Traduction de règles de construction des mots pour résoudre les problèmes d'incomplétude lexicale en traduction automatique Étude de cas

Bruno Cartoni

TIM/ISSCO – ETI – Université de Genève
40 bd du Pont-d'Arve, CH-1205 Genève
bruno.cartoni@eti.unige.ch

Mots-clefs – Keywords

Traduction automatique, morphologie constructionnelle, incomplétude lexicale

Machine Translation, constructional morphology, lexical incompleteness

Résumé – Abstract

Cet article propose d'exploiter les similitudes constructionnelles de deux langues morphologiquement proches (le français et l'italien), pour créer des règles de construction des mots capables de déconstruire un néologisme construit de la langue source et générer de manière similaire un néologisme construit dans la langue cible. Nous commençons par présenter diverses motivations à cette méthode, puis détaillons une expérience pour laquelle plusieurs règles de transfert ont été créées et appliquées à un ensemble de néologismes construits.

This paper presents a method which aims at exploiting constructional similarities between two morphologically-related languages (French an Italian), in order to create word-construction rules that can disassemble a constructed neologism and create in a similar way a constructed neologism into the target language. We present the main motivation for this method and describe an experiment for which transfer rules has been developed and applied to a group of constructed neologism.

1 Introduction

Le présent article s'inscrit dans un travail de recherche qui vise à exploiter les propriétés constructionnelles des mots néologiques construits pour résoudre l'incomplétude lexicale en traduction automatique (ci-après TA). Nous entendons exploiter ces propriétés à la fois pour l'analyse des mots inconnus et pour la génération d'une traduction possible de ces mots.

Cependant, une telle exploitation doit se faire au travers de moyens simples pour garantir la « portabilité » et une certaine efficacité indispensables à un tel système.

Deux hypothèses guident ce travail : premièrement, nous émettons l'idée que les constructions des néologismes sont suffisamment transparentes et peu ambiguës pour être analysées par des moyens simples nécessitant peu de ressources. Deuxièmement, nous pensons que les processus morphologiques sont suffisamment proches d'une langue à l'autre (du moins dans des langues de même famille) pour permettre d'envisager la traduction d'un processus de construction par un autre. Pour ce faire, nous proposons l'utilisation de règles bilingues de construction des mots. Leur élaboration nécessite l'étude approfondie des processus morphologiques à traiter. Dans cette optique, nous décrivons ci-après les différentes motivations qui sous-tendent cette approche (section 2), puis nous proposons un modèle de traitement des mots construits inconnus, que nous appliquons ensuite à une expérience (section 3) dans laquelle nous mettons au point un ensemble de règles bilingues de construction des mots exploitables en TA.

2 Motivation

2.1 Les lexiques informatisés et la néologie formelle

De nombreuses applications de TAL reposent sur un lexique qui contient les mots de la langue traitée (ou un sous-ensemble de ces mots), ainsi que certaines informations associées à ces mots (Sproat, 1992). La couverture et la qualité du lexique dépendent de l'application pour laquelle le lexique est élaboré (Arnold, *et al*., 1994). Or, l'absence d'un mot dans le lexique cause un certain nombre de problèmes, notamment en TA (Gdaniec, *et al.*, 2001). Les concepteurs de système de TA recourent alors à différentes stratégies pour faire face à l'incomplétude lexicale (comme la simple « transposition » du mot inconnu, sans le traduire). D'autres systèmes, comme Systran (Whitelock *et al.*, 1995), tentent au moins de deviner la catégorie grammaticale du mot inconnu en se basant sur sa terminaison. Ceci permet d'obtenir une analyse syntaxique plus correcte, même si la traduction n'en est pas plus réussie.

