Vers une exploitation des liens morphologiques multilingues

Cartoni, Bruno

Bruno Cartoni février 2003

Vers une exploitation des liens morphologiques multilingues

Mémoire présenté à l'École de Traduction et d'Interprétation en vue de l'obtention d'un diplôme d'études approfondies en

Traitement informatique multilingue

Directeur du Mémoire :	Prof. M. Daniel-King
1er juré :	Prof. S. Armstrong
2ème juré :	Mme P. Bouillon

- Alors, tu parles japonais avant de parler français ?

- Non. C'est la même chose

Pour moi, il n'y avait pas des langues, mais une seule et grande langue, dont on pouvait choisir les variantes japonaises ou françaises, au gré de sa fantaisie.

Amélie Nothomb, Métaphysique des tubes

Remerciements

Au terme de ce travail, nous aimerions vivement remercier toutes les personnes qui de près ou de loin ont contribué à sa réalisation.

Nos remerciements vont d'abord à Madame Margaret Daniel-King, qui a bien voulu diriger le présent travail, et à Madame Susan Armstrong qui a accepté d'être membre du jury. Une pensée chaleureuse pour Pierrette Bouillon, également membre du jury, pour son écoute, ses commentaires et ses encouragements. Sans elle, ce travail n'existerait pas.

De plus, la relecture et les corrections finales ont été assurée par un certain nombre de lecteurs anonymes, qu'ils en soient remerciés. Nous voudrions également remercier Christine Favre, psychologue et psycholinguiste, pour sa relecture attentive des sections sur la psycholinguistique, et Sylvain Neuvel, pour ces commentaires judicieux concernant la morphologie holistique. Un grand merci également à Dominique Couturier pour ses précieux conseils concernant la statistique et les méthodes de comptage de l'enquête du chapitre 3, ainsi qu'à Carole Tiberius pour ces références fournies dans les premières heures de ce travail.

Nous remercions également chaleureusement Ludovic Tanguy pour son aide à l'adaptation de Webaffix pour l'italien.

Nous remercions également les répondants de l'enquête, qui ont bien voulu donner un peu de leur temps pour participer à celle-ci.

Un merci chaleureux également à Bruna Novellas Vall et Maria Georgescul pour leur aide sur les lexiques espagnol, catalan et roumain.

Enfin, merci à tous les membres de l'équipe ISSCO/TIM, à Guy, à mes proches et à ma famille, pour leur soutien de tous les instants.

Sommaire

Introduction

Le traitement du lexique constitue une des pierres d'achoppement du traitement automatique de la langue. De nombreuses approches ont été étudiées pour formaliser et organiser les informations contenues dans le lexique (Ritchie *et al.*, 1992).

Le défi principal est l'expression des régularités morphologiques. En effet, si certains préfèrent une approche énumérative (consistant en l'énumération de toutes les formes de tous les mots), d'autres ont tenté d'adopter une approche polymorphique (Sproat , 1992). Cette dernière consiste en la généralisation maximale des informations morphologiques pour obtenir un module de traitement du lexique moins volumineux et plus puissant. Tout l'enjeu d'une telle approche est d'étudier l'expression des régularités pour en généraliser au maximum le contenu.

Dans le présent travail, nous étudierons la possibilité d'exploiter d'autres généralités, les liens morphologiques multilingues.

Des langues historiquement proches possèdent en effet de nombreux points communs, dans leur morphologie comme dans leur syntaxe et leur sémantique (Geysen , 1989). Par exemple, il existe une similitude intuitive entre les mots *possibilité* en français et *possibilità* en italien. Une telle similitude implique également qu'un locuteur francophone qui n'a que très peu de connaissances en italien pourra néanmoins déduire que le mot équivalent en italien du français *fraternité* sera sans doute *fraternità*.

La construction d'un lexique prenant en compte ces similitudes offre de nombreux avantages, notamment techniques, quant à la robustesse et à la taille de ce genre de lexique (en comparaison avec des lexiques monolingues pris séparément).

Dans le présent travail, nous nous proposons d'étudier les différentes possibilités d'exploitation de ces régularités.

Pour ce faire, nous ferons d'abord un tour d'horizon des différentes manières de traiter le lexique, et présenterons les différentes informations morphologiques et les autres considérations informatiques à prendre en compte (chapitre 1). Nous expliquerons ensuite les différents liens morphologiques multilingues qui existent entre le français et l'italien, en argumentant sur l'existence et l'intérêt de ces liens sur le plan linguistique et informatique (chapitre 2). Nous étofferons les considérations théoriques du chapitre 2 par une enquête dont nous présenterons les résultats au chapitre 3. Ensuite, nous présenterons un bref état de l'art de l'exploitation des liens morphologiques multilingues (chapitre 4).

Une fois posées les bases des différentes possibilités qu'offrent les liens morphologiques multilingues, nous présenterons, dans les chapitres 5, notre expérience pratique de construction d'une morphologique multilingue, en expliquant nos choix de formalisme et les différentes considérations qui ont dirigé cette construction.

Enfin, nous présenterons une petite évaluation quantitative et qualitative du lexique construit (chapitre 6), avant de tracer les grandes lignes d'éventuels développements futurs dans la conclusion.

La construction d'une morphologie multilingue n'est pas une fin en soi pour le présent travail mais une telle application pratique nous permet de mettre en pratique bon nombre de facteurs étudiés dans ce mémoire. Elle présente également un intérêt particulier pour souligner les grandes lignes d'investigations futures dans ce domaine.

Bonne lecture !

Chapitre 1 : Le lexique et l'expression des régularités

Dans cette première partie, nous passerons en revue les différents aspects du traitement du lexique et des informations inhérentes à celui-ci dans le cadre du traitement automatique des langues naturelles. Nous nous intéresserons plus particulièrement au traitement de la morphologie. Nous étudierons les différents types de morphologie, les phénomènes morphologiques et les problématiques qu'engendre leur traitement. Nous ferons également un détour vers la psycholinguistique pour voir dans quelle mesure cette discipline peut aider à la compréhension de ces phénomènes. Tous ces différents aspects seront exposés dans un contexte monolingue. Nous traiterons ensuite, dans le deuxième chapitre, chaque aspect sous un angle multilingue.

1. Le traitement du lexique

Le lexique est une ressource essentielle d'un système de traitement automatique de la langue. Il contient les mots de la langue traitée (ou un sous-ensemble de ces mots), ainsi que certaines informations associées à ces mots (Sproat, 1992, p. xi). Le lexique a deux buts : distinguer les mots incorrects des mots corrects et donner aux mots des informations nécessaires pour l'analyse syntaxique et sémantique.

De manière générale, le lexique constitue l'ensemble de toutes les entrées lexicales d'une langue. Une entrée lexicale est composée du mot et des informations lexicales qui lui sont liées. Ingria *et al* (ms) distinguent trois types d'informations.

- Les informations morphologiques et orthographiques qui expliquent comment les mots s'écrivent et sous quelles formes ils peuvent apparaître ;
- Les informations syntaxiques, en particulier la catégorie syntaxique du mot et éventuellement sa sous-catégorisation ;
- Les informations sémantiques, qui précisent le sens du mot.

Toutes ces informations peuvent être regroupées en deux groupes distincts, les informations intra-lexicales et inter-lexicales.

1.1. Les informations intra-lexicales

Ces informations sont celles qui spécifient le contexte dans lequel le mot apparaît. Elles incluent des informations de type sémantique et syntaxique.

Les informations syntaxiques regroupent les informations de catégories (noms, verbes, …) et les informations de sous-catégorisation. Ces dernières regroupent, par exemple, les information spécifiant les structures des compléments des verbes.

Les informations classées sous le label sémantique sont très variées. Certaines sont purement sémantiques, comme des pointeurs vers une ontologie sémantique, mais d'autres sont davantage syntaxiques. La distinction entre trait sémantique et syntaxique est bien souvent floue (Ingria, et al, p. 12) étant donné que ces deux informations apparaissent bien souvent dans une même structure complexe.

Les informations intra-lexicales sont dites « distributionnelles » dans le sens où elles spécifient le contexte des mots ainsi que les types de mot qui peuvent être associés à une entrée.

1.2. Les informations inter-lexicales

Ces informations sont celles qui lient les mots entre eux dans le lexique. On distingue différents types de liens :

a) Les liens morphologiques permettent de lier le mot à sa forme de base. Ils regroupent les informations flexionnelles et dérivationnelles :

   (i) Les premières déterminent la réalisation de surface (forme orthographique) d'un mot dans le processus de flexion, notamment pour les noms et les verbes. Ce genre d'informations relie, par exemple, la forme orthographique *mangeait* avec la forme de base *manger*.

   (ii) Les secondes relient une entrée lexicale à une autre sur laquelle elle est construite. Ce genre d'information exprime par exemple le lien entre *rapide* et *rapidement*.

   Nous étudierons plus en détail par la suite ces deux types de phénomènes.

b) Les liens sémantiques ou syntaxiques lient l'entrée lexicale avec ses informations sémantiques et syntaxiques. Ces liens peuvent être ambigus et donner lieu à plusieurs

analyses possibles. En effet, un mot peut correspondre à plusieurs analyses, et une analyse peut correspondre à plusieurs mots. On distingue trois cas de figure différents:

(i) analyse à catégories multiples: quand un mot orthographique est associé à plusieurs catégories (*porte* est soit la troisième personne du verbe *porter*, soit le nom, soit l'adjectif.)

(ii) analyse à sous-catégorisation multiples: quand un mot est associé à plusieurs sous-catégorisation (*commencer* est intransitif et transitif)

(iii) analyse sémantique multiple: qui advient quand un mot est associé à plusieurs interprétations sémantiques (*souris* est un animal et un matériel informatique)

Le principal problème réside donc dans le fait qu'il n'existe pas un lien 1 : 1 entre l'entrée lexicale et les informations associées. L'organisation du lexique est donc une affaire complexe et son traitement pose de nombreuses questions. Pour rendre le module lexical d'une application de TAL le plus exhaustif possible, on distingue deux approches pour traiter le lexique et donc surmonter ces difficultés.

1. l'approche énumérative (ou monomorphique) consiste à énumérer toutes les variantes de chaque mot (variante au niveau des flexions comme au niveau des catégories). Par exemple, il faudrait énumérer toutes les flexions de l'adjectif *beau* (*beau, beaux, belle, belles, bel*) ou toutes les sous-catégorisations du verbe *commencer* (*commencer, verbe transitif,* / *commencer, verbe transitif indirect, /commencer, verbe intransitif.*)
2. l'approche polymorphique tente d'exprimer les généralités. Cette approche se fonde sur une partition du système en deux parties : un lexique de base et un ensemble de règles chargées d'exprimer les informations relatives à ces formes. Ces deux parties permettront de produire un lexique complet. Par exemple, puisque la majorité des noms anglais forment leur pluriel en ajoutant un *–s* , il est plus économique d'énumérer uniquement les formes du singulier et de dériver les formes du pluriel par une simple règle reliant ces formes à la marque du pluriel (—*s*) (Sproat, 1992, p. xi).

> Tout l’enjeu de cette approche est donc d’avoir un formalisme qui décrit et qui organise au mieux ces deux parties et les liens qui les unissent.

Selon Sproat (1992, p.xi) l’approche énumérative est sans doute possible informatiquement étant donné qu’il n’existe actuellement presque plus aucune contrainte de stockage. Mais humainement le travail est presque impossible. L’ajout d’un nouveau mot dans un tel lexique implique l’énumération de toutes ses formes, avec la répétition, pour chaque forme, de toutes les informations intra et inter lexicales.

L'approche polymorphique présente de nombreux avantages. Premièrement, elle permet d'exprimer plus facilement la créativité. Par exemple le préfixe ‘pro-‘ peut s’attacher de manière très libre à de nombreux noms ou adjectifs. Ainsi, tous les noms propres peuvent potentiellement donner lieu à la création de mots composés du type *pro-Chirac*, *pro-Dreifuss*, etc. (Sproat, 1992, p.25). De plus, elle facilite l'ajout de nouveaux mots et permet donc une maintenance plus aisée. Ainsi, le néologisme *zapper* (Bouillon, *et al*, 1998, p.48) se conjuguera sur le modèle du verbe *aimer*. Son ajout dans le lexique consistera simplement à un ajout du radical et à un lien avec les mêmes flexions que le verbe *aimer*. Cette approche permet également, sur le plan informatique, de réduire considérablement la taille du lexique (Bouillon *et al*, 1998, p. 46,). Cependant, la généralisation pose le problème du traitement des exceptions. La régularité qui veut que le féminin des adjectifs français se forme par l’ajout d’un suffixe final (-e) est parfois contredite par des changements plus ou moins importants. Par exemple, *grec* donnera *grecque* (Bouillon *et al.* 1998, p.48). Nous y reviendrons plus en détail par la suite.

2. L’expression des généralités morphologiques

2.1. Pourquoi s’intéresser à la morphologie ?

Comme nous l’avons vu plus haut, l’approche polymorphique permet de réduire la place de stockage et facilite l’implémentation du module lexical. Elle tente donc d’exprimer les généralités qui existent entre les mots. Par exemple, plutôt que d’ajouter après chaque nom commun l’information syntaxique de la catégorie, comme le propose une approche énumérative, la démarche polymorphique implique de généraliser cette information et de la lier avec les mots concernés par cette information. L'expression de ce lien doit alors se faire de manière propre linguistiquement et informatiquement. Les liens morphologiques représentent donc un enjeu majeur pour le traitement du lexique. Ils possèdent en effet une certaine régularité qui permet une exploitation aisée. Ils sont très puissants et peuvent être facilement généralisables.

3. Qu’est-ce que la morphologie ?

3.1. Les fonctions de la morphologie

La morphologie étudie la structure des mots. Les mots sont vus comme une construction d’une ou plusieurs unités appelées morphèmes. Parmi les morphèmes, on distingue les morphèmes indépendants ou libres, qui peuvent être utilisés de manière isolée et les morphèmes liés, qui n’ont aucun sens pris isolément. Le phénomène morphologique a pour fonction de transporter des informations (Sproat, 1992, p. 19).

La qualité et la teneur de ces informations transportées par le phénomène morphologique diffèrent selon les langues.

3.1.1. Les différents types de morphologies dans les différentes langues

C'est au XIXème siècle que l'on a commencé à proposer une classification des langues en fonction de leur organisation grammaticale. En effet, si les mots et les morphèmes sont des concepts présents dans toutes les langues, la manière de les combiner varie selon les langues (O’Grady , 1996, p. 380).

De nombreux auteurs introduisent cette classification (Waksler (2000,p.228), Sproat (1992 , p.19) qui citait lui-même Bloonsfield, …). Elle n'est cependant pas unique ; il en existe d'autres, plus complexes et plus précises. De plus, il faut ajouter que les langues ne sont pas définies exclusivement par un seul type d'organisation, mais ont simplement tendance à accentuer tel ou tel type de structures.

Les langues sont donc divisées en quatre types principaux en fonction de la formation de leur mot : isolante, agglutinante, fusionnelle ou flexionnelle et polysynthétique (Sproat, 1992, p.19). Selon Walsker (2000, p.228), ces catégories sont fondées sur le nombre moyen de morphèmes par mot et le nombre d'informations contenues dans chaque morphème. Nous les passons en revue ci-dessous.

a) Les langues isolantes n'ont pas de formes liées (les affixes). Elles utilisent des mots invariables et non décomposables (les mots sont alors dits monomorphémique). Les informations grammaticales qui sont dans les autres langues indiquées par des affixes, sont dans ce cas signalées par le contexte sémantique ou pas du tout signalées. Un exemple typique de langue isolante est le mandarin.

Dans l'exemple suivant (Sproat , 1992, p.19), la phrase :

g• u bú ài ch• qu•ngcài

peut se traduire par

le chien n'aime pas manger des légumes

où

les chiens n'aimaient pas manger des légumes.

Le sens dépend du contexte. Dans l'exemple ci-dessus, « g• u » (chien) n'est pas marqué morphologiquement en terme de genre ou de nombre, et « ài » (aimer) n'est pas marqué ni en terme de personne ni de temps. Dans certaines de ces langues, le verbe est invariable et seul le pronom permet de savoir à quelle personne l'on se réfère. Le fonctionnement du verbe anglais est très proche de ce type de comportement. En effet, à l'exception de la troisième personne du singulier, c'est le pronom qui indique la personne du verbe, le verbe n'étant jamais marqué (*I eat, you eat, we eat*). De la même manière, la futurité est exprimée par un morphème libre : *I will leave* (O'Grady , 1996, p. 380).

b) Les langues polysynthétiques[1] expriment certains éléments (notamment syntaxiques) de manière morphologique alors que dans d'autres langues, ces éléments sont distincts. Les

[1] Waskler (2000, p. 231) parle de langues incorporantes

langues polysynthétiques (comme l'eskimo) mettent toute une phrase dans un même mot. Ces langues permettent par exemple la combinaison d'un prédicat avec un verbe.

Un exemple (Sproat, 1992, p. 21) de langue polysynthétique est le Yupik [2]:

qayáliyúlúni

signifie

il est très doué pour faire du kayak.

ou

yú signifie *il est très doué*
li signifie *faire*
et *qayá* signifie *kayak.*

Les langues isolantes et les langues polysynthétiques sont deux extrêmes et peu de langues sont purement isolantes ou polysynthétiques. La grande majorité des langues sont synthétiques, c'est-à-dire qu'elles permettent des mots de plusieurs morphèmes (O'Grady et al, 1996, p. 381). On distingue deux types de langues synthétiques :

c) Les langues agglutinantes, où les formes liées apparaissent dans les mots comme les maillons d'une chaîne. L'exemple couramment donné pour cette classe est le turc, qui agglutine un grand nombre de marques grammaticales de manière morphologique. Ces langues utilisent des affixes grandement productifs, chaque affixe possédant une seule fonction sémantique ou syntaxique. Les mots peuvent donc se former par une agglutination d'affixes. Dans ces langues, les radicaux et les affixes sont facilement séparables et peuvent être compris isolément. Par exemple, en turc, le complément du nom *des villages* est *køj-ler-in* où *køj* constitue la racine du mot (village), *ler* est la marque du pluriel et *in* est la marque du génitif. Un autre exemple plus parlant est celui du hongrois (Hachette encyclopédie, 2001), où :

ház-am	ma maison
ház-ad	ta maison
ház-am-bam	dans ma maison
ház-am-ból	hors de ma maison

d) Finalement, les langues flexionnelles utilisent une union de différents éléments dans une seule et même forme (on parle alors de morphème « portemanteau » car ils portent en eux plusieurs éléments syntaxiques et sémantiques). L'archétype même de la langue flexionnelle est le latin, mais aussi le grec, et toutes les langues latines dérivées. Dans ces langues, les

[2] Langue de la famille eskimo-aléoute parlée par environ 17'000 locuteurs répartis entre la Sibérie et l'Alaska.

racines et les affixes fusionnent littéralement. Contrairement aux langues agglutinantes, chaque morphème ne contient pas une seule information, mais peut en contenir plusieurs. Certaines langues utilisent comme flexion des voyelles qui viennent s'insérer à l'intérieur des mots. C'est par exemple (Hachette encyclopédie, 2001) le cas de l'arabe : *kateb* (celui qui écrit) et *kitab* (l'écrit, le livre).

Nous l'avons dit, cette typologie n'est pas stricte. Elle permet cependant de tirer certaines généralisations. Et en mettant l'accent sur tels ou tels phénomènes morphologiques propres à certaines langues, elle nous permet de définir les fonctions de l'outil que l'on voudra construire pour traiter la morphologie d'une langue particulière. Pour le présent travail, elle permet de confirmer la proximité morphologique du français et de l'italien, ces deux langues étant toutes deux considérées comme des langues flexionnelles.

3.1.2. Les différents phénomènes morphologiques

Communément, on divise le phénomène morphologique en trois fonctionnements distincts (Bouillon *et al.*, 1998, p.13):

- la morphologie flexionnelle
- la morphologie dérivationnelle
- la morphologie compositionnelle

Ces trois types de fonctionnement morphologique ne sont pas utilisés de la même manière suivant les types de langues. Les langues flexionnelles possèdent ainsi une morphologie flexionnelle et dérivationnelle très puissante et très productive en comparaison avec les langues isolantes qui n'ont pratiquement pas d'affixes.

La **morphologie flexionnelle** est caractérisée par l'ajout d'une marque exprimant des informations grammaticales (Bouillon et al, 1998) permettant par exemple de réaliser le pluriel ou la conjugaison des verbes. Les phénomènes de flexions occupent une place centrale dans l'analyse morphologique du fait de leur importance quantitative (Fuchs et al, 1993). La morphologie flexionnelle possède certaines caractéristiques:

1. elle est relativement systématique (Bouillon et al, 1998,): l'ajout d'un suffixe donné à un radical a les mêmes effets grammaticaux et sémantiques pour tous les radicaux. Par exemple, l'ajout du *-s* aux noms français conduit à la formation d'un pluriel ;

2. elle est très productive (Ritchie et al, 1992, p. 5) : un néologisme dans la langue prend automatiquement les mêmes règles d'affixation. Par exemple, un nouveau verbe prend automatiquement les conjugaisons de son groupe (cf. l'exemple de *zapper* susmentionné) ;
3. elle préserve la catégorie (Bouillon et al, 1998) : La catégorie grammaticale ne change pas dans le processus de flexion. Les verbes restent des verbes et les noms restent des noms.

Sproat (1992 , p. 22) met tout de même un bémol à ces distinctions. Ainsi, lorsqu'on dit que le fonctionnement de la flexion ne change pas la catégorie du mot, il est des cas où cette définition est tangente, comme par exemple dans la formation du participe passé en français, où le participe passé peut également être utilisé comme adjectif, et donc changer de catégorie syntaxique. A nouveau, la distinction entre ces phénomènes n'est pas complètement stricte.

La **morphologie dérivationnelle** est caractérisée par la formation de nouveau mot à partir de radicaux existants. Cette fois, le mot nouveau pourra appartenir à une autre catégorie grammaticale que sa racine. La morphologie dérivationnelle possède les caractéristiques suivantes:

1. elle est peu systématique (Bouillon et al, 1998): l'ajout du même affixe à différents radicaux n'a pas le même effet. Par exemple, l'ajout du préfixe *dé-* devant un verbe n'a pas toujours le même effet. Ainsi, le préfixe *dé-* peut avoir un sens privatif (*débaptiser*) ou un sens intensif (*démontrer*, *délaisser*) (Lehmann et al, 1998, p.135).
2. elle n'est pas que partiellement productive: Un nouveau mot ne prend pas forcément les mêmes dérivations que le reste du lexique. Nous avons par exemple *centre* qui a donné *centrage* et non pas **centrement* (Fuchs *et al.*, 1993 p. 87). Certaines dérivations sont cependant très productives , comme l'ajout du suffixe *-ment* après les adjectifs pour en faire des adverbes.
3. elle peut altérer la catégorie: l'ajout d'un suffixe change la catégorie grammaticale d'un mot.

On distingue plusieurs types de transformations dérivationnelles en fonction des catégories qu'elles transforment. En voici une liste, tirée de Sproat (1992, p. 33 et ss.) [3],

[3] cette liste n'est pas exhaustive mais Sproat renvoie lui-même à Marchand (1969)

- La **morphologie nominale déverbalisante** qui provoque la création d'un nom (généralement agentif) à partir d'un verbe (*manger* →*mangeur*) .
- La **morphologie adjectivale déverbalisante** qui crée un adjectif à partir d'un verbe. C'est par exemple le cas du suffixe *–able* : *manger*→ *mangeable*.
- La **morphologie désadjectivale**, qui transforme l'adjectif en un nom. C'est notamment le cas du suffixe *–ité* qui transforme *facile* en *facilité*.
- La **morphologie dénominalisante**, qui transforme le nom en un adjectif, comme dans le cas du suffixe anglais *–less* qui donne le sens opposé au nom en formant un adjectif signifiant *manquant de…* Par exemple : *arm* → *armless.*
- Les **morphologies préfixales** qui ajoute une valeur au mot en ajoutant un préfixe (c'est le cas notamment de *pseudo-*, *anti-*, *pro-*.)

Toutes ces formes de morphologie dérivationnelle n'ont pas la même productivité ni la même régularité. Ainsi, autant le suffixe *anti-* peut se retrouver dans un grand nombre de formation en indiquant à chaque fois le même trait sémantique, autant le préfixe *–eur* n'a pas la même productivité et n'apporte pas toujours la même valeur sémantique (*manger* → *mangeur*, qui fait l'action de manger, mais *aller* ne peut pas donner **alleur*. De même, on ressent bien la présence de ce suffixe dans *facteur* sans pouvoir en comprendre précisément le sens).

La **morphologie compositionnelle** couvre essentiellement la formation des mots composés.

Alors que les morphologies flexionnelle et dérivationnelle consistent en l'attachement d'un affixe à une forme de base, la morphologie compositionnelle concatène deux formes de base. La réalisation de la composition est différente selon les langues. En français, elle peut se faire soit par le trait d'union (*abat-jour*), soit par le phénomène de soudure (*autoroute*) ou encore par le phénomène de figement (*grand magasin*, *pomme de terre*). On remarque une hésitation entre ces trois procédés. Ainsi, en français, on écrira « portefeuille » mais « porte-monnaie », « enjeu » mais « en-cas » (Lehmann et al, 1998 p. 170).

L'analyse d'un mot composé est très difficile car le lien sémantique entre les deux mots est implicite.

Comparons par exemple, la signification du mot anglais *house* dans les trois exemples ci-dessous (Sproat, 1992, p.40):

dog house : **house** where a **dog** lives
tree house : **house** in a **tree**
fire house : **house** were **fire**fighters stay

Presque toutes les langues disposent d'une morphologie compositionnelle. Mais la manière de l'utiliser et la productivité de ce phénomène diffèrent selon les langues. Par exemple, la plupart des langues germaniques ont très peu de contraintes pour former des mots composés de plusieurs formes de base, comme en allemand par exemple (Sproat, 1992 p. 42):

Lebensversicherungsgesellschaftsangestellter
Leben+s+versicherung+s+gesellschaft+s+angestellter
Employé d'une compagnie d'assurance vie

3.2. Qu’est-ce qui se combine et comment ?

L’unité de base du phénomène morphologique est donc le morphème. Le morphème est une unité significative minimale (Lehmann, 1998, p. 107), récursive de part son sens et sa forme (Neuvel et Singh, sous-presse). Sproat (1989 , p. 65) propose une définition formelle du morphème :

> « Un morphème est en fait composé d’une paire d’éléments définitoires dont le premier élément représente les catégories morphologiques (structures syntaxiques et sémantiques) exprimées par le morphème, et le second élément représente sa forme phonologique avec des informations exprimant la manière dont le morphème s’attache à la base. »

Par exemple, dans :

(a)<<ETAT, adj/n>, -ité>
(b) <<PLURIEL, n/n> , -s>

(a) représente la structure du suffixe -*ité*, dont il sera question par la suite. Elle contient sa forme phonologique (-*ité*) ainsi que les informations syntaxiques et sémantiques du morphème en question. Ainsi, -*ité* transforme un adjectif en nom (*adj/n*) exprimant ainsi l’état (ETAT) dérivé de l’adjectif. De même, pour l’exemple (b) où le suffixe dont la forme phonologique est –*s* ne change pas la catégorie de la base (*n/n*) mais implique aux niveaux syntaxique et sémantique un changement du singulier au pluriel (PLURIEL).

Ces morphèmes sont généralement décrits en deux sortes: les morphèmes libres qui sont capables d'agir comme des mots isolés, et les morphèmes liés, qui ne peuvent opérer que lorsqu'ils sont liés à un autre morphème. Ces derniers s'appellent généralement des *affixes,* mais leur dénomination varie en fonction de leur place dans la concaténation. Le morphème qui joue le rôle central du mot (généralement un morphème libre) est appelé *radical.*

Un mot peut être formé de un ou plusieurs morphèmes.

Pour analyser la structure d’un mot, et donc identifier les unités significatives minimales, on procède généralement à une substitution ou commutation paradigmatique (Lehmann, 1998, p. 107). Ainsi, pour pouvoir segmenter le mot *embarquement*, on procède aux substitutions suivantes :

a) *embarquement*/*débarquement*, donne les éléments *em-* et *dé-*
b) *embarquement*/*emprisonnement* donne les éléments *barque* et *prison*
c) *embarquement*/*embarcation* donne les éléments *–ement* et *–ation*.[4]

Les éléments sont considérés comme valides dès lors qu'on les retrouve combinés à d'autres avec le même sens et avec la même forme (Lehmann, 1998, p, 107). Un tel procédé implique également la reconnaissance de certaines variations dans les éléments quand ils sont combinés entre eux. Ainsi, *barque-* devient *barc-* et est alors considéré comme un allomorphe (Lehmann, 1998, p. 107).

Il existe également plusieurs manières de combiner les morphèmes entre eux[5] :

- préfixation (devant le mot)

  préfixe *re-* : *re + commencer = recommencer*
  (Fuchs, 1993, p. 86)

- suffixation (derrière le mot)

  suffixe *-able* : *mang + able = mangeable*
  (Bouillon, 1998 p. 49)

- infixation (dans le mot)

  En langue Ulwa[6], l'infixe *-ka* marque le possessif de la troisième personne.
  suulu = chien
  suu + ka - lu = suukalu = son chien

- circonfixation (autour du mot)

  circumfixe *ge-* et *-t* en allemand pour former le participe passé

  machen → *ge + mach + t = gemacht*

- la morphologie zéro. Quand on ajoute rien. Par exemple, en anglais pour passer d'un nom à un verbe

  Let's table the motion
  I need to xerox this article

- morphologie soustractive: quand on enlève des morphèmes

  économique + e = économique

[4] Exemple tiré de Lehmann (1998, p. 107)
[5] Cette classification ainsi que les exemples sont tirés de Sproat (1992, p. 44 et ss), sauf informations contraires.
[6] Langue misumalpane de l'est du Nicargua parlée par environ 400 personnes

- la réduplication : dans certaines langues, les pluriels s'effectuent en re-duplicant le morphème de base. C'est par exemple le cas de l'indonésien, où :

orang	*homme*
orang orang	*hommes*

Sproat évoque également le fait que, dans certains cas, la combinaison de morphèmes est dirigée par des données phonologiques. Ainsi, en italien, le suffixe pluriel du féminin '-e' fait varier la forme orthographique de certains radicaux, pour en garder le même son :

tasca (f.sg) → tasc **+ h** + e (f. pl.)

Dans le même ordre d'idée, le phénomène d'haplologie est assez fréquent en français. Ce phénomène consiste en la suppression d'un segment répété après la transformation morphologique (Lehmann 1998 : p. 112). Ainsi, l'adjectif « féminin » est dérivé en *féminité* et non pas en **féminninité*.

Bien que les définitions données ci-dessus quant à l' « identité » du morphème semble claires et précises, et que certaines variations soient suffisamment motivées, il survient, dans l'analyse morphologique, des « restes », c'est-à-dire des segments peu employés ou dépourvus de sens, et dont l'identification pose certains problèmes (Lehmann, 1998, p. 108). A ce propos, Lehmann (1998, p. 108) cite De Saussure : « En rapprochant des mots tels que *coutelas*, *fatras*, et *canevas*, on a le vague sentiment que *–as* est un élément formatif, propre aux substantifs sans qu'on puisse le définir plus exactement ».

3.2.1. La morphologie holistique[7]

Jusqu'à maintenant, nous avons décrit la vision générale de la morphologie, basée sur les similarités (un morphème est un élément récurent de part sa forme et son sens). D'autres théories récentes remettent en cause ce postulat, en affirmant que l'étude de la morphologie basée sur les similarités engendre la création de type d'unité spécifique indésirable (Singh et Neuvel, sous-presse 2) pour décrire certaines unités découvertes par l'analyse morphologique. Selon eux, « la frontière entre accident phonologique et élément morphologique » n'est pas toujours évidente, et cela force le linguiste à traiter des éléments qui ne répondent pas toujours à la définition du morphème, mais qui doivent tout de même être pris en compte (Neuvel et Singh (sous-presse 2)

[7] Dénomination française de la théorie « Whole Word Morphology », proposée par Sylvain Neuvel.

citent à ce sujet un certain nombre de type d'unités récoltés dans la littérature, comme les semi-suffixes ('*lun*-' dans *lundi* (Marchand, 1969)) les affixoïdes ('*géo*-' dans *géographe*, *géologie* (Stein 1977)).

