have had a sore throat for four days	I have had my sore throat for the past four	i have had a sore throat for four days
el dolor de garganta empieza hace dos días	my sore throat has started for two days	2	3	2	my throat started to be sore two days ago	My sore throat started two days ago	my sore throat started two days ago
he consultado un médico hace dos días	i visited the doctor two days ago	4	3	4	I saw the doctor two days ago	I saw the doctor two days ago	i saw the doctor two days ago
he consultado un médico para el dolor de garganta	i visited the doctor for my sore throat three days	4	2	4	yes i saw a doctor three days ago	I saw a doctor three days ago	yes i saw a doctor three days ago
me duele la garganta hace tres días	i have had a sore throat for three days	4	3	4	i have had a sore throat for the last three days	I have had a sore throat for the past three days	I have had a sore throat since three days ago
tengo treinta y nueve grados de fiebre	yes my temperature is one hundred three	3	4		my fever was at thirty five point five	My fever was thirty-nine point five	i had a fever of thirty nine point five degrees

el estómago me duele desde hoy	i have experienced the pain in the stomach since	4	4	3	my stomachache started today	I have had a stomachache since	i have only had a stomachache since today
el resultado del examen ha sido una infección viral	the results of the test were a viral sore throat	3	4	1	the test result showed a viral infection	The result of the test was viral infection	the test results indicated a viral infection
fiebre ha empezado ayer	fever started yesterday	4	4	4	my fever started yesterday	My fever started yesterday	the fever started yesterday
la penicilina me da alergias	i am allergic to penicillin	4	4	4	i am allergic to penicillin	I am allergic to penicillin	i'm allergic to penicillin
el dolor ha empezado la semana pasada	the pain started during the last week	3	4	4	the pain started last week	The pain started last week	the pain started last week
el dolor de garganta ha empezado la semana	my sore throat started during the last week	3	4	4	my throat started to feel sore last week	My sore throat started last week	the sore throat began last week
el médico presenta este síntoma días	the doctor experiences the symptoms days	1	1	2	i have had a headache for more than a day	I have had a headache for over a day	i have had a headache for more than a day
me duele la garganta desde hace más de un mes	i have had a sore throat for more than one month	4	4	4	i have had a sore throat for more than a month	I have had a sore throat for over a month	i have had a sore throat for more than a month
me duele la cabeza desde hace más de una semana	i have experienced the pain in the head for more	4	4	4	i have had the pain for more than a week	I have had the pain for over a week	i have had the pain for more than a week
me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	4	4	4	i have had a sore throat for more than a week	I have had a sore throat for over a week	my throat has been sore for more than a week
me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	4	4	4	i have had a sore throat for more than a week	I have had a sore throat for over a week	my throat has been sore for more than a week
me duele desde hace más de una semana	i have experienced the pain for more than one	3	4	4	for more than a week	For over a week	more than a week
me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	4	4	4	i have had a sore throat for more than a week	I have had a sore throat for over a week	my throat has been sore for more than a week
me duele la garganta desde hace más de una semana	i have had a sore throat for more than one week	4	4	4	i have had a sore throat for more than a week	I have had a sore throat for over a week	my throat has been sore for more than a week
me duele desde hace cuatro días	i have experienced the pain for four days	3	3	4	my throat has been hurting for the last four	I have had my sore throat for the past four	i have had a sore throat since four days ago
me duele desde hace cuatro días	i have experienced the pain for four days	3	3	4	my throat has been hurting for the last four	I have had my sore throat for the past four	i have had a sore throat since four days ago
me duele desde hace cuatro días	i have experienced the pain for four days	3	3	4	my throat has been hurting for the last four	I have had my sore throat for the past four	i have had a sore throat since four days ago
me duele desde hace dos días	i have experienced the pain for two days	3	4	4	my throat has been hurting for the last two	I have had my sore throat for the past two	my throat has been sore since two days ago

[illegible]

me duele la garganta	i have a sore throat	3	4	3	my throat hurts	My pain is in the throat	the pain is in my throat
me duele la garganta	i have a sore throat	3	4	3	my throat hurts	My pain is in the throat	the pain is in my throat
me duele la garganta	i have a sore throat	3	4	3	my throat hurts	My pain is in the throat	the pain is in my throat
me duele la garganta cada día	i have a sore throat every day	3	3	3	no my throat hurts every day	No the pain is not less severe as my throat	my throat is sore every day
me duele la garganta desde hace un mes	i have had a sore throat for one month	3	4	3	my throat started to hurt a month ago	The pain started over a month ago	my throat has been sore since a month ago
me duele muy a menudo	i very often experience a pain	4	4	4	it hurts very often	The pain hurts often	i have pain very frequently
me duele un poco	i experience a bit of pain	4	3	4	it is a slight kind of pain	The pain is only a little	it hurt a little
he estado recientemente en contacto con la persona	i recently visited the person	1	1	2	my head feels very hot	My head feels hot	my head feels hot
he tenido mucha tos	i had a severe cough	3	4	3	yes i have been coughing a lot	I have had a lot of cough	i have been coughing a lot
tengo la nariz muy tapada	i feel severely congested	3	4	3	yes i have a lot of congestion	Yes I have a lot of congestion	i have a lot of congestion
el dolor es muy fuerte	the pain is very severe	3	4	4	the pain is very strong	The character of the pain is very strong	the pain is very strong
el dolor es muy fuerte	the pain is very severe	3	4	4	the pain is very strong	The character of the pain is very strong	the pain is very strong
el resultado ha sido negativo	the results were negative	4	4	4	the results were negative	The results were negative	the test results were negative
el resultado del examen ha sido negativo	the results of the test were negative	4	4	4	the result was negative	The result was negative	the result was negative
no	no	4	4	4	no	no	no
no	no	4	4	4	no	no	no
no	no	4	4	4	no	no	no
no	no	4	4	4	no	no	no

me duele la garganta desde hace horas	i have had a sore throat for hours	2	2	2	i am not sure how long i have had a sore throat	I am not sure for how long I have had my sore	i am not sure how long i have had a sore throat
el dolor ha empezado	no the pain started	3	3	2	no it hasn't gone down	No I am not better	it has not subsided
he tomado medicina	no i took medication	4	4	1	no i haven't taken any medicine for the pain	No I am not taking any medicine for the pain	no i have not taken any medicine for the pain
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
no me acuerdo	i don't remember	4	4	4	i don't remember	i don't remember	i don't remember
me duele	no i experience a pain	1	1	1	i don't have a stomach ache	I don't have a stomachache	i don't have a stomachache
me duele la garganta	no i have a sore throat	1	1	1	no my throat doesn't hurt	No I don't have a sore throat	no my throat doesn't hurt
el dolor es muy fuerte	the pain is very severe	1	1	1	the pain isn't very strong	The character of the pain is not very strong	the pain isn't very severe
he estado recientemente en contacto con alguien que	i was in contact with someone the evening	1	1	1	i haven't been in contact with anyone with strep	No I have not been in contact with anyone	no i haven't been in contact with anyone recently who
presento este síntoma	no i experience the symptoms	1	2	1	no i don't have a runny nose	No I don't have a runny nose	no i don't have a runny nose

