



Chapitre d'actes

2009

Open Access

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

Un système de traduction automatique paramétré par des atlas dialectologiques

Scherrer, Yves

How to cite

SCHERRER, Yves. Un système de traduction automatique paramétré par des atlas dialectologiques. In: Actes de TALN 2009. Senlis (France). [s.l.] : [s.n.], 2009. p. 10.

This publication URL: <https://archive-ouverte.unige.ch/unige:4696>

Un système de traduction automatique paramétré par des atlas dialectologiques

Yves Scherrer
LATL, Université de Genève
Rue de Candolle 5
1211 Genève 4, Suisse
yves.scherrer@unige.ch

Résumé. Contrairement à la plupart des systèmes de traitement du langage, qui s’appliquent à des langues écrites et standardisées, nous présentons ici un système de traduction automatique qui prend en compte les spécificités des dialectes. En général, les dialectes se caractérisent par une variation continue et un manque de données textuelles en qualité et quantité suffisantes. En même temps, du moins en Europe, les dialectologues ont étudié en détail les caractéristiques linguistiques des dialectes. Nous soutenons que des données provenant d’atlas dialectologiques peuvent être utilisées pour paramétrer un système de traduction automatique. Nous illustrons cette idée avec le prototype d’un système de traduction basé sur des règles, qui traduit de l’allemand standard vers les différents dialectes de Suisse allemande. Quelques exemples linguistiquement motivés serviront à exposer l’architecture de ce système.

Abstract. Most natural language processing systems apply to written, standardized language varieties. In contrast, we present a machine translation system that takes into account some specificities of dialects : dialect areas show continuous variation along all levels of linguistic analysis, and textual data is often not available in sufficient quality and quantity. At the same time, many European dialect areas are well studied by dialectologists. We argue that data from dialectological atlases can be used to parametrize a machine translation system. We illustrate this idea by presenting the prototype of a rule-based machine translation system that translates from Standard German into the Swiss German dialect continuum. Its architecture is explained with some linguistically motivated examples.

Mots-clés : Traduction automatique, dialectes, langues proches, langues germaniques.

Keywords: Machine translation, dialects, closely related languages, Germanic languages.

1 Introduction

La plupart des systèmes de traduction automatique ont comme objet des variétés de langues standardisées, écrites, et parlées par un grand nombre de locuteurs. Cette tendance s’explique d’une part par l’accès plus aisé à des ressources linguistiques, et d’autre part par une dissémination (ou commercialisation) plus facile du produit final.

Contrairement à ces systèmes, nous présentons ici un système de traduction automatique qui traduit vers des dialectes. Cette particularité pose quelques défis intéressants. Comme les dialectes

sont des variétés linguistiques parlées, il n'y a pas de règles orthographiques obligatoires, et peu de données écrites en général. Ensuite, les différents dialectes forment un continuum, avec des zones de transition (Chambers & Trudgill, 1998). Ces propriétés empêchent l'utilisation efficace de modèles standard de traduction automatique. Par contre, on peut déceler deux facteurs simplificateurs. D'une part, les dialectes sont en général dans une relation historique proche avec une langue standardisée. Une stratégie intéressante consiste donc à se baser sur des outils de traitement linguistique existants pour cette dernière. D'autre part, les dialectes ont été étudiés (du moins en Europe) de manière systématique par des dialectologues, et leurs résultats peuvent être exploitées dans un système de traduction automatique.

Cet article esquisse le développement d'un système de traduction automatique basé sur les règles, qui traduit de l'allemand standard vers n'importe quelle variété du continuum dialectal de la Suisse allemande. Le choix d'une aire large de variétés-cible implique que les règles de transfert ne sont pas appliquées de manière uniforme à travers cette aire, et que différentes variantes peuvent se chevaucher dans les zones de transition. Ces phénomènes seront modélisés à l'aide de données fournies par les atlas dialectologiques.

La section suivante résume la situation linguistique dans la partie germanophone de la Suisse, et présente également les données disponibles. Ces facteurs déterminent les choix méthodologiques que nous avons faits (et que nous détaillons en section 3). La section 4 illustre l'architecture du système en discutant quelques phénomènes linguistiques intéressants. Nous concluons la présentation par une revue d'articles de recherche connexes (Section 5), et par quelques réflexions sur des extensions futures du système (Section 6).