Au lieu de partir du postulat de similarité pour définir le lien entre deux mots (*completely* et *directly* sont construit sur le même procédé, un suffixe *–ly*), la théorie de la morphologie holistique part du principe que la morphologie d'une langue réside uniquement dans les différences. De ce point de vue, deux mots sont considérés comme morphologiquement liés s'ils diffèrent exactement de la même manière que deux autres mots de la même langue (Neuvel et Singh, sous-presse). Ainsi, ce qui est intéressant dans les mots *rapidement* et *stupidement,* ce n'est pas qu'ils se ressemblent (à cause du suffixe *–ment*), mais plutôt parce que la différence entre *rapide* et *rapidement* est exactement la même que entre *stupide* et *stupidement.*

Plus formellement, la théorie postule que :

> « Etant donné 4 mots (w1, w2, w3, et w4), w1 est considéré comme morphologiquement lié à w2 si et seulement si les propriétés formelles, sémantiques et grammaticales qui les distinguent, distinguent également w3 et w4 »[8]

De ce postulat, on peut créer une « stratégie morphologique », bidirectionnelle, qui permettra de formaliser la relation entre les mots. Par exemple, si l'on considère les verbes et les noms anglais suivants (Neuvel et Singh, sous-presse):

receive$_v$ reception$_n$
conceive$_v$ conception$_n$
deceive$_v$ deception$_n$

et que l'on formalise les différences comme suit :

receive$_v$ reception$_n$ /…ive/$_v$ • /… eption/$_n$
conceive$_v$ conception$_n$ /…ive/$_v$ • /… eption/$_n$
deceive$_v$ deception$_n$ /…ive/$_v$ • /… eption/$_n$

on remarque aisément que les différences entre chaque paire de mots sont exactement les mêmes. On peut donc en extraire une stratégie morphologique :

/X ive/$_v$ • /Xeption/$_n$

où X représente une variable.

[8] Neuvel et Singh (sous-presse)

De plus, pour éviter la génération de **beleption* $_{n}$ à partir de *believe* $_{v}$, les auteurs de la théorie pensent qu'il est toujours préférable de contraindre la variable en considérant comme une constante les similarités observées dans les paires de mots. Dans l'exemple ci-dessus, il sera donc nécessaire d'assigner la contrainte de la présence du 'c' dans la variable X.

Ce type d'analyse permet de ne pas utiliser des unités morphologiques difficiles à définir, et amène à conclure que le mot est la seule unité signifiante nécessaire à l'analyse morphologique.

Cette théorie est encore très récente. Néanmoins, ses auteurs promettent que l'individualisation des paires de différences permettra la définition de puissants algorithmes pour l'acquisition de la morphologie en permettant la génération automatique de nombreux dérivés, facilitant ainsi l'implémentation de nombreux systèmes en TAL, particulièrement dans l'enrichissement des lexiques.

3.3. Que faut-il prendre en compte dans la construction d'un module morphologique ?

En général, on considère que la fonction première d'un lexique informatisé est d'associer une chaîne de caractère représentant la forme des mots avec différentes informations rendant possible la distribution dans la phrase et une certaine compréhension du texte. D'un point de vue informatique, une organisation économique du lexique implique d'une part la capacité de faire des généralisations pour une classe de mots et d'autre part, d'exprimer les exceptions de cette généralisation s'appliquant à ces classes et aux sous-classes de mots. Ritchie *et al.* (1992, p.2) ajoutent qu'il faut, lors de la construction d'un module lexical, garder en tête des considérations à la fois pratique et théorique. Ainsi, ils élaborent quelques lignes directrices nécessaires à cette construction. Parmi les plus intéressantes, citons notamment :

- Les descriptions linguistiques doivent être déclarées sous forme de règles abstraites et non pas directement sous forme informatique.
- Il faut réfléchir pour savoir quels concepts sont inhérents au programme lui-même, et quels sont ceux qui sont davantage linguistiques.
- Il faut également élaborer une description linguistique aussi élégante que possible et proche des standards linguistiques courants.
- Les règles de formalisme doivent être définies précisément.

- Toutes les définitions de ce formalisme doivent être faciles d'utilisation et d'implémentation

D'un point de vue linguistique, la construction d'un module lexical prenant en compte l'exploitation des régularités morphologiques dépend également d'un certain nombre de facteurs. Selon Sproat (p. 83), la première question que l'on doit se poser avant de construire un outil de ce type est de savoir ce que l'on veut en faire et quelles sont les informations dont on a besoin. La question suivante est évidemment de savoir quelles langues et quelles particularités de celles-ci doivent pouvoir être traitées, pour savoir de quel phénomène le programme devra rendre compte (flexion, dérivation ou composition ?).

Il nous faudra ensuite étudier les différents morphèmes à traiter, les informations véhiculées par ces derniers, et les différentes combinaisons de morphèmes possibles dans les langues que nous voulons traiter. Ces différentes combinaisons soulèvent trois problèmes :

1. les contraintes : Un des problèmes soulevés par Sproat (1992) concerne les contraintes à mettre sur les morphèmes. Il ne suffit pas en effet d'énumérer les morphèmes et de penser que la concaténation se fera sans problème.

On distingue deux contraintes principales :

a) Les contraintes de restriction : les affixes doivent avoir certaines restrictions, car ceux-ci ne s'attachent pas avec tout. Par exemple, la concaténation d'un suffixe -able avec un verbe donnera un adjectif, sauf dans certains cas : "courable". Ces contraintes relèvent souvent de la sémantique du radical.

Sproat parle également des contraintes d'ordre (1989, p.83). La première contrainte d'ordre concerne par exemple le fait qu'un suffixe, par essence, ne peut se mettre que derrière un mot et qu'un préfixe ne peut se placer que devant un mot.

b) Les contraintes d'ordre de transformation et des arguments de cette transformation.

Par exemple, le mot *motorisabilité* est formé de trois suffixes *–iser*, *–able* et *–ité*. Pourquoi ces suffixes se déclarent-ils dans cet ordre ? Et bien tout simplement parce qu'une des propriétés du suffixe –iser est de faire un verbe à partir d'un nom, que le suffixe –able crée un adjectif à partir d'un verbe et que le suffixe –ité crée un nom à partir d'un adjectif.

2. Les changements orthographiques : Nous avons vu que même si certains phénomènes semblent très productifs, ils contiennent parfois des irrégularités. Par exemple, l'ajout d'une marque du féminin « -e » en français, engendre des variations orthographiques qui ne sont pas toujours régulières (Bouillon et al, 1998, p. 49) :

Par exemple (Bouillon et al, 1998 p, 50), l'ajout du -e provoque parfois le dédoublement de la dernière consonne du lemme :

paysan + e → paysanne

mais peut également provoquer la suppression d'une lettre :

économique + e → économique

De plus, ces changements orthographiques (Kay, 1993, cité par Bouillon et al, 1998, parle de changements morphographémiques) ne sont pas réguliers et ne dépendent pas de la forme de base. Ainsi, l'ajout d'une lettre provoqué par la marque « -e » du féminin amène des changements différents selon les cas.

Par exemple (Bouillon et al, 1998 p, 50) :

grec → gre**cqu**e
blanc → blanc**h**e
turc → tur**que**

Le programme devra donc être capable de formaliser les généralités morphologiques tout en pouvant gérer facilement et correctement les irrégularités et les contraintes.

3.4. Les applications informatiques de la morphologie

Le traitement de la morphologie, nous l'avons dit, est particulièrement utile pour gérer de manière optimale des ressources lexicales. Les applications informatiques de la morphologie sont variées. Sproat (1992, pp. 1 et ss.) décrit les nombreux domaines nécessitant un traitement morphologique. Il identifie quatre grands domaines : (a) les applications en langue naturelle (regroupant l'analyse, la génération, la traduction automatique, les outils annexes aux dictionnaires informatisés et la lemmatisation), (b) les applications pour la parole (système de lecture à partir d'un texte, reconnaissance vocale), (c) les applications liées au traitement de texte (correcteur d'orthographe, système de saisie de texte) et à la recherche documentaire. Il ajoute à cette liste (1992, p. 14) les outils d'enseignement de la morphologie qui ont été développés pour les étudiants apprenant une langue et pour les étudiants en linguistique. Daille *et al.* (2002, p. 224) mentionnent également les applications dans le domaine de la terminologie (acquisition de terme, détection des variations de terme, etc.).

> « Le rôle joué par la morphologie dans les applications de traitement automatique des langues peut aussi être expliqué par le fait

> que les modèles linguistiques sophistiqués ne sont pas toujours nécessaires et que leurs performances sont souvent améliorées par un simple traitement des phénomènes morphologiques. »[9]

Toutes ces applications nécessitent des modules plus ou moins complexes pour exploiter les connaissances morphologiques. Bien souvent, le but visé par l'application conditionne le type de connaissances et d'informations morphologiques nécessaires à celle-ci.

Dans le cadre de ce travail, la question de la finalité de l'outil ne sera que très peu abordée. Nous réfléchirons sur la manière d'aborder le lexique multilingue sur un plan avant tout théorique et les outils à disposition conditionneront les types d'informations exploitables pour la partie pratique.

3.5. Les preuves psycholinguistiques

Le lien entre un système informatique de traitement de la morphologie et le traitement de celle-ci par le cerveau humain mérite d'être posé. Selon Sproat (1992, p. 19), même si de nombreux systèmes sont construits sans aucune préoccupation de cet ordre, il est intéressant de comparer les approches purement informatiques et le comportement du cerveau humain.

Bien qu'une morphologie informatique n'ait pas forcément pour but de retranscrire l'activité du cerveau humain, il reste intéressant de prendre ce dernier en comparaison. De plus, il n'y a aucune garantie que le cerveau humain fonctionne de la manière la plus efficace (Sproat, 1989, p.110). Cependant, lorsqu'un constructeur de morphologie informatique tentera de formaliser et de traiter un fait linguistique, une analyse du fonctionnement du cerveau humain pourra lui donner des pistes pour résoudre les difficultés de traitement de certains phénomènes linguistiques et appuyer peut-être certaines options choisies. Sans compter que le cerveau humain est de loin plus robuste que tous les systèmes construits jusque là (Sproat : 1989 p. 111).

De nombreux travaux ont été effectués pour tenter de comprendre l'organisation du lexique dans le cerveau. Loin de nous la prétention d'en donner un panorama exhaustif, mais nous en citerons quelques-uns qui nous semblent pertinents.

Chaque locuteur possède un lexique interne, c'est-à-dire « un ensemble de représentations correspondant aux unités signifiantes de sa langue »(Caron, 1989, p.72). Il s'agirait d'une sorte de dictionnaire, liste d'entrées lexicales, dont chacune comporterait les informations nécessaires pour comprendre, identifier et utiliser ces unités (Caron, 1989, p.72). D'une manière extrêmement simplifiée, on peut dire que le lexique est divisé en deux niveaux de

[9] Daille *et al.* 2002 p. 230.

représentations : l'un qui stocke les formes et l'autre qui stocke les significations. De plus, étant donné que les mots se présentent sous deux formes (écrite et orale), il existerait deux formes de stockage des mots (De Groot, 1998). Donc, les informations contenues dans ce lexique seraient divisées en trois groupes : la forme phonologique, les propriétés morphologiques et syntaxiques, les informations de sens (Caron, 1989, p.72)[10]. L'accès lexical, c'est-à-dire la manière dont le locuteur comprend les mots et les produit, dépend surtout de l'effet de fréquence (plus un mot est fréquent, plus son accès est rapide) et de l'effet d'amorçage (un mot est plus facilement retrouvé s'il est précédé d'un mot qui lui est sémantiquement proche).

De nombreuses expériences ont montré que les relations morphologiques sont présentes dans le cerveau, à la fois quant il s'agit de lien morphologique flexionnel et de lien morphologique dérivationnel (Burani *et al.*, 1992, p.361). L'existence d'un véritable analyseur morphologique uniquement utilisé pour traiter des mots nouveaux ou peu fréquents pour le locuteur est semble-t-il désormais une évidence (Libben, 2000, p. 213). Cependant, à la question de savoir s'il existe un véritable analyseur morphologique qui entre en action pour tous les mots, les avis restent partagés. Depuis près d'un quart de siècle, deux théories s'affrontent. D'une part, l'hypothèse d'une décomposition morphologique selon laquelle les mots sont d'abord découpés en morphème avant que la racine soit soumise au lexique (opération morphologique pré-lexicale) et d'autre part, l'hypothèse « pleine-liste » (full-list hypothesis) qui soutient que toutes les formes sont présentes dans le lexique et que s'il existe un phénomène d'analyse morphologique, il intervient uniquement après le passage dans le lexique « pleine-liste » (Libben, 2000, p.213).

Mais ce sujet reste un vaste domaine. En effet, le traitement du lexique peut dépendre de nombreux facteurs (forme oral ou forme écrite, production ou compréhension du mot). Les recherches actuelles se penchent sur des phénomènes morphologiques précis, et leurs conclusions varient suivant le phénomène. Par exemple, certains travaux mettent en évidence que les mots donnent lieu à un amorçage sur le radical (c'est-à-dire que leur analyse par le lexique interne commence sur le radical) uniquement s'ils sont sémantiquement transparents (c'est à dire que leur sens est immédiatement construit à partir de leur composition). D'autres chercheurs émettent l'hypothèse que cette division entre lemme et flexion ne se fait que pour les mots plus rares. Les mots courants sont identifiés immédiatement, quelle que soit leur forme, fléchie ou lemmatisée.

[10] D'autres auteurs, comme Levelt (cité par Caron) pensent que ces deux derniers types d'informations sont ce qu'il propose d'appeler le "lemme", et qu'elles sont indépendantes de la forme phonologique, ce qui pourrait expliquer le phénomène des lapsus.

La variété de modèles proposés ne permet pas pour l'heure de se calquer sur la psycholinguistique pour rechercher les preuves de telle ou telle construction informatique. Le besoin de rationalité de la linguistique informatique empêche la construction de différentes architectures en fonction de différents types de mots. Cependant, il est intéressant de noter que ce sont le même genre de questions que les deux disciplines viennent à se poser. Ainsi, le conflit hypothèse pleine-liste vs. décomposition morphologique fait écho aux deux approches mono- et poly-morphique, et la distinction de traitement entre les différents phénomènes morphologiques se retrouvent dans les deux disciplines.

4. Conclusion

La morphologie est une partie importante du lexique et son traitement permet la rationalisation et l'optimisation des ressources lexicales. Comme nous l'avons vu, elle est au centre de nombreuses applications informatiques. Elle constitue par conséquent un vaste domaine d'étude, tant pour la rationaliser que pour la formaliser de manière optimale. En effet, si l'exploitation des régularités possède de nombreux avantages en terme de place et de maintenance, les irrégularités présentes dans les règles générales constituent un véritable défi tant linguistique qu'informatique. Malgré cela, l'approche polymorphique possède suffisamment d'avantage pour être favorisée dans la plupart des cas.

C'est dans ce même souci de rationalité et d'optimisation que nous allons maintenant étudier la possibilité d'exploiter et de gérer les régularités présentes entre deux langues distinctes.

Chapitre 2 : Le lexique et l'expression des régularités multilingues

Dans la première partie, nous avons exploré les enjeux du traitement d'un lexique monolingue. Nous traiterons à présent, les mêmes problématiques pour un lexique multilingue. Après un passage en revue des fondements d'un tel lexique, nous tenterons d'évoquer les problématiques du traitement de la morphologie multilingue en nous attardant sur la notion de cognats. Nous essayerons de motiver un tel lexique d'un point de vue informatique, linguistique et psycholinguistique. Nous tracerons ensuite les grandes lignes pour la construction de ressources lexicales de ce type, à partir d'exemples tirés de notre travail pratique.

1. Le traitement du lexique : un lexique multilingue

Nous l'avons vu, le lexique est une partie essentielle du traitement automatique de la langue. Dans de nombreuses applications, le module lexical prend une place importante. Ce module permet de reconnaître les mots incorrects des mots corrects et d'attribuer à chaque mot un certain nombre d'informations (intra ou inter lexicales). Dans une application qui regroupe plusieurs langues, il est normalement d'usage de construire plusieurs modules séparés contenant chacun les informations relatives à une langue déterminée.

L'organisation d'un module lexical est une question importante relevant de plusieurs facteurs, notamment la place de stockage et la facilité de maintenance et d'ajout de nouveau mot (Sproat, 1992, p. xii). C'est en suivant ces considérations que l'approche polymorphique a été privilégiée. Cette approche implique la partition du système en deux parties, l'une se chargeant des formes dites de base et l'autre contenant des règles qui vont permettre de produire à partir des bases un lexique dérivé. Cette partition permet une exploitation maximale des généralités rencontrées au sein du lexique. L'enjeu des différentes formalisations du lexique est d'organiser ces deux parties et de les relier entre elles pour pouvoir exploiter au mieux les généralisations. Nous posons ici la question de savoir, si, dans une application qui devra gérer, à un moment ou à un autre, deux langues différentes, il serait possible d'avoir un seul et même module. Jusqu'à présent, peu d'études ont été faites sur les possibilités de partager des informations à des niveaux de description linguistique autres que celui de la sémantique (Tiberius, 2000, p 701). Comme le cite Carole Tiberius (2000, p 701), quelques systèmes ont déjà exploré ce partage dans différents niveaux de description linguistique, comme la syntaxe - grammaire d'unification multilingue de Kameyana (1998), ou le lexique - le projet Polylex (Cahill et Gazdar, 1999, dont nous reparlerons plus tard) et le projet GREG (Kilgariff et al, 1999).

2. Une morphologie multilingue

Au sein du lexique, les généralisations sont possibles sur plusieurs plans. Comme nous l'avons expliqué dans la première partie de ce travail, les régularités morphologiques sont, dans certaines langues, suffisamment puissantes et productives pour permettre une exploitation et une généralisation optimales. Le but de ce travail est d'exploiter ces régularités dans un contexte multilingue en construisant un module lexical capable de traiter plusieurs langues à la fois.

Intuitivement, on sait que deux langues historiquement proches possèdent un grand nombre de similitudes, tant dans leur syntaxe, leur morphologie et leur sémantique. Nous partirons du postulat qu'un locuteur francophone n'ayant aucune notion d'italien pourra, uniquement, grâce à la forme graphique, reconnaître l'équivalent français de *possibilità* et de *maturità.* De même, nous pensons qu'il pourra, après avoir eu confirmation de la justesse de sa déduction, générer lui-même le mot italien *maternità* à partir de sa connaissance unique du mot français *maternité.*

La tentative de généraliser la morphologie à un niveau multilingue a déjà été effectuée pour les langues germaniques dans le cadre du projet Polylex. Nous tenterons, dans ce projet, de construire un lexique capable de traiter l'italien et le français dans la même architecture, en exploitant au maximum les régularités morphologiques.

Etant donné que les informations morphologiques « transportent » des informations (Sproat, 1992, p.19) il sera également nécessaire de se pencher sur l'aspect commun des informations entre ces deux langues. Mais penchons nous tout d'abord sur l'aspect commun au niveau lexical.

2.1. La notion de cognats

Dans de nombreux domaines, les similitudes formelles (et particulièrement les similitudes au niveau lexical) entre les langues ont déjà été abordées par l'étude du phénomène des cognats. Selon Lavaur et al,(1998) « le terme *cognats* s'applique à tous les mots qui ont une forme orthographique et phonologique proche de leur équivalent de traduction dans une autre langue ». Il nuance cependant en disant que « le degré de similitude entre les équivalents de traduction considérés comme cognats n'est pas établi de manière stricte dans la littérature. » Quand les cognats sont proches et dans leur phonologie et dans leur orthographie, ils sont dit « cognats parfaits » et lorsque leur distance est plus importante, ce sont alors des « non-cognats ». Mais la définition ne s'arrête pas à une parenté orthographique et phonologique. Dans un sens strict, et d'un point de vue linguistico-historique, les cognats sont des mots de deux langues proches qui se sont développés à partir du même mot d'origine (Kondrak, 2001, p. 103). Mais dans un sens

plus flou, les cognats sont des mots similaires en terme de signification et de forme. Cette dernière définition ne prend pas en compte la dimension étymologique. Les phénomènes d'emprunt ont marqué toutes les langues et influencé leur lexique. Et ces emprunts alimentent les phénomènes de cognation. Dans le présent travail, nous utiliserons donc le terme cognat dans sa dimension synchronique, et nous définirons des paires de mots cognats uniquement par des constats de similitudes de forme orthographique et morphologique. Sur le plan sémantique, la définition du cognat reste également floue. Kreif (2001, p. 843) se pose la question de savoir « à partir de quel niveau d'écart sémantique doit-on rejeter un cognat ? ». Ne trouvant aucun réponse satisfaisante à cette question, il finit par éluder le problème en prenant « un parti pris restrictif », qui sera suffisant pour son application[11]. Nous ferons de même en n'évoquant la problématique du sens que très brièvement.

La différence entre cognat parfait, semi-cognat et non-cognat représentera la question principale de ce travail. En effet, si tout l'enjeu d'un module de traitement de la morphologie est de pouvoir exploiter au maximum les similitudes tout en traitant facilement les irrégularités, l'enjeu d'une morphologie multilingue est de pouvoir formaliser le lien morphologique existant entre deux mots cognats, qu'ils soient parfaits ou non, tout en laissant la possibilité de traiter dans le même module les paires de mots pas du tout apparentés sur le plan morphologique. Habituellement, en TAL, on considère comme cognats tous les mots extraits d'un corpus bilingue ayant au moins quatre lettres consécutives communes (Kreif, 2001, p. 833). Kreif (*ibid.*) estime que l'on ne peut pas s'arrêter aux radicaux de deux mots pour déterminer leur cognation. Deux mots peuvent en effet être apparentés par le biais de préfixes ou de suffixes. Dans certains cas, la cognation de suffixe ou de préfixe pourrait être prise en compte pour apparenter deux mots.

L'analyse formelle des similitudes entre les langues peut donc prendre en compte la notion de cognats comme référence. Mais l'exploitation des cognats pour la construction d'une morphologique multilingue ne suffit pas à motiver complètement une telle démarche.

[11] les recherches de Kreif sont menées dans le cadre de l'utilisation des cognats pour améliorer les résultats d'un alignement. Nous les évoquerons brièvement dans le chapitre 4.

3. Les motivations du traitement multilingue de la morphologie

Nous passons à présent en revue les différentes motivations pour le présent travail, en commençant par les motivations informatiques, puis en décrivant les différentes motivations linguistiques.

3.1. Les motivations informatiques

Comme nous l'avions évoqué dans la première partie de ce travail, l'approche polymorphique est largement adoptée dans les systèmes de traitement de la morphologie (Sproat,1992, p.25). Elle permet un gain important sur le plan du stockage informatique, et une maintenance plus aisée. Cet argument vaut également pour les systèmes multilingues. Tiberius (2000, p.707) explique en effet que dans une approche multilingue, une information commune aux langues traitées n'est déclarée qu'une seule fois. La redondance est alors réduite et le stockage est plus transparent et concis. Ceci permet également une plus grande cohérence des lexiques (mêmes informations, mêmes étiquettes, etc.) Elle ajoute qu'une morphologie multilingue peut facilement être étendue à une autre langue proche de celles déjà traitées. Cet argument est également linguistique, et nous y reviendrons plus tard.

3.2. Les motivations linguistiques

Ce que nous classons sous le terme de motivation linguistique peut être en fait divisé en plusieurs axes : (a) un axe historico-linguistique apportant des éléments de preuve à la proximité formelle des langues, (b) un axe informatico-linguistique justifiant l'utilité d'une telle morphologie en TAL, et (c) un axe psycholinguistique, qui compare un module multilingue avec le fonctionnement du cerveau des individus bilingues, comparaison qui selon Sproat (1989, p. 19) reste toujours du plus grand intérêt. Nous étofferons ces trois axes par un axe expérimental (d) qui permettra de voir dans quelle mesure une inférence existe bel et bien entre deux langues.

3.2.1. Les motivations historico-linguistiques

Le premier point commun de ces deux langues que nous voulons traiter ensemble est qu'elles dérivent du latin. Malgré leur développement plus ou moins indépendant, Marie Hédiard (1989, p. 225) souligne les « nombreuses similitudes qui existent entre le français et l'italien, aussi bien sur les plans phonologique et lexical que morphologique et syntaxique ». Parmi ces similitudes, intrinsèquement dépendantes de leur racine commune, nous pouvons citer :

a) Elles font partie du même type de langues, les langues flexionnelles, qui utilisent l'union de différents éléments dans une seule et même forme.

b) Elles possèdent les mêmes phénomènes morphologiques de flexions, de dérivations et de compositions et nous pouvons dire qu'à peu de choses près, ces phénomènes sont tout aussi productifs dans les deux langues.

c) Les procédés de combinaison sont les mêmes. Elles ont toutes deux des suffixes et des préfixes qui se concatènent à la forme de base.

d) Les deux langues possèdent les mêmes informations lexicales.

En suivant la distinction de *Ingria et al.*(*ms*), voici une ébauche de comparaison de ces informations.

Les informations syntaxiques :

Etant donné que ces deux langues sont issues de la même langue, elles ont toutes les deux conservé une structure syntaxique assez identique. Une comparaison de deux manuels de grammaire[12] pour les deux langues ainsi que la consultation des normes EAGLES [13] pour l'étiquetage nous permettent de noter les points suivants.

Au niveau des catégories syntaxiques, elles possèdent toutes deux les mêmes catégories, par exemple le nom, le pronom, le verbe, l'adjectif, l'adverbe. Elles possèdent également les deux la même distinction de genre (masculin et féminin), de nombre (le singulier et le pluriel), le même nombre de personnes (excepté la forme polie qui se forme en français sur la deuxième personne du pluriel et en italien sur la troisième personne du singulier). Elles possèdent également des systèmes de temps et de modes très proches, bien que grammaticalement, la concordance des temps soit différente dans les deux langues.

[12] Dardano M. et Trifone P. La lingua italiana, 1985 et Grevisse M. Le Bon Usage, 1980
[13] Le rapport EAGLES est disponible sur http://www.ilc.pi.cnr.it/EAGLES96/home.html

Les informations morphologiques :

Au niveau morphologique, les deux langues possèdent un phénomène de flexion assez similaire. Il existe pour les deux langues des marques du pluriel et du féminin qui s'ajoutent aux formes de bases. Bien que ces flexions entraînent, en italien comme en français, des changements morphographémiques différents de la forme de base, les deux langues ont une morphologie dite concaténatoire. Elles possèdent des noms et des adjectifs qui ne se déclinent qu'en nombre, et d'autres qui se déclinent en genre et en nombre. Les flexions verbales se lient également par concaténation. D'un point de vue dérivationnel, on retrouve dans les deux cas des capacités à opérer des transformations nominales déverbalisantes (Fr : *manger* →*mangeur*, It : *mangiare* → *mangiatore*), des transformations adjectivales déverbalisantes (Fr : *manger* → *mangeable*, It : *mangiare* → *mangiabile*) des transformations désadjectivales (Fr : *possible* → *possibilité*, It : *possibile* → *possibilità*) et des opérations de préfixation et de suffixation (Fr : *possible* → *impossible*, It : *possibile* → *impossibile*). La proximité des phénomènes morphologiques entre les deux langues est facilement reconnaissable par la comparaison de deux livres de grammaires. Mais cette proximité a également été scientifiquement prouvée par quelques recherches. Citons notamment Nammer (2001, p. 295) qui a identifié des opérateurs morphologiques dont le processus constructionnel est identique, opérateurs qu'elle considère comme équivalent (*-able*, *-ité* et *-iser* pour le français, et *-bile*, *---ità*, et *-izzare* pour l'italien). Mais nous reviendrons sur cette étude de Nammer dans le chapitre 4.

Les informations sémantiques :

Les informations sémantiques seront pour l'instant mises de côté et évoquées brièvement par la suite. En effet, comme nous l'avions déjà évoqué en parlant de la notion de cognat, il peut exister un écart sémantiques plus ou moins important entre deux mots formellement apparentés. Par exemple, le verbe français *fermer* et le verbe italien *fermare* remplissent les critères de cognation évoqués plus haut. Cependant, l'écart sémantique est très important, le mot italien ne signifiant pas *fermer*, mais *s'arrêter*. Un autre exemple est le mot *confetti* qui est un cognat parfait entre les deux langues, mais qui prend, en italien, la signification de dragée. Les correspondances formelles associées à des écarts sémantiques importants sont plus connues sous le nom de « faux amis ». Dans le présent travail, nous n'utiliserons que quelques types sémantiques pour caractériser chaque mot traité et explorer la possibilité de représenter d'une manière multilingue le trait sémantique. Mais le but de ce travail reste l'exploitation des liens

morphologiques multilingues uniquement, dans l'optique de la construction d'un lexique morphosyntaxique sans prendre en compte la sémantique.

Les différences entre les langues :

Une fois individualisées les similitudes entre les langues, il sera bon de réfléchir aux différences qui existent entre elles. En effet, même si le but de ce projet et de tous les systèmes morphologiques basés sur une approche polymorphique est de généraliser les similitudes, il faut néanmoins être conscient des irrégularités morphologiques pour pouvoir (1) évaluer d'une manière objective l'étendue d'un lexique multilingue, (2) prendre conscience des problèmes que posent ces différences, (3) et enfin, gérer de manière propre et efficace ces différences pour pouvoir les inclure dans le lexique. Nous tâcherons, tout au long de ce travail, de garder en tête ces différentes considérations.

L'influence mutuelle des langues :

L'autre aspect intéressant dans le domaine historico-linguistique est d'étudier l'influence que ces deux langues ont eue l'une pour l'autre. Le phénomène de l'emprunt est courant dans de nombreuses langues. Le terme *emprunt* désigne tout élément d'une langue provenant d'une autre langue (Lehmann, 1998, p. 6). Certains emprunts sont dits assimilés lorsqu'ils se « moulent » dans les systèmes phonétique, orthographique et morphologique de la langue empruntante (Lehmann, *ibid*). Par exemple *beafsteack*, mot anglais est devenu *bifteck* en français, et *look* a donné *relooker* en français. Le lexique ne provient donc pas uniquement de la langue d'origine mais il s'enrichit également en empruntant des mots aux autres langues.

> « le lexique français ne s'est pas contenté de développer son héritage latin de toutes les façons possibles, mais […] il a parfaitement su tirer parti de ses contacts avec les usages linguistiques de ses voisins en les adaptant et en les intégrant à ses propres structures [14]».

Dans le cas qui nous concerne plus particulièrement, Henriette Walter[15] ajoute que :

> « Et à travers les siècles, les relations entre la France et l'Italie n'ont pas cessé de provoquer des emprunts de mots, dans un sens comme dans l'autre ».

Plus généralement, les phénomènes d'emprunts et de mondialisation linguistique ont fait l'objet d'études variées. Citons par exemple l'étude d'Henriette Walter (2001), qui a réussi à individualiser 1225 mots qui sont « communs » à onze langues européennes actuelles (provenant

[14] Walter, 1997, p.137
[15] Walter, *ibid.*

de famille linguistique assez différente, comme le grec, le français, le néerlandais et le finnois). Elle définit comme « communs » des mots ayant le même étymon (racine étymologique) et des terminaisons propres à chacune des langues. De plus, ces mots étaient très souvent proches au niveau sémantique.

Citons également le travail de Raymond Geysen qui établit des règles de correspondance entre les mots de sept langues européennes.

> « Etant donné que, par suite de la régularité des transformations progressives du latin aux langues romanes qui en dérivent, les mots latins ont dû passer en français, italien, espagnol (passage vertical), conformément à des lois (phonétiques) précises et rigoureuses, il s'en suit, logiquement et expérimentalement, qu'il doit être possible aussi de passer spontanément d'une langue romane à une autre (passage horizontal) suivant des règles empiriques de correspondance aussi sûres que constantes. [16]»

Cette influence mutuelle entre les langues permet donc de motiver davantage encore la proximité formelle qui existe entre les langues. Il devient alors possible d'envisager les langues comme des entités non-closes, qui s'influencent mutuellement. Ces influences créent des liens à plusieurs niveaux, dont la formalisation semble être relativement productive.

3.2.2. Les motivations informatico-linguistiques

Si la place de stockage évoquée plus haut est un argument purement informatique, d'autres arguments, propres à la linguistique informatique, doivent être cités. Premièrement, une morphologie multilingue est, par essence, plus facilement extensible à une autre langue proche (Tiberius, 2002, p. 707). C'est ainsi que l'estuary english[17] a pu être aisément ajouté au module conçu dans le cadre du projet Polylex (Evans, 1996, cité par Tiberius, 2002 p 704), que nous verrons en détail plus loin. Deuxièmement, la robustesse d'un tel lexique peut permettre de palier le manque de certain mot dans le lexique (*lexical incompleteness*). En effet, en exploitant les informations des deux langues, il sera possible de déduire une entrée lexicale manquante d'une des deux langues, en inférant sa construction à partir de son équivalent dans l'autre langue. Cette déduction peut être incorrecte mais représente la meilleure supposition possible (Tiberius, 2002, p. 707). Enfin, une telle approche peut rendre compte de manière excellente des divergences et des similitudes entre les langues, de leurs évolutions depuis leur racine commune. Ceci peut être d'un grand intérêt pour la recherche linguistique.