Les mots inconnus des lexiques informatisés posent donc un vrai problème. Parmi ces mots inconnus, on trouve un nombre important de noms propres. Toutefois, la majeure partie des mots inconnus a pour origine la créativité lexicale des langues. Celle-ci relève de plusieurs procédés, comme celui de la *néologie formelle* qui consiste en la création de nouveaux mots à partir de matériaux lexicaux préexistant (Gaudin *et al.*, 2000). D'un point de vue quantitatif, la néologie formelle est souvent décrite comme la plus productive et la plus utilisée. Ainsi, par exemple, Cabré (2002) constate que dans la presse catalane, ce procédé est à l'origine de 75 % des néologismes. Elle note également que au sein de la néologie formelle, la préfixation représente la ressource néologique la plus importante (32,3 %), suivie de près par la suffixation (24,7 %).

2.2 Le traitement de la morphologie constructionnelle

La préfixation et la suffixation, source principale de la néologie formelle, font partie des sujets d'étude de la morphologie constructionnelle, qui étudie la formation des mots construits, c'est-à-dire les mots dont le sens est prédictible et entièrement compositionnel par rapport à leur structure interne (Corbin, 1987). La formalisation d'un processus constructionnel passe par l'écriture de règles de construction des mots (Corbin, *ibid.*) (ci-après RCM) soumises à des contraintes permettant de décrire le processus concerné.

La morphologie constructionnelle reste souvent décrite comme irrégulière, et donc difficilement généralisable, et partant non exploitable en TAL, même si un récent article (Dal 2002) remet en cause cette réputation d'irrégularité. Comme le soulignent Dal, *et al*., (sous presse), le TAL s'intéresse peu à la morphologie constructionnelle, sans doute « parce qu'il voit [dans les données constructionnelles] des phénomènes imprévisibles qui échappent en grande partie au calcul ». Dans de nombreux systèmes de traitement des langues, les informations constructionnelles sont utilisées à la seule fin d'étiquetage morphosyntaxique (Dal, *et al., ibid.*). En TA, l'utilisation des propriétés de la morphologie constructionnelle est rare, sans doute parce que ces applications nécessitent également une opération de génération, qui semble plus difficile pour les mots construits inconnus (Gdaniec, *et al.*, 2001).

Parce que la morphologie constructionnelle est souvent à l'origine de la formation de nouveaux mots, certaines recherches se sont déjà penchées sur le traitement automatique de la néologie formelle, principalement dans des buts de génération (voir par exemple Namer *et al*, 2000). Pour le présent travail, nous partons de l'hypothèse que les néologismes formels doivent être sémantiquement transparents pour être compris des locuteurs. Ainsi, la construction de néologismes ne devrait pas être aussi imprévisible et irrégulière que la morphologie constructionnelle en général, ce qui devrait faciliter le traitement informatique.

Avant de décrire une expérience pratique utilisant certaines propriétés de la morphologie constructionnelle, nous étudions les possibilités de transfert de ces RCM d'une langue à une autre, de façon à pouvoir exploiter ces propriétés dans le traitement des néologismes en TA.

2.3 La traduction des règles de construction des mots

La deuxième hypothèse qui sous-tend notre approche concerne l'exploitation des similitudes morphosémantiques entre deux langues proches pour inférer (ou deviner) l'équivalent d'un néologisme dans l'autre langue. Nous partons du principe que la proximité des lexiques de deux langues proches d'un point de vue morphologique pourrait être exploitée en TA. Ainsi, deux langues morphologiquement proches possèdent des similitudes au niveau des procédés de construction morphologique. Certaines études ont déjà montré la relative proximité entre le français et l'italien (Namer, 2001). Cependant, même si ces deux langues possèdent le même « fonds lexical commun », des divergences se sont développées avec le temps, donnant notamment lieu au phénomène des faux amis. La production néologique semble néanmoins s'effectuer d'une manière plus régulière et avec une relative similitude dans les différentes langues, particulièrement dans les domaines scientifiques et techniques où la mondialisation et les échanges internationaux sont importants, influençant par là même leur vocabulaire. Dans ces domaines, les emprunts ou calques de constructions morphologiques s'effectuent en

mettant la nouvelle unité lexicale à la « sauce » morphologique de la langue empruntante (Gaudin *et al.*, 2000).

