Constance et variabilité de l'incomplétude lexicale

Bruno Cartoni

Université de Genève – TIM/ISSCO ETI
bruno.cartoni@eti.unige.ch

Résumé

Cet article propose, au travers des résultats de différentes expériences sur la couverture des lexiques informatisés, de montrer que l'incomplétude lexicale est un phénomène constant dans tous les lexiques de TAL, mais que les mots inconnus eux-mêmes varient grandement selon les outils. Nous montrons également que la constance de cette incomplétude est étroitement liée à la créativité lexicale de la langue.

Mots-clés : lexique informatisé, incomplétude lexicale, mots inconnus, typologie.

Abstract

Through various experiments on computational lexica, we show that lexical incompleteness is a regular phenomenon across NLP lexica, but that the unknown words themselves vary strongly according to the individual lexicon. We also demonstrate that the regularity of incompleteness is closely related to lexical creativity within individual language.

Keywords: computational lexicon, lexical incompleteness, unknown words, typology.

1. Introduction

Cet article s'inscrit dans le cadre d'un projet de recherche (dont les premiers résultats sont décrits dans (Cartoni, 2005), qui vise à la construction d'outils pour gérer les mots inconnus en traduction automatique (ci-après TA) et, particulièrement, les mots néologiques construits. Les expériences décrites dans le présent article permettent d'analyser l'incomplétude des systèmes de TA et de situer la place des néologismes dans cette problématique. En effet, même si de nombreux projets abordent déjà, d'une manière quantitative et qualitative, les mots inconnus des lexiques informatisés, notre propos est ici de démontrer que (1) l'incomplétude lexicale est présente dans tous les lexiques informatisés, mais de manière variable, (2) que l'alimentation du lexique n'est pas en soi une solution à l'incomplétude lexicale, et (3) que l'on peut tout de même dégager des constantes dans les mots inconnus. Celles-ci, une fois individualisées, nous permettront de formaliser des règles applicables à la traduction de certains types de mots inconnus.

Nous présentons dans cet article une série d'expériences sur des ensembles de mots inconnus visant à démontrer ces trois assertions, et nous concluons en esquissant les grandes lignes des possibilités de traitement des mots inconnus. Nous commençons tout d'abord par dresser un bref aperçu des quelques recherches qui ont déjà abordé ces problématiques.

2. État de l'art

De nombreuses recherches ont déjà proposé une caractérisation de l'incomplétude lexicale dans le but de déterminer où investir les efforts pour améliorer les ressources lexicales

(Ren *et al.*, 1992 ; Dister *et al.*, 2004 ; Maurel, 2004). Les conséquences de l'incomplétude lexicale varient en effet en fonction de l'outil et de la place accordée au lexique. En TA, l'absence d'un mot dans le lexique pose un double problème, d'analyse tout d'abord, puis de génération (Gdaniec, 2001).

Mais quelle que soit l'application, la couverture des lexiques se mesure généralement en nombres de mots inconnus de ce lexique. Par exemple, dans son étude sur un corpus de textes journalistiques, (Maurel, 2004) obtient un taux de 4 % de mots inconnus si l'on considère l'ensemble des occurrences du texte, et 13 % si l'on ne tient compte que du nombre de mots distincts. La couverture lexicale est donc rarement parfaite et dépend du lexique et du type de textes traités.

D'un point de vue qualitatif, de nombreuses recherches ont également caractérisé les mots inconnus (Ren *et al.*, 1992 ; Dister *et al.*, 2004 ; Maurel, 2004). Typiquement, toutes s'accordent à diviser les mots inconnus en trois grands ensembles : (i) les mots découlant des phénomènes de créativité lexicale, (ii) les noms propres et (iii) les mots erronés, même si, au sein de ces trois grands groupes de mots inconnus, les subdivisions peuvent varier de manière importante selon les auteurs. Le repérage automatique des noms propres s'avère assez aisé (par l'application de routines basées sur la majuscule, (Maurel, 2004), du moins pour les langues où celle-ci est un indice suffisant). Dans les expériences décrites ici, nous avons exclu les noms propres, qui représentent une problématique bien différente, pour nous pencher uniquement sur les deux autres groupes ((i) et (iii)).