2 La situation linguistique en Suisse allemande

Bien que la Suisse soit un pays relativement petit, elle présente une configuration linguistique plutôt complexe. La Constitution fédérale suisse définit l'allemand, le français, l'italien et le romanche comme langues officielles. L'aire linguistique allemande (64% de la population suisse) est la plus étendue et couvre, grosso modo, les deux tiers du territoire suisse à partir du Nord-Est.

Il est communément admis que la configuration socio-linguistique de la partie germanophone est un cas-modèle de diglossie, c'est-à-dire un environnement dans lequel deux variétés linguistiques sont utilisées de manière complémentaire dans des contextes fonctionnellement différents. En Suisse allemande, les dialectes sont utilisés dans les situations de communication orale, tandis que l'allemand standard est employé presque exclusivement dans la communication écrite. A la différence d'autres aires diglossiques, le choix de la variété linguistique ne dépend pas du statut social des locuteurs, mais uniquement du moyen de communication. Il s'ensuit que tous les Suisses allemands de naissance sont bidialectaux : ils maîtrisent leur dialecte local et l'allemand standard (et comprennent en plus les dialectes des autres régions de Suisse allemande).

Des données écrites en dialecte ont toutefois été produites dans différents contextes : il existe une littérature dialectale en Suisse allemande, et des enregistrements sonores ont été transcrits à des fins de recherche scientifique. De plus, on écrit souvent en dialecte dans les médias électroniques (SMS, E-mail, *chat*). Cependant, ces données sont très hétérogènes et proviennent de différents dialectes. Elles sont donc difficilement utilisables dans des approches basées sur des corpus. En outre, les besoins de traduction sont faibles dans une société diglossique, et des corpus parallèles sont donc quasi inexistantes.

La classification des dialectes suisse allemands se base en général sur des frontières administratives, confessionnelles et topographiques. Bien que ces frontières aient influencé les dialectes à des degrés divers, la classification qui en résulte ne correspond pas toujours à des réalités linguistiques. Notre modèle ne se base sur aucune classification à priori ; nous concevons l'aire dialectale suisse allemande comme un continuum au sein duquel certains phénomènes montrent des frontières plus nettes que d'autres.

Depuis le début du 20^{ème} siècle, le suisse allemand fait l'objet de recherches dialectologiques. A côté d'études de phénomènes dialectaux isolés, trois projets de grande envergure (couvrant tout le territoire germanophone en Suisse) ont été initiés. Dans le domaine de la lexicographie dialectale, le *Wörterbuch der schweizerdeutschen Sprache* est en élaboration depuis 1881 et devrait être terminé en 2020. Le *Sprachatlas der deutschen Schweiz* (SDS) couvre les différences phonétiques, morphologiques et lexicales sous la forme d'un atlas ; la collection des données et la publication de ses 8 volumes se sont échelonnées de 1939 à 1997 (Hotzenköcherle *et al.*, 1962 1997). L'absence de données syntaxiques dans le SDS a conduit à un projet successif appelé *Syntaktischer Atlas der deutschen Schweiz* (SADS), un atlas initié en 2000 et s'approchant de la publication. Dans notre projet, nous utilisons des données du SDS, et des résultats préliminaires du SADS.

3 Architecture du système et méthodologie

La diglossie en Suisse allemande suggère le développement d'un système de traduction multimodal, de l'allemand standard écrit vers le suisse allemand oral. Si cela peut constituer un objectif de recherche à long terme, cet article traite de la traduction vers du dialecte écrit.