Cette hypothèse soulève quelques questions théoriques auxquelles il est difficile d'apporter une réponse franche, au moins à ce stade de nos recherches. Par exemple, même si tout locuteur de deux langues morphologiquement proches sent très bien que le préfixe de réitération italien *ri–* peut être traduit par *re–* en français (comme l'attestent les paires *ricominciare*$_{it}$/*recommencer*$_{fr}$, *rifare*$_{it}$/*refaire*$_{fr}$, etc.), peut-on pour autant considérer que l'utilisation de ces deux procédés est régulière, et donc exploitable en TA ? Et, s'il existe des exceptions, comment les définir pour éviter des analyses ou des générations incorrectes ? Ces questions restent ouvertes, mais l'expérience ci-dessous propose quelques pistes de recherche pour découvrir les régularités morphosémantiques entre deux langues et ainsi extraire des règles bilingues de construction des mots.

3 Expérience

L'exploitation des régularités constructionnelles pourrait donc apporter une solution à certains problèmes d'incomplétude lexicale en TA. Ainsi, dans un cadre plus large, nous proposons un système de transfert d'information morphosémantiques d'une langue à l'autre de façon à générer un néologisme construit en langue cible à partir des informations reconnues dans le néologisme de la langue source.

En présence d'un mot inconnu du lexique d'un système de TA, notre modèle propose donc les étapes suivantes : **(1)** analyse des néologismes selon des RCM ; **(2)** traduction de la base des mots construits analysés en (1) ; **(3)** transfert des informations morphosémantiques ; **(4)** construction d'un néologisme construit grâce à la base traduite en (2) et aux informations provenant de l'étape (3). Ce « transfert » des informations (réalisé en 3) nécessite la mise en correspondance des RCM du français et de l'italien, un peu à l'image des règles de transfert syntaxique de certains systèmes de TA. Dans notre approche, cette mise en correspondance est formalisée par un ensemble de RCM bilingues permettant de décrire différents procédés de construction.

Cependant, même si nos recherches visent, à terme, une exploitation à large échelle dans un système de TA, nous concentrons pour l'instant nos efforts sur des unités lexicales hors contexte. Dans ce même esprit, nous limitons les ressources d'analyse à celles contenues dans un système de TA. Il serait, en effet, peu approprié de proposer des méthodes d'analyse des mots inconnus qui utilisent des procédés gourmands en place et en connaissances, et qui péjorerait ainsi les performances du système de TA en termes de vitesse et de taille. Nous présentons ci-après les différentes étapes de l'élaboration de telles règles, et leurs exploitations *in vivo*.

3.1 Le système dérivationnel choisi

Pour cet article, nous nous sommes limité à l'étude d'un système dérivationnel présent dans nos deux langues de travail, l'italien et le français.

En italien, il existe un préfixe verbal *r(i)–* (parfois sous la forme de *re–,*) qui correspond à la forme syntaxique de « de nouveau » (Dardano, 1978). Ce préfixe est un des plus productif, étant donné qu'il peut potentiellement être associé à tous les verbes (Dardano, *ibid.*). Cette productivité semble particulièrement intéressante pour ce travail. En effet, dans la mesure où ce préfixe peut potentiellement être associé à tous les verbes, il pose un réel problème d'exhaustivité pour les lexiques informatisés. La RCM bilingue qui le traitera peut donc s'avérer très rentable en termes de gain en qualité de traduction.

Dardano (*ibid.*) associe ce procédé de préfixation à celui du préfixe *de–*, en émettant l'hypothèse qu'une action qui doit être **re**-faite, a dû être auparavant **dé**-faite. Il propose alors une série paradigmatique comme *stabilizzare* → ***de****stabilizzare* → ***ri****stabilizzare*. Il note également que cette série possède la capacité de produire une série de noms déverbaux (comme dans *stabilizzazione* → *destabilizzazione* → *ristabilizzazione*).