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
sí	yes	4	4	4	yes	yes	yes
la penicilina me ha dado alergias	i was allergic to penicillin	4	1	1	yes i am allergic to penicillin	yes i am allergic to penicillin	yes i'm allergic to penicillin
el resultado ha sido positivo	yes the results were positive	4	4	4	yes the result was positive for	yes the result was positive for	yes the results were positive for streptococcus
me han realizado una prueba rápida por estreptococo la	yes the doctor carried out a rapid strep test during	3	3	3	yes the doctor did a rapid strep test last week	yes the doctor did a rapid strep test last	yes the doctor did a rapid strep test last week
he consultado un médico para el dolor de garganta	i visited the doctor for my sore throat two days ago	4	4	4	yes i saw a doctor two days ago	yes i went to see the doctor for my sore	yes i saw a doctor for my sore thorat two days ago
he consultado un médico hace días	i visited the doctor days ago	3	4	3	yes i went to the doctor two days ago	yes i went to the doctor two days ago	yes i went ot the doctor two days ago
me han realizado una prueba rápida por estreptococo hace	a rapid strep test was carried out two days ago	3	4	3	yes the doctor did a rapid strep test two days ago	yes the doctor did a rapid strept test two	yes the doctor did a rapid strep test two days ago
he estado en contacto con alguien que tiene el dolor de	i was in contact with someone two days ago	4	4	4	yes i was in contact with someone with a sore	yes i was in contact with someone with a sore	yes i was in contact with someone who has a sore
he estado en contacto con alguien que tiene el dolor de	yes i was in contact with someone two weeks ago	4	1	4	yes two weeks ago i was in contact with someone	yes i was in contact with someone with a throat	yes i was in contact with someone who has a throat
he consultado un médico hace una semana	i visited the doctor one week ago	4	4	4	yes i visited a doctor one week ago	yes i visited a doctor a week ago	yes i went to a doctor a week ago
he consultado un médico para el dolor de garganta	i visited the doctor for my sore throat one week ago	4	4	4	yes i saw a doctor for my sore throat one week	yes i went to see the doctor for my sore	yes i saw a doctor for my sore throat three days ago
he consultado un médico hace una semana	i visited the doctor one week ago	4	4	4	yes i visited a doctor one week ago	yes i visited a doctor a week ago	yes i went to a doctor a week ago
he consultado una enfermera hace una semana	yes i visited a nurse one week ago	4	4	4	yes i saw a nurse one week ago	yes i saw a nurse a week ago	I saw a nurse a week ago

he consultado una enfermera hace tres días	yes i visited a nurse three days ago	3	4	3	yes i saw a nurse three days ago	yes i consulted a nurse three days ago	yes i had an appointment with a nurse three days ago
he estado en contacto con alguien que tiene el dolor de	yes i was in contact with someone who had a sore	4	3	3	yes i have been in contact with someone	yes i have been in contact with someone	yes i have been in contact with someone who has a
he estado en contacto con alguien que tiene el dolor de	yes i was in contact with someone who had a sore	4	3	3	yes i have been in contact with someone	yes i have been in contact with someone	yes i have been in contact with someone who has a
he estado en contacto con alguien que tiene el dolor de	yes i was in contact with someone who had a sore	4	3	3	yes i have been in contact with someone	yes i have been in contact with someone	yes i have been in contact with someone who has a
este síntoma disminuye un poco	yes a bit of the symptoms is decreasing	2	2	2	yes i have had a slight fever	yes i have had a bit of fever	yes i have a slight fever
he tomado medicina	yes i took medication	4	4	4	yes i have taken medicines	yes i have taken some medicine	yes i took medicine
he consultado un médico hace tres días	yes i visited the doctor three days ago	4	1	1	yes i saw a doctor three days ago	yes i saw a doctor for my sore throat three	yes i saw a doctor three days ago for my sore throat
la amoxicilina me da alergias	i am allergic to amoxicillin	4	4	4	yes i am allergic to amoxillin	yes i am allergic to amoxicillin	yes i am allergic to amoxicillin
fiebre ha empezado gradualmente	yes fever started gradually	3	3	3	yes the fever has developed gradually	yes the fever has risen gradually	yes i have had a fever that came gradually
me han realizado una prueba rápida por estreptococo la	a rapid strep test was carried out during the last	3	3	3	yes the doctor did a strep test last week	yes the doctor did a strep test last week	yes the doctor did a strep test last week
me duele	yes i experience a pain	4	3	3	yes i have a headache	yes my head hurts	yes my head hurts
me duele	yes i experience a pain	4	1	1	yes i have vomiting	yes it hurts	yes i have been vomiting and my stomach hurts
me duele desde hace más de una semana	yes i have experienced the pain for more than	4	3	3	yes i have had a sore throat for over a week	yes my throat has been hurting for over a week	yes my throat has been hurting for more than a week
me duele la cabeza	yes i experience a pain in the head	4	3	3	yes i have a headache	yes i have a headache	yes i have a headache
me duele la cabeza desde hace algunos días	yes i have experienced the pain in the head for	4	4	4	yes i have had a headache for a few days	yes i have had a headache for the last	yes i have had a headache since a few days ago
me duele la garganta	yes i have a sore throat	4	4	4	yes i have a sore throat	yes i have a sore throat	yes i have a sore throat
me duele la garganta	yes i have a sore throat	4	4	4	yes i have a sore throat	yes i have a sore throat	yes i have a sore throat
me duele la garganta	yes i have a sore throat	4	4	4	yes i have a sore throat	yes i have a sore throat	yes i have a sore throat