3. Données d'expérimentation

Pour mener à bien les différentes expériences présentées ci-dessous, nous avons individualisé des ensembles de **mots inconnus** en confrontant des données textuelles à des lexiques de référence, qui consistent donc en des listes de **mots connus**.

S'agissant des corpus, nous disposons d'un corpus italien (les éditions du mois de février 1992 de *Il Sole 24 ore*, journal d'actualité quotidien) et d'un corpus français (les éditions du mois de juillet 1993 du journal *Le Monde*), tous deux publiés par ELRA, (corpus MLCC, 1997).

Pour les lexiques, nous avons dans un premier temps utilisé les bases lexicales italiennes et françaises construites dans le cadre du projet Multext[1], et utilisées dans la chaîne d'étiquetage de l'étiqueteur Tatoo[2]. D'un point de vue quantitatif, la base française contient 279 007 formes fléchies, et la base italienne 739 000 (cette forte différence provient de la génération des formes verbales cliticisées en italien).

Nous avons, dans un deuxième temps, évalué la couverture lexicale d'outils de TA commerciaux en confrontant notre corpus français[3] à leur lexique (Système 1 et Système 2). Pratiquement, nous avons fait traduire notre corpus en anglais, puis nous avons analysé les mots inconnus listés par les systèmes.

Dans la suite, nous décrivons les résultats des expériences visant à confirmer les trois hypothèses mentionnées plus haut. Les deux premières expériences relatées tentent d'évaluer la variabilité de l'incomplétude lexicale (sections 3 et 4), et la troisième série d'expériences s'attache à trouver des constantes dans les ensembles de mots inconnus (section 5).

[1] The Multext project : http://www.lpl.univ-aix.fr/projects/multext/

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

[3] Ces deux outils ne traitent pas la langue italienne, du moins dans les versions dont nous disposons.

4. De la variabilité de l'incomplétude lexicale

Au-delà de la simple quantification, nous avons cherché à savoir si l'incomplétude lexicale était un phénomène variable. Dans cette section, nous comparons tout d'abord l'incomplétude lexicale des deux lexiques de langue différente (l'italien et le français), puis celle de trois lexiques différents (Mmorph, Système 1, Système 2) pour le même corpus français. En comparant l'incomplétude lexicale de deux bases lexicales confrontées à deux corpus de langues différentes, on constate une disparité entre les **pourcentages de mots inconnus** des deux lexiques italien et français, comme le montre le tableau ci-dessous :

	Mmorph IT	Mmorph FR
Occurrences	1 667 227	908 953
Occurrences inconnues	208 464	84 675
% d'occurrences inconnues	12,50 %	9,32 %

Tableau 1. Occurrences inconnues des lexiques

Cette disparité est peut-être due à plusieurs facteurs de conception des lexiques, notamment au fait que le lexique français a été davantage utilisé (et donc amélioré) dans le cadre d'autres travaux de recherche, mais elle est confirmée ci-dessous par la comparaison de l'incomplétude lexicale des lexiques de TA.

En effet, en comparant ce lexique français (Mmorph) avec d'autres dictionnaires incorporés à des outils de TA commerciaux (Système 1 et Système 2) auxquels nous avons confronté le corpus du journal *Le Monde* (qui compte 908 953 occurrences). Le tableau ci-dessous présente les proportions d'occurrences inconnues des trois systèmes.

	Mmorph	Système 1	Système 2
Occurrences inconnues	84 675	49 067	47 06
% d'occurrences inconnues	9,32 %	5,40 %	5,17 %

Tableau 2. Mots inconnus des trois systèmes

Il est intéressant de constater la différence entre le lexique de Mmorph et les lexiques dits « commerciaux », qui s'explique par le fait que Mmorph est un lexique élaboré dans le cadre de projets bien précis. En revanche, il est intéressant de noter la grande similitude de proportions de mots inconnus dans les deux outils commerciaux.

Comme nous l'avons mentionné plus haut, un des trois grands groupes de mots inconnus est celui des noms propres, qui est une problématique dont la résolution est un champ d'investigation à part entière. En excluant donc les noms propres[4] et en réduisant les listes d'occurrences inconnues en listes de formes uniques, nous obtenons trois ensembles de mots inconnus des trois systèmes testés (MI 1, MI 2 pour les systèmes 1 et 2, et MI M pour Mmorph), qui peuvent être analysés les uns par rapport aux autres. Nous présentons dans le diagramme ci-dessous les résultats de cette analyse croisée.