Pour le français, un système dérivationnel équivalent semble exister. Le préfixe *re–* (et ses allomorphes, *r–* , et *ré–*) signifie également la répétition de l'action décrite par la base verbale (Rey-Debove, 2004). Le lien souligné pour l'italien par Dardano (1978) entre les préfixes *r(i)–* et *de–* est également présent dans le système morphologique du français (Huot, 2001), avec le préfixe *dé–* (qui prend aussi les formes dés– et des–), comme l'attestent les paires ***dé****faire*/***re****faire*, ou ***dé****coudre*/***re****coudre*. Enfin, les séries de noms déverbaux cités par Dardano (*ibid*) (*stabilizzazione* → *destabilizzazione* → *ristabilizzazione*) se retrouvent en français : *stabilisation* → *déstabilisation* →*restabilisation*.

3.2 Élaboration des règles de construction des mots bilingues

La linguistique informatique s'inspire largement de la linguistique descriptive ou de la linguistique théorique, mais elle doit bien souvent faire des compromis liés aux problèmes d'implantation (Tzoukermann *et al*., 1997). Dans notre cas, ce sont les contraintes liées à la portabilité et à la rapidité de traitement des systèmes de TA qui guident notre approche. Ainsi, nous implémentons des RCM bilingues basées uniquement sur les chaînes de caractères et à l'aide d'expressions régulières.

La RCM bilingue présentée dans la Figure 1 porte sur le traitement de verbes préfixés en *ri–* en italien et leur traduction en français. Les contraintes portent sur les bases, qui doivent appartenir au lexique de référence (L_{it} et L_{fr}) et qui doivent être la traduction l'une de l'autre. Une règle identique a été créée pour le préfixe *de–*. De plus, étant donné que nous ne travaillons que sur la forme orthographique des mots, les règles sont répétées pour chaque allomorphe (*r–*, *re–*, …). Notons également que faute de place, nous ne mentionnons pas dans les règles les changements morphographémiques résultant de la concaténation de la base et du préfixe qui ont été définies pour la partie française.

$$\text{IT} \begin{pmatrix} \text{X/VERBE} \Rightarrow \textbf{ri}\text{/PREF [Y/VERBE]} \\ \text{Y/VERBE} \in L_{it} \end{pmatrix} = \text{FR} \begin{pmatrix} \text{X'/VERBE} \Rightarrow \textbf{re}\text{/PREF [Y'/VERBE]} \\ \text{Y/VERBE} \in L_{fr} \end{pmatrix}$$

où : Y/VERBE = Y'/VERBE (équivalent de traduction)

Figure 1: RCM bilingue pour le préfixe italien *ri–*

La figure 2 présente la RCM bilingue des noms déverbaux préfixés par *ri–*, procédé étudié par Dardano (1978). Pour contraindre la règle, nous choisissons les suffixes les plus fréquents des noms déverbaux (–zione, –mento, –aggio). Une règle similaire est également construite pour le préfixe *de–* suivi d'un nom.

IT (X/NOM => **ri**/PREF [Y/NOM]
Y/NOM = [a-z]*zione/i | [a-z]*mento/i | [a-z]*aggio/i
Y/NOM ∈ L_{it}) = **FR** (X'/NOM => **re**/PREF [Y'/NOM]
Y/NOM = [a-z]*tion/s | [a-z]*ment/s | [a-z]*age/s
Y'/NOM ∈ L_{fr})

où : Y/NOM = Y'/NOM (équivalent de traduction)

Figure 2 : RCM bilingue pour le préfixe italien *ri–* sur une base nominale

Si l'on ne peut, à ce stade, affirmer avec certitude que la préfixation français en *re-* est la traduction de la construction italienne en *ri–*, il n'en reste pas moins que notre expérience donne des résultats encourageants, que nous présentons dans la suite.