Comme nous l'avons évoqué ci-dessus, les données textuelles ne sont pas suffisantes en quantité et en qualité pour être utilisées dans des modèles classiques de traduction automatique statistique. En revanche, nous comptons sur l'utilisation de données dialectales linguistiquement interprétées (sous forme d'atlas dialectologiques ou d'études théoriques), ce qui favorise une approche traditionnelle basée sur des règles de transfert. L'approche basée sur les règles a deux inconvénients majeurs, qui sont néanmoins atténués par la configuration linguistique que nous avons choisie. Premièrement, la multitude des règles de transfert et leurs interactions mutuelles sont difficiles à maîtriser ; mais dans notre cas, le nombre de règles est limité par la proximité linguistique entre l'allemand standard et les dialectes suisse allemands. Deuxièmement, cette approche nécessite des dictionnaires bilingues avec des informations linguistiques précises ; nous comptons l'éviter par la dérivation en temps réel de correspondances lexicales en utilisant des règles de correspondances phonétiques. Là encore, la proximité linguistique nous permet d'emprunter cette voie.

Le système proposé utilise une approche de transfert classique et consiste en trois opérations : l'analyse syntaxique du texte en langue source, le transfert, et la génération du texte en langue cible. Les opérations de transfert et de génération impliquent l'utilisation de cartes dialectologiques.

3.1 Analyse

Le texte source en allemand standard est analysé phrase par phrase, à l'aide de l'analyseur syntaxique *Fips* (Wehrli, 2007). Il en résulte une représentation structurée des phrases, contenant l'analyse en constituants, des informations morphologiques et de lemmatisation ainsi qu'une structure prédicat-arguments. Cette représentation structurée est basée sur un format XML, ce qui nous permet de créer des variations du système. Par exemple, il sera possible de remplacer *Fips* avec d'autres outils d'analyse, ou d'utiliser directement des phrases analysées provenant d'une banque d'arbres allemande.

3.2 Transfert

Le module de transfert modifie le document XML résultant de l'analyse. Il contient trois types de règles de transfert : les règles lexicales, syntaxiques et phonétiques.¹

Les règles syntaxiques sont déclenchées en fonction de configurations syntaxiques spécifiques, indépendamment des mots qui se trouvent dans le texte. Le SADS se base sur 118 questions, dont plusieurs traitent du même phénomène. Nous estimons donc qu'une couverture satisfaisante nécessitera une quarantaine de règles ; à l'heure actuelle, nous avons implémenté une dizaine de règles, dont celle illustrée dans la section 4.1.

Les règles phonétiques ont une portée plus locale et s'appliquent à des mots d'origine étymologique commune. Elles permettent de remplacer des phonèmes (ou graphèmes) spécifiques. Les données utilisées par (Scherrer, 2007), portant sur un seul dialecte suisse allemand, suggèrent qu'environ 17% des mots² sont identiques, et qu'environ 75% peuvent être inférés à l'aide de règles phonétiques. Pour l'implémentation, nous visons un ensemble qui dépasse légèrement les 50 règles utilisées dans (Scherrer, 2007) ; ces règles sont sélectionnées parmi les 300 cartes phonétiques du SDS. La section 4.2 en donne un exemple.

Les règles lexicales servent à traduire les correspondances non reliées étymologiquement, ainsi que de nombreux mots à haute fréquence (mots grammaticaux et lexicaux) dont les idiosyncrasies ne permettent pas un traitement adéquat avec des règles phonétiques. Les transformations lexicales se répercutent souvent sur le contexte : le changement d'un lexème nominal peut déclencher un changement de genre sur son article,³ et le changement d'un lexème verbal peut déclencher un changement dans sa structure argumentale. Nous sélectionnons les règles lexicales en fonction des données présentes dans les atlas et en fonction de la fréquence d'utilisation des mots selon (Ruoff, 1981).

Quelques-unes de ces règles sont générales et s'appliquent uniformément à l'aire dialectale suisse allemande entière. Mais la plupart des règles présentent plusieurs variantes régionales. Ainsi, on associe à chaque règle un ensemble de cartes probabilistes. Chaque carte représente l'étendue géographique d'une variante de règle. A chaque point, la somme des probabilités des variantes d'une règle est 1. Cette architecture tient donc compte de zones de transition où plusieurs variantes sont utilisées simultanément. Dans les zones de transition, le système ne proposera donc pas une traduction unique, mais plusieurs traductions avec des probabilités

¹Ces règles sont implémentées en *Python* et utilisent la fonctionnalité *XPath* du module *lxml*.