[4] Par l'application d'une routine basée sur la majuscule.

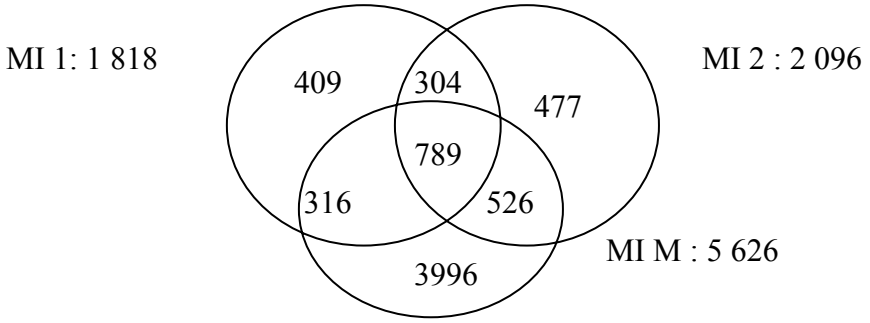


Figure 1. Répartition des trois ensembles de mots inconnus

La première constatation est une forte dispersion des mots inconnus entre ces trois ensembles, où beaucoup de mots inconnus ne le sont que d'un seul lexique. De plus, l'intersection centrale montre que seules 789 formes sont inconnues des trois systèmes à la fois. On pourrait également voir dans ce chiffre uniquement la présence de mots erronés, et donc absents de tous les lexiques. Or, comme nous le verrons à la section 5.3.2, les mots erronés ne représentent que 33,5 % de cet ensemble. En tous les cas, ces chiffres montrent une grande variabilité dans les mots inconnus, et donc, dans le contenu des lexiques.

Nous venons de le voir, même si l'incomplétude lexicale est un phénomène présent dans les lexiques testés, elle reste très variable suivant les cas. Non seulement la proportion de mots inconnus varie, mais ces mots inconnus ne sont pas les mêmes, même dans deux lexiques provenant de systèmes conçus pour la même tâche. Cette variabilité dans la conception des lexiques nous amène à nous poser la question de la possibilité d'un lexique exhaustif. En effet, si le manque d'exhaustivité des lexiques est un phénomène connu, souvent présenté comme un état de fait (et expliqué en invoquant les nombreuses contraintes matérielles et pratiques qui pèsent sur la construction des lexiques), il nous a semblé pertinent d'évaluer la possibilité théorique d'un lexique exhaustif.

5. De la possible exhaustivité d'un lexique ?

Sans aborder la question du caractère infini du lexique des langues, la problématique de l'alimentation des lexiques mérite d'être abordée, car la première solution à l'incomplétude lexicale reste l'alimentation. Bien que naïve, cette solution doit être envisagée dans l'absolu, même si certains aspects pratiques la remettraient rapidement en cause. Dans cette section, nous exposons les résultats d'une expérimentation visant à montrer que l'incomplétude lexicale est constante et inéluctable.

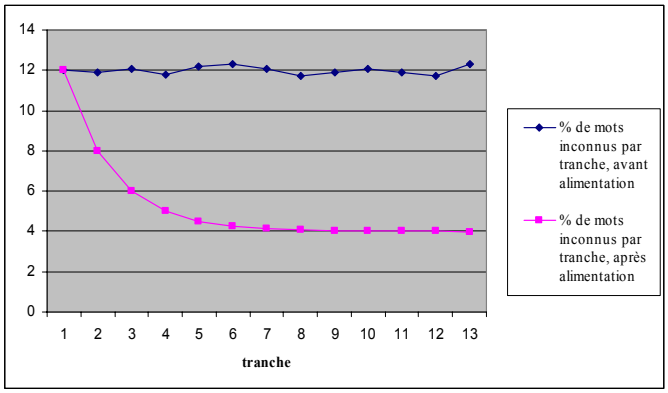
5.1. Déroulement de l'expérience

L'idée principale de l'expérience est de découper un corpus en tranches et de le soumettre à l'analyse lexicale tranche après tranche. Entre chaque tranche, nous ajoutons à notre lexique les mots inconnus de la tranche précédente avant d'analyser la tranche suivante avec ce lexique nouvellement alimenté. Nous obtenons ainsi, pour chaque tranche, deux données. D'une part, la proportion des mots inconnus dans celle-ci avant alimentation, et d'autre part la proportion des mots inconnus après alimentation du lexique par les mots inconnus des tranches précédentes.