---

[16] Geysen Raymond (1990)
[17] nom donné à la forme de l'anglais parlé principalement au sud-ouest de l'Angleterre.

3.2.3. Les motivations psycholinguistiques

Dans toutes applications du traitement automatique de la langue, il est intéressant de se pencher sur la motivation linguistique du sujet traité (Sproat, 1989, p.19). Dans le cadre d'une exploitation des liens morphologiques multilingues, les motivations peuvent être trouvées dans les différentes études faites sur l'organisation du lexique dans le cerveau d'individus bilingues.

Nous l'avions dit dans la première partie, il semblerait que, dans un cerveau monolingue, le lexique soit stocké à deux niveaux de représentation, l'un contenant les significations et l'autre contenant les formes (De Groot, 1998 p. 208). Dans un cerveau bilingue, toute la question est de savoir si cette disposition est multipliée par deux ou s'il y a partage d'information entre les lexiques mentaux propres à chacune des deux langues. Il semblerait que le niveau de signification soit indépendant des langues, et donc unique, et que les formes des mots sont « en général stockées de façon spécifique à chacune des langues (à l'exception peut-être des mots formellement apparentés, cognats en anglais […] [18].»

Le modèle présenté également par De Groot (1998, p. 310) contient donc trois niveaux : deux niveaux pour les formes spécifiques à chaque langue et un niveau pour la représentation conceptuelle, indépendant des langues. Mais ces niveaux sont reliés entre eux d'une manière plus ou moins importante, notamment dans le cas de mots formellement apparentés. Il n'en reste pas moins qu'il est difficile de « comprendre le traitement [des mots cognats] [19]». Le traitement de ces mots dépend de la taille du sous-ensemble lexical contenant les cognats entre deux langues. Et cette taille dépend de la proximité formelle entre les deux langues en présence (Lavaur, 1998, p. 330).

Beaucoup de recherches restent à faire dans le domaine (Lavaur 1998, p. 337), mais ces quelques résultats semblent tout de même confirmer l'existence de liens formels entre les mots de deux langues différentes. Les différentes études citées plus haut insistent également sur les différents paramètres qui conditionnent l'organisation des lexiques chez les individus bilingues (période d'acquisition de la deuxième langue, relation affective avec celle-ci, …). D'un point de vue linguistique et informatique, ces paramètres n'entrent pas en compte.

[18] De Groot, 1998 p. 308
[19] Lavaur 1998, p. 337

3.2.4. Expérience

Nous l'avons dit, intuitivement on constate une réelle proximité de forme entre les lexiques italiens et français. Cette intuition reste purement subjective et nous avons jugé bon de mener à bien une petite expérience pour prouver cette proximité et voir dans quelle mesure la reconnaissance et la génération de mot d'une langue à l'autre étaient bel et bien réel. Cette expérience fera l'objet d'un chapitre entier dans ce mémoire (cf. chapitre 3).

Après avoir décrit les différentes motivations, il faut maintenant nous pencher sur la méthodologie pour construire un lexique multilingue.

4. Que faut-il prendre en compte dans la construction d'un module morphologique multilingue ?

En suivant la 'méthodologie' proposée par Sproat (1989, p.83), il faut se poser de nombreuses questions. La première chose est de savoir à quoi l'on destine le module lexical qu'on envisage de construire. A ce stade de notre recherche, nous en resterons à une approche théorique mais nous évoquerons quelques pistes dans le dernier chapitre. De la décision de l'application de notre module dépend également le type d'information dont nous aurons besoin. Nous allons prendre le problème dans le sens inverse. Nous nous demanderons quelles sont les informations 'inter linguistiques' généralisables et nous verrons ensuite comment inclure les autres informations.

Nous passons en revue cette problématique selon deux approches différentes. Dans la section 4.1, nous verrons comment formaliser les liens morphologiques multilingues en suivant l'approche traditionnelle basée sur le morphème. Dans la section 4.2, nous étudierons les mêmes processus sous l'angle de la morphologie holistique.

4.1. L'exploitation du morphème

Pour traiter les liens morphologiques multilingues, nous nous sommes posé la question de savoir quelles unités de base nous allions utiliser. La segmentation en morphème s'effectue par substitution ou commutation (Lehmann, 1998, p. 107). Dans un contexte multilingue, nous avons donc effectué la même opération mais avec des paires de mots bilingues. Ainsi, pour identifier les éléments contenus dans *possibilità* et *complessità*, nous pouvons effectuer les comparaisons suivantes :

a) *possibilità/possibilité* donne les éléments *possibil* , *–ité* et *-ità*
b) *impossibilità/impossibilité* donne en plus l'élément *im-*
c) *complessità/complessité* donne les éléments *compless-* et *complex-*.

Un tel procédé permet d'envisager les similitudes formelles à un niveau plus profond, en individualisant les éléments formellement identiques et ceux légèrement différents. Par exemple, en (a), le procédé de substitution nous permet d'individualiser une forme commune aux deux langues *'possibil'*. Cette forme commune reste cependant peu motivée linguistiquement parlant. En (b), on distingue un préfixe parfaitement cognat (*im-*) et en (c), ce sont deux formes (*compless-* et *complex-*) qui peuvent être identifiées, deux formes dont la proximité formelle est plus qu'évidente, malgré une petite variation (*ss/x*).

Les exemples ci-dessus montrent que les relations formelles entre les langues se retrouvent à différents endroits dans le mot. Geysen (1990, p. 8) identifie trois types de régularités entre les langues:

- régularités de formes identiques (certains mots sont identiques dans plusieurs langues, par exemple *confetti* en français, comme en italien) ;
- régularités de radicaux communs, par exemple, ***possibi****lité /****possibil****ità* ;
- régularités dans la différence orthographique (le '*x*' du français est souvent changé en '*ss*' en italien : *se**x**e /se**ss**o, comple**x**e/ comple**ss**o*).

L'exploitation des régularités des radicaux communs représentait le premier enjeu de ce travail. Si l'on part du principe qu'une forme de base est un radical, auquel on ajoute un suffixe ou un préfixe, il semblait intéressant de se pencher ensuite sur les morphèmes liés comme les suffixes et les préfixes. Dans ce travail, nous nous sommes intéressé tout d'abord à formaliser la régularité potentielle entre les mots de même base mais dont le suffixe est différent. Pour restreindre l'application, nous nous sommes penchés sur la formalisation de trois types de mots : les mots terminés par les suffixes *–ité/-ità* (et leur dérivé *–té/-tà* et *-eté/-età*), ceux terminés par les suffixes *–tion/-zione* et ceux terminés par les suffixes *-teur/-tore*. A ces deux suffixes, nous ajouterons le traitement de leurs flexions, du pluriel pour les trois cas, et du féminin pour le cas des suffixes *-teur/-tore*. Nous étudierons également la possibilité d'inclure le processus de dérivation préfixale du suffixe *im-*. Nous nous pencherons également sur les variations internes aux mots et à leur traitement.

Nous décrivons maintenant les différentes similitudes au niveau des suffixes (4.1.1), des radicaux (4.1.2) et des préfixes(4.1.3), ainsi que la problématique de la flexion du pluriel(4.1.4).

4.1.1. Les suffixes cognats ou semi-cognats

Penchons nous tout d'abord sur les similitudes entre les suffixes.

En expliquant la notion de cognat, nous avions évoqué le fait que la notion de « cognation » était assez floue. En nous basant sur cette notion, nous avons postulé que les paires *liberté/libertà* et *possibilité/possibilità*, étaient des cognats presque parfaits. En suivant la distinction faite par Geysen, on se rend compte que les deux mots de chaque paire ont des radicaux identiques en forme (cognats parfaits) et que seul le suffixe change. Et ces suffixes, *(-té/-tà, -ité/-ità*) peuvent à leur tour être considérés comme des semi-cognats.

Dans un contexte multilingue, nous pouvons reprendre le schéma de Sproat pour le morphème, mais cette fois-ci pour des morphèmes semi-cognats.

<ETAT, adj/n, -ité/-ità>

Dans cet exemple, le morphème semi-cognat *-ité/-ità* est un suffixe permettant de transformer un adjectif en nom pour en exprimer l'état. Ces deux morphèmes sont semi-cognats de part leur forme et leur fonction dans le processus dérivationnel. Les deux premiers éléments de cette description sont inter-linguistiques, étant donné qu'ils s'appliquent aux deux langues en présence. Seule la forme varie légèrement. Ce suffixe semi-cognat s'attache par concaténation à une base commune aux deux langues.

Nous pouvons appliquer un traitement similaire aux noms qui possèdent une flexion du féminin, comme *accompagnateur- accompagnatrice /accompagnatore- accompagnatrice* et *accusateur - accusatrice* / *accusatore – accusatrice.* A nouveau, ces mots ont des formes de base cognats, et seul les suffixes varient. Mais de par leur similitude, les deux suffixes du masculin sont des cognats semi-parfaits. On peut donc les représenter de la manière suivante :

<AGENT , -teur/tore>

On remarque également que les suffixes du féminin sont quant à eux des suffixes cognats parfait (*-trice/-trice*), ce qui fait que les mots possédant ce suffixe et dont la base est commune ont une forme du féminin parfaitement cognate (*accompagnatrice/accompagnatrice accusatrice/accusatrice*). Les formes de base auxquelles sont liées les suffixes (*possibil-*, *accompagna-*, ou encore *matur-*) ne représentent que la partie du surface commune aux deux langues. Elle n'est pas très motivée linguistiquement parlant. Les radicaux sont généralement

considérés comme des formes indépendantes (morphème libre) (Sioufi, 1999 : p. 45). Mais dans ce cas, ils seront donc considérés comme des morphèmes liés.

4.1.2. Les variations internes

Certains radicaux ne sont pas des cognats parfaits, mais des semi-cognats (*complexité/complessità*, *traducteur/traduttore*). En étudiant très superficiellement un certain nombre de ces mots, et en suivant les théories de Geyser qui postulait une régularité dans ces différences orthographiques, nous nous sommes également penché sur la possibilité de traiter les changements internes aux radicaux. Dans l'exemple de *complexité/complessità*, la variation interne porte sur le '*x*' qui devient '*ss*', et dans l'exemple de *traducteur/traduttore* la différence porte sur la paire '*ct*' qui devient '*tt*'. La différence orthographique repérée entre *x/ss* se retrouve dans de nombreuses paires de mots (*sexe/sesso*, *annexe/annesso*, …), de même pour la variation *ct/tt* (*conducteur/conduttore*, *acteur/attore*). Le fait que ces différences soient fréquentes est un argument de plus à la validité d'une généralisation de cette variation. Cependant, informatiquement parlant, elle reste assez problématique à formaliser. Mais nous reviendrons sur ce point quand nous aborderons la partie pratique de ce travail.

4.1.3. Les préfixes cognats ou semi-cognats

Le préfixe est lui aussi généralisable. En effet, nous avons remarqué que les paires *possibilité/possibilità* et *maturité/maturità* pouvaient se dériver en *impossibilité /impossibilità* et *immaturité/ immaturità*. Le préfixe *im-* semble être un cognat parfait et peut donc être représenté de la sorte

<OPP, n/n im->

Dans cet exemple, la fonction d'opposition <OPP> est commune aux deux langues, et la forme graphique du préfixe également.

En résumé, l'exploitation des suffixes et des préfixes cognats ou semi-cognats permet donc une rationalisation des informations. Les mots évoqués ci-dessus peuvent donc être représentés ainsi :

<OPP, -im> <BASE> (<sem= état> <cat=n>
suffixe = -ità <lang=it>
suffixe = -ité <lang=fr>)

Pour les mots dont le suffixe est –teur/-tore, la représentation prend une forme un peu plus complexe :

<BASE> (<sem=agent> <cat=n>
<genre=m> (suffixe = -tore <lang=it>
suffixe = -teur <lang=fr>)
<genre=f> (suffixe = -trice <lang=it>
suffixe = -trice <lang=fr>))

Dans les exemples ci-dessus, les éléments sans informations de langue sont des éléments communs aux deux langues traitées. Les éléments marqués par des attributs langues (<lang=it> ou <lang=fr>) sont spécifiques à chaque langue. Concernant les suffixes féminins –trice/-trice, le fait qu'ils soient cognats parfaits aurait pu motiver un groupement dans la même structure. Cependant, ils sont les deux dérivés de la forme du masculin, qui dans ce cas, n'est pas un cognat parfait mais un semi-cognat. De plus, dans un souci de cohérence, il est important de faire la disjonction au même endroit pour les quatre cas, permettant un gestion plus aisée de la flexion du pluriel.

4.1.4. Les flexions du pluriel

Pour assurer une certaine complétude à notre lexique, il a également fallu se pencher sur les flexions du pluriel. Par exemple, pour le cas de la paire *possibilité/possibilità* les flexions du pluriel sont caractérisées par l'ajout du 's' final en français, et par l'absence de marque de pluriel pour les noms finissants par une voyelle accentuée en italien (voyelle tonique) (Dardano, 1985, p. 116). Pour le cas de *accompagnateur/accompagnatore* le pluriel du mot français consiste également en l'ajout du 's' final et le pluriel du mot italien consiste en la transformation du 'e'

en ‘i’. Donc, les marques du pluriel ne peuvent pas être considérées comme cognates et elles seront traitées individuellement pour chaque langue. De plus, dans les langues concaténatoires, ce sont les terminaisons du mot qui bien souvent influencent l’utilisation de telle ou telle marque du pluriel. Et dans les exemples donnés ci-dessus, étant donné que les suffixes sont disjonctifs (une pour l’italien, une pour le français), il semble plus logique de traiter les pluriels après le traitement de ceux-ci, une fois que la langue est spécifiée. De plus, le fait de traiter les pluriels de manière indépendante peut permettre l’ajout d’un mot non cognat, en le reliant uniquement à la flexion de sa langue. Ainsi, les mots totalement non-cognats, comme *maison* et *casa* trouveront tout de même leur place dans un tel lexique, en étant relié aux informations du pluriel, respectivement du français et de l’italien.

4.2. Une morphologie multilingue basée sur la morphologie holistique

A l’inverse de l’exploitation des morphèmes décrits plus haut, la théorie de la morphologie holistique pourrait être une autre possibilité pour découvrir et formaliser les similitudes entre les mots. Dans le cadre restreint de ce travail, nous ne ferons que survoler les possibilités qu’offre cette théorie, étant donné sa relative nouveauté et son manque de formalisme encore disponible.

Si l’on reprend l’hypothèse de départ de cette théorie, qui veut que :

> « Etant donné 4 mots (w1, w2, w3, et w4), w1 est considéré comme morphologiquement lié à w2 si et seulement si les propriétés formelles, sémantiques et grammaticales qui les distinguent, distinguent également w3 et w4 »[20]

nous pourrions facilement imaginer la même opération avec des paires de mots bilingues. Ainsi, étant donné le mot *possibilità*, celui-ci peut être formellement apparenté au mot *possibilité* si et seulement si les propriété formelles, sémantiques et grammaticales qui les distinguent, distinguent également *maturité* et *maturità*.

Formellement, les différences entre les paires de mots suivantes :

possibilità	possibilité
maturità	maturité
identità	identité

[20] Neuvel et Singh (sous-presse)

pourraient être formalisées comme suit :

possibilità	possibilité	/…ità/ • /…ité/
maturità	maturité	/…ità/ • /…ité/
identità	identité	/…ità/ • /…ité/

où l'on remarque aisément que les différences sont les mêmes et qu'une stratégie morphologique pourrait être définie ainsi :

/Xità/• /Xité/

où X est une variable. Pour contraindre la stratégie, une étude plus poussée des paires de mots correctes ou incorrectes devraient être effectuée pour définir des contraintes.

Dans un contexte multilingue, cette nouvelle approche a le mérite de ne pas se heurter aux écueils d'éléments indésirables et linguistiquement peu motivés. Par exemple, la mise en évidence de radicaux parfaitement cognats est une pure vision de l'esprit qui n'est pas toujours motivé linguistiquement. En effet, nous avions individualisé *liber-* et *possibil-* des paires *liberté/libertà* et *possibilité/possibilità* uniquement sur le critère de similitude de leur forme, sans qu'ils soient pour autant des racines communes du point de vue linguistique ou étymologique.

5. Conclusion

Les langues possèdent donc entre elles des similitudes plus ou moins importantes qu'il est intéressant d'exploiter. Dans le cadre du traitement automatique des langues naturelles, l'exploitation des liens morphologiques multilingues semble en effet apporter de nombreux avantages. Mais même si d'un point de vue théorique ces similitudes et la régularité de ces similitudes sont désormais évidentes, la construction d'un lexique multilingue reste une problématique importante et soulève de nombreuses questions sur un plan pratique. L'exploitation des cognats, (parfaits ou non), semble être une piste intéressante pour la construction d'un tel système. Par la suite, après avoir fait un état de l'art des systèmes actuels exploitants les liens morphologiques multilingues, nous nous pencherons sur la réalisation pratique d'un tel système, qui nous permettra de mettre en évidence les difficultés inhérentes au sujet. Mais avant cela, nous décrivons dans le chapitre suivant une expérience réalisée dans le but d'évaluer l'existence de l'inférence entre l'italien et le français.

Chapitre 3: Expérience sur l'inférence des mots cognats français/italien

1. Introduction

Dans le chapitre 2, nous avions évoqué les motivations informatiques et linguistiques d'un lexique multilingue. Parmi les motivations linguistiques, nous avions parlé de psycholinguistique, de linguistique informatique et de linguistique historique. Ces trois motivations peuvent être considérées comme théorique et nous avons ressenti le besoin d'étoffer ces théories par une expérience plus pratique. En effet, si la théorie apporte de nombreuses motivations quant à la proximité formelle entre les lexiques de certaines langues, une expérience pratique pourra nous aider à quantifier l'inférence d'un mot d'une langue à partir d'une autre langue par des sujets humains. L'implication de sujets humains dans une telle démarche fait s'apparenter l'expérience au domaine de la psycholinguistique, même si ce ne sont pas les processus mentaux qui seront étudiés, mais uniquement le résultat de ces processus.

Dans le projet Polylex, l'inférence semblait être donnée comme acquise. En effet, dans la présentation de ce projet, ses concepteurs maintenaient que :

> « Si un locuteur anglophone qui sait un peu de japonais et regarde dans un dictionnaire bilingue un mot peu commun en anglais et qu'il ne le trouve pas, il ne peut plus rien faire. Mais si un allemand qui sait un peu de néerlandais ne peut pas trouver un mot dans son dictionnaire allemand/néerlandais, il pourra sans doute deviner ce que pourrait être l'équivalent néerlandais, quelles seront ses propriétés syntaxiques, comment ces flexions fonctionnent et comment il doit être écrit et prononcé.[21] »

Le but de l'expérience qui va suivre est de montrer d'une manière un peu plus systématique que tout sujet humain est capable d'inférer de manière intuitive un mot d'une langue étrangère à partir d'un mot de sa propre langue, particulièrement quand les langues sont étroitement liées.

[21] The Polylex Web page, http://www.cogs.susx.ac.uk/lab/nlp/polylex/

2. Procédure

2.1. Panel

Nous avons donné l'enquête à un panel de 12 personnes francophones, agés entre 20 et 60 ans, et n'ayant aucune connaissance de l'italien. La première question de l'enquête portait sur les langues qu'ils connaissaient, notamment le latin. Cette question nous semblait importante pour voir si ce facteur influençait les réponses. Cependant, la connaissance du latin s'est révélé trop difficile à quantifier. En effet, certains répondants l'avaient appris dans leur scolarité mais ne l'avaient pas utilisé depuis. D'autres semblaient montrer une certaine aisance avec la langue latine.

En ce qui concerne les autres langues parlées, ce paramètre nous a au contraire semblé pertinent, étant donné que tous les répondants semblaient les connaître de manière suffisamment active et avancée. Dans le cadre de cette enquête, nous nous sommes donc essentiellement concentrés sur la connaissance d'une autre langue latine. Sur les 12 répondants, 7 ne connaissaient aucune autre langue latine, et 5 connaissaient une autre langue latine, dans les 5 cas, l'espagnol.

2.2. L'enquête

Nous avons divisé l'enquête proprement dite en deux parties. Dans la première, nous avons demandé aux répondants de traduire spontanément en français des mots donnés en italien. Cette partie tentait de prouver l'inférence d'une langue étrangère (l'italien) par des sujets francophones. Les mots étaient présentés sous forme de listes aux répondants, regroupés selon leur catégorie syntaxique.

La deuxième partie, à l'inverse, demandait aux répondants de traduire des mots français en italien. Une liste de paires de mots français-italien formellement apparentés leur était fournie (comme par exemple, les paires : *posizione* / *position*, *settore* / *secteur*, *conformità* / *conformité*). Il s'agissait de voir si l'inférence à partir d'observation de paires attestées se faisait aussi dans le sens inverse de celui de la première partie. Plus précisément, nous voulions voir si les régularités des changements morphographémiques (du préfixe ou de la forme de base) étaient spontanément inférées par les répondants.

Ainsi, le processus de la première partie pourrait s'apparenter à celui de l'analyse en linguistique informatique (reconnaissance de forme et attribution d'information pour en traduire le sens (Bouillon *et al*, 1998, p. 6) et celui de la deuxième partie s'approche d'avantage de la

génération (prise en compte de différents éléments pour produire une forme (Bouillon et al, *ibid.*)).

2.3. Constitution du corpus

Pour la partie concernant l'inférence dans l'analyse, et afin d'obtenir une liste de mots suffisamment objective, nous avons extrait des mots d'un corpus composé de trois articles tiré du journal italien *La Repubblica* du 3 septembre 2002, composé en tout de 1250 mots et préalablement étiqueté à l'aide de l'étiqueteur morphologique Mmorph de l'ISSCO[22].

De cette première liste, nous avons individualisé quatre groupes de mots en fonction de leur catégorie: les adjectifs, les noms, les verbes et les adverbes. En effet, ces catégories sont considérées comme des mots pleins (au niveau sémantique), par opposition au mots vides (préposition, déterminant, etc.). Ils constituent également la plus grande partie du lexique. C'est également dans ces 4 catégories que s'effectuent les différents processus morphologiques (flexion, dérivation, composition). Nous avons ensuite extrait plus ou moins aléatoirement un certain nombre de mots pour chacune des catégories choisies. L'extraction aléatoire permettait de choisir notre corpus de base, sans aucune considération subjective. Les résultats obtenus pourront donc également donner un ordre d'idée sur le nombre de cognats existant entre l'italien et le français.

De plus, nous avons annoté les mots pour savoir s'ils étaient sous leur forme lemmatisée ou non. Pour les adverbes, nous les avons divisés en deux groupes, les adverbes formés sur des adjectifs et les adverbes simples. Cette distinction nous permettra de voir si ce paramètre influence également le degré d'inférence.

Nous sommes donc arrivés au corpus d'enquête suivant:

	lemmes	Non-lemmes	Total
Noms	22	17	39
Adjectifs	16	18	34
Verbes	15	15	30

	Simple	Formés sur un adjectif	
Adverbes	8	9	17
		Total	120 mots

[22] Mmorph, Petitpierre et Russel, 1995. : Automate à deux niveaux développé à l'ISSCO et disponible à l'adresse suivante http://www.issco.unige.ch/tools/

Pour la deuxième partie de l'enquête concernant l'inférence dans la génération, nous avons fourni aux répondants une liste de paires de mots français/italien plus ou moins formellement apparentés. Ces mots ont été choisis en fonction de plusieurs critères. Certains avaient des suffixes semi-cognats, mais la base était parfaitement cognate (*posizione–position*, *associazione–association*, *conformità–conformité, continuità–continuité*) et d'autres, avaient un suffixe et une base semi-cognats mais les changements morphographémiques de la base étaient potentiellement généralisables (*amministrazione–administration*, *coerenza–cohérence*, *complesso–complexe*, *competenza–compétence*, *settore–secteur*, *conduttore–conducteur*, *attrice–actrice*, *credibilità–crédibilité*).

Nous avons ensuite fourni une autre liste de mots français en demandant aux répondants de déduire, après avoir observé attentivement la liste de paires de mots évoquées ci-dessus, les mots italiens qu'ils pensaient être équivalents aux mots français.

3. Dépouillement

Pour dépouiller les résultats de la première partie de l'enquête, nous avons multiplié le nombre de mots par le nombre de répondants. Nous obtenons ainsi un nombre absolu de réponses (au total, 1440 mots). Le dépouillement a été dirigé par un certain nombre de considérations (voir point 3.1 ci-dessous).

La deuxième partie de l'enquête a été dépouillée selon le même principe. Nous avions proposé 12 mots aux 12 répondants, le nombre de réponses absolues était donc de 144. Nous avons également dirigé notre dépouillement selon un certain nombre de questions (voir point 3.2 ci-dessous).

3.1. Les différents axes de dépouillement de la première partie

Pour la première partie de l'enquête, le premier point intéressant concernait la proportion de mots reconnus et de mot non-reconnus (section 4.1). Dans un premier temps, nous avons considéré comme mot reconnu les réponses données qui montraient que la base du mot avait été correctement inférée, sans prendre en compte la justesse du suffixe. Ainsi, pour le mot italien *osservatore*, les réponses *observateur* et *observatoire* ont été considérées comme correctes. De même, pour le mot italien *direttore*, et les réponses *directeur* et *direction.* Dans le même ordre d'idée, nous n'avons pas pris en compte les erreurs relatives aux flexions du pluriel ou du féminin (par exemple, la réponse *disposition* était considérée comme correcte, même si elle

n'avait pas la marque du pluriel que demandait le mot italien *disposizion**i***). Un mot était donc jugé correct si sa forme de base était juste, même si une marque du pluriel ou du féminin manquait.

Les réponses vides ou les réponses ne correspondant pas du tout au mot italien proposé ont reçu l'étiquette "non reconnu".

Ce dépouillement nous permettait de connaître la proportion de mots reconnus c'est-à-dire, le nombre de cas où l'inférence entre deux entrées lexicales existait, même partiellement.

La deuxième question consistait à être plus précis dans le jugement de la reconnaissance, pour savoir parmi les mots reconnus ceux qui l'étaient complètement et ceux qui l' étaient partiellement. Nous avons donc, parmi les mots étiquetés comme corrects, fait une différence entre les mots reconnus complètement (racine, suffixe, flexion) et les mots dont un de ces éléments était erroné. Ainsi, les paires *osservatore/observatoire* et *direttore/direction* ont été jugées "pas entièrement correctes". Par contre, les paires *osservatore/observateur* et *direttore/directeur* ont été jugées totalement correctes.

Une fois individualisées les proportions de mots reconnus complètement et partiellement et les mots non reconnus, nous avons étudié quels paramètres pourraient influencer cette reconnaissance, et donc cette inférence. Nous avons donc regardé de plus près les trois paramètres suivants :

1. Catégories plus reconnues qu'une autre (point 4.2 plus bas) ;
2. Lemme vs. non-lemmes (point 4.3 plus bas) ;
3. La connaissance d'une autre langue (point 4.4 plus bas).

3.2. Les différents axes de dépouillement de la deuxième partie

Concernant la deuxième partie de l'enquête, nous avons choisi d'analyser les résultats en fonction de deux critères distincts. Nous avons tout d'abord cherché à savoir si le suffixe semi-cognat avait été facilement assimilé par les répondants et donc "spontanément" reproduit lors de la traduction (les résultats de ce premier dépouillement se trouvent au point 5.1). Ensuite, il s'agissait de voir si les changements morphographémiques internes (facilement généralisables comme *a**mm**issione–a**dm**ission*) étaient facilement reconnaissables par les locuteurs d'une autre langue, et facilement reproduisibles (voir point 5.2 pour les réponses à cette question).

3.3. Marges d’erreurs

Cette étude n’a aucune prétention statistique, et cela pour plusieurs raisons. Premièrement, le dépouillement des réponses ayant été effectué manuellement, il reste assez subjectif. Deuxièmement, le nombre de mots fournis est assez limité pour ne pas allonger le temps de réponse au questionnaire. Troisièmement, le nombre de répondants est lui aussi très restreint et assez homogène. Il nous avait semblé intéressant de voir si la connaissance d’une autre langue romane influençait les réponses et c’est dans ce but que nous avions demandé aux répondants s’ils connaissaient une autre langue. Malheureusement, tous ceux qui ont répondu par l’affirmative ont déclaré connaître l’espagnol, ce qui rend le panel trop homogène.

Tout ceci amoindrit la valeur statistique d’une telle étude. Cependant, elle nous permettra tout de même de dégager une certaine tendance, d’évaluer les paramètres qui influencent les réponses et de tracer ainsi les grandes lignes d’une étude à plus grande échelle.

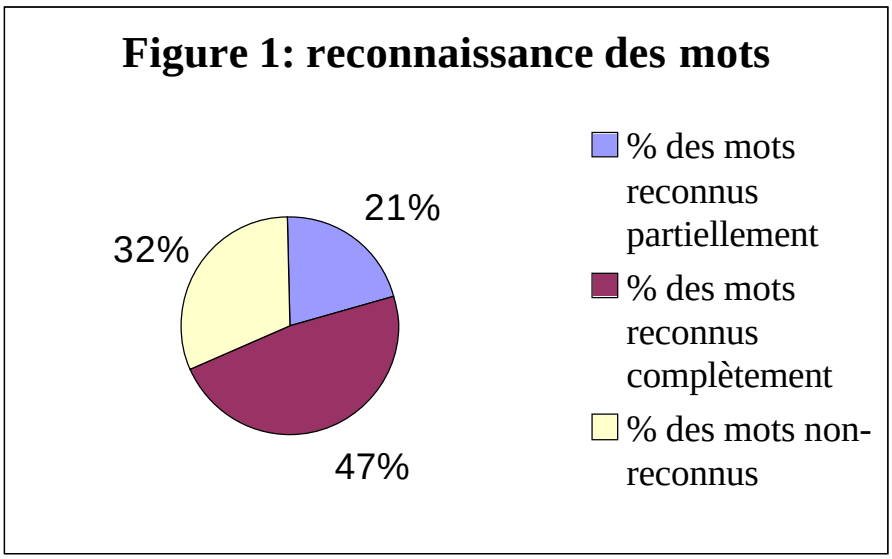
Dans la suite, nous présentons les résultats des deux parties de l’enquête puis tirons quelques conclusions.

4. Résultats de la première partie sur l'analyse

Dans cette section, nous présentons les résultats du dépouillement de la première partie de l'enquête. Nous montrons les différents taux de reconnaissance en fonction des paramètres évoqués plus haut.

4.1. Résultats généraux

La figure 1 reprend le taux de reconnaissance entre des mots, partiellement et totalement.



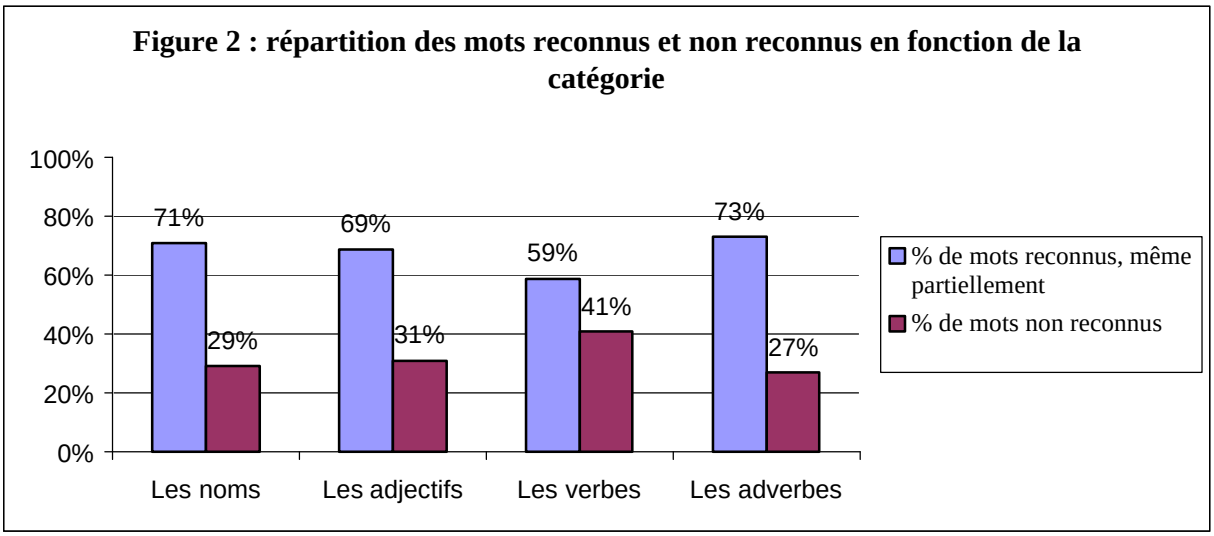
Ces premiers résultats montrent clairement qu'une grande majorité de mots sont globalement reconnus (68 %) et que 32 % des mots ne sont pas du tout reconnus. De plus, il est intéressant de souligner, déjà à ce niveau superficiel d'analyse, que 47 % des mots sont complètement reconnus.