3.3 Le corpus de mots inconnus

La presse écrite étant un terrain particulièrement fertile pour la production néologique (Pruvost *et al*., 2003), nous avons utilisé comme corpus textuel un recueil de textes italiens publié par ELRA, (corpus MLCC, 1997), contenant les éditions du mois de février 1992 du quotidien italien *Il Sole 24 ore*[1]. Ce corpus contient 1,88 millions d'occurrences.

Dans l'étude empirique des phénomènes de néologie, la découverte de nouveaux mots ne peut se faire qu'à partir d'un lexique de référence, tant la notion de nouveauté est difficile à définir. Pour obtenir une liste de néologismes, nous avons donc confronté le corpus à un lexique de référence qui a joué le rôle de corpus d'exclusion (Gaudin *et al.*, 2000). Ce lexique est celui d'un analyseur morphosyntaxique, qui entre dans un processus complet d'étiquetage morphosyntaxique (*Tatoo*[2]). De cette première confrontation, nous obtenons une liste de 225 075 unités lexicales inconnues de notre lexique de référence, ce qui correspond à environ 12 % du nombre total d'occurrences. Evidemment, ces mots inconnus ne sont pas tous des néologismes. Nous affinons notre liste en excluant les noms propres, qui représentent une part importante de l'incomplétude lexicale en TA (cf. plus haut). Pour ce faire, nous appliquons une simple routine basée sur les majuscules (à l'instar de Maurel, 2004), et nous obtenons un nombre total de mots inconnus potentiellement néologiques de 90 260 occurrences (environ 4,8 % du corpus).

3.4 Résultats et évaluations

Nous avons donc appliqué à notre corpus de mots inconnus les RCM bilingues proposées plus haut, à la fois pour analyser les mots inconnus construits selon le système dérivationnel

---

1 http ://www.ilsole24ore.com/

2 The ISSCO Tagger Tool : http://issco-www.unige.ch/staff/robert/tatoo/tatoo.html

choisi, et pour générer la traduction en français de ces néologismes construits. Nous présentons ci-dessous les résultats des étapes d'analyse et de génération. Notons que la traduction des bases a été effectuée par le système de traduction automatique Systransoft©[3].

3.4.1 Analyse des mots préfixés

Cette première étape consiste analyser les mots construits en utilisant la partie italienne de la RCM bilingue, en recensant semi-automatiquement les dérivés dans notre corpus de mots inconnus. La partie automatique consiste à rechercher les mots qui correspondent aux règles établies plus haut, avec la contrainte que la base soit présente dans le lexique de référence. De cette extraction automatique, nous obtenons en sortie une paire formée du dérivé néologique et de sa base, elle-même accompagnée de son analyse dans le lexique (`"riorganizzare" = "organizzare" Verb […]`). Etant donné que cette automatisation peut générer une certaine quantité de bruit, dû au fait que l'analyse s'effectue uniquement sur la forme orthographique du mot, nous vérifions ensuite manuellement les couples dérivé/base extraits automatiquement. Le tableau 1 présente les résultats de l'analyse des préfixes verbaux (et de leurs allomorphes), accompagnés du nombre de formes « lemmatisées ». Par « lemme », il faut ici entendre la forme lemmatisée de la base du dérivé.

Préfixes verbaux	Occurrences	Lemmes	Erreur
ri–	508	63	5
r–	37	4	2
re–	96	9	0
de-	36	10	6

Tableau 1: analyse des verbes préfixés

Le nombre d'erreurs est calculé sur les formes lemmatisées. Ces erreurs correspondent à des cas où le sens du mot construit n'est pas compositionnel, comme dans **debuttare* = *de* + *buttare*[4]. Elles peuvent paraître fréquentes (13 cas sur 86) mais elles concernent uniquement des formes qui ne sont pas néologiques et qui constituent donc d'avantage des lacunes du lexique de référence utilisé pour cette expérience. D'ailleurs, tous les mots décomposés incorrectement sont connus des systèmes de TA classiques, qui reposent sur des lexiques plus exhaustifs.