Dans chaque tranche, nous comparons ensuite le pourcentage de mots inconnus avant et après alimentation, pour voir si celui-ci diminue et dans quelle proportion. En effet, une diminution

linéaire tendrait à montrer que l'exhaustivité est possible dans l'absolu, car nous pourrions potentiellement parvenir à 0 % de mots inconnus, alors qu'une diminution exponentielle montrerait qu'il existe un « seuil », une sorte de limite au-delà de laquelle il est vain de vouloir être exhaustif. Le schéma ci-dessous tente de modéliser l'hypothèse.

Figure 2. Schématisation du résultat de l'expérience



Nous avons donc tenté l'expérience avec nos deux lexiques (français et italien) et nos deux corpus, tels que décrits plus haut. Nous avons également intégré un certain nombre de paramètres pour des raisons tant pratiques que théoriques. Premièrement, nous n'avons pas pris en compte les noms propres, parce qu'ils représentent une problématique différente de la créativité lexicale. Deuxièmement, nous avons découpé nos deux corpus en 26 tranches plus ou moins égales, car nous disposions pour l'italien d'un corpus déjà divisé de la sorte. Troisièmement, nous avons présenté les tranches dans trois ordres de passage différents, pour éviter tout biais provenant de l'ordre de passage et de la disparité des tranches (en termes de richesse de vocabulaire, de proportion de mots inconnus, etc.). Les résultats présentés sont donc des moyennes. Enfin, notons également que nous avons ajouté dans le lexique uniquement les formes telles qu'elles se présentaient dans le corpus, sans lemmatisation, et par conséquent, sans génération de toutes les flexions possibles. Nous pensons que cette option méthodologique n'influence que très peu la diminution des mots inconnus. Tout au plus, la génération des formes fléchies ne ferait qu'accélérer la diminution de mots inconnus au fil des tranches, mais ne modifierait pas la tendance générale, en tout cas pour les langues en présence.

5.2. Résultats

Pour chaque tranche, nous avons calculé l'écart entre le pourcentage de mots inconnus avant et après alimentation. Les deux graphiques ci-dessous présentent l'accroissement de l'écart moyen pour les deux langues. Clairement la tendance dans la progression des écarts est la même dans les deux corpus.