Nous affinons à présent notre dépouillement en montrant les différents taux de reconnaissance en fonction de plusieurs paramètres, en commençant par celui la catégorie syntaxique.

4.2. Taux de reconnaissance en fonction de la catégorie syntaxique

Dans la figure 2 ci-dessous, nous mettons en parallèle la reconnaissance des mots en fonction de leur catégorie syntaxique. Nous avons additionné les pourcentages de reconnaissance partielle et complète.

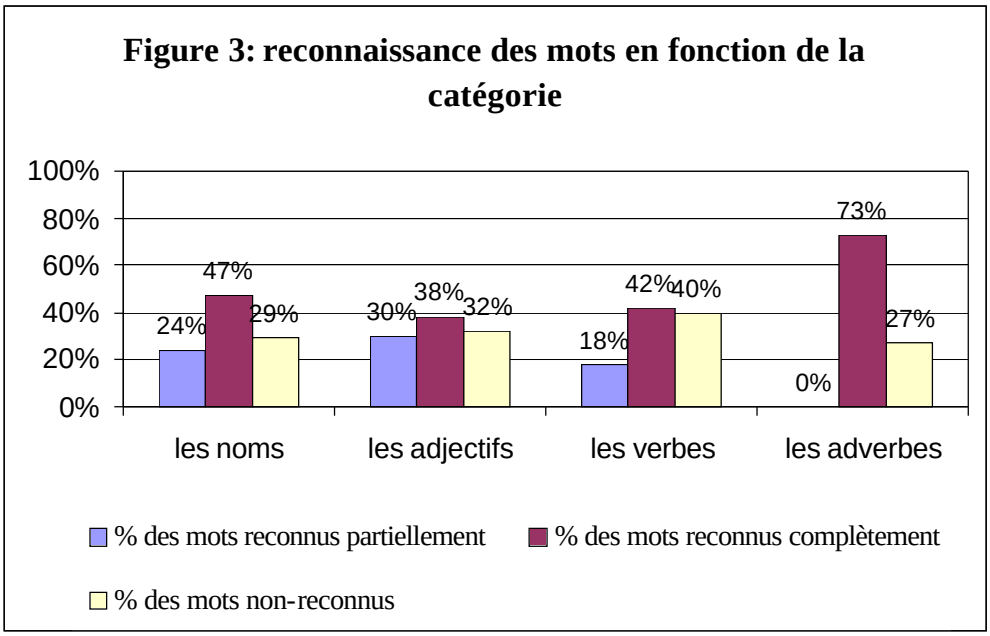
Figure 2 : répartition des mots reconnus et non reconnus en fonction de la catégorie



La figure 2 montre clairement que toutes les catégories n'ont pas le même taux de reconnaissance. Alors que les noms, les adjectifs et les adverbes ont tous un taux de reconnaissance qui se situe autour de 70 %, les verbes atteignent seulement 59 % de reconnaissance.

En affinant le dépouillement et en distinguant la reconnaissance exacte de la reconnaissance approximative, on obtient les résultats suivants:

Figure 3: reconnaissance des mots en fonction de la catégorie



D'une manière générale, nous remarquons que toutes les catégories sont majoritairement reconnues complètement.

Concernant les adverbes, la première constatation évidente est qu'ils sont soit reconnus complètement, soit pas du tout reconnus. Nous reviendrons sur ce point lorsque nous évoquerons le phénomène des flexions (Section 4.3). Au niveau de la reconnaissance partielle, ce sont à nouveau les verbes qui sont le moins reconnus. Il semblerait que, tout comme les adverbes mais dans une moindre mesure, les verbes aient une tendance à être soit complètement reconnus, soit pas du tout.

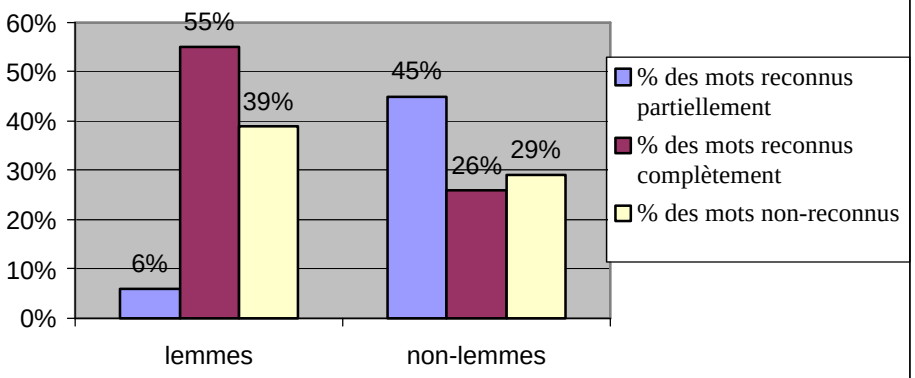
A ce premier niveau d'analyse, une première conclusion s'impose sur l'importance du taux de reconnaissance parmi ces mots qui, rappelons-le, ont été choisis de manière aléatoire. Ces premiers résultats confirment l'existence d'une inférence, même partielle, entre ces deux langues proches. De plus, la catégorie semble influencer les résultats. En effet, si la reconnaissance des mots semble identique pour les noms, les adjectifs et les adverbes, il semblerait que les verbes soient moins facilement reconnus.

Nous affinons à présent notre dépouillement pour voir quels autres paramètres pertinents influencent les taux de reconnaissance.

4.3. Comparaison des résultats en fonction du type de mot (lemmes ou non-lemmes)

Pour affiner nos résultats, nous avons d'abord effectué quelques statistiques en fonction de la lemmatisation du mot

figure 4: reconnaissance de mot en fonction de leur lemmatisation

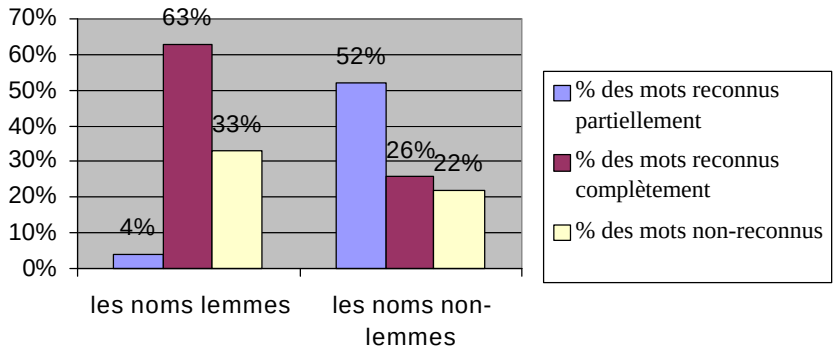


La figure 4 ci-dessus ne concerne que les mots qui peuvent apparaître sous une forme fléchie (c'est-à-dire, les noms, les verbes et les adjectifs). Nous verrons par la suite le cas particulier des adverbes. Nous voyons clairement dans cette figure ci-dessus que la forme du mot influence la reconnaissance : les mots lemmatisés sont davantage reconnus complètement que les mots non-lemmatisés. Il semblerait donc que la terminaison des mots (flexion ou dérivation) complique la reconnaissance complète. Par exemple, le mot pluriel italien *esperienze*, n'a pas été reconnu complètement. La plupart des répondants ont reconnu qu'il s'agissait du mot français *expérience*, mais n'ont pas fléchi le mot correctement dans sa forme du pluriel.

Voyons à présent si ce paramètre de la lemmatisation a une influence similaire pour chaque catégorie.

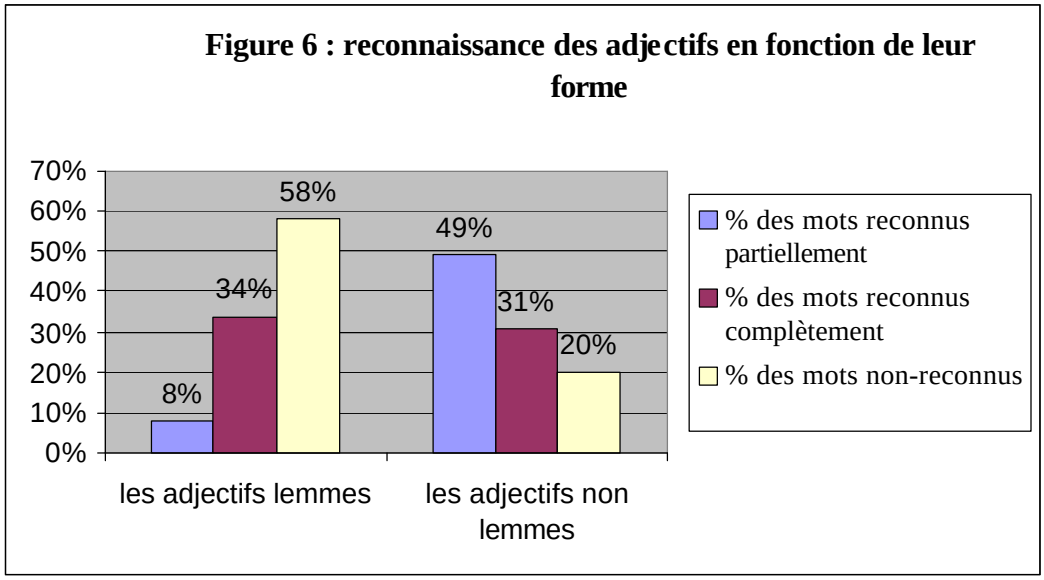
Pour les noms, nous obtenons les résultats suivants :

Figure 5: Reconnaissance des noms en fonction de leur forme



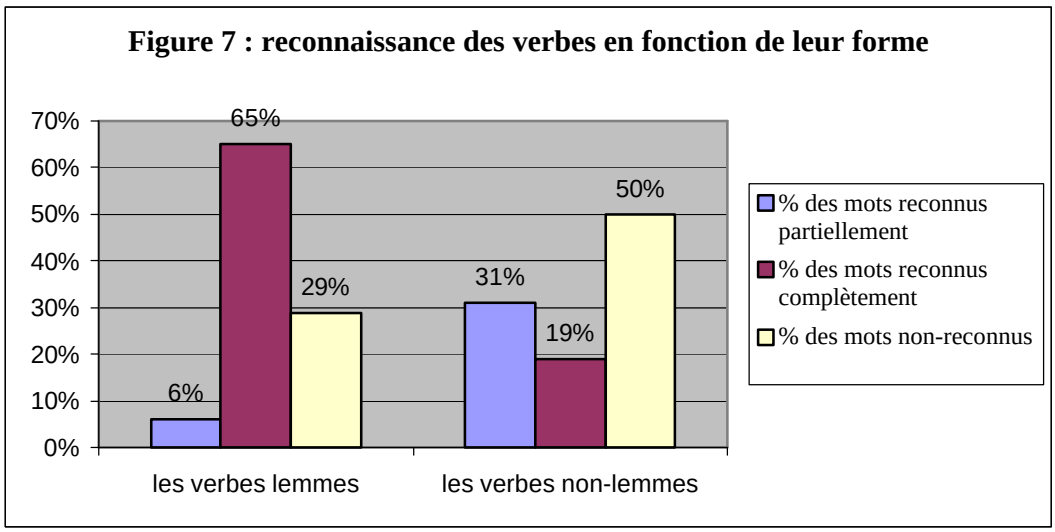
Ce graphique confirme l'influence de la forme du mot sur la reconnaissance totale ou partielle. En effet, la plupart des mots lemmatisés sont reconnus complètement. Il semblerait que la forme de base permette une reconnaissance complète plus aisée. En revanche, pour les mots fléchis, c'est la reconnaissance partielle qui est la plus importante. En effet, les répondants reconnaissaient bien la forme de base, mais la flexion ou la dérivation du mot n'était pas souvent reconnue (uniquement dans 26% des cas). Etant donné que notre distinction entre complètement reconnu et partiellement reconnu portait essentiellement sur la reconnaissance de la forme de base (cf. les exemples concernant *disposizioni / disposition* ou *dispositions*), cette différence de reconnaissance nous mène à dire que les flexions sont plus difficiles à reconnaître. Ce point rejoint notre appréciation linguistique des flexions des deux langues (chapitre 2), où nous avions mis en évidence que les deux systèmes de flexions n'avaient pas une grande cognation.

Les adjectifs:



Tout comme pour les noms, les adjectifs présentent des types de reconnaissance différents en fonction de leur formes morphologiques. Mais la différence entre les deux taux de reconnaissance complète est moins grande que pour les noms (seulement 3% - 34% moins 31%). En revanche, la différence des pourcentages de non-reconnaissance est la même pour les noms et les adjectifs. On voit en effet qu'il y a moins de mots non-lemmes non reconnus que de mots lemmes non reconnus.

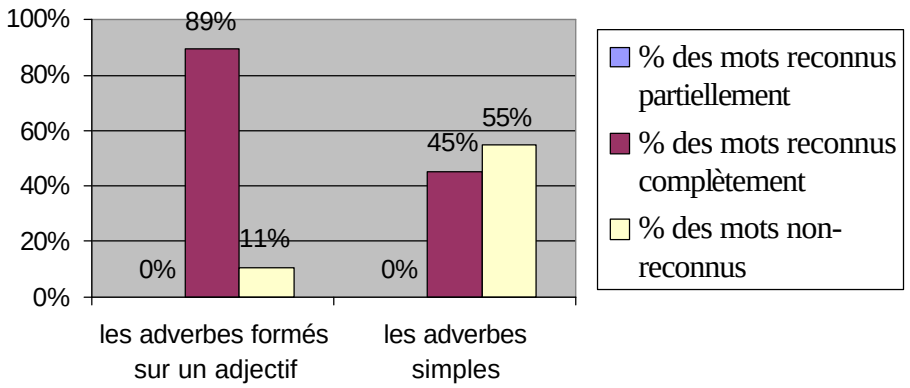
Les verbes :



Ici, la différence est encore plus marquante. En effet, les verbes lemmatisés sont davantage reconnus que les non-lemmatisés. La flexion du verbe semble donc jouer un rôle dans la reconnaissance de verbes fléchis, peut-être encore plus important que pour les noms et les adjectifs.

Les adverbes:

Figure 8: reconnaissance des adverbes en fonction de leur forme



Dans la figure 8 ci-dessus, on remarque que le type d'adverbe influence également le taux de reconnaissance. La première constatation porte sur le taux de reconnaissance partielle et complète. En effet, les adverbes sont soit totalement reconnus, soit pas du tout reconnus. Les adverbes simples sont moins fréquemment reconnus que les adverbes formés sur un adjectif. Nous reviendrons sur ce phénomène dans notre analyse (section 6.2).

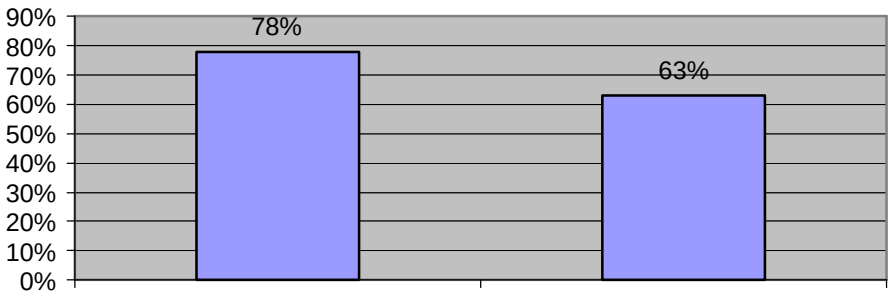
Dans la suite, nous poursuivons notre analyse en nous intéressant aux types de répondants, pour voir si ce paramètre influence également les résultats.

4.4. Comparaison des résultats en fonction du type de répondant

Nous avons également comparé les réponses des répondants connaissant une autre langue latine (plus loin, les romanistes) avec celles des répondants ne connaissant aucune autre langue latine (plus loin, les non-romanistes) pour voir si ce paramètre influence les réponses.

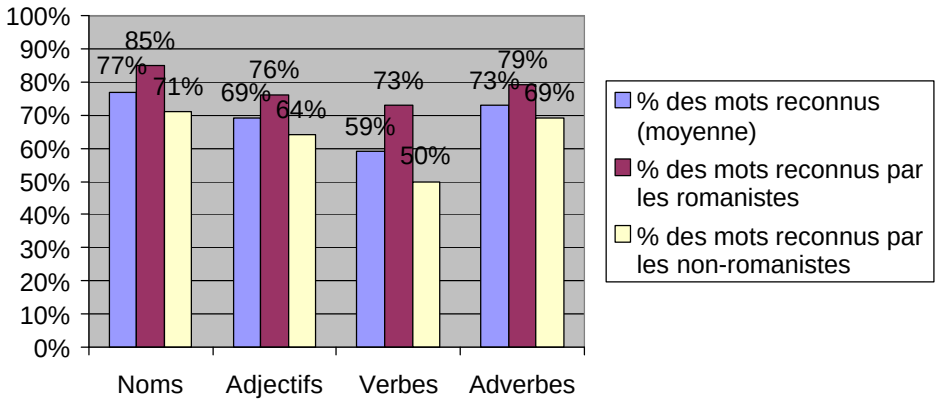
D'une manière générale, la différence est assez bien marquée, comme le montre la figure 9 où nous avons additionné la reconnaissance globale et la reconnaissance partielle:

Figure 9: comparaison du taux de reconnaissance en fonction du type de répondant



Nous avons également mis en parallèle les taux de reconnaissance totale des romanistes, des non-romanistes, et la moyenne des deux (figure 10).

Figure 10: taux de reconnaissance par catégorie



Cette figure montre que la différence entre les réponses des romanistes et des non-romanistes se retrouve pour chaque catégorie de manière plus ou moins identique. Les romanistes ont un taux de reconnaissance toujours au-dessus de la moyenne, à l'inverse des non-romanistes qui sont toujours en dessous de la moyenne. Mais l'écart entre les deux taux de reconnaissance montre une variation en fonction de la catégorie. En effet, si l'écart entre romanistes et non romanistes est de 12 % pour les noms et les adjectifs (et seulement de 10% pour les adverbes), celui pour la catégorie des verbes se creuse pour atteindre les 23 %. Tout comme pour le taux de reconnaissance absolu par l'ensemble des répondants (Figure 2), c'est donc la catégories des verbes qui semble se détacher des autres. En revanche, les taux de reconnaissance en fonction des catégories sont plus ou moins semblables aux taux de reconnaissance remarqués plus haut, quand nous ne faisions aucune distinction entre les répondants.

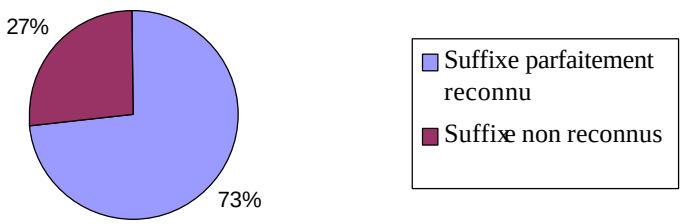
5. Résultats de la partie sur la génération

Dans cette deuxième partie de l'enquête, nous avions donc proposé un certain nombre de paires de mots avec différents suffixes considérés comme formellement similaires, comme *continuità/continuité*. Nous avions ensuite demandé aux répondants de traduire eux-mêmes en italien des mots français, en se basant sur les exemples donnés. Nous avons distingués deux axes de dépouillement : nous avons tout d'abord analysé les réponses en nous intéressant à la génération des suffixes (voir 5.1), puis nous avons analysé les réponses concernant la génération des formes de bases (voir 5.2).

5.1. Résultats de la génération des suffixes semi-cognats

Les paires de suffixes semi-cognats ont été majoritairement bien interprétées pour la génération, comme le montre la figure 11 ci-dessous.

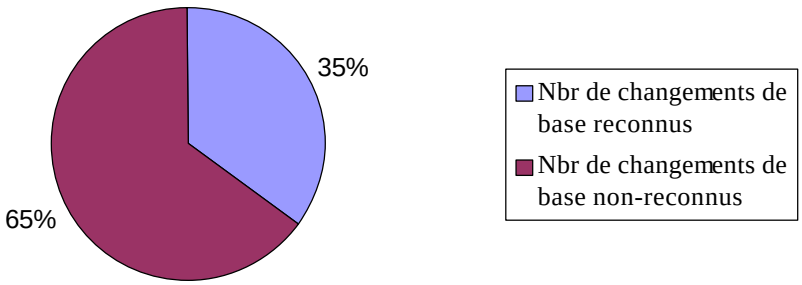
Figure 11: Génération des suffixes



5.2. Résultats de la génération des bases semi-cognates

Pour les mots avec de petits changements morphographémiques internes, nous présentons les résultats dans la figure 12.

Figure 12 : Génération du changement dans la base



Cette figure 12 montre clairement que les changements internes sont moins identifiés par les répondants que les suffixes (cf. figure 11). Mais ce ne sont que des valeurs absolues. Il est donc intéressant de voir quels sont les changements les mieux reconnus. Ainsi, sur les 12 répondants, 9 (75 %) ont reconnu le changement de 'ct' en 'tt' dans la paire *traduc**ct**eur-tradu**tt**ore*, et 8 (67 %) l'effacement de l'accent aigu dans la paire *pr**é**sidence-pr**e**sidenza*. Ces deux proportions sont nettement au-dessus de la moyenne de reconnaissance de ces changements (35 %). Les autres changements ont été reconnus par moins de 25 % des répondants.

Pour cette partie pour la génération, nous ne présentons pas les résultats des romanistes et des non-romanistes, étant donné qu'ils étaient très variés et peu cohérents. Il semblerait en effet que le petit nombre de mots proposés ne permette pas de dégager une tendance entre ces deux types de répondants.

6. Analyse générale des résultats

6.1. L'influences des catégories dans l'analyse

Nous l'avons vu, les catégories influencent la reconnaissance des mots. Si l'on en reste à ce critère, il est difficile de tirer une vraie conclusion de cette constatation. Cette tendance montre-t-elle davantage de proximité formelle entre les langues au niveau des adjectifs et des noms, et beaucoup moins au niveau des verbes ? Une étude aussi restreinte ne permet pas de le dire. Cependant, cette différence de reconnaissance entre les catégories reste un fait important à étudier plus à fond.

6.2. Les différences entre les lemmes /non-lemmes dans l'analyse

Les résultats montrent aussi que le paramètre de la lemmatisation des mots proposés aux répondants a significativement influencé les réponses. Si l'on compare le taux de reconnaissance complète des mots lemmatisés (55%) à celui des mots non-lemmasisés (26 %), on constate une différence importante. De plus, cette différence semble être inversée si l'on compare les deux taux de reconnaissance partielle (6 % de mots lemmatisés et 45 % des mots non-lemmatisés). Ceci semble être facilement explicable.

Il semblerait en effet que ce qui constitue la différence entre les lemmes et les non-lemmes, c'est-à-dire les flexions, soit moins formellement apparenté que le lemme de ces mots. Par exemple, les mots italiens *abusi* et *conferenze* ont des flexions du pluriel (pluriel masculin pour *abuso* en *–i* et pluriel féminin pour *conferenza* en *–e*) qui ne sont formellement pas proches des flexions du pluriel français (*-s* pour la majorité des cas). Ainsi, pour un même mot, il semblerait que la forme lemmatisée soit plus facilement reconnaissable de manière complète que la forme non-lemmatisée. Dans notre étude, étant donné que les mots donnés lemmatisés et non-lemmatisés n'étaient pas identiques, la différence de taux entre les deux types de reconnaissance n'est pas inversement proportionnelle.

Cette non-cognation des flexions se retrouve d'une manière encore plus flagrante pour la catégorie des verbes. En effet, les verbes lemmatisés sont reconnus complètement dans 65 % des réponses alors qu'ils ne le sont qu'à 19 % quand ils ne sont pas lemmatisés. Les flexions verbales sont, semble-t-il, encore moins transparentes que celles des autres mots. Par exemple, la

forme verbale *desiderino* (en français, *désirer*) est constituée d'un lemme *desid* et du suffixe de la troisième personne du pluriel du subjonctif présent *–erino* qui est peu formellement apparentée à son équivalent dans le système de conjugaison français (*-èrent*).

Cette comparaison entre lemme et non lemme a également été étendue aux adverbes où nous avons comparé les différents taux de reconnaissance entre adverbes de forme simple et adverbes dérivés d'un adjectif par un processus de dérivation. Ici aussi, nous avons remarqué une différence importante dans les taux de reconnaissance.

Ainsi, aucun adverbe, qu'il soit de forme simple ou dérivé d'un adjectif n'a été reconnu partiellement. Donc, les adverbes étaient, dans notre échantillon, soit reconnus totalement, soit pas reconnus du tout. En revanche, entre les deux taux de reconnaissance, on assiste à une différence très importante (89% de reconnaissance pour les adverbes dérivés et seulement 45 % pour les adverbes simples).

Pour expliquer ce phénomène, il nous a semblé bon d'étudier la cognation des suffixes de dérivation adverbiale. En effet, le suffixe italien de formation de l'adverbe sur un adjectif est formellement très proche de celui du français (*-mente* en italien et *–ment* en français). Partant du principe que les adjectifs constituent une catégorie très fréquemment reconnue (voir figure 3, où 68 % des adjectifs étaient reconnus, même partiellement), il n'est pas étonnant que les adverbes formés sur les adjectifs le soit également. Cependant, ce résultat dépend du taux de reconnaissance de l'adjectif sur lequel l'adverbe est formé. Dans notre échantillon, ces adjectifs étaient donc majoritairement reconnus, mais une étude plus poussée pourrait confirmer cette relation entre adverbe et adjectif.

Parmi les adverbes de formes simples, nous l'avons dit, seuls 45 % sont reconnus. Parmi ceux-ci, certains étaient formellement proches du français (par exemple *quando*) et d'autres l'étaient beaucoup moins (comme *sopra*, *sempre*).

6.3. Les différences entre les répondants (romanistes/non-romanistes) dans l'analyse

Le fait que les répondants connaissent ou non une autre langue latine semble influencer également le taux de reconnaissance. Comme nous l'avions remarqué dans la figure 9, le taux de reconnaissance totale est nettement supérieur chez les romanistes (78 %) que chez les non-romanistes (63 %). Cette tendance se retrouve pour toutes les catégories de mots (cf. figure 10)

et avec plus ou moins les mêmes différences entre ces catégories que dans les résultats des répondants en général.

6.4. La reconnaissance des cognats et semi-cognats (base ou suffixe) dans la génération

Dans cette partie de l'expérience, le premier point important à souligner est la forte 'génération' des suffixes semi-cognats par les répondants (73 %). En revanche, la généralisation des changements internes des mots et leur réutilisation est beaucoup plus faible. Bien que pour cette reconnaissance des bases semi-cognates les différences soient extrêmement fortes entre chaque changement pris individuellement (certains sont largement assimilés et d'autres pas du tout), ce taux de génération reste bien moindre que celui des suffixes semi-cognats.

7. Conclusion

Une telle étude devrait évidemment être menée à une plus large échelle pour avoir des résultats plus probants statistiquement parlant. Ainsi, le panel des personnes interrogées et les nombres de mots soumis devraient être plus importants et plus hétérogènes. Cependant, les nettes tendances qui se dégagent de ces quelques résultats nous permettent de tirer certaines conclusions.

La première conclusion, évidente, est qu'il existe un lien formel entre les deux langues et que les locuteurs d'une de ces deux langues semblent en être conscients. Le taux de reconnaissance est encourageant pour notre tentative de formaliser ce lien.

La deuxième conclusion porte sur la cognation des flexions. Alors que les formes lemmatisées sont plus facilement reconnues, les formes non-lemmatisées le sont beaucoup moins. Ceci est encore plus flagrant pour la catégorie des verbes dont les flexions semblent être moins transparentes que celles du pluriel ou du féminin. Tout ceci semble donc être un argument pour traiter les flexions individuellement et pour concentrer nos efforts de formalisation des liens sur les formes de base.

La troisième conclusion concerne la connaissance d'une autre langue romane. l'inférence fonctionne d'avantage chez les romanistes que chez les autres. Faut-il y voir un argument pour la

formalisation des liens morphologiques non pas entre deux langues, mais entre plusieurs langues latines ? La question mérite d'être posée.

Enfin, concernant la faculté de génération des formes correctes, les bons résultats que nous avons trouvés dans l'utilisation adéquate des suffixes semi-cognats sont également encourageants pour envisager une formalisation des dits-suffixes. L'inférence semble bel et bien exister sur ce point. En revanche, la généralisation des changements morphographémiques internes ne semble pas être uniformément comprise et semble dépendre du changement en question. Cependant, nous ne pouvons tirer une telle conclusion de manière aussi tranchée à partir d'une étude sur un si petit panel, de mot et de répondants.

Dans les chapitres suivants, nous présenterons tout d'abord un état de l'art des différentes expériences déjà réalisées pour l'exploitation des liens morphologiques multilingues (chapitre 4), puis nous décrirons l'ébauche du lexique que nous avons élaboré dans le cadre de ce travail (chapitre 5).

Chapitre 4 : Exploitation des régularités morphologiques multilingues : Etat de l'art

1. Introduction

Bien que l'exploitation des régularités morphologiques multilingues reste un domaine marginal, elle a conduit à plusieurs applications dans le domaine du TALN, ouvrant la voie à de nombreux axes d'investigation. Dans ce chapitre, nous présentons trois projets, qui exploitent de manière plus ou moins grande, les régularités morphologiques multilingues. Le premier, le projet Polylex (point 2), est un projet à large échelle qui vise à construire un lexique multilingue. Dans les deux autres, l'exploitation des régularités morphologiques multilingues est le point central, mais l'application est plus restreinte. Nous parlons tout d'abord le projet GédéFrIt (point 3), dont le but est la génération automatique de dérivés morphologiques en parallèle (italien et français). Nous mentionnons ensuite les recherches de Olivier Kreif (point 4) qui se penchent sur l'utilisation des mots cognats pour améliorer l'alignement de corpus bilingue.

2. Le projet Polylex

2.1. But du projet

Le projet Polylex[23] propose la création d'un module lexical capable de gérer plusieurs langues à la fois. Partant du constat que les langues germaniques possèdent des similitudes au niveau morphologique, morphophonologique et syntaxique, le projet tente de les capturer et de les exprimer au sein d'un même lexique. Ce projet constitue donc un large champs d'investigation pour individualiser les généralités morphologiques multilingues, et organiser ces généralités dans un outil rationnel et robuste.

Le formalisme choisi est celui de l'approche hiérarchique des lexiques d'héritage. Dans ces hiérarchies, les informations les plus générales apparaissent au sommet de la hiérarchie et plus une information est spécifique, plus elle se retrouvera à un niveau inférieur. Les classes inférieures héritent donc toujours des classes supérieures, mais peuvent être infirmées par une information plus spécifique à un niveau inférieur. De plus, le principe d'héritage multiple (ou

[23] Les exemples sont tous tirés de The Polylex Web pages, http://www.cogs.susx.ac.uk/lab/nlp/polylex/, sauf en cas d'indications contraires.

orthogonal multiple) permet à un nœud d'hériter de plusieurs hiérarchies en même temps (nous reparlerons plus en détail des lexiques d'héritages lorsque nous décrirons notre propre application au chapitre 5).