me duele mucho	yes i experience a severe pain	3	3	3	yes i have a big headache	yes i have a very bad headache	yes i have a very bad headache
me duele mucho la cabeza	yes i experience a severe pain in the head	4	3	3	yes my head hurts a lot	yes i have a very bad headache	yes i have a very bad headache
me duele mucho la cabeza	yes i experience a severe pain in the head	4	3	3	yes my head hurts a lot	yes i have a very bad headache	yes i have a very bad headache
me duele un poco	yes i experience a bit of pain	1	1	1	yes i have had a slight fever	yes it hurts a bit	yes i have had a fever and i ache a little
tengo los ganglios muy inflamados	i have severely swollen glands	3	4	4	yes i have very swollen glands	yes my glands are very swollen	yes my glands are very swollen
la cabeza me ha mucho dolido	i experienced a severe pain in the head	4	3	3	yes i have had a big headache	yes i have had a very bad headache	yes i have had a bad headache
he tenido mucho moco	i had a runny nose	3	3	3	yes i have had a very runny nose	yes i have had a very runny nose	yes i have had a very runny nose
me duele mucho la garganta	i have a severe sore throat	3	4	4	yes i have a very sore throat	yes i have a very sore throat	yes my throat is very sore
me duele mucho la cabeza	i experience a severe pain in the head	4	3	3	yes i have had a big headache	yes i have had a very bad headache	yes i have had a bad headache
tengo mucho moco	i have a runny nose	3	3	3	yes i have (a) very runny nose	yes i have a very runny nose	yes my nose is very runny
he tenido mucha fiebre	i had a severe fever	3	4	4	yes i have had a high fever	yes i have had a very high fever	yes i have had a high fever
tengo mucho moco	i have a runny nose	3	3	3	yes there is a lot of mucus	yes i have a very runny nose	yes my nose is very runny
me han realizado el examen	a test was carried out	1	3	3	yes the doctor also did a rapid strep test	yes he also did a rapid strep test	yes the doctor also did a rapid strep teest
la amoxicilina me da alergias	yes i am allergic to amoxicillin	4	4	4	yes i am allergic to amoxicillin	yes i am allergic to amoxicillin	yes i am allergic to amoxicillin
la amoxicilina me da alergias	yes i am allergic to amoxicillin	4	4	4	yes i am allergic to amoxicillin	yes i am allergic to amoxicillin	yes i am allergic to amoxicillin
la penicilina me da alergias	yes i am allergic to penicillin	4	4	4	yes i am allergic to penicillin	yes i am allergic to penicillin	yes i am allergic to penicillin
la penicilina me da alergias	yes i am allergic to penicillin	4	4	4	yes i am allergic to penicillin	yes i am allergic to penicillin	yes i am allergic to penicillin
me duele la cabeza	yes i experience a pain in the head	3	4	4	yes my head aches	yes i have a headache	yes i have a headache

tengo fiebre	yes i have fever	1	3	3	yes i have a fever	yes i have a fever	yes i have a fever
tengo fiebre	yes i have fever	1	3	3	yes i have a fever	yes i have a fever	yes i have a fever
tengo fiebre	yes i have fever	1	3	3	yes i have a fever	yes i have a fever	yes i have a fever
tengo la nariz tapada	yes i feel congested	3	3	3	yes my nose is congested	yes i have a blocked nose	yes i have had a stuffy nose
tengo la nariz tapada	yes i feel congested	3	3	3	yes my nose is congested	yes i have a blocked nose	yes i have had a stuffy nose
tengo mucha fiebre	yes i have a severe fever	3	4	4	yes i have a high fever	yes i have a very high fever	yes i have a high fever
tengo mucha tos	yes i have a severe cough	3	4	4	yes i am coughing a lot	yes i cough a lot	yes i am coughing a lot
tengo mucho moco	yes i have a runny nose	3	3	4	yes there is a lot of mucus	yes i have a very runny nose	yes my nose is very runny
tengo tos	yes i cough	1	3	3	yes i am coughing a little	yes i have a bit of a cough	yes i have a bit of a cough
tengo tos	yes i cough	1	4	4	yes i have a cough	yes i have a cough	yes i have a cough
tengo treinta y ocho grados de fiebre	yes my temperature is one hundred one degrees	3	4	4	yes i have a 38-degree fever	yes i have a fever of thirty eight degrees	yes i have a fever of thirty eight degrees
tengo una erupción cutánea	yes i have a rash	4	4	4	yes i have a rash	yes i have a rash	yes i have a rash
tomo medicina la noche	yes i take medication the evening	3	3	1	yes i took medicine last night	yes i took some medicine for my pain	yes i took medicine last night
tomo medicina para el dolor de garganta	yes i take medication for my sore throat	1	4	3	i am taking medicine for my sore throat/i am	i took some medicine for my sore throat	i took medicine for my sore throat
me duele un poco la garganta	i have a bit of a sore throat	4	4	3	yes my throat is a little sore	yes i have a bit of a sore throat	yes my throat is a little sore
el dolor mejora un poco	a bit of pain is decreasing	1	3	3	yes the pain is a little less severe	yes the pain is a little less severe	yes the pain is a little less severe
tengo un poco de tos	i have a bit of a cough	4	4	3	yes i have a slight cough	yes i have a bit of a cough	yes i have a bit of a cough
tengo un poco los ganglios inflamados	my glands are slightly swollen	4	4	3	yes my glands are a little swollen	yes my glands are slightly swollen	yes my glands are a little swollen