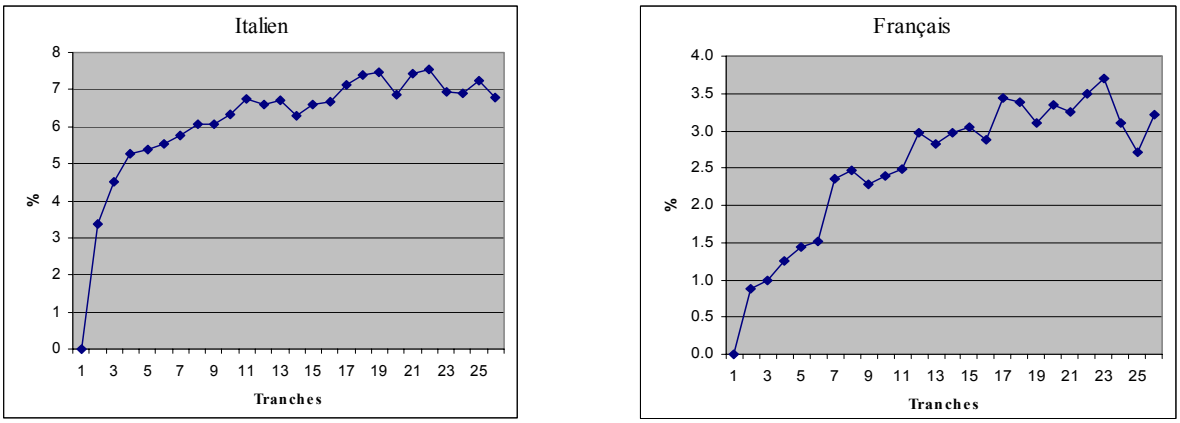


Figure 3. Progression des écarts avant/après alimentation

Ces courbes montrent nettement une tendance de type logarithmique, même si la progression pour le français semble moins régulière, sans doute à cause d'une plus forte disparité entre les proportions de mots inconnus dans les différentes tranches. Pour les deux langues, on assiste tout d'abord à une augmentation rapide de l'écart (la différence entre les mots inconnus avant et après alimentation est importante) ; puis, cette augmentation de l'écart s'amenuise pour tendre ensuite vers l'aplatissement. La première augmentation est explicable par le fait que, dans ces premières étapes, ont été ajoutés un certain nombre de mots qui étaient peut-être des lacunes des lexiques de référence et qui, surtout, étaient fréquents dans les différentes tranches du corpus. Par la suite, la tendance à l'aplatissement de la courbe tendrait à montrer que l'on atteint un seuil fatidique de mots inconnus uniques. Ainsi, plus le lexique est alimenté, moins l'écart entre le pourcentage initial de mots inconnus et le pourcentage après alimentation progresse.

Cette expérience montre qu'une alimentation du lexique avec des mots inconnus n'aurait pas forcément de conséquences sur la couverture future du lexique, étant donné qu'il y a de fortes chances pour que ces mots nouvellement intégrés au lexique n'apparaissent plus dans les textes soumis à ces lexiques. Cette unicité des mots inconnus est sans doute à mettre en relation avec la part importante de néologismes dans les mots inconnus.

6. De la constance des mots inconnus

Si, dans un premier temps, nous avons montré que l'incomplétude lexicale était un phénomène qui pouvait varier grandement selon les lexiques, nous venons de voir qu'elle semble être également un phénomène relativement constant. Cette constance est sans doute liée aux éléments qui composent ce phénomène : les mots inconnus. Dans cette section, nous proposons un certain nombre de résultats de quantification et de typologie qui permettent de voir plus précisément à quoi ressemble un mot inconnu. Rappelons que nous avons volontairement exclu les noms propres de nos analyses.

Pour mener à bien ces quantifications et cette typologie, nous disposons donc de plusieurs ensembles de mots inconnus, issus des expériences évoquées ci-dessus. Tout d'abord, nous disposons de trois ensembles de mots inconnus des trois lexiques français décrits à la section 3 (MI M, MI 1 et MI 2). Nous disposons également d'un ensemble qui contient tous les mots inconnus « communs » aux lexiques de deux applications commerciales de TA (MI 1 $\cap$ MI 2, que nous nommons dans la suite EXP 1). Ce groupe de mots inconnus, bien que réduit (1093 formes), est particulier, car il s'agit d'un échantillon qui regroupe (et peut-être qui représente plus fidèlement) les mots inconnus des outils de TA. Enfin, nous disposons

de deux autres ensembles (EXP 2-FR -194 mots et EXP 2-IT -488 mots) qui sont issus de la dernière tranche des corpus (italien et français) de l'expérience décrite à la section 4. Les mots contenus dans ces deux ensembles n'étaient pas présents dans les tranches précédentes du corpus, et sont donc essentiellement des hapax.

6.1. Les hapax

Le phénomène des hapax (mots présents une seule fois) semble être un caractère typique des mots inconnus en comparaison avec le nombre d'hapax présent dans un corpus textuel. Ainsi, si le pourcentage d'hapax est de 43 % dans notre corpus textuel français, il s'élève à 57 % dans l'ensemble MI M, et à 75,7 % dans les ensembles MI 1 et MI 2. Ces chiffres, et surtout l'écart par rapport à un corpus textuel, tendent à montrer le caractère unique des mots inconnus.