Le langage de programmation choisi est le langage DATR (R. Evans et G. Gazdar, 1996, pp. 167-216). Dans celui-ci, les informations sont organisées sous la forme de nœuds reliés entre eux. Un nœud peut correspondre à n'importe qu'elle représentation linguistique (un phonème, un morphème, un mot, une classe, etc.). Ainsi, dans l'exemple ci-dessous (Bouillon *et al.*, 1998, p. 223) :

Verb :
<syn cat> = = verb
< > = = Null

Intransitif :
<sujet catégorie> = = sn
<sujet cas> = = nominatif
<objet > = = aucun
< > = = Verb

partir
< > = = Intransitif

le nœud *Verb* contient l'information de catégorie de la classe exprimée sous forme de structure de traits (*<syn cat> = = verb*, où l'attribut *syn cat* (catégorie syntaxique) prend la valeur *verb*), le nœud *Intransitif* contient toutes les informations relatives à cette classe et le nœud lexical *partir* hérite de toutes les propriétés de nœuds supérieurs. La relation d'héritage est exprimée par la déclaration *< > = = Intransitif*. Quand celle-ci contient la valeur *Null* la classe est alors la plus haute dans la hiérarchie et n'hérite donc de rien. Après compilation du lexique, le verbe *partir* reçoit ainsi la structure de trait suivante :

partir
<syn cat> = = verb
<sujet catégorie> = = sn
<sujet cas> = = nominatif
<objet > = = aucun

2.2. Approche choisie

Pour représenter les liens morphologiques multilingues existants entre les trois langues germaniques étudiées dans ce projet (l'allemand, l'anglais et le néerlandais), les concepteurs de Polylex procèdent à une décomposition en morphèmes et en phonèmes, permettant une

généralisation des deux unités de décomposition (C. Tiberius et R. Evans, 2000). Ainsi, chaque mot est décomposé comme suit :

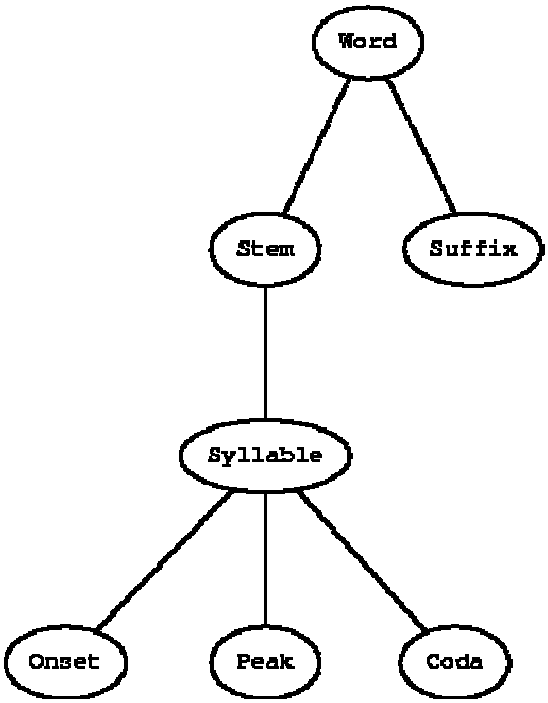


Figure 1

La figure ci-dessus montre que le mot (*Word)* est tout d'abord décomposé en morphèmes : la racine (*Stem*) suivie du suffixe (*Suffix*) puis, la racine est à nouveau décomposée en syllabes, qui seront elles-mêmes décomposées en trois phonèmes, l'onset (en français, l'attaque), le peak (le noyau) et la coda.

Par exemple, les mots *Fish* (En), *Fisch* (De) et *Vis* (Nl) (*poisson*), n'ont pas de suffixe et leur racine ne contient qu'une syllabe. Cette dernière est alors décomposée comme dans l'exemple ci-dessus et ces trois mots sont représentés de la manière suivante[24] :

Pour l'anglais fish : f * I S
Pour l'allemand Fisch : f * I S
Pour le néerlandais Vis : v * I s

Dans l'exemple ci-dessus, les trois phonèmes correspondent respectivement à l'onset, au peak et à la coda. L'astérisque indique l'accent tonique du mot.

[24] La structure phonologique de ces mots est exprimée avec l'alphabet Sampa, (http://www.phon.ucl.ac.uk/toc/set_eresource.htm)

2.3. Application de cette décomposition dans un lexique d'héritage

Dans l'exemple ci-dessous (Tiberius, ms), nous montrons un exemple de description des trois mots précédemment cités dans un lexique d'héritage de type DATR.

ML_Noun :
< > = = Syllable
<phn noun> = = « <phn root> »
<syn cat> = = noun

Syllable :
<phn root> = = « <phn syll> »
<phn syll> = = « <phn onset> » « <phn rhyme> »
<phn ryhme> = = « <phn peak> » « <phn coda> »
< > = = Null

Les deux premières classes (*ML_Noun* et *Syllable*) sont les deux classes supérieures du lexique. La classe *Syllable* décrit la concaténation entre les différents phonèmes composant le mot. La racine (*root*) correspond à la syllabe (*syll*) qui est elle-même formée d'une attaque (*onset*) et d'un *rhyme* (comme le déclare l'équation *<phn syll> = = « <phn onset> » « <phn rhyme> »*). Celle-ci n'hérite d'aucune classe (< > = = Null), elle est donc la classe la plus haute dans cette hiérarchie. La classe *ML_Noun* quant à elle, permet de déclarer les informations catégorielles propres au nom (*<syn cat> = = noun*).

A la suite de ces deux classes supérieures, on trouve les classes permettant de gérer l'entrée lexicale concernant les trois mots qui nous intéressent. La première classe est la classe multilingue contenant les informations communes aux trois langues.

ML_Fish :
< > = = ML_Nouns
<phn onset> = = f
<phn peak> = = 'I'
<phn coda> = = 'S'

Elle contient les informations qui la relient à la classe supérieure (dans ce cas *ML_Nouns*) et les informations multilingues de l'onset (<phn onset> = = f) , du peak (<phn peak> = = 'I') et du coda (<phn coda> = = 'S'). Ces informations sont considérées comme 'par défaut', et elles pourront être infirmées par des informations inférieures dans la hiérarchie. Elles donnent en quelque sorte le cadre phonologique du mot.

Les trois autres nœuds (*Dutch_Vis*, *German_Fisch*, *English_Fish*) sont des nœuds monolingues qui contiennent uniquement les informations qui diffèrent de celles contenues dans

le nœud multilingue. Ainsi, le nœud *Dutch_Vis* ci-dessous décrit le mot néerlandais *Vis* en déclarant uniquement les variations par rapport au nœud multilingue :

Dutch_Vis :
< > = = ML_Fish
<phn onset> = = v
<phn coda> = = s

Le nœud *German_Fisch* pour le mot allemand Fisch ne contient aucunes informations différentes de la classe multilingue. Il ne contient donc que l'information qui lui permette d'hériter des propriétés de celle-ci.

German_Fisch :
< > = = ML_Fish

Enfin, le nœud English_Fish contient quant à lui une seule information différente concernant la valeur de son *coda*.

English_Fish :
< > = = ML_Fish
<phn coda> = = s

Ainsi, après compilation du lexique, chaque entrée recevra une structure de traits contenant toutes les informations résultant de l'unification des différentes classes. Par exemple, le mot anglais *fish* aura la structure de traits suivante :

Fish :
<syn cat> = = noun
<phn onset> = = f
<phn peak> = = 'I'
<phn coda> = = s

Le fait d'utiliser les lexiques d'héritages permet de formaliser différentes informations linguistiques dans le même lexique. Cette approche présente l'avantage de ne déclarer qu'une seule fois chaque information pour plusieurs mots de plusieurs langues. Le lexique est alors plus cohérent et son stockage davantage réduit. De plus, un mot peut hériter de la hiérarchie traitant des généralisations morpho-syntaxiques, de celle traitant de la sémantique, etc.

2.4. Les différentes architectures

Dans l'exemple ci-dessus, nous avons décrit le cadre général de l'exploitation des liens morphologiques multilingues. Mais dans la pratique, ce projet a également donné lieu à de nombreuses discussions sur la manière la plus appropriée de généraliser les informations multilingues.

L'idée de base était de considérer un nombre *n* de hiérarchies monolingues pour un nombre n de langues distinctes, mais liées, et de transformer le tout en un lexique multilingue qui

comprendrait un nombre de *n+1* hiérarchies où : (a) la hiérarchie supplémentaire contiendrait toutes les informations que les langues *n* ont en commun et (b) un nombre *n* de hiérarchies monolingues qui n'exprimeraient que les paramètres particuliers de chaque langue.

La question est donc de savoir quelle méthodologie choisir pour mettre en commun les informations multilingues. Carole Tiberius (2002, p. 702) décrit différentes architectures possibles. Elle opère une première distinction entre les modèles non-paramétrables et les modèles paramétrables.

2.4.1. Les modèles non-paramétrables

Les modèles non-paramétrables consistent en une gamme de hiérarchies monolingues reliées entre elle par une hiérarchie multilingue supérieure qui regroupe les informations communes aux hiérarchies monolingues.

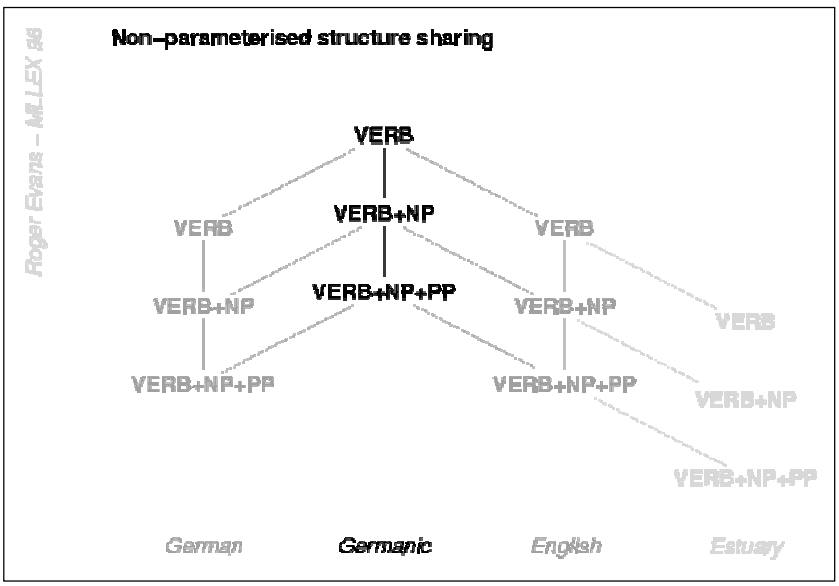


Figure 2

Dans la figure 1 ci-dessus, tirée de l'exposé de Roger Evans[25], la structure en noir est la structure multilingue et les structures grisées représentent les structures monolingues. Ces structures sont appelées non-paramétrables car la langue n'est pas explicitement utilisée comme un paramètre. Chaque hiérarchie concerne en effet une seule langue ou les informations communes à toutes les langues, mais l'information de langue n'est pas nécessairement déclarée (Tiberius, *ibid.*).

[25] Roger Evans, Exploiting inheritance in multilingual lexicon

2.4.2. Les modèles paramétrables

La deuxième architecture possible, les modèles paramétrables, intègre toutes les langues dans la même hiérarchie. Dans ces modèles, les paramètres de langue permettent de déclarer pour chaque partie du lexique quelles informations sont valables pour quelles langues. Ainsi, schématiquement, l'architecture paramétrable peut se représenter comme dans la figure 2 ci-dessous :

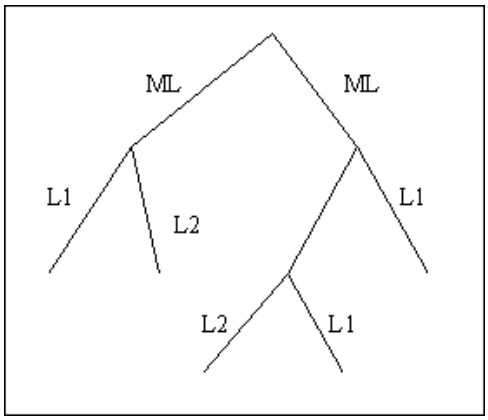


Figure 3

Dans la figure 2, les branches du lexique propres à une langue sont étiquetées L1 ou L2, alors que les branches générales et communes aux langues sont étiquetées ML. On remarque que, suivant la philosophie des lexiques d'héritage, plus une information est spécifique, plus elle est située bas dans la hiérarchie. Ainsi, les informations propres à une langue sont plus spécifiques que les informations multilingues, et sont donc situées aux bouts des branches.

Le modèle paramétrable a donné lieu à de nombreuses recherches pour tenter d'affiner cette représentation. Mais les différentes options qui en découlent sont étroitement liées aux formalismes utilisés[26]. Nous y reviendrons dans le chapitre 5 où nous décrivons notre expérience de construction d'un tel lexique, et où nous évoquerons les différentes options choisies dans ce cadre.

2.5. L'intérêt du projet

Polylex constitue pour le présent travail une source de référence inestimable. C'est en effet, à notre connaissance, le projet le plus important en terme d'activité de recherches effectuées dans ce cadre. Nous avons pu profiter des nombreuses réflexions effectuées autour de ce projet pour pouvoir définir de manière adéquate le présent travail. Comme nous le verrons

[26] Pour plus d'information sur les options de ce modèle dépendant du langage DATR, voir Tiberius Carole, 2001, et l'exposé de Roger Evans, (*op. cit.*)

dans le chapitre 5, notre partie pratique (la construction d'un lexique multilingue) s'inspire largement de ce projet, notamment pour ce qui est du formalisme choisi. Nous n'avons cependant pas décomposé nos mots cognats en phonèmes, nous limitant aux morphèmes pour des questions de temps et de connaissances.

3. GédéFrIT

3.1. Le but de GédéFrIt

Le projet GédéFrIt (Namer 2001) est une extension du projet Gédérif (Namer *et al.* 2000), qui consiste en la génération de néologismes morphologiquement construits sur la forme X-iser et X-able. Le but de ce premier projet était de générer, sur la base de 2570 lemmes présents dans les dictionnaires, le maximum de néologismes absents des dictionnaires mais attestables par la suite dans des documents. La langue produit en effet constamment des néologismes morphologiquement et sémantiquement construits, dont le sens peut être prédictible à partir de leur structure, mais qui sont par définition absents des dictionnaires (Namer, 2001, p. 287).

Le but du projet GédéFrIt était d'appliquer les mêmes méthodes de génération et de vérification que pour le projet Gédérif mais à deux langues morphologiquement proches, et le tout en parallèle à partir de paire de mots italien-français considérés comme la traduction l'un de l'autre. Ainsi, si l'on prend deux mots qui sont équivalent de traduction et qu'on leur applique les mêmes processus constructionnels avec des suffixes jugés équivalents, on obtiendra deux dérivés qui seront également la traduction l'un de l'autre.

Par exemple, à partie du verbe *tartiner* les concepteurs du projet ont pu dériver l'adjectif *tartinable* et le nom *tartinabilité*, deux néologismes absents des dictionnaires mais attestés par la suite en corpus. De même, pour le verbe italien *spalmare*,(en français *tartiner*) et en appliquant les mêmes règles de dérivation avec des suffixes propres à l'italien, (*-bile* et *-ità*) , ont été générés l'adjectif *spalmabile* et le nom *spalmabilità* , tout deux également néologismes absents des dictionnaires mais attestés en corpus.

3.2. Motivations

Les motivations de ce projet sont donc au nombre de trois. La première est de mesurer le pouvoir générateur de néologismes des suffixes employés, l'étape de validation permettant de quantifier le nombre de néologisme générés correctement. La deuxième motivation concerne la vérification de l'existence de la même créativité lexicale entre les deux langues. Enfin, la

troisième motivation concerne le parallélisme supposé qui existerait entre deux langues dans l'utilisation de certains suffixes formellement apparentés. Ainsi, même si formellement on ressent une très forte similitude entre certains suffixes comme *-ité* et *-ità*, il était intéressant de générer en parallèle des mots formés grâce à ceux-ci pour quantifier les similitudes dans leur pouvoir générateur.

3.3. Résultats

Une fois que la traduction des couples primitifs est garantie (comme *tartiner* et *spalmare* dans l'exemple cité plus haut), les auteurs du projet constatent une forte ressemblance entre les opérateurs *–able*, *-ité*, et *-iser*, et ceux qu'ils avaient considérés comme équivalents en italien. Ainsi, parmi tous les mots générés automatiquement en parallèle sur la base de paires de mots équivalents de traduction, 36,5 % pour le français et 45,7 % pour l'italien ont été trouvés dans les documents[27]. Cette étude a également montré le haut degré de similitude de leur pouvoir générateur. En effet, de nombreux mots de base engendrent le même nombre de formes en français et en italien. Ainsi, si le mot de base italien (B_{it}) est considéré comme la traduction d'un mot de base français (B_{fr}), alors les mots générés sur ces deux bases par les suffixes jugés équivalents seront également la traduction l'un de l'autre, et leur analyses constructionnelles seront isomorphiques (cf. l'exemple de *tartiner* et *spalmare* et leur dérivés). Les auteurs en tirent la conclusion que l'italien et le français sont donc deux langues morphologiquement très proches, ce qui justifie la génération et l'analyse en parallèle à partir d'un système d'analyse linguistiquement contraint pour les deux langues.

3.4. L'intérêt du projet

Ce projet a l'avantage pour nous de confirmer la cognation des suffixes ainsi que leur mode opératoire. Cependant, le projet GédéFrIt ne prend pas en compte les similitudes au niveau des bases, mais uniquement au niveau des suffixes.

4. L'exploitation des cognats pour l'alignement de texte

De nombreuses autres études ont été menées en TALN[28] pour exploiter les similitudes multilingues notamment dans le cadre de l'alignement de corpus bilingue en utilisant les cognats.

[27] Pour des résultats plus détailler concernant ce projet, se référer à Namer (2001)
[28] citons notamment l'article de Simar et al, 1992

Olivier Kreif envisage notamment d'utiliser comme point d'encrage pour l'alignement de corpus bilingues des mots formellement apparentés. Dans l'expérience menée, une phase de préparation a donné lieu à une détermination manuelle des cognats. Dans ce travail, l'auteur a tenté de définir plus précisément la méthode habituelle qui consiste à considérer comme apparentés deux mots ayant au minimum 4 lettres communes (calcul de n-gramme).

Dans cette optique, il explore différentes voies d'appariement de mots, dont nous résumons les grandes lignes. Premièrement, il s'est basé sur l'existence de liens étymologiques entre les mots. A ce sujet, il évoque le fait que ce ne sont pas seulement les radicaux qui peuvent avoir un lien étymologique, mais également les suffixes ou les préfixes. Pour cette application de marquage de point d'encrage, le fait que des mots ne soient apparentés que par leurs affixes n'enlève rien à leur utilité. Ainsi, pour reprendre l'exemple de Namer cité plus haut, la paire *spalmabilità/tartinabilité* n'a pas de lien étymologique au niveau du radical (*spalma* et *tartin*), mais selon Kreif, la mise en évidence de la cognation entre les suffixes *-abilità* et *-abilité* serait également un critère pour l'appariement de ces deux mots comme des points d'encrage pour l'alignement. De plus, il propose d'utiliser des « systèmes de conversion graphémique ad hoc permettant d'obtenir des équivalences telles que *administration→ amministrazione*, avec un minimum de règles » (Kreif 2001, p. 842).

L'auteur de l'étude précise également que ce genre de recherche devrait être effectué pour plusieurs paires de langues même non apparentées, pour des textes de spécialité. Comme le suggère Kreif, « le vocabulaire spécialisé est souvent […] forcé par les standards internationaux à utiliser [des mots formellement proches], basés essentiellement sur le fond gréco-latin. »

Cette étude montre une autre application cognats qui semble prometteuse. Elle a l'avantage pour nous de mettre en évidence la cognation des préfixes et des suffixes permettant, dans le cadre de l'alignement de bi-textes, d'apparier deux mots en fonction de leurs affixes, même si leur base n'est pas cognate. Elle ouvre ainsi d'autres perspectives pour l'utilisation concrète des liens morphologiques multilingues.

5. Conclusion

Dans ce chapitre, nous avons tenté de donner un aperçu de différents projets, en cours ou achevés, qui utilisaient de manière plus ou moins grande les liens morphologiques multilingues. Nous l'avons dit, le projet Polylex reste notre référence pour le présent travail, mais les deux autres projets nous ont permis de valider certaines suppositions, notamment quant à la proximité formelle des deux langues de notre travail en général, et des suffixes en particulier.

Dans le chapitre suivant, nous décrivons l'expérience pratique que nous avons menée pour mettre en œuvre dans un lexique les généralités multilingues précédemment identifiées.

Chapitre 5: Exploitation des liens morphologiques multilingues : utilisation concrète

1. Introduction

Nous l'avons vu dans le chapitre 2, il existe un lien formel entre le français et l'italien. Ce lien se retrouve à de nombreux niveaux linguistiques, et particulièrement dans la morphologie de ces deux langues. L'expérience décrite dans le chapitre 4 a également prouvé que ces liens existaient et que l'inférence entre ces deux langues était bel et bien réelle.

Dans le chapitre 1, nous citions Sproat (1992, pp. 1 et ss.) qui décrivait les nombreuses applications informatiques de la morphologie. Après avoir décrit, dans les chapitres précédents, les différents liens morphologiques existants, nous présentons maintenant une application de ces liens. En nous inspirant du projet Polylex, décrit au chapitre 4, nous avons construit une ébauche de lexique informatisé qui suit une approche polymorphique. Ce type de lexique pourrait s'appliquer à de nombreuses applications en langue naturelle, et contient toutes les informations nécessaires à l'analyse et à la génération. Cependant, notre lexique ne se veut pas une fin en soi. En effet, la conception d'un lexique à large échelle reste une entreprise de longue haleine nécessitant de nombreuses ressources. Dans le cadre de ce travail, notre lexique représente une tentative de formalisation concrète, pouvant être utilisé comme point de départ pour réfléchir plus avant aux différentes possibilités d'exploitation des liens morphologiques multilingues.

Dans ce chapitre, nous commencerons par présenter le formalisme choisi (les lexiques d'héritage), en décrivant certain de ces aspects les plus saillants (héritage multiple, héritage par défaut), puis nous montrerons les différentes particularités de l'outil, principalement celles utilisées pour la construction de notre lexique multilingue. Ensuite, nous détaillerons notre implémentation d'ELU pour la morphologie multilingue.

2. Formalisme utilisé

2.1. Pourquoi les lexiques d'héritage ?

Dans la première partie de ce mémoire, nous avions évoqué Sproat (1992) qui citait les différentes considérations à prendre en compte lors de la construction d'un lexique. Loin de nous

l'idée de construire ici un programme informatique précis pour traiter un lexique multilingue. Dans le cadre restreint de ce travail, nous tenterons plus modestement d'utiliser un outil déjà existant qui permet l'implémentation de lexiques multilingues.

La première question que nous nous sommes donc posée était de savoir quel formalisme et quel outil utiliser. Cette question était évidemment conditionnée par les outils à disposition. Nous avions en effet le choix entre un automate à deux niveau (Mmorph[29]) et un lexique d'héritage (ELU[30], Environnement linguistique de grammaire d'Unification). Alors que Mmorph est un automate à deux niveaux particulièrement conçu pour le traitement de la morphologie, ELU est en fait un ensemble d'outils particulièrement destinés à la recherche en linguistique informatique. Il offre notamment la possibilité d'écrire des grammaires d'unification pour l'analyse et la génération, ainsi que des règles de transfert pour la traduction automatique. Sa composante lexicale se fonde notamment sur la théorie des lexiques d'héritage. Ces deux formalismes sont donc bien différents, et notre choix s'est porté sur celui des lexiques d'héritage contenus dans ELU, et cela pour plusieurs raisons. La première d'entre elles était le fait que le projet dont nous nous inspirions depuis le début (le projet Polylex, décrit au chapitre 4) avait également utilisé ce formalisme. Prendre une même approche nous permettait de nous inspirer des différents choix opérés dans ce projet. De plus, les nombreux problèmes d'architecture avaient déjà été étudiés dans le cadre de ce projet et la littérature constituait pour nous une source non négligeable de références. Enfin, la méthode des lexiques d'héritage permet de gérer non seulement la morphologie, mais également la syntaxe et la sémantique, ce qui laisse la porte ouverte à des généralisations multilingues à tous les niveaux linguistiques. Cependant, le projet Polylex utilise le formalisme DATR, qui diffère sous certain aspect de l'outil ELU. Nous avons donc dû adapter certains éléments propres à cet outil.

2.2. Les lexiques d'héritage

Parmi les différents types de techniques qui permettent une approche polymorphique, on trouve les lexiques d'héritage. Le lexique d'héritage consiste en une hiérarchie, c'est-à-dire une ensemble de nœuds reliés par des arcs (Bouillon et al, 1998). Les nœuds sont associés à des objets ou à des ensembles d'objets et les arcs définissent les relations qui les lient (Bouillon, *ibid.*). Les liens correspondent à la relation *est un* (is a). A chaque nœud peuvent être associées

[29] Mmorph, Petitpierre et Russel, 1995. Outil développé à l'ISSCO et disponible à l'adresse : http://www.issco.unige.ch

des informations qui sont propagées aux nœuds inférieurs par la technique d'héritage. L'exemple ci-dessous représente un héritage simple:

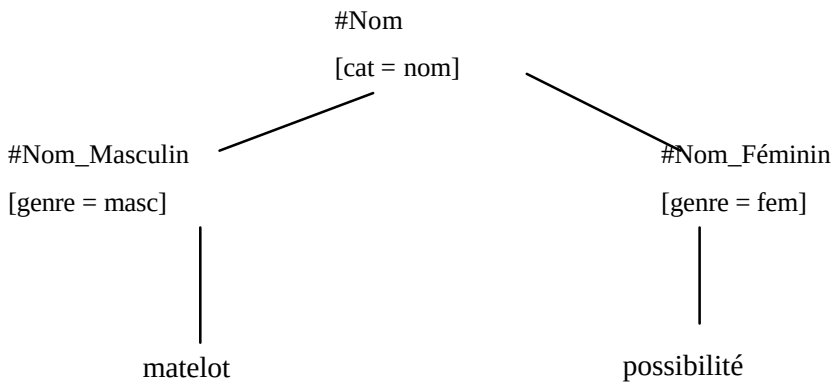


Figure 1

Dans la figure 1, le nœud lexical *matelot* hérite des propriétés contenues dans la classe #Nom_masculin (informations de genre), ainsi que des propriétés contenues dans la classe #Nom (information de catégorie). Sa structure de traits est alors composée de toutes les propriétés de tous les nœuds transmises par héritage (*matelot* est donc un *nom* de genre *masculin*). Le nœud lexical *possibilité* hérite quant à lui des propriétés contenues dans #Nom_Féminin et de celles contenues dans #Nom (*possibilité* est donc un *nom* de genre *féminin*).

Rappelons que l'approche polymorphique a pour but d'organiser le lexique de manière économique en capturant toutes les généralisations entre les mots. Sur le plan informatique, un module lexical doit également être capable de gérer les exceptions, très fréquentes dans le lexique. Ainsi, les nœuds d'un lexique d'héritage sont des classes représentant des informations plus ou moins générales où les nœuds inférieurs héritent des informations contenues dans les nœuds supérieurs.

On distingue trois types d'héritages qui permettent de gérer différents cas de généralité et d'exception au sein du lexique :

→ l'héritage simple, où un nœud fille hérite des propriétés d'un seul nœud mère à la fois (comme dans l'exemple ci-dessus) ;

---

[30] ELU, Estival D, 1990.

→ l'héritage multiple où un nœud fille peut hériter de plusieurs nœuds mères à la fois ;

→ l'héritage par défaut où un nœud hérite des propriétés du nœud supérieur sauf s'il contient une information contradictoire.

Si l'héritage simple est facile à mettre en œuvre, les autres types d'héritage posent toujours un certain nombre de problème « architecturaux ». Nous décrivons plus en détail ci-dessous les techniques d'héritage multiple et d'héritage par défaut.

2.2.1. Héritage multiple

La technique de l'héritage multiple permet de faire hériter un nœud de plusieurs classes supérieures à la fois. Cette structure convient particulièrement bien pour traiter en même temps des informations syntaxiques et morphologiques. Ces deux types d'informations sont en effet indépendants les uns des autres, et ne peuvent donc pas être traités dans la même structure. Dans la figure 2 ci-dessous, tiré de Bouillon (1998), nous présentons une partie d'un lexique d'héritage qui gère les propriétés syntaxiques et morphologiques des verbes français :

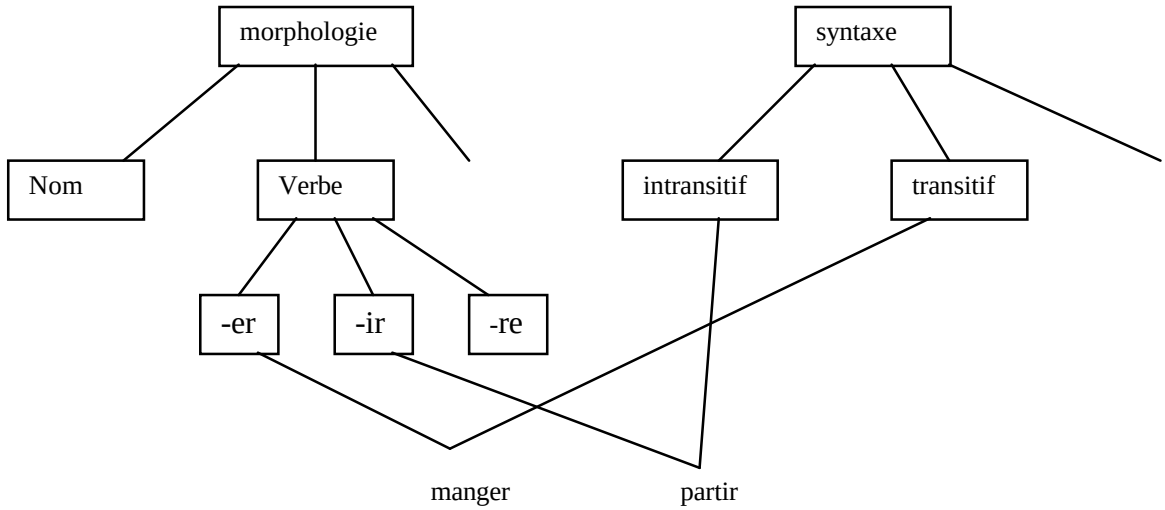


Figure 2

Dans la figure 2, les deux verbes *manger* et *partir* héritent des propriétés contenues dans les deux hiérarchies *syntaxe* et *morphologie*. Si les informations provenant de deux classes sont toujours disjointes (ou non-contradictoires), le système est alors 'orthogonal'. Dans ce cas, aucun conflit ne survient, et le résultat des informations héritées par défaut des deux classes supérieures ne variera pas en fonction de l'ordre dans lequel ces défauts s'appliquent.

Dans la construction d'une telle architecture, il faut donc faire attention à ne pas avoir d'ambiguïté entre les deux hiérarchies, où deux informations contradictoires pourraient se trouver, et générer ainsi une contradiction d'information. Pour empêcher de telles contradictions,

chaque système utilise une méthode différente ; nous verrons plus loin que ELU a opté pour une déclaration d'un ordre de priorité dans l'héritage de plusieurs classes.

2.2.2. Héritage par défaut

Nous l'avons dit, les exceptions sont nombreuses dans les langues naturelles. Par exemple, il est souvent souhaitable de déclarer une information propre à une classe et de l'infirmer dans de rares cas où elle est contredite. Ainsi, il est possible de placer dans la classe exprimant les propriétés générales des verbes une équation du genre 'aux=no' (c.-à-d. les verbes ne sont pas des auxiliaires) et de placer une information contradictoire 'aux=yes' pour les quelques cas où cette information est valable.

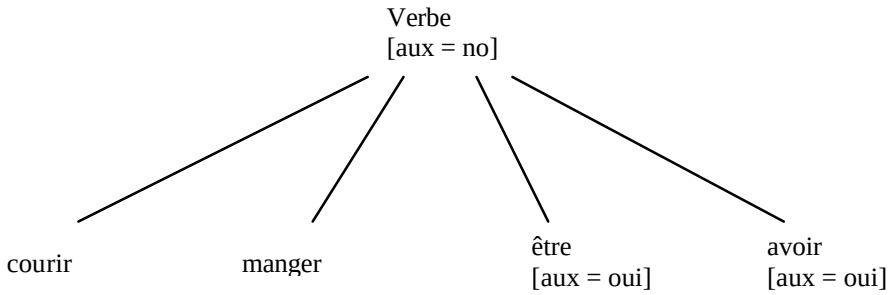


Figure 3

Cette capacité « d'encoder des propriétés rares des éléments du lexique est extrêmement attractive d'un point de vue linguistique » (Russell 1992). Ainsi, sur le plan architectural, il en découle que plus une propriété apparaît bas dans la hiérarchie, plus cette propriété est exceptionnelle.

Nous présentons maintenant le lexique d'héritage ELU que nous avons utilisé dans ce travail, et nous abordons les différentes techniques proposées par cet outil pour gérer les différents types d'héritage évoqués ci-dessus.