²Les statistiques portent sur des types de formes fléchies.

³Il peut également y avoir des modifications à plus longue distance, par exemple au niveau des pronoms anaphoriques, ou dans certains dialectes au niveau des adjectifs prédicatifs.

différentes. Le choix du dialecte cible est effectué par l'utilisateur. Le système traduit donc toujours vers un seul dialecte spécifié par ses coordonnées.

Les cartes pour les règles phonétiques et lexicales sont compilées à partir de données du SDS. Malheureusement, ces cartes ne sont pas disponibles en format numérique et doivent être numérisées à la main. Pour les règles syntaxiques, les cartes sont créées par l'auteur à partir des données préliminaires du projet SADS.

3.3 Génération

Les fichiers XML obtenus par le module de transfert contiennent les structures syntaxiques et les racines des mots selon le dialecte choisi par l'utilisateur. L'objectif du module de génération est donc de créer les formes fléchies sur la base des racines dialectales et des traits morphologiques contenus dans la structure XML. Les règles de génération peuvent aussi être géoréférencées, c'est-à-dire associées à des cartes de probabilités. Cette approche nous permet de tenir compte des divergences régionales dans les paradigmes flexionnels. Actuellement, notre système contient les paradigmes des noms, des verbes et des adjectifs réguliers ainsi que des déterminants ; ces règles ne sont pas encore géoréférencées. Ultérieurement, le module de génération tiendra également compte des phénomènes de liaison et d'élision.

4 Quelques règles de transfert

Dans cette section, nous illustrons l'architecture de notre système de traduction automatique avec deux exemples de règles de transfert utilisant des cartes de probabilités.

4.1 Marquage prépositionnel du datif

En allemand standard, le datif est réalisé morphologiquement dans la flexion du déterminant, parfois également dans la flexion du substantif. Le datif sert (entre autres) au marquage des objets indirects (1). Dans la plupart des dialectes suisse allemands, la situation est analogue. Cependant, dans certains dialectes de Suisse centrale, les objets indirects ne sont pas simplement marqués par le datif morphologique, mais également par une préposition (Seiler, 2002). Selon la région, c'est la préposition *a* ou *i* qui est utilisée (2).⁴ Par contre, les datifs à l'intérieur d'un syntagme prépositionnel ne sont pas affectés par ce marquage supplémentaire : (3) en allemand standard, (4) en dialecte de Suisse centrale. La Figure 1 montre les distributions géographiques des trois variantes de marquage du datif.

- (1) *Er schreibt der Mutter.*
il écrit la-DAT mère
'il écrit à la mère'

- (2) *Er schribt a/i de Mueter.*
il écrit à/dans la-DAT mère

⁴Ces deux prépositions sont les équivalents formels des prépositions allemand standards *an* 'à' et *in* 'dans', mais ont perdu leur sens locatif.

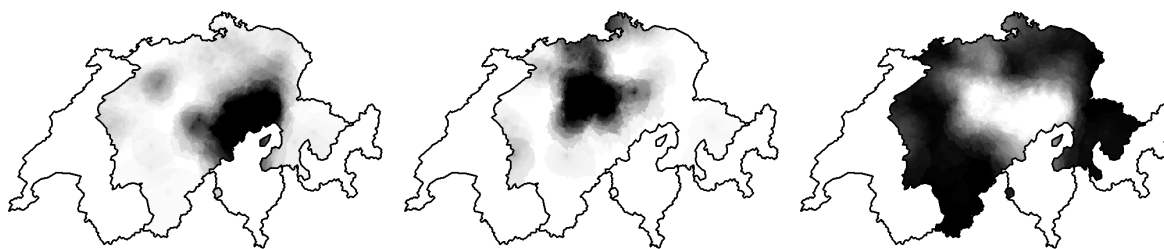


FIG. 1 – Distribution du marquage prépositionnel du datif en Suisse allemand (SADS I/4, variante la plus naturelle). Les zones noires représentent des probabilités élevées, les zones blanches des probabilités basses. A gauche, la distribution du marquage avec la préposition *a* ; au milieu, la distribution du marquage avec la préposition *i* ; à droite, les zones sans marquage prépositionnel du datif.