me duele un poco la cabeza	i experience a bit of pain in the head	3	4	3	yes i have a slight headache	yes i have a bit of a headache	yes i have a bit of a headache
tengo un poco de tos	i have a bit of a cough	4	4	3	yes i have a slight cough	yes i have a bit of a cough	yes i have a bit of a cough
he consultado una enfermera	i visited a nurse	4	4	4	yes i have seen a nurse	yes i saw a nurse	yes i say a nurse
fiebre ha empezado siempre	yes fever always started	2	1	2	yes i have had a fever	yes i have had a fever	Yes I have had a fever
tengo mucha fiebre	i have a severe fever	3	4	4	i have a high fever	my fever is very high	my fever is very high
tengo tos	i cough	4	1	1	i have a cough	i have a cough	I have a cough
tengo tos desde hace más de tres días	i have coughed for more than three days	4	4	1	i have had a cough for more than three days		i have a cough since more than three days ago
tengo tos desde hace un mes	i have coughed for one month	3	1	1	i have had a cough for a month		i have had a cough since over a month ago
tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	1	1	3	i have a temperature of 39 degrees	i have a temperature of thirty nine degrees	i have a fever of thirty nine degrees
tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	1	1	4	i have a temperature of 39 degrees	i have a fever of thirty nine degrees	i have a fever of thirty nine degrees
tengo treinta y nueve grados de fiebre	my temperature is one hundred three degrees	1	1	3	i have a temperature of 39 degrees	yes i have a fever of thirty nine degrees	i have a fever of thirty nine degrees
tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	3	3	4	i have a temperature of 38 degrees	i have a fever of thirty eight degrees	i have a fever of thirty eight degrees
tengo una infección viral	i have a viral sore throat	1	1	1	i have a viral infection	i have a viral infection	i have a viral infection
tengo nueve grados de fiebre	my temperature is nine degrees	1	1	4	i had a temperature of 39 degrees	i have a fever of thirty nine degrees	i have a fever of thirty nine degrees
tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	3	3	3	i have a temperature of 38 degrees	i have a fever of thirty eight	i have a fever of thirty eight degrees
he tenido treinta y ocho grados de fiebre	my temperature was one hundred one degrees	3	3	4	my temperature was 38 degrees	my temperature was thirty eight degrees	i have had a fever of thirty eight degrees
tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	3	3	3	i have a temperature of 38 degrees	i have a fever of thirty eight	i have a fever of thirty eight degrees
tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	3	3	4	my temperature is 38 degrees	my temperature if thirty eight degrees	i had a temperature of thirty eight degrees

tengo treinta y ocho grados de fiebre	my temperature is one hundred one degrees	3	3	3	my temperature is 38 degrees	my temperature is thirty eight degrees	i have a temperature of thirty eight degrees
tengo un poco de tos	i have a bit of a cough	4	4	3	i have a slight cough	yes i have a bit of a cough	i have a slight cough
el estómago me duele un poco	i experience a bit of pain in the stomach	3	4	3	i have a slight stomachache	yes i have a bit of a stomachache	i have a bit of a stomachache
tengo un poco de fiebre	i have a bit of fever	4	4	3	i have a slight cough	yes i have a bit of a cough	i have a slight cough
tengo un poco de tos	i have a bit of a cough	4	4	3	i have a slight cough	yes i have a bit of a cough	i have a slight cough
me duele desde hace una semana	i have experienced the pain for one week	3	4	3	i have had the pain for a week	i have had the pain for a week	i have had the pain for a week

Annexe 9 : Questionnaires distribués aux utilisateurs lors de l'EXPQ1

Questionnaire pour les patients

1 – Preguntas generales

1. Sexo: ☐ Masculino ☐ Femenino
2. Edad: _____ años
3. País de origen: _____

Tiene hijos

☐ sí ☐ no

Estoy acostumbrado a trabajar con una computadora

☐ sí ☐ no

Ya he usado un sistema de reconocimiento del habla

☐ sí ☐ no

2. – Comparación de los sistemas

MedSLT 1 – se puede usar frases enteras (con verbo)

MedSLT 2 –no se puede usar frases enteras, solo frases incompletas

Prefiero MedSLT 1 o MedSLT 2 (*marque el sistema que le conviene más*) por las razones siguientes:

MedSLT 1	Completamente en desacuerdo	En desacuerdo	Ns / Nc ¹	De acuerdo	Completamente de acuerdo
Es fácil aprender a usar el sistema MedSLT 1	☐	☐	☐	☐	☐
Las instrucciones de cómo usar MedSLT 1 son claras	☐	☐	☐	☐	☐
Las instrucciones sobre cómo formular las respuestas son muy útiles	☐	☐	☐	☐	☐
La ventana de ayuda es útil para aprender a formular las respuestas	☐	☐	☐	☐	☐

¹ Ns / Nc = No sabe / No contesta

MedSLT 2	Completamente en desacuerdo	En desacuerdo	Ns / Nc	De acuerdo	Completamente de acuerdo
Es fácil aprender a usar el sistema MedSLT 2	☐	☐	☐	☐	☐
Las instrucciones de cómo usar MedSLT 2 son claras	☐	☐	☐	☐	☐
Las instrucciones sobre cómo formular las respuestas son muy útiles	☐	☐	☐	☐	☐
La ventana de ayuda es útil para aprender a formular las respuestas	☐	☐	☐	☐	☐

3. – MedSLT en general

	Completamente en desacuerdo	En desacuerdo	Ns / Nc	De acuerdo	Completamente de acuerdo
Es fácil aprender a usar el sistema MedSLT	☐	☐	☐	☐	☐
MedSLT es suficientemente rápido	☐	☐	☐	☐	☐
Me ha gustado el diseño de la interface	☐	☐	☐	☐	☐
Me gustaría tener este sistema si no hay un intérprete disponible en el hospital para interpretar	☐	☐	☐	☐	☐
Un sistema como MedSLT puede mejorar la atención a los pacientes hispano hablantes	☐	☐	☐	☐	☐
Un sistema como MedSLT mejora la comunicación entre pacientes y médicos	☐	☐	☐	☐	☐

4. – Como intérprete que tiene la costumbre de interpretar para pacientes del hospital “Children’s” ¿qué cree que opinarán los padres de niños enfermos sobre MedSLT?:

	Completamente en desacuerdo	En desacuerdo	Ns / Nc	De acuerdo	Completamente de acuerdo
Para los padres MedSLT es fácil de usar	☐	☐	☐	☐	☐
Los padres se sentirán demasiado limitados por no poder contestar libremente	☐	☐	☐	☐	☐
Para los pacientes, el sistema mejora la comunicación entre pacientes y médicos	☐	☐	☐	☐	☐
Los pacientes estarán satisfechos con el sistema	☐	☐	☐	☐	☐

5. Comentarios técnicos

	Completamente en desacuerdo	En desacuerdo	Ns / Nc	De acuerdo	Completamente de acuerdo
No me gustó usar un micrófono y auriculares	☐	☐	☐	☐	☐
No me gustó compartir el ratón con el médico	☐	☐	☐	☐	☐
No me gustó deber apretar el botón antes de hablar	☐	☐	☐	☐	☐

Me gustó:

No me gustó:

Questionnaire Médecin

Part 1- Background Demographics

4. What is your status as a physician?
- ☐ Resident PGY Year_____
 - ☐ Attending Years in practice_____
5. What is your gender?
- ☐ Male
 - ☐ Female
6. Do you speak a foreign language?
- ☐ Yes
 - ☐ No
7. About what proportion of your patients and families have limited English proficiency? _____%