3. L'outil ELU

Comme DATR, le formalisme ELU[31] est un lexique d'héritage basé sur l'unification de structure de traits. L'unification est un mécanisme au centre de nombreux formalismes de traitement des langues (Bouillon et al, 1998, p. 89). Elle est similaire à l'union dans la théorie des ensembles, à la différence qu'elle échoue en cas de valeurs contradictoires pour un même attribut. Cette unification de structures permet d'obtenir une représentation globale et cohérente du résultat d'une construction (qu'elle soit lexicale ou syntaxique). Une structure de traits est une description d'un objet effectuée par une énumération de ces caractéristiques significatives et se présente comme un ensemble de paires attributs - valeurs. Dans les exemples cités plus haut, le mot *matelot* prend par exemple la structure de trait suivante :

matelot $\begin{pmatrix} \text{cat} = \text{nom} \\ \text{genre} = \text{masc} \end{pmatrix}$

Cette structure de traits contient la paire 'cat=nom' où l'attribut 'cat' reçoit la valeur 'nom' et la paire 'genre=masc' où l'attribut 'genre' reçoit la valeur 'masc'. Elle est le résultat de l'unification des deux structures de traits, contenues dans les deux nœuds supérieurs de la hiérarchie et héritées par le nœud *matelot*.

Dans le cadre de l'outil ELU, l'unification est décrite comme « un choix attractif comme mécanisme informatique de base d'un tel système, pour de nombreuses raisons qui sont désormais bien connues. Ses propriétés algébriques permettent une sémantique déclarative et « monotone » pour le langage qui est indépendante du programme qui l'interprète » (Russell 1992, p. 311).

3.1. Organisation générale et mode de déclaration[32]

Tout comme les autres outils se basant sur la technique d'héritage (cf. DATR), ELU est un ensemble de 'classes' (ou nœuds) reliées entre elles par des liens d'héritages. Chacune d'elles est

[31] ELU, Environnement linguistique d'unification développé à l'ISSCO (cf. Johnson and Rosner (1989), Russell *et al.* (1990), Estival *et al.* (1990))
[32] Tous les exemples et la description de l'outil sont tirés de l'article « a practical approach to Multiple Default Inheritance for unification-based Lexicon » (Russel, 1992) auquel nous vous renvoyons pour de plus amples informations.

une collection de structures d'équations qui encode une information commune à un ensemble de mots. Chaque classe peut être liée à une ou plusieurs classes supérieures.

Les entrées lexicales sont elles-mêmes des classes et chaque information qu'elles contiennent est spécifique à un seul mot.

3.2. Un exemple

L'exemple suivant illustre l'entrée lexicale du verbe anglais *walk* et ses deux classes supérieures, *# Class Intransitive* et *#Class*.

Exemple (a) :

```
# Class Intransitive ( )
     <subcat> = [Subj]              <Subj cat> = np

#Class Verb ( )
     <aux> = no                     <cat> = v
     |
     <form> = <stem> && ed                <tense> = past
     |
     <agr> = sg3                    <tense> = present
     <form> = <stem> && s
     |
     <agr> = ~sg3            <tense>= present
     <form> = <stem>

# Word walk (Intransitive Verb)
     <stem> = walk
```

3.3. Déclaration de la classe

Une définition de classe est en fait une directive pour le compilateur ('#Class' pour les classes non-lexicales et '#Word' pour les classes lexicales). Cette directive est suivie du nom de la classe, d'une liste des classes supérieures directes (qui peut être vide), d'une gamme d'équations principales (par défaut), qui peut être vide et d'une ou plusieurs (ou aucune) gammes d'équation de variantes séparées par une barre verticale (|). Dans l'exemple (a), la première classe s'appelle *Intransitive* et n'hérite d'aucune classe ('( )') . Dans la suite, nous étudierons les composantes de ces classes les unes après les autres, en nous referant toujours à l'exemple (a).

3.4. Déclaration de (ou des) classe(s) supérieure(s)

La déclaration de classes supérieures consiste en une liste de noms des classes directement supérieures à la classe en question. Si une classe hérite de plusieurs classes, leurs noms doivent

être déclarés de manière ordonnée, ce qui permet d'introduire une précédence dans l'héritage, résolvant ainsi les problèmes liés à l'héritage multiple d'informations contradictoires. Ainsi, les classes déclarées le plus à gauche sont dites 'plus spécifiques' et leurs informations héritées auront la précédence sur les autres informations des autres classes. Dans l'exemple (a), la classe lexicale *#Word walk* hérite des deux classes supérieures déclarées entre parenthèses *#Class Intransitive* et *#Class Verb.*

La déclaration des classes supérieures peut également être vide si la classe en question n'a pas de classe supérieure. Elle est alors la classe la plus générale du lexique, et elle n'hérite d'aucune information. C'est notamment le cas des classes *#Class Intransitive* et *#Class Verb* dans l'exemple (a).

3.5. Les équations principales

La première déclaration consiste en une ou plusieurs équations, appelée également équation principale ou équation par défaut. Cette équation peut également être vide. Elle contient des informations par défaut qui s'appliqueront à toutes les classes héritant de cette classe. Ces équations sont des variables indépendantes, dans le sens où elles peuvent être contredites par une information d'une classe inférieure. Dans ce cas, c'est l'information spécifique de la classe inférieure qui se substituerait à l'information par défaut. Par exemple, dans (a), la classe *#Class Verb* a pour équation principale <*aux = no* > et <*cat = v*> (pour déclarer que les mots héritant de cette classe sont des verbes (*v*) non auxiliaires (*aux=no*)). Si dans cet exemple nous voulions ajouter le verbe *to be* (auxiliaire), nous pourrions, dans sa classe lexicale, déclarer l'équation principale <*aux=yes*>. Cette information plus spécifique serait alors prise par défaut et se substituerait à l'équation principale de la classe supérieure.

3.6. L'équation des variantes

A la suite des déclarations des informations par défaut (qui peuvent être vides), il est possible de déclarer une ou plusieurs équations représentant des variantes 'intra-classe', qui correspondent à des alternatives du même niveau conceptuel. Ces équations dans les variantes sont des contraintes absolues, par rapport à celles contenues dans l'équation principale, ce qui signifie qu'elles ne peuvent pas être contredites par des informations plus spécifiques. Si ces variantes provoquent un conflit avec des informations d'une classe plus spécifique, l'unification échoue. Ainsi, dans la classe *#Class Verb* de l'exemple (a), trois variantes ont été déclarées pour

la formation du passé (<form> = <stem> && ed, <tense> = past) et pour celle des formes du présent, à la troisième personne (<agr> = sg3, <form> = <stem> && ed) et aux autres personnes (<agr> = non_sg3 <form> = <stem>). Ces trois variantes sont séparées par une barre verticale (|).

3.7. La concaténation des chaînes

La construction et la décomposition de mots complexes sont effectuées par l'opérateur de concaténation de chaîne '&&' dans une équation du genre :

Chaîne = base && Suffixe

Ce genre d'équation unifie la chaîne avec le résultat de la concaténation de la base et du suffixe. Ainsi, pour la formation de la forme du passé dans l'exemple (a), la déclaration *<form> = <stem> && ed* permet la concaténation de la forme de base (*<stem>*) avec le suffixe *ed*. La même méthode est utilisé pour la concaténation du *s* de la troisième personne du singulier au présent *(<form> = <stem> && s*).

3.8. La disjonction

Dans les structures de traits atomiques, il est possible d'opérer une disjonction. Elle est particulièrement utile pour des attributs ayant deux possibilités de valeurs. Par exemple, le mot français *nez* peut être à la fois singulier ou pluriel, ce qui peut se représenter de la manière suivante (Bouillon et al, 1998) :

[nombre = singulier/pluriel]

L'unification aura alors lieu avec d'autres structures de trait contenant soit la valeur 'singulier' soit la valeur 'pluriel', soit les deux valeurs.

3.9. La négation

Tout comme la disjonction, la négation est définie sur une structure de traite atomique. Elle est déclarée par le signe '~'. Elle permet de déclarer une valeur par son contraire, évitant ainsi de fastidieuses énumérations. Ainsi, pour la formation de toutes les formes du présent autres que la troisième personne, la variante correspondante dans l'exemple (a) déclare, pour l'attribut de personne, *<agr> = ~sg3*, ce qui signifie toutes les personnes, sauf la troisième personne du singulier.

3.10. Les macros

ELU dispose également d'un système de macro très puissant permettant d'éviter la redondance de certaines déclarations et d'effectuer des changements morphographémiques. Une macro (aussi appelée 'abstraction relationnelle') est composée d'un nom et d'une équation (ou une structure de traits), le tout étant défini au début du lexique dans la section 'Define'. La macro peut alors être appelée dans le lexique en déclarant son nom précédé d'un '!'.

Par exemple, la macro

accord_Nom(<* syn accord genre> <* syn accord nombre><* syn accord pers>)

comporte un nom 'accord_Nom' et un certain nombre d'arguments qui correspondent aux différents attributs. Chaque fois que l'on a besoin de déclarer les informations de genre, de nombre et de personne, au lieu d'écrire les trois paires attributs-valeurs, il suffira d'appeler la macro et de lui donner les trois valeurs, comme dans l'exemple ci-dessous :

```
#Word matelot ( )
!accord_Nom (masc, sing, 3)
```

Une macro peut également contenir une équation, et des variables, qui permettent d'effectuer des changements morphographémiques. Dans ce cas, la macro définie dans la section 'Define' contient la forme du mot telle qu'elle est avant la transformation, comme dans l'exemple ci-dessous pour la transformation de 'au' en 'll' dans la formation de *chamelle* à partir de *chameau*.

Ainsi, la macro ci-dessous

```
Ajout_Au(Base, S1)
    Base = S1 && au
    S1 = _
```

représente la forme avant transformation, où le mot (Base) doit être formé de n'importe quelle chaîne de caractère (*S1*) suivie de la chaîne *au* .

Par la suite, dans le lexique, pour effectuer cette transformation, on appellera la macro en définissant les résultats de la transformation.

```
#Class Fem_ell(Nom_Adj_Fem)
    !accord_Ngenre (f)
    !Ajout_Au (<base>, S1)
    Femform = S1 && ll
```

Ainsi, dans cette classe, on appelle la macro (*!Ajout_Au (<base>, S1)*) et l'on définit le résultat de la transformation (*Femform = S1 && ll*) en créant une nouvelle forme (la forme du féminin *Femform*) et en conservant la chaîne de caractère (*S1*), mais en y ajoutant un chaîne *ll* à la place de la chaîne *au*.

Nous espérons, dans cette section, avoir donné un aperçu suffisamment complet de l'outil ELU et du traitements de lexique d'héritage. Nous décrivons dans la suite notre implémentation de cet outil pour le traitement de la morphologique multilingue. Nous verrons ainsi dans la pratique l'application des différentes fonctionnalités décrites ci-dessus.

4. Un lexique d'héritage multilingue, application

Dans le chapitre 2, nous avions mis en évidence les différents phénomènes multilingues que nous voulions tenter de traiter à savoir les suffixes semi-cognats, les formes de base cognates et semi-cognates, et les préfixes semi-cognats ou parfaitement cognats. Puis, dans le chapitre 3, nous avons présenté une expérience dont les résultats nous ont permis de quantifier et de prouver l'existence des liens morphologiques multilingues. Pour réfléchir plus avant aux possibilités de formalisation de ces différents liens, et pour pouvoir en imaginer une application concrète, nous avons donc construit à l'aide de l'outil ELU (cf. point 3 de ce chapitre) un lexique d'héritage permettant de traiter en même temps la morphologie de l'italien et du français. Nous présentons dans ce chapitre les différentes parties de ce programme, en commençant par une description de l'architecture générale puis en traitant individuellement chaque nœud et leur relation.

4.1. Architecture générale

Avant de commencer à construire un lexique d'héritage multilingue, il faut avant tout réfléchir à l'organisation générale de celui-ci, et mettre en évidence ce que l'on veut exprimer et à quel endroit de la hiérarchie les informations doivent se trouver.

Dans le cadre du projet Polylex, différentes options architecturales avaient été étudiées pour formaliser au mieux les liens multilingues. Dans le chapitre 4, nous avions décrit les deux principales approches, le modèle non-paramétrable et le modèle paramétrable. Alors que le premier consistait uniquement en un accolement de deux hiérarchies monolingues (le plus souvent pré-existantes) et en l'adjonction d'une hiérarchie supérieure reliant les deux hiérarchies monolingues, le deuxième modèle vise davantage à recréer une hiérarchie pouvant gérer en même temps les deux langues souhaitées. De part cette 'recréation', ce modèle implique une plus grande réflexion préliminaire.

La formation des mots dans un lexique d'héritage doit en effet avoir une certaine cohérence. Le projet Polylex avait individualisé des similitudes phonologiques. Pour cette raison, les concepteurs du projet avaient effectué un découpage des mots en se basant sur les propriétés phonologiques de chaque mot, et en individualisant les phonèmes communs. Dans le deuxième chapitre, nous avons individualisé des similitudes étudiées entre l'italien et le français sur un plan morphologique plutôt que phonologique, comme les bases et les affixes semi- ou totalement cognats. L'expérience (décrite au chapitre 3) a également montré que les mots présentés étaient

plus facilement reconnus lorsque leur base était cognate, même si leurs suffixes différaient partiellement ou totalement. Pour toutes ces raisons, mais également par manque de temps et de connaissances, nous nous sommes limités à décomposer les mots en morphèmes, suivant par là une approche purement morphologique.

Nous voulons donc exprimer les généralités de formation des mots à partir d'une base commune, et établir le processus constructionel de cette base. Par exemple, à partir de *possibil-* (forme commune aux deux langues : la base), nous pouvons former deux formes lemmatisées avec l'ajout des suffixes - *possibilité/possibilità* (suffixes proches au niveau morphologique, et donc contenus dans la même classe). Puis, à partir de ces formes lemmatisées, il est possible de lui ajouter un préfixe: *impossibilité/impossibilità*, et ensuite, à partir de ces deux formes, de dériver les formes du pluriel et du singulier. C'est seulement dans un deuxième temps que nous avons voulu également pouvoir gérer les changements morphographémiques internes des bases semi-cognates. Nous voudrions par exemple pouvoir traiter dans notre lexique des paires de mot du genre de *complexité/complessità*, où le changement morphographémique interne (x → ss) semble être utilisable pour d'autres paires (comme dans la paire *sexe/sesso*) (cf. chapitre 2, point 4.1.2).

Nous reprenons dans la figure 4 le schéma proposé par les concepteurs de Polylex en le modifiant pour qu'il corresponde à notre démarche de construction basée sur la décomposition préfixe-base-suffixe.

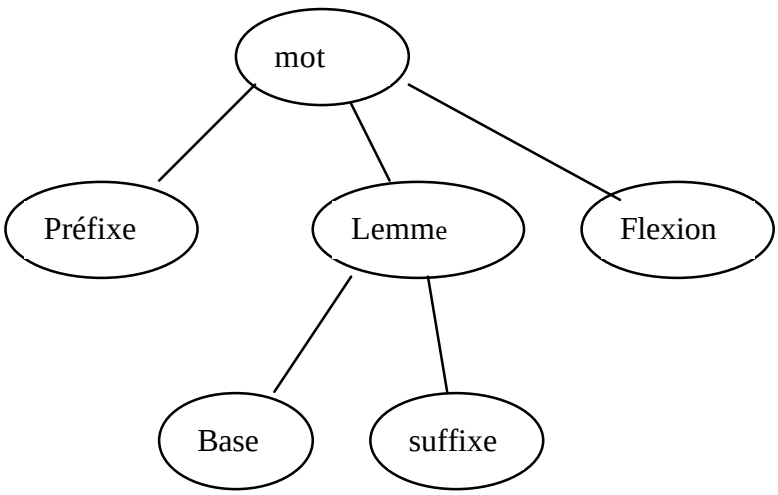


Figure 4

Dans cette figure 4, un mot comme *impossibilités* est composé d'un préfixe (optionnel) *im-* , et d'un lemme *possibilité* et d'une flexion (également optionnelle) *–s*. Le lemme est lui-même composé d'une base *possibil-*, et d'un suffixe *-ité.* C'est en suivant ce principe de construction que nous avons cherché à identifier quelle partie pouvait être utilisée pour une généralisation multilingue. Evidemment, dans le cadre d'une morphologie monolingue, le lemme n'a pas besoin d'être décomposé, mais dans le cadre d'une morphologie multilingue où nous voulons exploiter les similitudes des bases et des suffixes, il nous fallait opérer cette décomposition. Ceci a engendré une nouvelle problématique, celle d'une base linguistiquement peu motivée. Nous en reparlerons dans le point 5.1.

Cette approche décompositionnelle où les informations communes se retrouvent à tous les niveaux de la formation du mot nous a fait choisir un modèle paramétrable, dans lequel nous pouvons mettre en commune toutes les informations communes aux deux langues.

En nous fondant sur le processus de construction décrit ci-dessus, nous décrivons maintenant l'architecture générale.

Au sommet de la hiérarchie, la classe des noms est la classe la plus haute de la hiérarchie étant donné que cette hiérarchie s'occupe des noms. En plus de contenir les informations catégorielles (<* cat> = n), elle contient également un certain nombre de variantes exprimant les concaténations permettant de former les pluriels et les singuliers dans les deux langues. Comme nous le verrons par la suite, ces concaténations sont dépendantes des langues. Directement en dessous d'elle se trouve la classe de la formation par dérivation des opposés, qui permet la concaténation ou non des éventuels préfixes (permettant de former *impossibilité* sur *possibilité*). Pour permettre plus de flexibilité, ces deux classes contiennent uniquement des informations de concaténation, les préfixes ou les suffixes n'étant pas déclarés à ce niveau.

En dessous de ces deux classes qui peuvent être considérées comme un tout, on trouve toute la classe de formation des formes lemmatisées à partir de la forme commune de base. Dans cette classe, on retrouve les informations de concaténations entre la forme de base (communes aux deux langues) et les suffixes propres à chaque langue. A nouveau, les suffixes ne sont pas déclarés à ce niveau, car ces classes s'occupent uniquement de concaténer les valeurs de la base et celles des suffixes.

En dessous des classes de formation se trouve un certain nombre de classes utilisées pour déclarer les valeurs des suffixes et des préfixes qui seront concaténés par les classes supérieures régissant les concaténations.

Il est important que le choix du suffixe du pluriel, et donc la déclaration de la classe concernée, se fasse au niveau des classes des suffixes de type *–ité / –ità*, étant donné que ces derniers formeront la fin du mot, et conditionneront donc le choix du pluriel. Mais nous reparlerons plus bas de l'organisation de ces 'petites classes'.

Schématiquement, la hiérarchie pourrait être représentée ainsi :

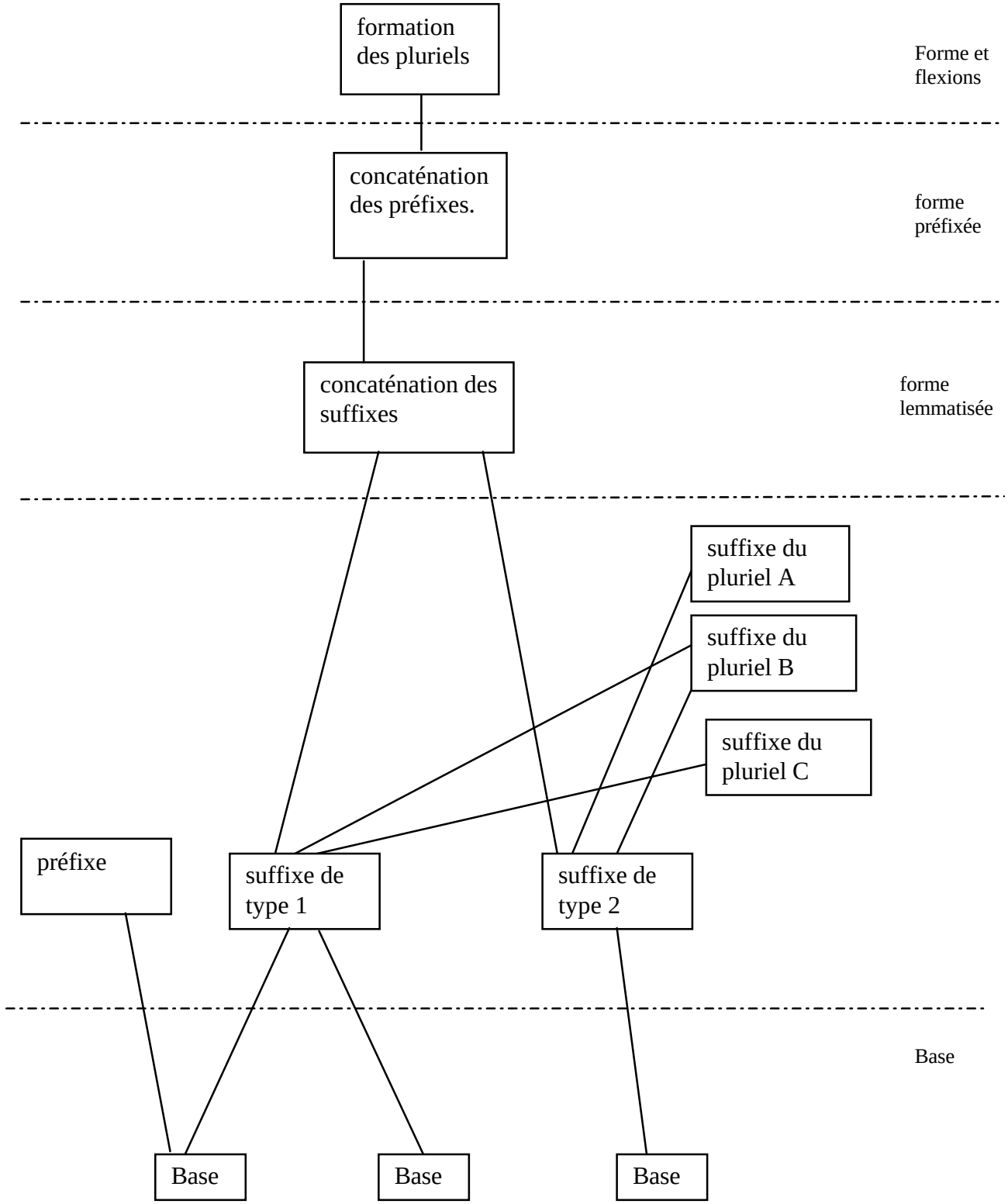


Figure 5

Dans la figure 5, les indications à droite décrivent la partie du mot qui est formée à chaque niveau. Normalement, dans un modèle paramétrable tout comme dans une hiérarchie d'héritage, plus une information est déclarée en bas dans la hiérarchie, plus celle-ci est spécifique. Notre approche diffère sur ce point. En effet, nous voulions exploiter la cognation de certaine base, mais le formalisme ELU impose la déclaration de la base dans les feuilles de l'arbre. C'est pourquoi les bases se retrouvent donc dans les entrées lexicales, au niveau inférieur de la hiérarchie, bien qu'elles ne soient pas des informations plus spécifiques.

Une fois cette organisation définie, il faut, en construisant le lexique, se pencher sur le type d'informations à exprimer, tout en gardant en tête une certaine modularité pour la construction de ce type de lexique.

4.2. Description des différentes classes

Prenons maintenant nœud par nœud, en partant du bas, les différentes informations nécessaires pour chaque classe. La place des classes dans la hiérarchie constitue également un enjeu important. En effet, non seulement elles sont conditionnées par la nécessité d'orthogonalité de l'héritage multiple mais également par les déclarations des classes supérieures. En effet, nous avons vu que chaque classe contient la déclaration de la ou des classes directement supérieures. Cette déclaration exprime donc un lien précis entre le contenu de la classe inférieure et celui de la classe supérieure. Nous verrons dans plusieurs cas que ceci revêt une importance particulière.

4.2.1. La forme de base

La forme de base est définie dans une classe lexicale de ce type :

```
#Word possibil (IM SuffITEITA)
<*sem> = etat
```

Comme toutes les classes d'un lexique d'héritage, elle contient un titre (dans notre cas 'Word', par opposition à 'Class' pour les classes non-lexicales) suivi du nom de la classe qui dans une classe lexicale est en fait la base de formation des mots. Entre parenthèses vient ensuite la déclaration des classes supérieures directes. Dans ce cas, la classe #Word possibil héritera des propriétés de la classe IM (qui contient la déclaration du préfixe *im-* pour la formation de

impossibilité/impossibilità) et la classe SuffITEITA (qui contient la déclaration des deux suffixes semi-cognats –*ité/-ità*).

La classe contient également un certain nombre d'attributs-valeurs qui sont spécifiques à ce mot. C'est le cas du type sémantique (<* sem> = état)[33]. C'est par cet attribut également que l'on peut contraindre certaines formations si la base ne forme pas toute les dérivations proposées par les lexiques. Nous reparlerons de ces contraintes par la suite.

La déclaration des classes supérieures directes contient donc plusieurs classes. Il s'agit alors d'héritage multiple.

Comme nous l'avons expliqué dans la première partie de ce chapitre, l'héritage multiple permet d'exprimer les liens qui existent entre une classe et différents éléments de la hiérarchie. La contrainte qui survient avec les héritages multiples c'est que, dans l'outil ELU, la déclaration des classes supérieures doit se faire de manière ordonnée. Cette 'précédence de classe' implique que les classes supérieures à une classe données doivent se situer au même niveau de la hiérarchie si ces différentes classes appartiennent à la même hiérarchie.

L'héritage multiple orthogonal conditionne particulièrement l'architecture générale du lexique. Voyons à présent les classes situées au niveau supérieur. Ces classes concernent les classes de déclaration des suffixes et des préfixes.

4.2.2. Les classes de déclaration des préfixes

Comme nous venons de le mentionner, nous avons voulu formaliser la formation de *impossibilité* sur *possibilité.* Même si la classe de concaténation entre la forme de base (*possibil-*) et le préfixe *im-* est situé à un niveau supérieur de la hiérarchie, la déclaration du préfixe doit se situer à un niveau bien inférieur. C'est en effet la base qui définit si le mot peut prendre un préfixe pour former un dérivé ou non et quelle est la forme lexicale de ce suffixe. Il fallait donc que cette information de préfixe soit directement héritée par les mots des classes lexicales. Ainsi, juste au-dessus des classes lexicales, nous trouvons la classe suivante :

[33] Nous avions dit ne pas vouloir introduire de sémantique lexicale dans ce lexique purement morphosyntaxique, mais il nous paraissait intéressant d'exprimer au moins un type sémantique, quand celui-ci est commun au deux langues.

```
#Class IM()
    |
    Preff = im
```

La classe ci-dessus, introduite par la déclaration '#Class' qui indique qu'il s'agit d'une classe non lexicale, contient une information qui donne la valeur à l'attribut Preff (pour préfixe), dans notre cas, *im-*. Remarquons également que cette classe n'hérite d'aucune autre, pour éviter des problèmes d'orthogonalité. En effet, les classes inférieures à celle-ci hériteront évidemment de plusieurs classes en même temps. Cet héritage multiple nécessite donc une 'orthogonalité' pour que l'unification fonctionne. Cependant, il paraît difficile de rattacher la classe des préfixes à un autre point de la hiérarchie. Nous avons donc préféré créer une hiérarchie indépendante pour ce type de classe. C'est pourquoi, cette classe n'hérite de rien et constitue donc le sommet d'une hiérarchie. Le même genre de classe a également été défini pour le préfixe *il-*, préfixe parfaitement cognat qui se retrouve par exemple dans la paire *illégalité/illegalità*

```
#Class IL()
    |
    Preff = il
```

4.2.3. Les classes de déclaration des suffixes

Egalement directement au-dessus des classes lexicales se trouvent les classes de déclaration des différents suffixes. A nouveau, la dépendance entre la base et les classes des suffixes nécessitait leur rapprochement dans la hiérarchie, étant donné que c'est dans les classes des formes de base que se décide quel suffixe doit être associé.

Chaque classe contient une paire de suffixes semi-cognats, un pour l'italien et l'autre pour le français. Par exemple, nous avons déclaré la classe suivante :

```
#Class SuffITEITA (PLU_A_IT  PLU_S_FR  CONCA2)
!accord_Ngenre (f)
Suff_it = ità
Suff_fr = ité
```

Cette classe contient tout d'abord son nom (#Class SuffITEITA) suivi de la déclaration de ses classes supérieures. Ces classes supérieures sont la classe de concaténation (#CONCA2) ou

les suffixes seront concaténés à la forme de base, et les classes de déclaration des pluriels. Ces classes de pluriel, dont nous reparlerons, sont dépendantes des langues, le phénomène n'étant pas considéré comme cognat. De plus, il est important qu'il soit situé en dessus de la classe de déclaration des suffixes, étant donné que les suffixes, par essence situés à la fin des mots, conditionne les choix de la forme du pluriel.

Parmi les informations internes de cette classe, on retrouve l'attribut genre (défini par un macro) qui dans ce cas a pour valeur le féminin. Cette valeur est en effet commune aux deux langues et est conditionnée par le suffixe. Les deux suffixes semi-cognats sont également déclarés, un pour chaque langue.

D'autres classes de déclaration des suffixes ont également été créées. Une concerne l'isomorphe du suffixe *-ité/-ità* c'est-à-dire *–té-/tà*. Un processus de transformation des isomorphes suffixales pourrait être également envisagé, étant donné que le même phénomène se retrouve dans les deux langues (suppression du 'i'). Une autre classe concerne d'autres suffixes semi-cognats (*-ation/-azione*) pour la formation de paires de mot du genre *information/informazione.*

```
#Class SuffTION(PLU_EI_IT PLU_S_FR Conca2)
!accord_Ngenre (f)
Suff_it = azion
Suff_fr = ation
```

La classe ci-dessus est construite sur le même modèle que la classe #SuffITEITA sauf que le suffixe italien est sous sa forme lemmatisée, et ceci pour une raison précise. Alors que la flexion du pluriel français s'effectue dans de nombreux cas par l'ajout de la consonne 's', l'italien forme son pluriel par la transformation de la voyelle finale (en générale 'o' en 'i', 'a' en 'e' et 'e' en 'i'). Il est alors d'usage en linguistique informatique, de prendre une forme de base à laquelle on ajoute les différentes marques, du pluriel et du singulier. Dans la classe du suffixe *#Class SuffTION*, le suffixe français est déclaré en entier (correspondant à la terminaison du singulier à laquelle peut se concaténer la marque du pluriel 's') et le suffixe italien est déclaré sans voyelle finale, qui sera instantiée dans la classe de formation du pluriel par la concaténation soit de la terminaison du singulier (dans ce cas, *-e*) soit de la terminaison du pluriel (dans ce cas, *-i*). Notons également que cette forme spéciale du suffixe italien n'était pas nécessaire pour la classe des suffixes *–ité/-ità* étant donné que les mots se terminant par une voyelle accentuée ne prennent aucune marque du pluriel en italien (Dardano *et al.*, 1985, p.116).

Les deux classes citées ci-dessus traitent des mots qui ne possèdent qu'un genre et partant qu'une flexion en genre. Nous avons également tenté de formaliser les suffixations des mots possédant deux flexions de genre, pour le masculin et le féminin. C'est par exemple le cas de *aviateur/aviatore* et leur équivalent féminin *aviatrice/aviatrice*. Pour déclarer ces quatre suffixes, nous avons utilisé la classe ci-dessous:

```
#Class SuffTORE (PLU_S_FR PLU_EI_IT Conca2)
    |
    Suff_it = tric
    Suff_fr = trice
    !accord_Ngenre (f)
    |
    Suff_it = tor
    Suff_fr = teur
    !accord_Ngenre (m)
```

Dans cette classe, très similaire à la classe #SuffITEITA, nous avons affaire à une disjonction de variable qui porte sur le genre. Chaque variante est séparée par une barre verticale ('pipe') et contient une information de genre défini par une macro (!accord_Ngenre). A nouveau, si les suffixes français (Suff_fr) sont déclarés dans leur intégralité, les suffixes italiens sont tronqués pour permettre leur flexion dans les dernières voyelles. C'est cette raison qui conditionne également le fait que les suffixes *-trice/-trice*, bien que cognats parfaits, ne sont pas déclarés une fois dans une seule structure.

4.2.4. Les classes de déclaration des pluriels

Le dernier type de classe de déclaration nécessaire pour rendre ce lexique complet concerne les déclarations des pluriels. Ces classes se situent juste au-dessus des classes de concaténation, pour les raisons déjà invoquées. Elles prennent la forme suivante :

(i)
```
#Class PLU_EI_IT( )
        Suff_plu_it = i
        Suff_sing_it = e
```

(ii)
```
#Class PLU_S_FR()
        Suff_plu_fr = s
```

Comme nous pouvons le constater dans les deux exemples ci-dessus, les suffixes du pluriel sont dépendants des langues, pour les raisons déjà évoquées plus haut. Ainsi, il est nécessaire de déclarer dans (i) le suffixe du singulier (Suff_sing_it = e) et le suffixe du pluriel (Suff_plu_it = i) pour l'italien, et dans (ii) le suffixe pluriel du français (Suff_plu_fr = s), le singulier correspondant à la forme lemmatisée en français. Pour chaque suffixe du pluriel, il est nécessaire de créer une nouvelle classe.