- (3) *mit der Mutter*
avec la-DAT mère
- (4) *mit (*a/i) de Mueter*
avec la-DAT mère

La règle de transfert correspondante se base sur l'analyse syntaxique fournie par *Fips*. Une représentation schématique de l'analyse de (1) se trouve dans la Figure 2.⁵

D'abord, il faut tester si la règle de transfert s'applique à la phrase, ou en d'autres termes, si la phrase contient des DP au datif qui ne sont pas régis par une préposition. Si de tels syntagmes existent, le système charge les trois cartes et détermine les probabilités des trois variantes dans le dialecte-cible choisi par l'utilisateur. Si plusieurs variantes sont utilisées à cet endroit, chaque variante est appliquée à une copie différente de la structure de la phrase. Par exemple, pour la variante avec la préposition *i*, l'algorithme de transformation est le suivant : "A la position de chaque noeud DP au datif non gouverné par un PP, créer un nouveau noeud PP, lui attacher une tête prépositionnelle avec le mot *i*, et attacher le noeud DP original en tant que complément droit du PP." Enfin, on ajuste la probabilité de l'analyse. La structure modifiée, après application de la règle de transfert, est montrée dans la Figure 2.

4.2 Traitement du *-nd* final

Nous voudrions montrer ici un exemple de règle de transformation phonétique. Ce type de transformation est particulier à notre système et ne se retrouve pas dans des systèmes de traduction automatique génériques. Les transformations phonétiques seront exemplifiées à l'aide du mot allemand *Hund* 'chien', dont le groupe consonantique final *nd* est sujet à modification en suisse allemand. Les cartes de la Figure 3 montrent les quatre variantes de prononciation qui existent en Suisse allemande.

Les règles de transformation phonétique sont déclenchées après les règles syntaxiques, car elles opèrent de manière locale et n'interfèrent pas avec le contexte grammatical. En partant

⁵L'analyseur *Fips* suit l'hypothèse DP – le déterminant se trouve donc à la tête des syntagmes "nominaux", appelés plus précisément "syntagmes déterminants". Les étiquettes des noeuds finaux suivent les conventions du *Stuttgart-Tübingen-TagSet*.