Concernant les noms dérivés déverbaux, nous procédons de la même manière avec la RCM bilingue correspondante. Le tableau 2 propose un résumé des résultats de cette extraction et son évaluation. (Pour des raisons de place, nous ne mentionnons pas les formes qui n'ont donné aucun résultat – les noms préfixés avec l'allomorphe *r–* et le suffixe déverbal *–aggio* pour les trois préfixes).

---

[3] SYSTRAN S.A. http://www.systranet.com/systran/net (21 janvier 2005)

[4] ce qui correspondrait en français à **débuter* = *dé* + *buter*

Préfixe verbal	Suffixe déverbal	Occurrences	Lemmes	Erreurs
ri–	*–zione*	119	17	1
	–mento	201	10	0
re–	*–zione*	19	4	0
	–mento	16	1	0
de–	*–zione*	47	11	0
	–mento	0	0	0

Tableau 2: analyse des noms déverbaux préfixés

La vérification manuelle permet de trouver un taux de bruit presque nul. Une seule erreur où le sens du mot construit n'est pas compositionnel a été trouvée, et elle provient d'une erreur de typographie dans le corpus de départ.

De ces deux petites règles, pour les noms et les verbes, nous voyons qu'un nombre important d'unités lexicales peut être analysé d'une manière correcte d'un point de vue constructionnel (1065). Un tel résultat provient sans doute de plusieurs facteurs. Premièrement, les règles ont été soumises à des contraintes (simples mais importantes), ce qui évite un nombre important de mauvaises analyses (qui pourraient être dues par exemple à l'homographie des affixes). Mais le facteur le plus important est sans doute que les mots analysés sont uniquement néologiques (les erreurs ne se retrouvent que dans les mots non néologiques) et que l'analyse s'effectue en fonction du lexique de référence. Tout ceci tend à montrer que les mots construits néologiques sont rarement ambigus, ce qui vérifie, pour ce corpus en tout cas, notre première hypothèse.

Précisons cependant que, même si le bruit est quasiment inexistant, le silence serait également à évaluer. Toutefois, comme la contrainte la plus importante mise sur la règle concerne la présence de la base dans le lexique de référence, les mots non trouvés (le silence) sont forcément des mots dont la base est absente du lexique, et donc « intraitables » par le reste du processus de traduction.

3.4.2 Traduction des mots analysés

Après avoir analysé les mots construits et traduit les bases, nous avons « reconstruit » les équivalents de traduction en appliquant la partie française des RCM bilingues. Nous obtenons alors une liste de mots construits italiens et leur équivalent de traduction, comme *deresponsabilizzazione = déresponsabilisation*, *ridistribuzione = redistribution*, *riclassificare = reclassifier*. Evaluer la correction de la traduction ne nous semblait pas très pertinent, car une bonne traduction dépend également du contexte que nous n'avions pas. Nous avons donc évalué uniquement la correction du néologisme français. Cette évaluation doit se faire en fonction de la correction du néologisme par rapport à un sentiment linguistique propre à la langue française. Mais le sentiment linguistique reste un concept très flou. Les mots ont donc

été évalués sur une échelle de trois valeurs : correct, incorrect, incertain. Cependant, aucun mot n'a été jugé « incorrect ». Les résultats de l'évaluation sont résumés dans le tableau 3.