Part 2 – General questions

I am familiar with computers	☐ yes	☐ no
I have used a speech recognition tool before	☐ yes	☐ no
I frequently see patients with pharyngitis	☐ yes	☐ no

Part 3 – Usability questionnaire for MedSLT

		Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
1	It was easy to learn how to use MedSLT	☐	☐	☐	☐	☐
2	The instructions explaining how to use MedSLT were clear	☐	☐	☐	☐	☐
3	The instructions were clear on what the system can and cannot understand	☐	☐	☐	☐	☐
4	When the system did not understand me, the help module usually showed me another way to ask the question	☐	☐	☐	☐	☐
5	The response time for MedSLT was satisfactory	☐	☐	☐	☐	☐
6	I was able to ask enough questions to be sure of my diagnosis	☐	☐	☐	☐	☐
7	I would use this system if professional medical interpreters were not available	☐	☐	☐	☐	☐
8	I could easily integrate a system of this type into clinical practice	☐	☐	☐	☐	☐
9	I would be concerned about patient safety issues (such as medical errors) if I used this system with real patients	☐	☐	☐	☐	☐
10	The system could be a time saver for physicians, if medical interpreters are unavailable	☐	☐	☐	☐	☐
11	If the system were more sophisticated and could address most urgent pediatric care issues, I might regularly use it as a tool for communication with patients and families with limited English proficiency.	☐	☐	☐	☐	☐
12	The system has the ability to improve patient care.	☐	☐	☐	☐	☐
13	The system has the ability to improve the communication with patients.	☐	☐	☐	☐	☐

Part 4 – Satisfaction Questionnaire

Please indicate your level of satisfaction for each component below.

		Very Dissatisfied	Dissatisfied	Neither Satisfied nor Dissatisfied	Satisfied	Very Satisfied
1	How satisfied were you with the ability to take an accurate history?	☐	☐	☐	☐	☐
2	How satisfied were you with the amount of time it took to take the patient's history?	☐	☐	☐	☐	☐
3	How satisfied were you with the quality of the patient-physician communication?	☐	☐	☐	☐	☐
4	How satisfied were you with the ease of use of the system?	☐	☐	☐	☐	☐
5	Please rate your overall satisfaction with this system.	☐	☐	☐	☐	☐
6	How satisfied would patients and their families be with the system?	☐	☐	☐	☐	☐

Part 5. - Comparing the two versions of the system

MedSLT 1: The patient could only answer with short incomplete sentences (without verb).

MedSLT 2: The patient could answer both with full sentences and short incomplete sentences (without verb)

1. I preferred to interact with the patient using MedSLT 1 or MedSLT 2 (*circle correct choice*) for following reasons:

2. I think my patients would be more comfortable using MedSLT 1 or MedSLT 2 (*circle correct choice*) for following reasons:

Part 6. - Open-Ended Questions

1. What are the strengths of this system?

2. What are the weaknesses of this system?

3. Please add any other observations that you believe might be useful in the development and implementation of this system for use in clinical practice.

4. Which technical set-up would be best suited for your working environment (headsets, microphone, PDA, laptop...)?

Annexe 10 : Résultats du questionnaire de satisfaction concernant MedSLT en général

Nous reprenons dans cette annexe, les principales réponses au questionnaire de satisfaction qui portait sur MedSLT de manière générale, sans distinction entre V1 et V2. Nous avons saisi l'occasion de l'expérience conduite dans le cadre de cette thèse pour poser des questions sur la satisfaction de l'utilisation de *MedSLT* en général, sans faire de distinction entre les deux versions, afin d'obtenir des réponses sur des questions d'ordre plus général sur l'interface, mais aussi sur la possibilité d'utiliser un logiciel de ce type dans la pratique hospitalière. Le tableau 1 reprend les questions traitées.

1	C'est facile d'apprendre à utiliser le système MedSLT.
2	MedSLT est suffisamment rapide.
3	L'interface me plaît.
4	J'aimerais avoir un système de ce type à disposition s'il n'y a pas d'interprète disponible à l'hôpital.
5	Un système comme MedSLT peut améliorer les soins apportés aux patients hispanophones.
6	Un système comme MedSLT améliore la communication entre patients et médecins.

Tableau 1: Questions générales sur MedSLT

La section quatre contenait des questions relatives à l'adéquation de *MedSLT* par rapport à leur expérience comme interprète pour le public cible à l'hôpital pédiatrique de Dallas (tableau 2).

1	MedSLT est facile à utiliser par les parents.
2	Les parents se sentiront trop limités de ne pas pouvoir répondre librement aux questions.
3	Pour les patients, le système améliore la communication entre patients et médecins.
4	Les patients seront satisfaits du système MedSLT.

Tableau 2 : Questions liées à l'expérience des répondants comme interprète

La dernière partie contenait des questions d'ordre technique (tableau 3) suivies d'un espace pour apporter des suggestions ou des critiques sur l'aspect matériel de *MedSLT*.

1	Je n'ai pas aimé devoir utiliser un microphone et un casque.
2	Je n'ai pas aimé devoir partager la souris avec le médecin.
3	Je n'ai pas aimé devoir cliquer sur le bouton avant de parler.

Tableau 3 : Questions d'ordre technique

Pour répondre aux cinq parties décrites, le questionnaire prévoyait une échelle de Likert allant de 1-5, exprimée de la manière suivante en espagnol « Discordar fuertemente » (Complètement en désaccord, 1), « No de acuerdo » (Pas d'accord, 2), « No sabe/No contesta » (Ne se prononce pas, 3), « De acuerdo » (D'accord, 4), « Convenir fuertemente » (Complètement d'accord, 5).

Le questionnaire destiné aux médecins était formé de six parties. La première partie récoltait les informations démographiques et la deuxième des informations générales portant (1) sur l'expérience préalable avec un outil de RV, et (2) sur la fréquence avec laquelle le répondant reçoit des patients souffrant de maux de gorge. Dans la troisième partie, les médecins répondaient à des questions ayant trait à l'utilité de *MedSLT* (cf. tableau 4).