Les suffixes du pluriel seront concaténés dans la classe des Noms, dont nous reparlerons.

4.2.5. Les classes de concaténation des suffixes

La principale classe de concaténation concerne la concaténation entre les formes de base (définies comme potentiellement cognate entre les deux langues) et les suffixes semi-cognats déclarés dans les différentes classes précédemment citées.

```
#Class Conca2(OPP)
    |
    Form_lem = <base> && Suff_it
    <lang> = it
    |
    Form_lem = <base> && Suff_fr
    <lang> = fr
```

La classe Conca2 hérite directement de la classe #OPP, avant dernière classe de la hiérarchie où l'on pourra concaténer les formes lemmatisées et leurs éventuels préfixes. Elle est constituée d'une disjonction entre les deux concaténations possibles, une pour former le mot italien et l'autre pour former le mot français. Les attributs langues sont donc introduits à ce niveau par les structure (<lang> = it ou <lang> = fr).

4.2.6. Les classes de concaténation des préfixes

De même que pour les suffixes, la classe de concaténation des préfixes contient une disjonction.

```
#Class OPP(Nom)
    Form_lem = <base>
    Preff = vide
    |
```

```
Singform = Form_lem
|
Singform = Preff && Form_lem
Preff = ~vide
```

Cette disjonction est indépendante des langues. En effet, les deux langues possèdent un processus de préfixation identique, qui est donc exprimée dans cette classe. Mais cette dernière ne contient aucune déclaration du préfixe. Si le préfixe concerné est propre à une langue, il suffira de déclarer cette langue comme attribut du préfixe (dans la classe de déclaration des préfixes). Ainsi, étant donné que dans la classe OPP, le préfixe est concaténé sur la forme lemmatisée (*Form_lem*), et que celle-ci contient déjà une information de langue (attribuée dans une classe de concaténation du suffixe), si la langue du préfixe et de la forme lemmatisée sont différentes, l'unification échouera.

En revanche, dans ce cas, la disjonction concerne la présence ou non du préfixe. En effet, cette classe doit pouvoir gérer à la fois les mots ayant une préfixation (et donc un préfixe déclaré dans les classes inférieures) et les mots n'ayant pas de préfixe possible (comme *liberté/libertà*). C'est pour cela que nous avons utilisé dans cette classe les potentialités offertes par la règle du 'par défaut'. Comme nous l'avons vu plus haut dans ce chapitre, les informations situées directement en dessous du nom de la classe sont considérées comme 'par défaut', c'est-à-dire qu'elles ne s'appliquent seulement si aucune autre information ne vient la contredire dans les classes inférieures ou dans les variables de la classe. Dans notre cas, la valeur par défaut du préfixe (Preff) est 'vide'. Si l'attribut a déjà une valeur héritée d'une autre classe (comme par exemple *Preff= im*), cette information plus spécifique est alors héritée en priorité et permet la concaténation exprimée dans la deuxième variante où l'on a déclaré *Preff = ~vide* (le ~ signifiant la négation). En revanche, si aucun n'attribut n'a été donné au préfixe, (comme *liberté* qui n'hérite d'aucune classe d'assignation d'un préfixe), la valeur par défaut s'applique alors et c'est la première variante qui est prise en compte.

4.2.7. La classe de noms

La classe au sommet de la hiérarchie s'appelle la classe des *#Noms*. Elle permet simplement d'assigner la catégorie grammaticale à tous les mots du lexique (<* cat> = n). Elle contient également toutes les règles de concaténation des suffixes du pluriel.

```
#Class Nom ()
      <* cat> = n
      |
      <form> = Singform && Suff_sing_it
      !accord_Nnombre(sg)
      <lang> = it
      |
      <form> = Singform && Suff_plu_it
      !accord_Nnombre(pl)
      <lang> = it
      |
      <form> = Singform
      !accord_Nnombre(sg)
      <lang> = fr
      |
      <form> = Singform && Suff_plu_fr
      !accord_Nnombre(pl)
      <lang> = fr
```

Chaque variable (introduite par une barre verticale) a pour fonction de concaténer la forme de base avec le suffixe du pluriel déclaré précédemment, et apporte les informations de nombre correspondant (!accord_Nnombre).

4.2.8. La formation des bases légèrement différentes (les changements morphographémique)

Jusqu'à présent, nous avons décrit un lexique multilingue qui gère les mots dont la base est parfaitement cognate et dont les différents affixes sont semi-cognats. Mais nous avons également étudié la possibilité d'introduire des bases semi-cognates dans ce lexique, et de gérer les changements morphographémiques internes, comme celui de la paire *complexité/complessità* où le 'x' devient 'ss'. Etant donné que cette variation semble assez productive, il nous a paru intéressant de tenter d'en faire une classe à part entière, pour pouvoir profiter de ces changements pour différents mots. Nous avons créé des classes de déclaration de ces changements au premier niveau de la hiérarchie, juste au-dessus des classes lexicales. La position de ces classes de changement permet à la classe lexicale d'hériter directement de ces changements. De plus, ces classes, tout comme la classe de déclaration des préfixes, n'héritent de rien et constituent donc une hiérarchie à elle toute seule. Elle ne pose donc pas le problème de l'héritage multiple.

Concrètement, ces changements sont exprimés par le biais de macro (c.f. point 3.10 de ce chapitre). La macro est donc définie au sommet du lexique dans la partie '#Define Mac'. Dans le cas qui nous intéresse, elle est constituée d'un nom X_SS suivi, entre parenthèses, des différents attributs possibles.

X_SS(Base,S1,S2)
Base = S1 && **x** && S2

La macro est ensuite 'appelée' dans la classe X_SS pour définir la nouvelle forme de base pour l'italien (Form_base_it) :

#Class X_SS( )
!X_SS(<base>,S1,S2)
Form_base_it = S1 && **ss** && S2

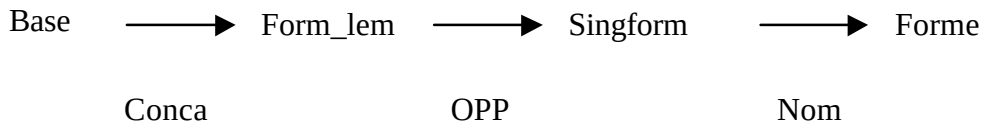
A noter également que cette macro oblige la formation d'une nouvelle variable **Form_base_it**. Ceci entraîne des petits changements structurels dans toutes les autres classes. Nous aborderons ce sujet dans le point suivant. Ajoutons que nous avons également utilisé ce type de macro pour gérer les changements morphographémiques présents dans les paires *légalité/legalità* où le *é* devient *e*, et dans les paires de types *conducteur/conduttore* où le *ct* devient *tt* .

4.2.9. Les problèmes de bases

Comme nous l'avons vu dans les différentes classes de concaténation que nous avons décrites ci-dessus, le processus de concaténation est effectué par une équation qui prend en compte plusieurs éléments. Par exemple, dans la classe de concaténation des suffixes, l'équation des concaténations est définie comme suit :

Form_lem = <base> && Suff_fr

L'équation contient à droite deux éléments, qui proviennent des classes inférieures à cette classe, et à gauche, une variable qui sera reprise dans les classes supérieures. Ici, le suffixe **Suff_fr** est défini dans la classe #SuffITEITA et l'élément <base> est prédéfinie par le programme comme étant le mot (la classe lexicale) qui hérite de cette classe. Dans notre architecture, le fait de construire un mot en plusieurs étapes nous a obligé à spécifier chaque fois une 'nouvelle' variable qui pourra être utilisée dans la concaténation suivante, comme le montre le schéma ci-dessous :



En dessous de chaque flèche, nous montrons quelle classe de concaténation procède à la création de la forme suivante. Ainsi, la *Base* contenue dans la classe lexicale (par exemple *possibil*) hérite des propriétés de la classe Conca où sont concaténés la *Base* et le suffixe, formant ainsi la forme lemmatisée **Form_lem** (*possibilité/possibilità*), et ainsi de suite.

Le problème réside dans le fait que cette architecture semble figée, dans le sens où l'ajout d'une nouvelle classe procédant à un quelconque changement entraînera nécessairement la création d'une nouvelle variable et il sera alors plus difficile d'intégrer cette nouvelle classe dans la hiérarchie.

Reprenons l'exemple de la macro X_SS précédemment citée. Elle formait une nouvelle variable pour la base italienne (**Form_base_it**) à partir de la forme de base définie dans la classe lexicale. Cette même classe lexicale hérite plus tard de la classe de concaténation des suffixes (#Conca) qui est normalement définie comme suit :

```
#Class Conca2(OPP)
    |
    Form_lem =  <base> && Suff_it
    <lang> = it
    |
    Form_lem = <base> && Suff_fr
    <lang> = fr
```

Le problème réside dans le fait que la forme lemmatisée (**Form_lem**) est construite sur la forme de base (<base>) provenant de l'entrée lexicale et des suffixes. Cette classe fonctionne donc très bien pour toutes les entrées lexicales sans changement morphographémique, et doit toujours fonctionner ainsi. Mais cette classe doit également fonctionner pour les 'nouvelles' variables générées par les classes de changements morphographémiques (**Form_base_it**). Pour résoudre ce problème, nous avons utilisé les potentialités fournies par la fonction 'par défaut' d'un lexique d'héritage. Nous avons donc modifié la classe comme suit:

```
#Class Conca2(OPP)
    Form_base_it = <base>
    |
```

```
Form_lem = Form_base_it && Suff_it
<lang> = it
|
Form_lem = <base> && Suff_fr
<lang> = fr
```

Dans cette nouvelle version de la classe #Conca2, les deux variables sont précédées par une valeur par défaut, qui se réalise seulement s'il n'y a pas d'élément contradictoire dans la hiérarchie. Ainsi, si une classe lexicale hérite directement de cette classe, elle prend la valeur **Form_base_it** déclarée dans cette classe. En revanche, si le mot a déjà reçu une valeur pour **Form_base_it**, (dans une classe de changement morphographémique par exemple), c'est cette valeur qui sera retenue.

Par exemple, une base semi-cognate comme *complex-* qui a été modifiée dans la classe de changements morphographémiques X_SS, et qui possède donc deux valeurs (base = *complex-* et **Form_base_it** = *compless-*) peut hériter de cette classe #Conca et les deux variantes s'effectuent sans problème. De même, une base cognate déclarée dans une classe lexicale (*possibil-* qui n'a qu'une seul forme pour les deux langues), pourra s'appliquer aux deux équations de concaténation grâce à la valeur par défaut, qui s'applique à la base.

Le même genre de problématique s'est posé lorsque nous avons envisagé la modularité de ce lexique multilingue.

4.2.10. Un lexique modulaire

Nous avions déjà évoqué le fait que cette morphologie doit pouvoir traiter les similitudes mais également les différences. Ainsi, pour assurer la complétude du lexique, il faut pouvoir également y insérer des mots pas du tout cognats et les rattacher aux informations le concernant en excluant les informations concernant l'autre langue. Idéalement, il faudrait que la classe lexicale d'un mot non cognat, comme par exemple *tagliatore* (le tailleur), puisse hériter des informations de concaténation du pluriel. Pour un cas comme celui-ci, il faut tout d'abord modifier légèrement la classe lexicale par rapport aux classes standard du lexique :

```
#Word taglia (SuffTORE)
<* sem> = person
<lang> = it
```

Le principal changement réside dans le fait que l'information de langue est directement donnée à ce niveau là. Ceci permet à cette classe d'hériter des classes des Suffixes en *–tore/-trice* sans pour autant hériter des suffixes français contenus dans cette classe.

Le même problème s'est posé pour les mots dont les seules informations pertinentes étaient celles de la formation du pluriel.

Par exemple, pour le mot français *chaise* qui n'a pas d'équivalent cognat en italien, nous voudrions utiliser les informations contenues dans la classe #Nom. Pour ce faire, nous faisons hériter directement la classe lexicale *chaise* de cette classe, ainsi que des petites classes de déclaration des flexions du pluriel (PLU_S_FR).

#Word chaise (PLU_S_FR Nom)
<* sem> = artefact
<lang> = fr

Mais dans cette situation, nous retrouvons le même problème que dans le cas de la nouvelle variable créée par la macro. En effet, dans la classe #Nom, toutes les formes sont formées par la concaténation de **Singform** (créée dans la classe OPP) et du suffixe du pluriel.

<form> = Singform && Suff_sing_it

Pour résoudre ce problème, nous avons également utilisé la fonction de l'héritage par défaut.

Ainsi, dans la classe #Nom ci-dessous :

```
#Class Nom ()
    Singform = <base>
    <* cat> = n
    |
    <form> = Singform && Suff_sing_it
    !accord_Nnombre(sg)
    <lang> = it                                    …
```

la variable **Singform** a pour valeur par défaut celle contenue dans l'attribut base (donc *chaise* dans notre cas) si cette variable est définie différemment dans une classe inférieure (ce qui est le cas pour tous les mots à base cognat et qui héritent de classe de concaténation par exemple).

Nous avons ajouté la même valeur par défaut dans chaque classe d'où potentiellement, un nom non-cognat pouvait hériter.

5. Commentaire et évaluation

Jusqu'à maintenant, nous avons décrit les différentes démarches d'élaboration de ce lexique d'héritage multilingue. Avant de passer à une évaluation quantitative à large échelle (dans le chapitre 6), nous abordons quelques problématiques engendrées par la construction d'une telle morphologie.

5.1. Problème de la base

Pour construire *possibilité* et *possibilità*, nous partons d'une base hypothétique commune aux deux langues, à laquelle nous ajoutons les suffixes propres à chacune de ces langues. Cette base, *possibil-*, ne correspond pas à l'adjectif sur lequel le substantif aurait pu se former. Linguistiquement, cette déclaration n'est donc pas très motivée. Informatiquement, elle reste définie comme l'unité commune aux deux langues, mais la question reste de savoir si cet argument purement pratique est valable. Cet écueil ressemble en tout point à ceux des théories décompositionnelles de la morphologie que nous avons évoqués plus haut quand nous avions introduit la théorie de la morphologie holistique. Ainsi, à force de vouloir toujours décomposer en unités minimales les mots d'une langue, on arrive toujours à des unités abstraites qui ne sont que des vues de l'esprit théoriques (Neuvel et Singh, sous-presse2).

Hélène Huot (1994, p. 68) revient également sur les problèmes de base, en comparant ce concept à celui des racines. En effet, si l'on décompose les verbes français *absorber* et *résorber*, on découvre une forme de base *sorb*, certes communes aux deux verbes, mais qui ne correspond à aucune autre forme du français moderne. En revanche, si *sorb* n'est pas une forme de base française, il peut être envisagé comme le reste d'une racine ancienne. Hélène Huot émet l'hypothèse que les grammairiens du siècle dernier ont été « conduits par la comparaison entre langues pour définir des racines indo-européennes de forme relativement abstraite » (Huot, *ibid*.). De plus, elle ajoute que « la confrontation avec les autres langues romanes conduit à rendre compte de parentés lexicales évidentes par des racines moins immédiatement liées aux unités lexicales de chaque langue » (Huot, *ibid*.). Dans cette optique comparatiste, notre base commune peut donc très bien être considérée comme cette racine commune relativement abstraite.

5.2. Problème de la précédence des formes

Nous avons également déjà parlé des classes de changements morphographémiques qui pouvaient gérer les petits changements dans les bases comme dans *complex-* et *compless-*. Pour

effectuer un tel changement, il faut évidemment partir d'un point A pour le transformer en B. Et à nouveau, d'un point de vue linguistique, il n'existe aucune motivation pour dériver l'italien du français, ou inversement. Il est en effet difficile de trancher entre une langue ou une autre pour le choix de cette base. La précédence de forme, consistant à déclarer comme forme de base, la base d'une langue X et d'en dériver la base de la langue Y reste donc un choix purement arbitraire et rendu nécessaire par le formalisme utilisé.

Si un développement futur de ce lexique tente d'appliquer ce lexique à une troisième langue (nous esquisserons les grandes lignes de cet ajout dans le chapitre 6), le choix de la forme de base entre les bases des trois langues en présence sera régi également par des principes d'économie de stockage. On pourra prendre comme base celle qui est partagée par au moins deux de ces langues (s'il y en a une), et produire, à partir de cette base, celle pour la troisième langue. A ce niveau, seules des considérations informatiques en terme d'économie ou des considérations linguistiques dépendant de la langue de travail du développeur nous permettront de prendre comme base une langue plutôt qu'une autre.

5.3. Difficulté de construction

Un des critères essentiels d'un lexique reste sa flexibilité et sa facilité d'implémentation. Dans notre cas, nous aborderons la question de la flexibilité dans le chapitre suivant en intégrant une troisième langue. En ce qui concerne la facilité d'implémentation, elle est fortement conditionnée par l'architecture choisie à l'origine pour pouvoir facilement être implémentée. Un test à large échelle devrait être fait pour évaluer cette faisabilité.

De plus, même s'il est vrai que l'outil ELU a été conçu de manière déclarative pour qu'il soit facilement implémentable par un linguiste ayant peu de connaissance informatique, le système même des lexiques d'héritages reste assez complexe à comprendre. Enfin, l'implémentation d'un lexique multilingue nécessite non seulement une bonne connaissance des lexiques des langues en présence, mais également des ressources importantes pour sa création à grande échelle. Il faut notamment disposer de ressources lexicales importantes dans les langues en présence pour pouvoir individualiser les paires de mots plus ou moins cognats, ainsi que les différents changements morphographémiques internes.

6. Conclusion

Dans ce chapitre, nous avons tout d'abord expliqué nos choix du formalisme utilisé, et proposé un tour d'horizon des différents aspects de l'outil utilisé. Nous avons ensuite exposé l'architecture et les différentes classes du lexique d'héritage construit pour ce travail. Ce lexique est assez restreint, et avait uniquement pour but de montrer la faisabilité d'une telle entreprise et de tracer les grandes lignes pour la création d'un tel lexique à plus large échelle. Dans le chapitre suivant, nous présentons une évaluation quantitative (sur le plan lexical et sur le plan de la place de stockage) et qualitative d'un tel lexique.

Chapitre 6 : évaluations quantitative et qualitative

1. Introduction

Dans ce chapitre, nous présentons plusieurs petites expériences qui évaluent, d'un point de vue quantitatif et qualitatif le lexique multilingue décrit dans le chapitre 5. D'un point de vue qualitatif, nous décrivons tout d'abord la faisabilité de l'ajout d'une ou plusieurs langues romanes, pour que le lexique puisse gérer plusieurs langues à la fois. Ensuite, d'un point de vue quantitatif, nous présentons de petites évaluations effectuées sur des ressources lexicales existantes et sur des corpus de plus large envergure pour pouvoir chiffrer le nombre de mots qui pourraient être traités par notre lexique. Enfin, nous mentionnons également la problématique de la place de stockage qui reste l'un des principaux arguments en faveur d'un traitement multilingue de la morphologie.

2. Evaluation qualitative: l'ajout d'une autre langue

Les concepteurs du projet Polylex[34] citaient, parmi les autres avantages d'un lexique multilingue, le fait que l'on pouvait facilement ajouter une autre langue formellement proche. Nous avons donc tenté d'ajouter quelques mots d'espagnol, pour voir comment réagissait l'architecture et quels changements étaient nécessaires. L'espagnol est également une langue latine et semble, formellement, être apparentée au français et à l'italien. Nous n'évoquerons ici aucune étude quant à la proximité de l'espagnol avec les autres langues, comme nous l'avons fait pour la paire français/italien.

Pour mener à bien l'expérience de l'ajout d'une autre langue, nous avons demandé à une personne hispanophone de traduire la liste de mot de notre lexique. Nous avons ainsi obtenu des triplets du type *maternité/maternità/maternidad*. Nous avons ensuite formalisé cette description de manière à l'intégrer dans notre lexique.

[34] The Polylex web pages, http://www.cogs.susx.ac.uk/lab/nlp/polylex/ , et chapitre 4 de ce mémoire.

2.1. Les principaux changements

Pour pouvoir générer les formes espagnoles à partir de la base commune, il faut tout d'abord déclarer le suffixe de cette langue dans la classe de déclaration des suffixes. Evidemment, le suffixe *–idad* est considéré comme un suffixe semi-cognat de *–ité* et de *–ità*, et sera donc déclaré dans la même classe. De plus, ce suffixe est présent dans des mots de genre féminin, ce qui constitue une généralisation multilingue de plus.

Nous obtenons ainsi la classe suivante:

```
#Class SuffITEITAIDAD (PLU_A_IT PLU_S_FR PLU_ES_ES Conca3)
!accord_Ngenre (f)
Suff_it = ità
Suff_fr = ité
Suff_es = idad
```

Dans cette classe, les trois suffixes (*-ité, -ità,* et *-idad*) se retrouvent déclarés ensemble. Comme nous l'avions déjà mentionné (chapitre 5 point 4.2.4), les suffixes conditionnent le type de pluriel. Par conséquent, les classes supérieures à cette classe sont celles de déclaration du pluriel (PLU …) et celle de la concaténation (Conca3).

Comme pour l'italien et le français, la classe de formation des pluriels espagnols (dans laquelle on déclare le suffixe du pluriel espagnol (Suff_plu_es = es)), a donc été ajoutée.

Puis, nous avons ajouté une variante dans la classe de concaténation (qui devient Conca3). Cette classe, qui concatène la forme de base et le suffixe, est alors décrite comme suit :

```
#Class Conca3(OPP)
    |
    Form_lem = <base> && Suff_it
    <lang> = it
    |
    Form_lem = <base> && Suff_fr
    <lang> = fr
    |
    Form_lem = <base> && Suff_es
    <lang> = es
```

Les formes lexicales des trois langues sont ainsi formées.

Le dernier ajout concerne la classe des #Noms, où l'on forme la forme des pluriels et des singuliers en concaténant les formes lemmatisées et les suffixes du pluriel. Cette classe nécessite également l'ajout d'une règle de concaténation pour traiter l'espagnol.

```
#Class Nom ()
      <* cat> = n
      |
      <form> = Singform && Suff_sing_it
      !accord_Nnombre(sg)
      <lang> = it
      |
      !accord_Nnombre(pl)
      <form> = Singform && Suff_plu_it
      <lang> = it
      |
      !accord_Nnombre(sg)
      <form> = Singform
      <lang> = fr/es
      |
      !accord_Nnombre(pl)
      <form> = Singform && Suff_plu_fr
      <lang> = fr
      |
      !accord_Nnombre(pl)
      <form> = Singform && Suff_plu_es
      <lang> = es
```

Alors que l'italien nécessite deux règles de concaténation (une pour le singulier et une pour le pluriel, comme nous l'avions expliqué dans le chapitre 5) l'espagnol fonctionne comme le français, du moins pour les mots qui nous intéressent ici. Ainsi, la forme du singulier correspond à la forme lemmatisée (Singform) et c'est seulement pour le pluriel que l'on procède à une concaténation avec le suffixe approprié *(<form> = Singform && Suff_plu_es*). Pour le singulier, nous avons donc simplement ajouté une disjonction dans la valeur de l'attribut langue, permettant à cette variante d'être valable pour le français et pour l'espagnol.

L'ajout d'une langue apparentée reste assez aisé, pour autant qu'une étude préliminaire des lexiques des langues en présence soit effectuée pour individualiser les différentes cognations. Une fois effectués les changements principaux, il ne reste plus qu'à ajouter les bases cognates pour pouvoir générer le lexique.

Cependant, l'ajout d'une quatrième langue provoquerait à nouveau une modification des règles dans leur ensemble. A chaque modification, il est donc nécessaire de garder en tête la totalité du lexique pour ne pas 'détruire' le travail réalisé.

Il nous faut également ajouter, que, comme pour la construction d'un lexique bilingue, les ressources lexicales et les connaissances des langues en présence sont primordiales pour la mise en évidence de bases communes et de changements morphographémiques internes généralisables.

2.2. L'utilité d'un tel ajout

Le fait de pouvoir ajouter à un tel lexique une autre langue présente de nombreux avantages. Le principal est de pouvoir intégrer, à moindre coût et de manière cohérente une langue pour laquelle des recherches en TALN et la constitution de lexiques informatisés n'ont pas encore fait l'objet de projet d'envergure. Nous pensons notamment à d'autres langues latines comme le catalan ou le romanche.

De plus, si la place de stockage est fortement diminuée dans un lexique gérant deux langues à la fois, elle devrait l'être d'autant plus dans un lexique pour trois langues, voir même quatre.

2.3. Encore une autre langue?

Après avoir demandé de traduire les différents mots de notre lexique à une personne hispanophone, nous avons également étendu l'expérience aux langues catalane et roumaine. Le résultat, dont nous présentons un résumé ci-dessous, semble parler de lui-même. En effet, les similarités morphologiques multilingues se retrouvent, à première vue, dans de nombreuses langues, comme le montre les figures suivantes :

a) les mots de suffixes –*ité* et leurs correspondants dans d'autres langues romanes

Français	Italien	Espagnol	Roumain	Catalan
complexité	complessità	complexidad	complexitate	complexitat
légalité	legalità	legalidad	legalitate	legalitat
maternité	maternità	maternidad	maternitate	maternitat
fraternité	fraternità	fraternidad	fraternitate	fraternitat
illégalité	illegalità	ilegalidad	ilegalitate	ilegalitat
impossibilité	impossibilità	imposibilidad	imposibilitate	impossibilitat
liberté	libertà	libertad	libertate	llibertat
possibilité	possibilità	posibilidad	posibilitate	possibilitat

b) les mots de suffixes –*teur* et leurs correspondants dans d'autres langues romanes

Français	Italien	Espagnol	Roumain	Catalan
acteur	attore	actor	actor	actor
collaborateur	collaboratore	colaborador	colaborator	colaborador
conducteur	conduttore	conductor	conduc•tor	conductor
traducteur	traduttore	traductor	traduc•tor	traductor

c) les mots de suffixes –*tion* et leurs correspondants dans d'autres langues romanes

Français	Italien	Espagnol	Roumain	Catalan
information	informazione	información	informa•ie	informació
libération	liberazione	liberación	liberalizare	alliberació

La première constatation repose sur les bases communes. Un certain nombre de mots ont une base commune et seul le suffixe varie, comme dans le cas de *maternité* qui devient *matern-ità* (It) , *matern-idad* (Es), *matern-itate* (Ro), *matern-itat* (Ca). On remarque également que le suffixe *–ité,* possède également un isomorphe dans les autres langues romanes (*-tà, -tad, -tate, -tat*), avec à chaque fois la suppression du *i*. Les préfixes d'opposition *im–* est quant à lui parfaitement cognat dans les cinq langues citées. Notons encore que les changements internes se retrouvent avec une certaine uniformité dans toutes les langues. Ainsi, les deux 's' de *possibilité*

sont présents en français, en italien et en catalan, alors qu'ils *deviennent* 's' en espagnol et en roumain.

Cette brève comparaison apporte encore un argument de plus à l'exploitation des régularités multilingues à travers plusieurs langues romanes apparentées et permet d'envisager l'extension de notre lexique bilingue à plusieurs autres langues.

3. Evaluation quantitative

Pour valider encore d'avantage la construction d'un tel lexique, nous avons également procédé à une évaluation quantitative des mots formellement proches dans ces deux langues.

3.1. Evaluation sur des ressources lexicales existantes

Pour évaluer le nombre de mots formellement apparentés et qui pouvaient être traités par les différentes règles de notre lexique d'héritage, nous nous sommes basés sur les ressources lexicales élaborées à l'ISSCO dans le cadre du projet Multext[35]. La figure 1 résume le nombre d'entrées que nous avions à notre disposition.

Français	Italien
24'792 entrées	46'300 entrées

Figure 1

Nous avons d'abord extrait de ces deux lexiques tous les mots avec un suffixe donné (par exemple, la paire *–ité/-ità*). Pour évaluer le nombre de mots ayant une base commune et un suffixe semi-cognat, nous avons ensuite transformé les mots italiens en remplaçant le suffixe par leur équivalent français, à l'aide de simples commandes Unix de substitution. Finalement, nous refaisions un contrôle de cette nouvelle liste dans le lexique français.

Par exemple, le mot *acerbità* était extrait du lexique italien, transformé dans une forme potentiellement française par substitution du suffixe (*acerbité*), puis contrôler dans le lexique français.

[35] Multiligual Text Tools and Corpora, Projet LRE 62-050, 1994-1996

Nous obtenions ainsi une liste du nombre de mots communs aux deux langues. Les chiffres sont résumés dans la figure 2, en fonction de chaque suffixe.

Paires de suffixes	Nombre de mot
ité/ità	128
té/tà	4
eté/età	12
teur/tore trice/trice	54
tion/zione	241
Total	439

Figure 2

Il faut évidemment ajouter que ces chiffres ne concernent que les formes du singulier, mais que notre programme est capable, nous l'avons vu, de gérer également les flexions du pluriel propres à chacun de ces suffixes.

Pour les changements morphographémiques internes, le calcul est beaucoup plus difficile étant donné que ces changements sont indépendants des suffixes considérés ici. L'évaluation du nombre de paires de mots pouvant bénéficier de ceux-ci sera forcément petite si nous nous limitons à les compter uniquement dans les classes des suffixes que nous avons. Ainsi, par exemple, pour les mots au suffixe *–ité/-ità*, nous avons trouvés quatre mots dont la base était commune et qui présentaient le changements morphographémique x→ss (*complexité/complessità, sexualité/sessualità, toxicité/tossicità, proximité/prossimità*).

Nous résumons dans la figure 3 les quelques expériences faites pour quantifier l'étendue de l'application de certaines règles morphographémiques.

	-ité/-ità	-teur/-tore	-tion/-zione
changement 'e' 'é'	27	10	65
changement 'tt' 'ct'	5	16	1
changement 'ss' 'x'	4	0	0

Figure 3

Ce petit nombre nous fait nous demander si une telle règle de changement vaut vraiment la peine. Il faudrait en effet, évaluer leur potentiel génératif avec un plus grand nombre de suffixes semi-cognats. De plus, une approche trilingue pourrait également répondre à cette question, étant

donné que ces changements se rencontrent entre d'autres paires de langues, comme nous l'avons vu dans la section précédente.

3.2. Evaluation sur des données "empirique"

La taille limitée de nos lexiques de référence ne nous donne qu'un bref aperçu du pouvoir 'généralisant' de nos règles. Nous avons donc voulu rechercher sur la base d'un plus grand corpus des candidats-mots dans nos deux langues de travail et pour nos suffixes, à savoir *–ité/-ità* , *-teur/-tore* (en partant du principe que le féminin de ces deux suffixes *-trice/—trice* est potentiellement généralisable) et *-tion/-zione*.

Cette base de candidats-mots devait présenter deux avantages. D'une part, elle n'était pas contrainte par nos ressources lexicales, forcément limitées, et d'autres part, elle devait contenir un grand nombre de néologismes, permettant ainsi de confirmer l'hypothèse émise par Namer (2001, cf. chapitre 4, point 3 de ce travail) au sujet du fort potentiel dérivationnel de nos suffixes et du parallélisme entre l'italien et le français, non seulement au niveau de la forme mais également au niveau de cette capacité à générer de nouvelles formes. Ces deux facteurs nous semblaient être essentiels à la validité d'un lexique multilingue.

Nous nous sommes donc servi d'un outil développé à Toulouse par Ludovic Tanguy, Webaffix (Tanguy et Hathout, 2002) qui permet de rechercher dans le corpus du Web des mots en fonction de leur suffixes. Evidemment, le corpus du Web est caractérisé par sa grande hétérogénéité (Tanguy et Hathout, *ibid*.) et il convient de prendre les résultats avec grande précaution. En effet, le corpus Web n'est pas contrôlé. Les nouveaux mots ainsi récoltés peuvent en fait n'être que des fautes d'orthographes, des mauvaises segmentations, ou encore de mots pas vraiment 'lexicaux' (noms de fichier, langages de programmation, codes informatiques, …).