6.2. Les ratio forme/lemme

Nos calculs sur les mots inconnus et les hapax ne portent que sur des formes graphiques différentes, et ne prennent pas en compte le fait que deux mots inconnus puissent faire partie du même paradigme flexionnel. Néanmoins, une évaluation sur les 1 000 premiers mots inconnus du français et de l'italien nous montre que ces 1 000 formes représentent 855 lemmes en italien et 867 lemmes en français. Cette proportion est sensiblement plus élevée que celle rencontrée dans un corpus textuel (citons notamment une étude de (Lebart *et al.*, 1994) qui trouvait une proportion de 9 309 lemmes pour 13 590 formes dans un corpus de discours politique, soit 689/1 000). Ceci confirme à nouveau la tendance à l'unicité des mots inconnus, unicité qui s'explique par le caractère même des mots inconnus. Outre les noms propres, les mots inconnus sont soit des erreurs (qui n'apparaissent donc normalement qu'une seule fois), soit des néologismes qui appartiennent majoritairement à des catégories ayant un paradigme flexionnel restreint (nom ou adjectif). C'est justement sur une typologie précise des mots inconnus que porte la suite de nos expériences.

6.3. Typologie

Outre les noms propres et les erreurs, une grande partie des mots inconnus sont généralement catégorisés dans le groupe « créativité lexicale », qui concerne tous les mots formés selon des procédés identifiables, ou empruntés à des langues étrangères (les emprunts). Dans ce groupe, une large place est occupée par les « néologismes », qui sont des mots nouveaux généralement construits avec des matériaux lexicaux existants.

Dans les tableaux ci-dessous, nous présentons, à côté des proportions de référence (ci-après RefTyp, tirées de (Maurel, 2004), une typologie des mots inconnus communs au deux lexiques des outils de TA évalués (EXP 1) et des mots inconnus dans les dernières tranches de l'expérience décrite dans la section 4, pour l'italien et le français (EXP 2-IT et EXP 2-FR). Ces deux derniers ensembles sont particuliers car ils sont constitués majoritairement d'hapax (un mot inconnu dans ces dernières tranches apparaît pour la première fois dans le corpus). Soulignons également que les catégories proposées par Maurel ont parfois été fusionnées par manque de place. Ci-dessous, nous nous penchons tout d'abord sur les mots inconnus issus de la créativité lexicale, puis nous évoquons les mots erronés et les mots inclassables.

6.3.1. Créativité lexicale

	RefTyp	EXP 1	EXP 2-IT	EXP 2-FR
Sigle et abréviation	15 %	0,8 %	3 %	0 %
Chiffres romains	9 %	0 %	0 %	0 %
Dérivés de noms propres	6 %	6,3 %	1 %	3 %
Néologismes	**17 %**	**20,2 %**	**32 %**	**25 %**
Onomatopées	2 %	0,2 %	0 %	0 %
Mots étrangers	13 %	19,3 %	11 %	4 %
Total	**62 %**	**46,8 %**	**47 %**	**32 %**

Tableau 3. Créativité lexicale

Entre RefTyp et EXP 1, la proportion de néologismes est sensiblement la même, et toujours majoritaire par rapport aux autres catégories. La proportion de mots étrangers y est aussi nettement plus importante que dans les mots inconnus de RefTyp. Ces proportions plus importantes dépendent également de la faible place occupée par les sigles et les chiffres romains dans les proportions de EXP 1.

Même si les ensembles RefTyp et EXP 2 ne sont pas forcément comparables, il est tout de même intéressant de souligner certaines différences de proportion. On constate d'abord une grande différence de proportion dans l'ensemble des mots inconnus néologiques, qui tendrait à prouver que l'ensemble des mots néologiques est plus ouvert et plus propice à l'extension. À l'inverse, si l'on compare les pourcentages de mots étrangers (et particulièrement ceux de RefTyp (13%) et EXP 2-FR (4%)) on constate une différence importante qui semble montrer que l'ensemble des mots étrangers n'est pas aussi infini que l'ensemble des néologismes. Evidemment, une typologie plus précise de chaque tranche (cf. expérience section 4) permettrait de confirmer cette tendance.

6.3.2. Les erreurs et les mots qui restent

Le dernier groupe de mots inconnus est celui des erreurs. La proportion d'erreurs reste constante dans les trois ensembles : 33,5 % dans EXP 1, 23 % dans EXP 2-IT et 24 % dans EXP 2-FR. Ceci est assez logique étant donné que les erreurs dépendent plus de la qualité du corpus que de la qualité des lexiques. On remarque aussi que la proportion d'erreurs est identique dans nos deux corpus italien et français, et dans le corpus de RefTyp. Notons également que nous avons individualisé une catégorie d'erreurs de segmentation, qui étaient compris dans la catégorie des erreurs de Ref Typ (EXP 1 : 14,4 %, EXP 2-IT : 4 %, EXP 2-FR : 9 %).