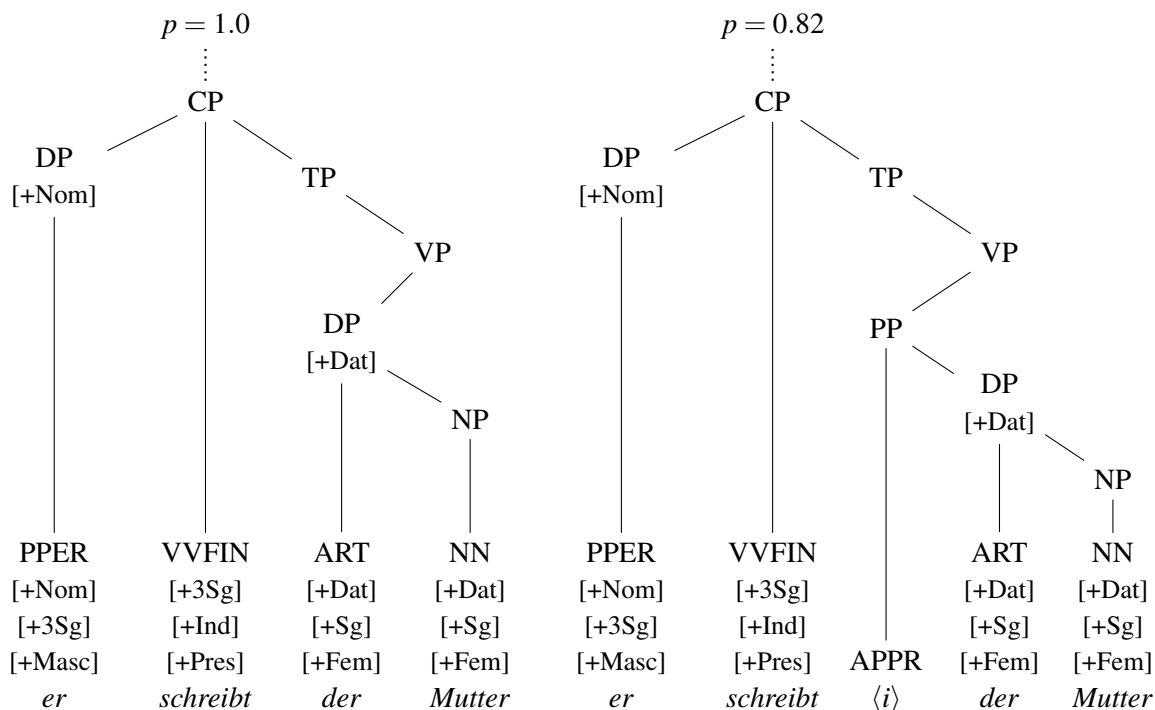


FIG. 2 – Structures syntaxiques des exemples (1) à gauche, et (2) à droite. Seule la préposition *i* est présentée dans sa forme dialectale, les autres mots seront transformés par d’autres règles. La probabilité est valable pour la ville d’Einsiedeln, où la variante avec la préposition *a* occupe les 18% restants.

de la racine du mot allemand standard, les différentes règles phonétiques sont appliquées séquentiellement afin de construire la racine dialectale.⁶ Ces règles dépendent évidemment de la structure phonétique du mot, mais aussi de sa partie du discours et de ses traits morphologiques.

Les transformations phonétiques opèrent d’une manière similaire aux autres types de transformations : le système détecte les contextes auxquels la règle s’applique, charge les cartes de probabilités des différentes variantes, extrait les probabilités des variantes correspondant aux coordonnées du dialecte-cible, et applique une expression régulière de substitution pour chaque variante, créant plusieurs copies selon les besoins.

La transformation phonétique de *nd* est assez régulière et s’applique, avec peu de changements dans la distribution géographique, à d’autres mots. Ceci veut dire que les mêmes cartes peuvent servir à prédire la prononciation d’autres mots comme *Hand* ‘main’ ou *gesund* ‘sain’. De telles généralisations limitent le nombre de cartes en exploitant la proximité linguistique entre les dialectes suisse allemands et la variété standard de l’allemand.

⁶Par exemple, une autre règle phonétique déterminerait la qualité de la voyelle du mot *Hund*.



FIG. 3 – Transformation du *nd* final dans le mot *Hund* (SDS II/120). La prononciation *nd* est maintenue dans la plupart des régions de Suisse allemande (carte en haut à gauche). *ng* ([ŋ] en écriture phonétique) est utilisé dans la région de Berne (carte en haut à droite). *n* est utilisé dans la région fribourgeoise et dans quelques régions du Sud-Est (carte en bas à gauche). *nt* est limité aux régions alpines (carte en bas à droite).

5 Travaux connexes

Le mémoire de (Forst, 2002) s'apparente probablement le plus à notre recherche. Il propose un système de traduction automatique de l'allemand standard vers le dialecte zurichois, basé sur le cadre théorique de la LFG. Toutefois, l'utilisation d'une théorie standard l'empêche de bénéficier de la proximité linguistique des deux variétés linguistiques traitées, et il ne tient pas compte de la diversité dialectale en Suisse allemande. Un autre projet de traduction automatique concerne le dialecte italien Veneto, qu'il est prévu de traduire vers l'anglais en utilisant l'italien standard comme langue pivot.⁷ Les dialectes suisse allemands ont aussi fait l'objet de recherches en synthèse de la parole (Siebenhaar *et al.*, 2004).

Un champ d'études connexe est l'analyse syntaxique de dialectes. (Vaillant, 2008) propose un modèle factorisé pour analyser les différents créoles antillais à base française. Son modèle est basé sur le framework TAG et utilise un paramètre numérique pour sélectionner le créole. La topographie de l'aire dialectale antillaise simplifie le modèle puisqu'il n'existe pas de zones de transition entre les différentes îles. (Chiang *et al.*, 2006) présentent une grammaire synchrone pour l'arabe standard moderne et le dialecte arabe du Levant. Leur approche est essentiellement basée sur les corpus, mais inclut des adaptations manuelles.