Webaffix permet donc de faire une requête généralisée sur un moteur de recherche[36] en utilisant comme critère un suffixe donné. L'outil interroge Altavista™ et récolte les 20 premières pages où un mot pertinent a été trouvé. Pour éviter le bruit, il utilise également une liste de mots attestés, permettant de ne pas récolter ce qui se trouve déjà dans le lexique de l'utilisateur, et une

[36] dans ce cas, Altavista™ (www.altavista.com) , qui est un des rares moteurs permettant de faire des recherches tronquées.

liste de triplets de lettres correspondant à tous les tri-grammes de début de mot d'une langue donnée. Ainsi, pour le suffixe *–ité* il lancera une requête du type: *aba*ité*, *abc*ité*, …, *zyt*ité*.

Nous avons donc fait fonctionner le programme pour chaque paire de suffixes intéressants du point de vue de notre évaluation. Nous avons trouvé les résultats suivants:

Paire de suffixes	Nombre de mots trouvés en français	Nombre de mots trouvés en italien
-ité/-ità	3804	4973
-teur/-tore	3994	3729
-tion/-zione	3236	5571

Figure 4

Les chiffres de la figure 4 représentent le nombre de mots (type) et non pas le nombre d'occurrences (token).

Nous avons ensuite, à partir de ces résultats, évalué le nombre de mots dont la base était parfaitement cognate, puis tenté quelques expériences d'application de changements morphographémiques internes.

Dans le tableau ci-dessous, nous présentons le résultat de cette première étape, à savoir le nombre de mots dont la base est parfaitement cognate.

Paire de suffixes	Nombre de mots dont la base est parfaitement cognat
-ité/-ità	405
-teur/-tore	248
-tion/-zione	162

Figure 5

Bien sûr, une vérification manuelle serait nécessaire pour éviter des écueils dus notamment au corpus, comme nous l'avons évoqué plus haut. Cependant, la comparaison de ces deux listes de mots et l'extraction des mots cognats permet d'exclure un grand nombre de mots qui seraient erronés.

Nous avons ensuite tenté d'appliquer automatiquement les changements morphographémiques internes pour évaluer, sur la base de notre nouveau corpus le nombre de mots pouvant être formalisés grâce à ces règles.

Nous résumons dans la figure 6 nos résultats :

	-ité/-ità	-teur/-tore	-tion/-zione
changement 'e'→ 'é'	110	69	59
changement 'tt'→'ct'	17	53	6
changement 'ss'→ 'x'	17	2	4

Figure 6

Evidemment, ces changements chiffrés ne reflètent pas toutes les possibilités. Non seulement, comme nous l'avions dit plus haut, chiffrer le potentiel générateur de ces changements nécessiterait de les appliquer à de nombreuses paires de mots avec plusieurs suffixes différents, mais, dans les exemples ci-dessus, un seul changement par mot et par type de suffixe a été testé. La combinaison d'un ou de plusieurs changements pourrait amener d'autres résultats.

Cette expérience effectuée à l'aide de Webaffix nous a permis d'avoir accès à un grand nombre de néologismes, qui par définition ne se trouvent pas dans les dictionnaires. De plus, il a confirmé le pouvoir générateur de ces néologismes dont Namer (2001) parlait dans son expérience sur GédéFrIt. Il ouvre la voie à une extension des lexiques de manière plus ou moins automatisée. L'accès à des corpus constamment mis à jour permet également d'avoir accès à un grand nombre de néologismes dans des domaines scientifiques notamment. Et ces domaines, de par leur internationalisation, ont une tendance à uniformiser leur vocabulaire sur des bases communes, le plus souvent latines ou grecques (Walter, 2001), ce qui offrirait encore plus de ressources pour un lexique multilingue.

3.3. Evaluation de la place de stockage

Les concepteurs du projet Polylex avançaient dans leur explication que l'un des principaux arguments pour gérer la morphologique dans un lexique multilingue était la réduction de la place de stockage, permettant une plus grande robustesse informatique de l'outil, par rapport à des systèmes classiques monolingues.

Nous abordons ici cette problématique, en faisant une distinction entre informations grammaticales et informations lexicales. En effet, nous l'avons dit, une approche polymorphique du lexique implique une partition entre les entrées lexicales et les informations généralisables qui

seront déclarées une seule fois. Dans notre lexique décrit au chapitre 6, nous distinguons donc les informations morphosyntaxiques des autres informations.

3.3.1. Les informations morpho-syntaxiques

Dans notre lexique, les informations morpho-syntaxiques ont l'avantage d'être déclarées une seule fois quand celles-ci sont communes aux langues en présence.

Ainsi, les informations catégorielles (cat = nom) et les informations de genres (genre=f/m) ne sont déclarées qu'une seule fois, quelque soit le nombre de langues décrites dans le lexique. Les informations de concaténation des préfixes ne doivent pas non plus être répétées, et cela grâce au fait que les opérations constructionnelles des langues romanes sont identiques (concaténation du préfixe à la forme de base). En revanche, les déclarations de concaténation de la base et du suffixe sont dépendantes des langues, étant donné que c'est cette construction qui réalise la forme de surface du mot propre à chaque langue.

3.3.2. Les informations lexicales

Les informations lexicales sont également, le plus possible, déclarées une fois. C'est notamment le cas des bases communes qui n'apparaissent qu'une fois et diminuent grandement la place de stockage. De plus, toutes les informations liées à cette base ne seront déclarées qu'une fois lorsqu'elles sont communes aux deux langues. De même, nous avons individualisé un suffixe, '*im-*' qui semble être parfaitement cognat dans plusieurs langues romanes. Il est donc également décrit une seule fois, tout comme certains types sémantiques.

Ces réflexions ne sont pas basées sur des données chiffrées mais elles mettent en évidence le nombre d'informations qui sont déclarées une fois plutôt que plusieurs fois. Cet avantage doit être relativisé par le fait que certaines déclarations sont nécessaires dans un cadre multilingue alors qu'elles seraient inutiles dans un lexique monolingue (comme la décomposition des mots en base + suffixe). A ce stade, il est donc difficile d'affirmer que la place de stockage est fortement diminuée dans un tel lexique. Cependant, si l'on considère l'éventualité d'extension à une troisième langue, (comme décrit dans la section 2 de ce chapitre) on peut s'autoriser à penser que la place de stockage d'un lexique multilingue pour trois langues serait bien moindre que pour trois lexiques pris séparément.

4. Conclusion

Dans ce chapitre, nous avons tenté d'exposer des arguments chiffrés qui motivent le bien fondé d'un tel lexique. Pour un nombre restreint de règles de formation de mots, nous avons trouvé un nombre assez important de mots à base cognate permettant une formalisation de ce type. Les descriptions linguistiques n'apparaissent alors qu'une fois, permettant une plus grande cohérence des lexiques des langues en présence. Les résultats concernant les changements morphographémiques internes peuvent paraître moins concluants, si l'on compare la place de stockage qu'ils représentent avec leur pouvoir générateur. Bien sûr, des tests à plus large échelle et sur un plus grand nombre de règles permettraient de prouver plus avant la validité d'un tel lexique. On peut toutefois considérer de manière générale qu'une règle tire sa raison d'être du fait qu'elle évite la redondance, ne serait-ce qu'une fois, d'une information ou d'un fait linguistique.

L'exploitation du pouvoir générateur de certains suffixes, comme l'ont montré les résultats de Webaffix, semble avoir un attrait pour d'éventuels développements futurs. Nous aborderons ces différents développements dans la conclusion.

Conclusion et perspectives d'avenir

Dans ce mémoire, nous avons tenté de tracer les grandes lignes de l'exploitation des liens morphologiques multilingues. Nous avons, sur la plan théorique, souligné la proximité formelle des langues, et sur un plan pratique, quantifié cette proximité entre ces deux langues. Nous avons ensuite proposé, dans le cadre d'un lexique d'héritage, une formalisation de ces liens. De ce travail, nous avons pu tirer quelques conclusions. Premièrement, le lien formel entre les deux langues romanes qui constituaient notre sujet d'étude a été prouvé, tant sur le plan théorique que pratique. De plus, la possibilité de formaliser dans une même structure ces liens pour pouvoir en tirer le meilleur parti en terme d'économie semble plus ou moins atteint. Un bémol concerne tout de même la difficulté pour le linguiste informaticien d'implémenter un tel lexique. Il doit en effet bénéficier d'importantes ressources lexicales et de connaissances linguistiques dans les langues concernées pour pouvoir extraire les mots cognats et semi-cognats. Mais toutes les expériences décrites dans ce travail ne constituent que des tentatives de formalisation et d'ébauche de lignes directrices qui doivent nous permettre de continuer à étudier, sous d'autres angles, les liens morphologiques multilingues.

Nous abordons à présent différents développements futurs qui pourraient être envisagés dans le domaine des lexiques multilingues.

1. Extension du lexique

Rappelons-le, le lexique multilingue que nous avons construit n'était pas une fin en soi. Cependant, il nous a permis de montrer que certaines régularités multilingues étaient facilement formalisables, à moindre coût et pour une place de stockage réduite. De plus, les expériences d'évaluation quantitative, particulièrement les recherches sur le web, ainsi que les conclusions de l'étude de Fiammetta Namer (2001) sur la potentialité dérivative de certains suffixes, nous montrent l'utilité d'un tel lexique. En effet, les néologismes et les dérivés dans un sens large sont un véritable défi pour le TALN. La réalisation de ressources lexicales exhaustives et cohérentes est souvent coûteuse. Dans ce cadre là, l'exploitation des liens morphologiques multilingues pour la création et l'acquisition d'un lexique pourrait se révéler particulièrement prometteuse.

Cependant, pour pouvoir gérer en parallèle plusieurs langues, des études poussées sur le parallélisme de leur morphologie seraient nécessaires. Des recherches devraient donc être

menées sur un plus grand nombre de suffixes au pouvoir générateur. L'identification des processus et des régularités reste un vaste sujet d'étude.

La deuxième voie d'extension est sans conteste l'ajout d'une ou de plusieurs autres langues formellement apparentées. Les figures présentées aux pages 114-115 nous autorisent à penser que des liens formels existent entre de nombreuses langues romanes, et que leur formalisation semble réalisable, comme nous l'avons suggéré dans le chapitre 6 lors de l'extension du lexique pour l'espagnol.

Mais dans ce cas également, les connaissances linguistiques et les ressources lexicales de toutes les langues restent un pré-requis important pour implémenter un tel lexique.

2. Utiliser un autre formalisme

Le formalisme utilisé dans le lexique décrit dans le chapitre 5, les lexiques d'héritage, a de nombreux avantages et présente des possibilités très intéressantes. Cependant, avant d'entamer une quelconque extension de notre lexique (sur le plan quantitatif ou dans l'ajout d'une langue), des expériences similaires mériteraient d'être effectuées avec d'autres formalismes basés sur l'approche polymorphique, comme les automates à deux niveaux, pour voir quel formalisme convient mieux au traitement de la morphologie multilingue.

3. Application de la morphologie holistique à la morphologie multilingue

Dans le chapitre 2, nous avions évoqué la possibilité d'appliquer la théorie de la morphologie holistique à la morphologie multilingue. Cette théorie, rappelons-le, se fonde essentiellement sur les différences et non sur les similitudes entre les mots pour pouvoir générer d'autres mots. Elle présente l'avantage de se baser essentiellement sur le mot, sans nécessiter un découpage de ceux-ci. Ceci permet donc de ne pas se retrouver en présence d'unité de base abstraite peu motivée linguistiquement.

En revanche, le travail préparatoire à un tel lexique (comme décrit plus haut) reste plus ou moins nécessaire, à la différence que les outils basés sur ces théories sélectionnent eux-mêmes les différences entre deux mots. A notre connaissance, un seul outil a été développé pour mettre

en œuvre la morphologie holistique : le Whole Word Morphologizer[37] . Cet outil permet, sur la base d'un texte étiqueté, d'extraire des paires de mots plus ou moins semblables, et d'en identifier les différences, pour pouvoir mettre au point des stratégies (Neuvel, 2000.) de formation des mots.

Un tel système pourrait nous permettre d'acquérir automatiquement des paires de mots bilingues et d'implémenter ainsi des stratégies de formation des mots. Les liens formels seraient ainsi identifiés de manière plus aisée et moins coûteuse.

4. Conclusion générale

De nombreuses possibilités d'étude s'ouvrent à la suite de ce travail. Même si nous ne pouvons affirmer que les liens multilingues représentent une solution idéale pour le traitement des lexiques, leur plein potentiel semble prometteur. Non seulement ils restent une option intéressante sur la plan pratique, dont de nombreuses applications pourraient tirer parti, mais ils offrent également un champs d'études théoriques sur la plan linguistique, psycholinguistique, historico-linguistique et informatique de grand intérêt pour des recherches futures.

[37] Neuvel (2000)

Bibliographie:

Page Web :

The Polylex Web page http://www.cogs.susx.ac.uk/lab/nlp/polylex/polylex.html
The DATR Web page http://www.cogs.susx.ac.uk/lab/nlp/datr/datr.html
Encyclopédie Hachette Multimédia, 2001,
http://fr.encyclopedia.yahoo.com/articles/sy/sy_271_p0.html
SAMPA computer readable phonetic alphabet : http://www.phon.ucl.ac.uk/toc/set_eresource.htm

Ouvrage :

Arcaini Enrico (2000), *Italiano e francese, un'analisi comparativa*, Turin, Paravia.

Bouillon Pierrette *et al.* (1998), *Traitement automatique des langues naturelles*, Bruxelles, Editions Duculot.

Bloomsfield L. (1933), *Language,* Holt, Rinehart and Winston.

Caron Jean, (1989) *Précis de psycholinguistique*, coll. Quadrige, Paris, , PUF.

Dardano M. et Trifone P. (1985), *La lingua italiana*, Bologne, Zanichelli.

Fuchs C. *et al,* (1993) *Linguistique et traitement automatique des langues*, Paris, Hachette Livre.

Geysen Raymond (1990) *Dictionnaire des formes analogues en 7 langues avec résumé de grammaire comparée,* Deuxième édition, Paris, Duculot.

Ritchie D. Graeme, *et al,* (1992) *Computational Morphology*, Cambridge, The MIT Press.

Grevisse M. (1980) *Le Bon Usage*, onzième édition, Paris, Duculot.

Lehmann A, et Martin-Berthet F., (1998), *Introduction à la lexicologie*, Paris, Dunod.

Marchand, H., (1969) *The categories and types in the present-day english Word formation*. C.H. Beck.

O'Grady William *et al.*, (1996) , *Contemporary Linguistics, An introduction*, 3ème édition, Londres et New York, Longman.

Siouffi G. et Van Raemdonck D., (1999) *100 fiches pour comprendre la linguistique*, Paris, Bréal.

Sproat Richard, (1992), *Morphology and Computation*, Cambridge, The MIT Press.

Walter Henriette, (1997) *L'aventure des mots français venus d'ailleurs,* Paris, R. Laffont.

Walter Henriette, (2001) *Honni soit qui mal y pense*, Paris, R. Laffont.

Articles :

Burani, Cristina et Laudanna Alessandro, (1992) « Units of Representation for Derived Words in the Lexicon », in *Orthography, Phonology, Morphology and Meaning*. pp. 361-376. R. Frost and L. Katz editors, Elsevier Science Publisher.

Daille B. et al, (2002), « Application of Computational Morphology », in *Many Morphology*, pp. 210-234, Ed. by Paul Bopucher, Cascadilla Press.

De Groot Annette M.B. (1998) *La représentation lexico-sémantique et l'accès lexical chez le bilingue*, in Psychologie française, pp. 297-312, n°43-4.

Roger Evans & Gerald Gazdar (1996) « DATR: A language for lexical knowledge representation » in *Computational Linguistics*, 22.2, pp. 167-216.

Evans Roger, (2001) *Exploiting inheritance in multilingual lexicons*, disponible à l'adresse : http://www.itri.bton.ac.uk/~Roger.Evans/papers/aisb96/slides.html

Hediard Maire, (1989) « Langues voisines, langues faciles ? », in *Analisi comparativa : francese/italiano.* pp. 27-35 A cura di Arcaini, Fourment e Lévy-Mongelli, Padova, Liviana editrice.

Ingria Robert *et al. Lexicon for Natural Language Processing*, manuscrit non publié.

Kameyama M. (1988) « Atomization in grammar sharing » in *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics* , pp. 194-203

Kay M. (1983), « When Meta-Rules are not Meta-Rules », in *Automatic Language Parsing*, pp.94-116 K. Sparck-Jones et Y Wilks (éd), Chischester, Ellis Horwood.

Kilgarriff A. *et al.* (1999) The GREG framework for multilingual valency lexicons. GREG deliverable 2.1, ITRI.

Kondrak Grzegorz (2001), « Identifying Cognates by phonetic and Semantic Similarity », in *NAACL-2001*, pp. 103-110, Pittsburgh, Carnegie Mellon University.

Kraif Olivier (2001), « Exploitation des cognats pour l'alignement » in *TALN* volume 42, pp. 832-864, n°3/2001.

Lavaur Jean-Marc et Font Noëlle, (1998) *Représentation des mots cognats et non cognats en mémoire chez les bilingues français-espagnol.* In Psychologie française, n° 43-4 1998, pp. 330-338, Paris.

Libben Gary et De Almeida Roberto G., (2002) « Is there a morphological parser ? » in *Morphology 2000*, pp. 213-226 Amsterdam, ed. S. Bendjaballah *et al*. editeur, John Benjamin.

Namer F. (2001), « Génération automatique de néologisme bilingues morphologiquement construits en français et en italien », in *TALN 2001*, pp. 281-296.

Namer F. et Dal. G. (2000), « GéDériF : automatic generation and analysis of morphologically constructed lexical ressources », in *Actes de LREC 2000*, pp. 1447-1454 Athènes.

Neuvel Sylvain et Sing R. (sous-presse) , Vive la différence ! What morphology is about. *Folia linguistica*, 35, 3-4

Neuvel Sylvain, (2000). « Whole Word Morpholgizer. Expanding the Word-Based Lexicon : A non stochastic computational Approach ». Presented at the *2nd International Conference on the Mental Lexicon*. Montréal, Canada, Octobre 19, 2000.

Neuvel Sylvain et Sing R, (sous-presse 2) *Quelles unités ?* Troisième Forum International de Morphologie, Lille 2002

Simar *et al*, (1992) « Using cognats to align sentences in bilingual corpora » in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, pp. 67-81

Tanguy L. et Hathout N. (2002) « Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web », in *Taln 2002*. pp. 245-254

Tiberius Carole, (2002) « How to buid a multilingual inheritance-based lexicon » in *LREC 2002*, Proceedings, volume II. pp. 701-708.

Tiberius Carole et Evans Roger, (2000) « Phonological feature based multilingual lexical description », in *Taln 2000*. pp. 347-356

Waksler Rachelle, (2000) « Morphological systems and structure in language production », in *Aspect of language production*, pp. 124-136 edited by Linda Wheeldon, East Sussex, Psychology Press.

Manuel :

Estival Dominique, (1990) *ELU User Manual*, Technical Report, Fondation Dalle Molle,

Annexes

Nous présentons ici en annexe notre lexique d'héritage complet pour le français et l'italien (annexe 1) puis notre lexique d'héritage pour les trois langues (annexe 2), tous les deux accompagnés de la liste des mots générés. Enfin, en annexe 3 se trouve une copie de l'enquête décrite au chapitre 3.

1. Annexe 1 : lexique d'héritage pour le français et l'italien

```
## GRAMMAIRE ELU

# Declare
   Form = form
   Identifier = base

# Restrict Cat Tete
   <* cat> = Cat
   <* syn tete> = Tete

# Sempaths <syn>
# Genpaths <syn>
# Define relation
# Lookup
   noms      <* cat> = n

#Define Mac
accord_Ngenre (<* syn accord genre>)
accord_Nnombre (<* syn accord nombre>)
accord_Npers (<* syn accord pers>)

X_SS(Base,S1,S2)
Base = S1 && x && S2
t_c(Base,S3)
Base = S3 && c
é_e(Base,S5,S6)
Base = S5 && é && S6

#Nlexicon noms

#Class Nom ()
        Singform = <base>
        <* cat> = n
        |
        <form> = Singform && Suff_sing_it
        !accord_Nnombre(sg)
        <lang> = it
        |
        !accord_Nnombre(pl)
        <form> = Singform && Suff_plu_it
        <lang> = it
        |
        !accord_Nnombre(sg)
        <form> = Singform
        <lang> = fr
        |
        !accord_Nnombre(pl)
        <form> = Singform && Suff_plu_fr
        !accord_Nnombre(pl)
        <lang> = fr
```

```
#Class OPP(Nom)
        Form_lem = <base>
        Preff = vide
        |
        Singform = Form_lem
        |
        Singform = Preff && Form_lem
        Preff = ~vide


#Class Conca2(OPP)
        Form_base_it = <base>
        |
        Form_lem = Form_base_it && Suff_it
        <lang> = it
        |
        Form_lem = <base> && Suff_fr
        <lang> = fr

#Class PLU_EI_IT( )
        Suff_plu_it = i
        Suff_sing_it = e

#Class PLU_A_IT ()
        Suff_plu_it = ""
        Suff_sing_it = ""

#Class PLU_S_FR()
        Suff_plu_fr = s
```

```
#Class SuffTORE (PLU_S_FR PLU_EI_IT Conca2)
        |
        Suff_it = tric
        Suff_fr = trice
        !accord_Ngenre (f)
        |
        Suff_it = tor
        Suff_fr = teur
        !accord_Ngenre (m)

#Class SuffITEITA (PLU_A_IT PLU_S_FR Conca2)
!accord_Ngenre (f)
Suff_it = ità
Suff_fr = ité

#Class SuffTION(PLU_EI_IT PLU_S_FR Conca2)
!accord_Ngenre (f)
Suff_it = azion
Suff_fr = ation

#Class SuffTETA (PLU_A_IT PLU_S_FR Conca2)
!accord_Ngenre (f)
Suff_it = tà
Suff_fr = té


#Class IM()
        |
        Preff = im
```

```
#Class IL()
        |
        Preff = il

#Class X_SS ()
!X_SS(<base>,S1,S2)
Form_base_it = S1 && ss && S2

#Class E_E ()
!é_e(<base>,S5,S6)
Form_base_it = S5 && e && S6

#Class T_C()
!t_c(<base>,S3)
Form_base_it = S3 && t

#Word chaise (PLU_S_FR Nom)
<* sem> = artefact
<lang> = fr
!accord_Ngenre (f)
#Word ac (T_C SuffTORE)
<* sem> = person
#Word traduc(T_C SuffTORE)
<* sem> = person
#Word ladron(PLU_EI_IT Nom)
<* sem> = person
<lang> = it
!accord_Ngenre (m)
#Word conduc (T_C SuffTORE)
<* sem> = person
#Word taglia (SuffTORE)
<lang> = it
<* sem> = person
#Word tartinabil(SuffITEITA)
<* sem> = etat
<lang> = fr
#Word possibil(IM SuffITEITA)
<* sem> = etat
#Word liber(SuffTETA)
<* sem> = etat
#Word matern(SuffITEITA)
<* sem> = etat
#Word complex(X_SS SuffITEITA)
<* sem> = etat
#Word légal(E_E IL SuffITEITA)
<* sem> = etat
#Word inform(SuffTION)
<* sem> = abstrait
#Word libér(E_E SuffTION)
<* sem> = etat

**************************
```

Liste de mots générés par le programme:

"acteur"
"acteurs"
"actrice"
"actrices"
"attore"
"attori"

"attrice"
"attrici"
"chaise"
"chaises"
"complessità"
"complexité"
"complexités"
"conducteur"
"conducteurs"
"conductrice"
"conductrices"
"conduttore"
"conduttori"
"conduttrice"
"conduttrici"
"illegalità"
"illégalité"
"illégalités"
"impossibilità"
"impossibilité"
"impossibilités"
"information"
"informations"
"informazione"
"informazioni"
"ladrone"
"ladroni"
"legalità"
"liberazione"
"liberazioni"
"libertà"
"liberté"
"libertés"
"libération"
"libérations"
"légalité"
"légalités"
"maternità"
"maternité"
"maternités"
"possibilità"
"possibilité"
"possibilités"
"tagliatore"
"tagliatori"
"tagliatrice"
"tagliatrici"
"tartinabilité"
"tartinabilités"
"traducteur"
"traducteurs"
"traductrice"
"traductrices"
"traduttore"
"traduttori"
"traduttrice"
"traduttrici"

2. Annexe 2 : lexique d'héritage avec les ajouts pour l'espagnol

```
## GRAMMAIRE ELU

# Declare
   Form = form
   Identifier = base

# Restrict Cat Tete
   <* cat> = Cat
   <* syn tete> = Tete

# Sempaths <syn>
# Genpaths <syn>
# Define relation

# Lookup
   noms      <* cat> = n

#Define Mac
accord_Ngenre (<* syn accord genre>)
accord_Nnombre (<* syn accord nombre>)
accord_Npers (<* syn accord pers>)
X_SS(Base,S1,S2)
Base = S1 && x && S2
t_c(Base,S3)
Base = S3 && c
é_e(Base,S5,S6)
Base = S5 && é && S6
```

```
#Nlexicon noms

#Class Nom ()
        Singform = <base>
        <* cat> = n
        |
        <form> = Singform && Suff_sing_it
        !accord_Nnombre(sg)
        <lang> = it
        |
        !accord_Nnombre(pl)
        <form> = Singform && Suff_plu_it
        <lang> = it
        |
        !accord_Nnombre(sg)
        <form> = Singform
        <lang> = fr/es
        |
        !accord_Nnombre(pl)
        <form> = Singform && Suff_plu_fr
        <lang> = fr
        |
        !accord_Nnombre(pl)
        <form> = Singform && Suff_plu_es
        <lang> = es
```

```
#Class OPP(Nom)
        Form_lem = <base>
        Preff = vide
        |
        Singform = Form_lem
        |
        Singform = Preff && Form_lem
        Preff = ~vide

#Class Conca3(OPP)
        Form_base_it = <base>
        Form_base_es = <base>
        |
        Form_lem = Form_base_it && Suff_it
        <lang> = it
        |
        Form_lem = <base> && Suff_fr
        <lang> = fr
        |
        Form_lem = Form_base_es && Suff_es
        <lang> = es

#Class Conca2(OPP)
        Form_base_it = <base>
        |
        Form_lem = Form_base_it && Suff_it
        <lang> = it
        |
        Form_lem = <base> && Suff_fr
        <lang> = fr
```

```
#Class PLU_EI_IT( )
        Suff_plu_it = i
        Suff_sing_it = e

#Class PLU_A_IT ()
        Suff_plu_it = ""
        Suff_sing_it = ""

#Class PLU_S_FR()
        Suff_plu_fr = s

#Class PLU_ES_ES()
        Suff_plu_es = es

#Class PLU_S_ES()
        |
        Suff_plu_es = s
        !accord_Ngenre (f)
        |
        Suff_plu_es = es
        !accord_Ngenre (m)
```

```
#Class SuffTORE3 (PLU_S_FR PLU_EI_IT PLU_S_ES Conca3)
        |
        Suff_it = tric
        Suff_fr = trice
        Suff_es = tora
        !accord_Ngenre (f)
        |
        Suff_it = tor
        Suff_fr = teur
        Suff_es = tor
        !accord_Ngenre (m)

#Class SuffITEITAIDAD (PLU_A_IT PLU_S_FR PLU_ES_ES
Conca3)
!accord_Ngenre (f)
Suff_it = ità
Suff_fr = ité
Suff_es = idad

#Class IM()
        |
        Preff = im

#Class IL()
        |
        Preff = il
        <lang> = it/fr
        |
        Preff = i
        <lang> = es
```

```
#Class X_SS ()
!X_SS(<base>,S1,S2)
Form_base_it = S1 && ss && S2

#Class E_E ()
!é_e(<base>,S5,S6)
Form_base_it = S5 && e && S6
Form_base_es = S5 && e && S6

#Class T_C()
!t_c(<base>,S3)
Form_base_it = S3 && t

#Word chaise (PLU_S_FR Nom)
<* sem> = artefact
<lang> = fr
!accord_Ngenre (f)

#Word ac (T_C SuffTORE3)
<* sem> = person
!accord_Ngenre (m)
#Word traduc(T_C SuffTORE3)
<* sem> = person
!accord_Ngenre (m)
#Word conduc (T_C SuffTORE3)
<* sem> = person
#Word matern(SuffITEITAIDAD)
<* sem> = etat
#Word un(SuffITEITAIDAD)
<* sem> = etat
#Word complex(X_SS SuffITEITAIDAD)
<* sem> = etat
```

#Word légal(E_E IL SuffITEITAIDAD)
<* sem> = etat

****

Liste de mots générés par le programme:

"acteur"
"acteurs"
"actor"
"actores"
"attore"
"attori"
"chaise"
"chaises"
"conducteur"
"conducteurs"
"conductor"
"conductora"
"conductoras"
"conductores"
"conductrice"
"conductrices"
"conduttore"
"conduttori"
"conduttrice"
"conduttrici"
"ilegalidad"
"ilegalidades"
"illegalità"
"illégalité"
"illégalités"
"legalidad"
"legalidades"
"legalità"
"légalité"
"légalités"
"maternidad"
"maternidades"
"maternità"
"maternité"
"maternités"
"traducteur"
"traducteurs"
"traductor"
"traductora"
"traductoras"
"traductores"
"traductrice"
"traductrices"
"traduttore"
"traduttori"
"traduttrice"
"traduttrici"
"unidad"
"unidades"
"unità"
"unité"
"unités"

3. Annexe 3 : enquête

Expérience sur l'inférence des cognates Italiens-Français.

Données personnelles:

Langue maternelle: _____

Autres langues: (merci d'inscrire toutes les langues que vous connaissez, même passivement)

_____
_____
_____
_____
_____

Avez-vous appris le latin ?: ☐ oui ☐ non

Partie 1: Traduction à vue

Dans cette première partie, nous allons vous présenter une liste de mots italiens. Il s'agira pour vous de deviner leur signification en inscrivant leur traduction en français dans l'espace prévu à cet effet. Nous vous demandons de remplir cette liste le plus rapidement possible. Traduisez les mots un à un, dans l'ordre dans lequel ils apparaissent. Si vous ne trouvez aucune traduction, laissez l'espace vide. Les mots sont regroupés en fonction de leur catégorie grammaticale.

A) Les noms:

Mot en italien	Mot en français
abusi	
osservatore	
trasporto	
affari	
tribunale	
emendamenti	
disposti	
tappa	
cambiamenti	
conferenze	
congiunte	
tempo	
consigli	
contributi	
taglia	
trasparenza	
richiesta	
respinto	
disposizioni	
vicino	
effetti	
esperienze	
volta	
ricerca	
adattamenti	
responsabilità	
revisione	

comunicazioni	
tenore	
capi	
trattato	
tutela	
droga	
vigore	
svolta	
accuse	
volontà	
voto	
direttore	

B) Les adjectifs

Mot en italien	Mot en français
assistita	
applicabili	
disposto	
arrivate	
dissennato	
intesi	
operative	
diverso	
equilibrato	
spente	
estero	
esperto	
successivi	
illecito	
illimitato	

invitati	
medesime	
doganale	
monetaria	
espresso	
necessaria	
globale	
netta	
opportune	
pubblico	
seguenti	
qualificato	
ragionevole	
raccolto	
risoluti	
grave	
rispettive	
suddetti	
breve	

Les verbes (attention, il y a des infinitifs, des participes et des formes conjuguées)

Mot en italien	Mot en français
deputato	
praticare	
scambio	
creare	
precisare	
incarica	
influenza	
sedare	

pronuncia	
desiderino	
delineare	
scopo	
può	
pregiudicare	
progetto	
segnare	
pubblico	
scontro	
presentare	
processo	
pressare	
deliberare	
determinati	
dare	
seguire	
punto	
trasferire	
trasportare	
processi	
trafficare	

Les adverbes

Mot en italien	Mot en français
abbondantemente	
sopra	
altrimenti	
come	
copiosamente	

così	
finalemente	
immediatamente	
sempre	
nervosamente	
perfettamente	
quando	
rapidamente	
appena	
altrettanto	
proprio	
vigorosamente	

Partie II:

Dans cette partie, nous allons vous présenter une liste de couple de mots, contenant chacun un mot français et sa traduction en italien. Nous vous présenterons ensuite une liste de mot français et se sera à vous de traduire ces mots en italien.

posizione - position
amministrazione - administration
associazione - association
coerenza – cohérence
complesso – complexe
competenza - compétence
settore – secteur
conduttore - conducteur
attrice - actrice
conformità – conformité
continuità - continuité
credibilità - crédibilité

A vous de jouer !!!!!

Mot en français	Mot en italien
collaboration	
communication	
consultation	
condition	
admission	
criminalité	
liberté	
complexité	
présidence	
traducteur	
collaboratrice	
possibilité	

Mille merci pour votre collaboration !!!