1	C'est facile d'apprendre à utiliser le système MedSLT.
2	Les instructions sur comment utiliser MedSLT sont claires.
3	Les instructions décrivaient clairement ce que le système peut et ce qu'il ne peut pas comprendre.
4	Lorsque le système ne me comprenait pas, le module d'aide me montrait une autre façon de poser la question.
5	Le temps de réponse de MedSLT est satisfaisant.
6	J'ai pu poser assez de questions pour être sûr(e) de mon diagnostic.
7	J'utiliserais ce système si aucun interprète médical professionnel n'est disponible.
8	Je pourrais facilement intégrer un système de ce type dans ma pratique médicale.
9	L'utilisation de ce système avec des patients réels soulève des problèmes de sécurité comme celui des erreurs médicales.
10	Le système pourrait faire gagner du temps aux médecins, lorsqu'aucun interprète n'est disponible.
11	Si le système était plus sophistiqué et pouvait être utilisé dans les cas d'urgence pédiatrique, j'utiliserais régulièrement ce système pour communiquer avec des patients et des familles ayant une maîtrise limitée de l'anglais.
12	Le système peut améliorer les soins apportés aux patients.
13	Le système peut améliorer la communication avec les patients.

Tableau 4 : Extrait sur l'utilité de MedSLT tiré du questionnaire pour les médecins

Dans la quatrième partie, les médecins se prononçaient sur leur degré de satisfaction, en particulier sur : (1) la possibilité de faire l'anamnèse du patient, (2) le temps mis pour l'effectuer, (3) la qualité de la communication patient-médecin, (4) la facilité d'utilisation et la satisfaction d'utilisation en général. Dans la partie cinq, nous avons demandé aux médecins de choisir quelle version ils préféreraient et pourquoi, même si pour cette catégorie d'utilisateur la couverture des versions ne change pas. Enfin, dans la partie six, nous avons posé des questions ouvertes sur les forces et faiblesses du système. Nous les avons aussi invités à proposer des

améliorations du système pour le rendre utilisable dans un environnement hospitalier et du type d'équipement matériel (portable, PDA, micro, casque...) serait le plus adapté selon eux.

Pour revenir à la première remarque concernant les insuffisances du système, il ressort des commentaires que les médecins n'ont pas pu poser toutes les questions voulues. Cette impression est toutefois aussi confirmée par le résultat médiocre de 2,64 sur une échelle de 5 pour cette question dans le questionnaire : « *I was able to ask enough questions to be sure of my diagnosis* » (J'ai pu poser assez de questions pour être sûr(e) de mon diagnostic). Cette question fait partie d'une série de questions posées uniquement aux médecins, comme la suivante : « Je pourrais facilement intégrer un système de ce type dans ma pratique médicale », qui obtient un score également assez bas de 2,64 sur une échelle de 5. Il ressort aussi de cette série de questions (Q6-Q11 dans le tableau 6.8) que l'utilisation de ce système n'est pas sans soulever certaines craintes au niveau de la sécurité : cinq répondants sur huit ont répondu ne pas être à l'abri d'éventuelles erreurs de traduction en utilisant un système de ce type. Cependant, selon les réponses des médecins, tous envisageraient l'utilisation de ce type de système s'il était plus sophistiqué pour communiquer avec des patients non-anglophones. Ils déclarent aussi qu'en l'absence d'un interprète ce système leur ferait gagner du temps.

Comme les utilisateurs patients de notre expérience sont interprètes pour l'hôpital de Dallas, nous avons posé des questions quant à leur propre expérience avec le système, mais aussi par rapport à la façon dont ils pensent que le système va être perçu par les véritables patients, population avec laquelle ils sont au contact au quotidien. La figure A met en parallèle ces résultats.

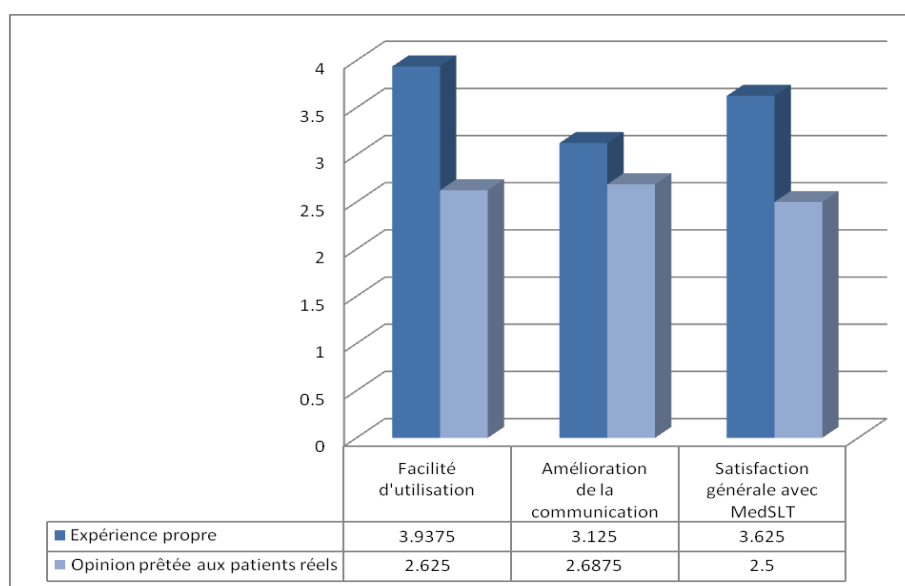


Figure A : Comparaison entre opinion propre et opinion prêtée aux patients réels

Il apparaît une nette différence entre l'opinion sur *MedSLT* par rapport à l'expérience propre des participants et celle qu'ils prêtent aux patients réels. L'argument le plus souvent invoqué est le niveau scolaire insuffisant des utilisateurs potentiels, qui ne seraient pas en mesure d'utiliser un logiciel de ce type. Les médecins ont une opinion similaire à ce sujet, puisque, selon eux, les patients véritables seraient seulement moyennement satisfaits (3,12 sur 5 sur l'échelle de Likert).

En outre, il est intéressant de constater qu'un certain nombre de remarques sont partagées par les médecins et les patients. Ainsi, les médecins répondent de manière générale dans le questionnaire que la couverture de *MedSLT* est encore trop limitée. Les patients aussi déclarent que le vocabulaire de *MedSLT* est encore trop limité (autant pour V1 que pour V2) par des remarques comme celles-ci : « Le système ne m'a pas permis d'expliquer complètement mes symptômes » ou « Le vocabulaire est très limité. ». Cependant, il n'en reste pas moins que médecins comme patients sont prêts à utiliser ce système en dernier recours, s'il n'y a pas d'alternatives. Enfin, dans les deux groupes, des répondants déclarent apprécier la facilité d'utilisation tandis que, selon d'autres, l'utilisation d'un tel système leur semble impossible, notamment dans le contexte d'une population peu lettrée et peu habituée à manipuler un ordinateur.