La traduction entre variétés dialectales peut être vue comme un cas de traduction entre langues proches. Dans ce domaine, on peut citer des travaux sur différentes langues slaves (Hajic *et al.*, 2003) ou sur les langues romanes de l'Espagne (Corbi-Bellot *et al.*, 2005).

Une sous-tâche de la traduction automatique est la création de correspondances lexicales bilingues, généralement extraits d'un corpus parallèle à l'aide de techniques d'alignement. Pour des langues proches, des approches basées sur la similarité phonétique peuvent être utilisées (Mann & Yarowsky, 2001; Kondrak & Sherif, 2006). À côté de méthodes d'apprentissage automatique, (Scherrer, 2007) utilise des modèles de similarité basés sur les règles afin d'induire des correspondances lexicales entre l'allemand standard et le dialecte bernois.

Du côté de la dialectologie, nous avons déjà mentionné les atlas dialectologiques que nous utilisons. Pour améliorer la couverture lexicale, il pourrait être intéressant d'utiliser des ouvrages de lexicographie dialectale. Si de tels dictionnaires ne sont pas disponibles en format numérique pour la Suisse, le projet *dbo@ema* prévoit de créer une base de donnée lexicale géoréférencée

⁷R. Delmonte, communication personnelle.

pour les dialectes autrichiens (Wandl-Vogt, 2008).

Les méthodes computationnelles sont aussi utilisées en dialectométrie, afin de calculer les différences entre différents dialectes à l'aide de mesures numériques objectives. Si ces différences peuvent être mesurées sur tout type de donnée linguistique, l'approche la plus pratique est de comparer des mots de différents dialectes avec des méthodes de distance d'édition (Nerbonne & Heeringa, 2001). Les dialectes peuvent ensuite être classifiés moyennant une analyse de cluster.

6 Perspectives et conclusion

Nous avons présenté le prototype d'un système de traduction automatique qui tient compte des spécificités de la traduction de l'allemand standard vers le suisse allemand. On peut distinguer deux idées principales dans ce modèle.

La première idée consiste à modéliser un continuum entier de variétés dialectales (plutôt que de se limiter à un ou deux dialectes prototypiques). En conséquence, nous proposons d'utiliser des données récoltées par des dialectologues afin de paramétrer les règles de transfert.

La seconde idée découle du fait que tous les dialectes suisse allemands partagent une base linguistique commune avec l'allemand standard. Comme des ressources linguistiques dialectales sont difficiles à obtenir ou à créer, nous proposons d'exploiter au maximum les ressources existantes pour l'allemand standard. Par exemple, beaucoup de correspondances lexicales ne sont pas introduites explicitement dans un lexique, mais sont dérivées automatiquement à l'aide de règles phonétiques.

Nous avons également présenté la nature ouverte de notre système et les possibilités d'extension qui résultent de ce choix. Peut-être l'extension la plus intéressante est celle de créer une banque d'arbres pour quelques dialectes suisse allemands, qui pourraient ensuite rendre possible l'utilisation de méthodes statistiques. En particulier, des modèles statistiques pourraient s'avérer plus robustes pour la direction de traduction allant des dialectes vers l'allemand standard.

En ce moment, le système présenté contient une vingtaine de règles de transfert. La moitié de ces règles s'applique à l'aire dialectale entière sans distinction géographique, l'autre moitié est géoréférencée. La couverture du système est donc encore extrêmement limitée, et une évaluation de la performance ne serait pas pertinente à ce stade du projet. Néanmoins, avec la couverture et la complexité croissante du système, des moyens d'évaluation doivent être considérés. Une possibilité est d'utiliser un texte parallèle existant dans plusieurs variétés dialectales. Un tel texte existe pour 70 points d'enquête, mais il est relativement court (environ 250 mots). Une autre possibilité est de recourir à une évaluation manuelle, en se basant sur un réseau d'informants. Cependant, il est difficile de prévoir la qualité d'une telle évaluation puisque les dialectophones ne sont pas habitués à juger des textes dialectaux, et encore moins des textes écrits. De toute façon, en l'absence d'un grand corpus pluridialectal, toutes les évaluations resteront qualitatives.