Procédés constructionnels	Lemmes	néologismes corrects	néologismes incertains
ri– + verbe	58	56	2
r– + verbe	2	2	0
re– + verbe	9	9	0
de– + verbe	4	4	0
ri– + nom	26	22	4
re– + nom	5	5	0
de– + nom	11	11	0

Tableau 3 : évaluation des néologismes en fonction du préfixe

Concernant les néologismes jugés incertains, nous avons utilisé le corpus du Web pour affiner encore l'évaluation. Ainsi, bien que des mots comme *recrocheter* ou *réassociation* aient paru étranges aux évaluateurs, le nombre d'occurrences et la pertinence des sources découvertes sur l'Internet permettent de valider ces créations dans tous les cas.

4 Conclusion et travaux futurs

Le petit nombre de mots étudiés dans cette expérience ne peut nous permettre de tirer des conclusions définitives et une expérience similaire sur d'autres corpus permettrait de confirmer les premières tendances constatées. Il n'en reste pas moins que cette expérience semble montrer que des règles de construction simples et précises donnent des résultats extrêmement fiables. De plus, le faible bruit provenant de l'analyse compositionnelle des néologismes est très prometteur.

D'un point de vue quantitatif, l'ensemble des règles proposées dans cet article a permis de traduire 1065 occurrences. Dans l'absolu, ce chiffre peut paraître faible en comparaison des 90 260 mots inconnus dans le corpus. Cependant, si la pertinence d'une telle règle est validée à plus large échelle et si l'on y ajoute d'autres règles similaires pour d'autres processus de construction réputés productifs, nous profiterions alors d'un mécanisme simple et moins chronophage en terme de gestion des ressources linguistiques pour la TA.

La validité de ces règles reste néanmoins à prouver, notamment pour d'autres processus constructionnels moins transparents et beaucoup moins décrit par la linguistique théorique. Et tout comme pour les règles syntaxiques dans le domaine de la TA, le traitement des divergences entre les deux langues (on dit *clonage* en français, mais *clonazione* en italien) est aussi un large objet d'étude.

Référence

Arnold D., Balkan L., Humphreys R., Meijer S., Sadler , (1994). *Machine translation: An introductory Guide,* Manchester, Blackwell.

Cabré T., Freixa, J., Solé E., (2002), A la limite des mots construits possible, Actes du *Forum de morphologie*, pp. 65-78.

Corbin D., (1987), *Morphologie dérivationnelle et structuration du lexique*, Tuebingen, Niemeyer.

Dal G. (2002). A propos d'une idée reçue, ou de la prétendue irrégularité de la dérivation, *Bulag,* Vol. 27, pp. 57-73.

Dal G., Hathout, N., Namer F., (*sous presse*), Morphologie Constructionnelle et Traitement Automatique des Langues : le projet MorTAL*, Lexique* Vol.16.

Dardano M., (1978), *La formazione delle parole nell'italiano di oggi*, Rome, Bulzoni.

Gaudin F., Guespin L., (2000), *Initiation à la lexicologie française*, Bruxelles, Duculot.

Gdaniec C., Manandise, E., McCord, M., (2001), Derivational Morphology to the Rescue: How It Can Help Resolve Unfound Words in MT. Actes de *MT Summit VIII*.

Huot, H., (2001). *Morphologie, Forme et sens des mots français*. Paris, Armand Colin.

Maurel, D. (2004). Les mots inconnus sont-ils des noms propres? Actes de *JADT 2004*, Louvain-la-Neuve.

Namer, F. (2001), Génération automatique de néologismes bilingues morphologiquement construits en français et en italien. Actes de *TALN 2001*. pp. 281-296.

Namer, F. (2000), GéDériF : Automatic Generation and Analysis of Morphologically Constructed Lexical Resources, Actes de *LREC 2000*, pp. 1447-1454

Rey-Debove J., Ed. (2004). *Brio*, Paris, Dictionnaire Le Robert.

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

Tzoukermann E., Jacquemin, C. (1997), Analyse automatique de la morphologie dérivationnelle et filtrage des mots possibles, Actes de *Forum de morphologie*, pp. 251-260.

Whitelock P., Kilby K, (1995) *Linguistic and computational techniques in machine translation system design,* London, UCL Press.