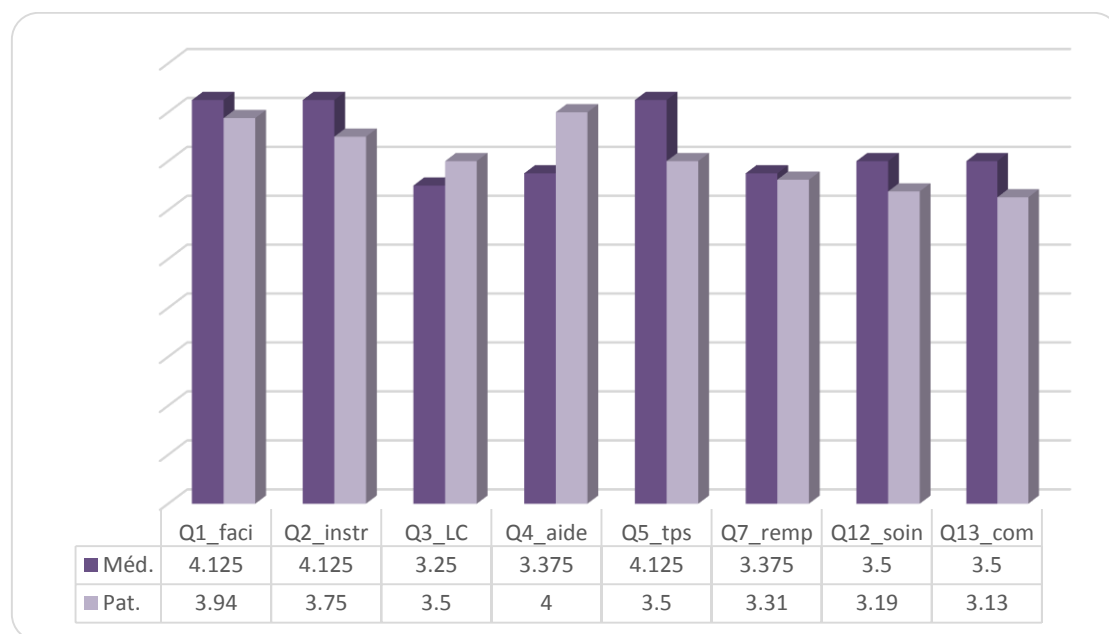


Figure B : Comparaison entre réponses des médecins et des patients

Si l'on compare les réponses aux questions posées aux deux groupes d'utilisateurs, on constate que, de manière générale, le score moyen pour les patients est légèrement inférieur pour toutes les questions, sauf pour les questions sur l'utilité des instructions de langage

contrôlé (Q3) et sur l'apport de l'aide en ligne (Q4), pour lesquelles les patients donnent de meilleurs scores. Sans comparer les deux versions, *MedSLT* jouit d'une bonne opinion concernant la facilité de l'utilisation du système avec 4,13 sur 5 pour les médecins et 3,94 sur 5 pour les patients.

Pour résumer les résultats obtenus s'agissant de l'évaluation de la satisfaction en général, on peut dire que les répondants médecins et patients ont plutôt une bonne opinion de *MedSLT* en général, mais qu'ils restent sceptiques quant à l'utilisation de ce système par de véritables patients.

Annexe 11 : Exemple de fichier d'évaluation Excel pour EXPQ2_S4-S6

Fluency and adequacy

Instruction 1: This evaluation is done in two stages. The evaluation starts with the page Fluency_evaluation_spa. Please rate the following sentences according to the scale given below. There is a picklist to help you filling the cells. Please rate these sentences as quickly as possible without referring to the second page Adequacy_eng_spa. Once you finished rating this page, please go to the second page Adequacy_eng_spa and follow the scale given for Adequacy on the second page.

Description for fluency	SCALE
Flawless spanish	5
Good spanish	4
Non-native spanish	3
Disfluent spanish	2
Incomprehensible spanish	1

TARGET	FLUENCY
el dolor disminuye	
desde cuándo tiene el dolor	
los antibióticos le dan alergia	
tiene tos	
vomita	
cuándo ha empezado el dolor de garganta	
tiene alergia a algunos antibióticos	
ha mucho tenido moco	
cuál es la intensidad del dolor en la garganta	
tiene una erupción cutánea	
alguien tiene el dolor de garganta	
ha consultado un médico	

le han realizado un examen	
desde cuándo tiene tos	
el estómago le ha dolido	
le han realizado la prueba rápida por estreptococo	
ha tenido la nariz tapada	
ha consultado un médico esta semana	
tiene la nariz tapada	
desde cuándo tiene fiebre	
tiene tos	
ha tenido fiebre	
tiene fiebre cuando tiene el dolor	
tiene alergia a alguna medicina	
ha tenido alergia a alguna medicina	
toma medicina para el dolor	
le ha dolido la cabeza	
desde cuándo le ha dolido la garganta	
desde cuándo le duele la garganta	
tiene fiebre	
ha mucho tenido moco	
le duele la garganta desde hace algunos días	
ha consultado un médico hace dos días	
ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo en la semana pasada	
ha tomado medicina	
ha tomado medicina para el dolor de garganta	
ha consultado un médico	
desde cuándo tiene la nariz tapada	
alguien en su entorno tiene el dolor de garganta	
tiene la nariz tapada	
ha tenido tos	
ha tenido alergia a alguna medicina	
ha vomitado	
ha consultado un médico	

ha consultado un médico	
ha estado cansado	
ha consultado un otro médico	
ha tomado medicina	
cuántos grados tiene	
qué le han diagnosticado	
cuál es la intensidad del dolor en la garganta	
cuántos grados ha tenido	
cuándo empieza el dolor de garganta	
ha estado en contacto con esa persona	
tiene alergia a algunos antibióticos	
desde cuándo le duele la cabeza	
tiene alergia a alguna medicina	
desde cuándo le duele la cabeza	
ha tenido alergia a algunos antibióticos	
tiene alergia a alguna medicina	
ha tenido alergia a alguna medicina	
tiene alergia a alguna medicina	
tiene alergia a algunos antibióticos	
tiene alergia a alguna medicina	
tiene alergia a alguna medicina	
ha tenido tos	
cuántos grados ha tenido	
ha estado en contacto con alguien que tiene el dolor de garganta	
cuál ha sido el resultado	
ha consultado recientemente un médico	
el estómago le duele	
tiene mucho una infección de la garganta por estreptococo	
tiene fiebre	
tiene el dolor	
tiene fiebre	
cuál ha sido el resultado del examen	
el dolor disminuye	
ha consultado un médico	