Actuellement, la tâche la plus importante est d'étendre la couverture du système selon les critères esquissés ci-dessus. Nous allons sélectionner un ensemble de phénomènes dialectaux importants afin de limiter le travail de préparation et de numérisation des données. Ainsi, nous espérons atteindre une couverture satisfaisante malgré nos ressources limitées.

Références

- CHAMBERS J. K. & TRUDGILL P. (1998). *Dialectology*. Cambridge : Cambridge University Press, 2 edition.
- CHIANG D., DIAB M., HABASH N., RAMBOW O. & SHAREEF S. (2006). Parsing Arabic dialects. In *Proceedings of EACL'06*, p. 369–376, Trento.
- CORBI-BELLOT A. M., FORCADA M. L., ORTÍZ-ROJAS S., PÉREZ-ORTIZ J. A., RAMÍREZ-SÁNCHEZ G., SÁNCHEZ-MARTÍNEZ F., ALEGRIA I., MAYOR A. & SARASOLA K. (2005). An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In *Proceedings of EAMT'05*, p. 79–86, Budapest.
- FORST M. (2002). La traduction automatique dans le cadre formel de la LFG – un système de traduction entre l'allemand standard et le zurichois. In *Publications du CTL*, volume 41. Université de Lausanne.
- HAJIC J., HOMOLA P. & KUBON V. (2003). A simple multilingual machine translation system. In *Proceedings of the Machine Translation Summit XI*, p. 157–164, La Nouvelle-Orléans.
- R. HOTZENKÖCHERLE, R. SCHLÄPFER, R. TRÜB & P. ZINSLI, Eds. (1962-1997). *Sprachatlas der deutschen Schweiz*. Berne : Francke.
- KONDRAK G. & SHERIF T. (2006). Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proceedings of the ACL'06 Workshop on Linguistic Distances*, p. 43–50, Sydney.
- MANN G. S. & YAROWSKY D. (2001). Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL'01*, Pittsburgh.
- NERBONNE J. & HEERINGA W. (2001). Computational comparison and classification of dialects. In W. VIERECK & H. GWOSDEK, Eds., *Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics*, number 9, p. 69–83. Alessandria : Edizioni dell'Orso.
- RUOFF A. (1981). *Häufigkeitwörterbuch gesprochener Sprache*. Tübingen : Niemeyer.
- SCHERRER Y. (2007). Adaptive string distance measures for bilingual dialect lexicon induction. In *Proceedings of ACL'07, Student Research Workshop*, p. 55–60, Prague.
- SEILER G. (2002). Prepositional dative marking in Upper German : a case of syntactic microvariation. In S. BARBIERS, L. CORNIPS & S. VAN DER KLEIJ, Eds., *Syntactic Microvariation*, volume II. Amsterdam : Meertens Institute Electronic Publications in Linguistics.
- SIEBENHAAR B., FORST M. & KELLER E. (2004). Prosody of Bernese and Zurich German. what the development of a dialectal speech synthesis system tells us about it. In P. GILLES & J. PETERS, Eds., *Regional Variation in Intonation*, p. 219–238. Tübingen : Niemeyer.
- VAILLANT P. (2008). Grammaires factorisées pour des dialectes apparentés. In *Actes de TALN'08*, p. 159–168, Avignon.
- WANDL-VOGT E. (2008). An der Schnittstelle von Dialektwörterbuch und Sprachatlas : Das Projekt "Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema)". In S. ELSPASS & W. KÖNIG, Eds., *Germanistische Linguistik 190-191. Sprachgeographie digital. Die neue Generation der Sprachatlanten*, p. 197–212. Hildesheim : Olms.
- WEHRLI E. (2007). Fips, a "deep" linguistic multilingual parser. In *Proceedings of the ACL'07 Workshop on Deep Linguistic Processing*, p. 120–127, Prague.