vomita	
le realizan un examen	
le duele la garganta hoy	
le han diagnosticado una infección de la garganta por estreptococo	
cómo es el dolor	
ha estado recientemente en contacto con alguien que tiene el dolor de garganta	
ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo	
le duele la cabeza	
desde cuándo tiene el dolor	
qué antibiótico le da alergia	
le han realizado un examen	
cuántos grados ha tenido	
cuándo ha tomado antibióticos	
cuál es la intensidad del dolor	
le han realizado la prueba rápida por estreptococo	
ha estado en contacto con alguien que tiene el dolor de garganta	
tiene la nariz tapada	
ha tenido una erupción cutánea	
ha consultado un médico esta semana	
ha tomado medicina	
el dolor disminuye	
le han realizado un cultivo de la garganta	
cuál ha sido el resultado del examen	
ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo	
desde cuándo le duele el estómago	
dónde le duele	
ha tomado medicina para el dolor	
ha tomado medicina para el dolor	

ha tenido fiebre	
ha consultado un médico para el dolor de garganta	
ha consultado un médico	
ha tomado antibióticos	
ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo en la semana pasada	
vomita	
dónde le duele	
ha tomado la penicilina	
le han realizado la prueba rápida por estreptococo	
medicina le da alergia	
qué le da alergia	
cuántos grados tiene	
desde cuándo tiene el dolor	
cuál es la intensidad del dolor	
tiene mucho moco	
cuántos grados tiene	
ha consultado un otro médico	
le han realizado un cultivo de la garganta	
le duele la garganta	
ha estado en contacto con alguien que tiene el dolor de garganta en la semana pasada	
ha consultado hoy un médico	
ha mucho tenido moco	
qué medicina toma	
tiene la nariz tapada	
le han realizado un examen	
le han realizado a menudo la prueba rápida por estreptococo	
el estómago le duele	
ha estado en contacto con alguien que tiene una infección de la garganta por estreptococo	
desde cuándo le duele la garganta	

tiene una infección de la garganta por estreptococo	
cuál es la intensidad del dolor	
le ha dolido la cabeza	
ha tenido fiebre	
le duele la garganta	
toma medicina	
desde cuándo le duele la garganta	
cuándo ha estado en contacto con una persona que tiene una infección de la garganta por estreptococo	
cuándo ha consultado un médico	
consulta un médico desde hace esta semana	
el dolor está aliviado cuando tiene fiebre	
el dolor disminuye	
toma medicina	
le han realizado un cultivo de la garganta	
ha consultado un médico	
tiene mucho moco	
le duele la cabeza	
cómo es el dolor	
ha estado en contacto con alguien que tiene el dolor de garganta	
cuál ha sido el resultado del cultivo de la garganta	
el resultado ha sido positivo	
tiene tos	
el estómago le duele	
el resultado ha sido positivo	
ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo	
ha estado recientemente en contacto con alguien que tiene una infección de la garganta por estreptococo	
le duele la garganta	

le han realizado un cultivo de la garganta	
tiene tos	
le duele la garganta	
tiene una erupción cutánea	
cuándo ha consultado un médico	
ha tenido tos	
tiene náuseas	
cuántos grados tiene	
desde cuándo le duele la garganta	
le han realizado la prueba rápida por estreptococo la noche	
cómo es el dolor	
ha estado en contacto con alguien que tiene el dolor de garganta	
ha tomado antibióticos	
tiene los ganglios inflamados	
el resultado del examen ha sido positivo	
desde cuándo le duele la garganta	
ha tenido fiebre	
tiene los ganglios inflamados	
ha tenido una erupción cutánea	
ha consultado recientemente un médico	
cuándo ha consultado una enfermera	
cuál ha sido el resultado del examen	
le ha dolido la garganta	
hola	
cómo es el dolor	
ha tomado medicina para el dolor	
cuál ha sido el resultado del examen	
le duele más a menudo	
ha estado en contacto con una persona enferma	
le han realizado la prueba rápida por estreptococo	
ha consultado un médico para el dolor de garganta	

ha estado en contacto con alguien que tiene el dolor de garganta	
tiene mucho moco	
le han realizado un examen	
ha consultado un otro médico	
el dolor disminuye	
toma medicina	
desde cuándo le duele la garganta	
el dolor disminuye	
desde cuándo tiene mucho moco	
cuándo ha empezado el dolor	
desde cuándo le duele la garganta	
le han realizado un examen	
le duele menos a menudo	
ha estado en contacto con alguien que tiene el dolor de garganta	
tiene tos ayer	
desde cuándo tiene fiebre	
cuándo ha empezado el dolor	
cuál ha sido el resultado	
tiene fiebre	
qué medicina ha tomado	
cuántos grados ha tenido	
le han realizado un examen	
el dolor es severo	
alguien en su entorno tiene el dolor de garganta	
cómo es el dolor	
el estómago le duele	
el estómago le ha dolido	
desde cuándo le duele la cabeza	
le han realizado la prueba rápida por estreptococo	
is the pain getting better	el dolor disminuye



UNIVERSITÉ
DE GENÈVE

FACULTÉ DE TRADUCTION
ET D'INTERPRÉTATION

Déclaration attestant le caractère original du travail effectué

J'affirme avoir pris connaissance des documents d'information et de prévention du plagiat émis par l'Université de Genève et la Faculté de traduction et d'interprétation (notamment la *Directive en matière de plagiat des étudiant-e-s*, le *Règlement d'études de la Faculté de traduction et d'interprétation* ainsi que l'*Aide-mémoire à l'intention des étudiants préparant un mémoire de Ma en traduction*).

J'atteste que ce travail est le fruit d'un travail personnel et a été rédigé de manière autonome.

Je déclare que toutes les sources d'information utilisées sont citées de manière complète et précise, y compris les sources sur Internet.

Je suis conscient-e que le fait de ne pas citer une source ou de ne pas la citer correctement est constitutif de plagiat et que le plagiat est considéré comme une faute grave au sein de l'Université, passible de sanctions.

Au vu de ce qui précède, je déclare sur l'honneur que le présent travail est original.

Nom et prénom :

Starlander Marianne

Lieu / date / signature :

Genève, le 08.03.2016

M. Starlander

Ce formulaire doit être dûment rempli par tout étudiant ou toute étudiante rédigeant un travail substantiel et remis à l'enseignant ou l'enseignante.