En effectuant la typologie décrite ci-dessus, nous sommes resté avec un nombre (certes faible) de mots inclassables, qui ne sont pas vraiment néologiques et qui ne rentraient pas non plus dans les autres catégories. Dans EXP 1, 51 mots (4,64 %) entraient dans cette catégorie (comme *allers, bravos, consorts, pleuvent, sandwiches, truchement,* etc.), 93 mots dans EXP 2-IT (19 %) et 68 mots dans EXP 2-FR (35 %). Une partie de ces mots pourraient être considérée comme des néologismes flexionnels (Pruvost *et al*, 2003), c'est-à-dire des formes fléchies incorrectes ou rares, mais possibles dans certains contextes. Mais plus généralement, ce phénomène montre, selon nous, que les facteurs « inhérents à la langue » (c'est-à-dire la

présence dans les textes de noms propres ou de néologismes) ne sont pas les seuls responsables de l'incomplétude lexicale. En effet, la constitution de lexiques étant un travail humain, elle reste une entreprise qui peut avoir ses faiblesses. C'est sans doute ces quelques cas marginaux que nous avons rencontrés.

7. Conclusion

Dans cet article, nous avons souligné le caractère à la fois constant et variable du phénomène de l'incomplétude lexicale. Nous avons tout d'abord montré que, non seulement les proportions de mots inconnus différaient selon les lexiques, mais aussi que la couverture de ces derniers est assez différente, même pour deux lexiques à même visée (systèmes de TA). Nous avons également souligné que l'incomplétude lexicale était un phénomène infini, à cause d'un seuil infranchissable de mots inconnus, seuil qui est sans doute le reflet de la constante créativité lexicale de la langue. Du point de vue des mots inconnus, nous avons également montré quelques constances, comme leur caractère unique (reflété notamment par l'important nombre d'hapax présents parmi eux). De plus, nous avons montré que la majeure partie des mots inconnus était constituée de néologismes et que cet ensemble était « extensif ».

La résolution des mots inconnus est donc une problématique importante pour le TALN. Outre les mots erronés et les emprunts, qui constituent des problématiques à part entière, nous avons constaté que la majorité des néologismes n'était pas le fait de création *ex-nihilo* mais le résultat de règles de construction internes à la langue. Nous envisageons donc, dans la suite de nos recherches, la création d'outils permettant de gérer les mots construits dans les systèmes de traduction automatique. Pour ce faire, ces outils devront être composés de règles de construction des mots bilingues (Cartoni, 2005) qui pourront analyser tous les mots nouveaux construits qui peuvent apparaître dans les textes, et proposer une traduction.

Références

CARTONI B. (2005). « Traduction de règles de construction des mots pour résoudre les problèmes d'incomplétude lexicale en traduction automatique Étude de cas ». In *Actes de RECITAL 2005* : 565-574.

DISTER A., FAIRON C. (2004). « Extension des ressources lexicales grâce à un corpus dynamique ». In *Actes de L'analyse des données textuelles : de l'enquête aux corpus littéraires*. Lexicometrica. http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema7/Texte-Dister.pdf.

GDANIEC C., MANANDISE E., MCCORD M. (2001). « Derivational Morphology to the Rescue: How It Can Help Resolve Unfound Words ». In *Actes de MT Summit VIII* : 127-131.

LEBART L., SALEM A. (1994). *Statistique textuelle*. Dunod, Paris.

MAUREL D. (2004). « Les mots inconnus sont-ils des noms propres ? » In *Actes de JADT 2004*. Presse universitaires de Louvain, Louvain-la-Neuve.

PRUVOST J., Sablayrolles J-F, (2003). *Les néologismes*. PUF, Paris.

REN X., PERRAULT, F. (1992). « The Typology of Unknown Words: An experimental Study of Two Corpora ». In *Actes de Coling 92* : 408-414.

SILBERZTEIN M. (1993). Dictionnaires électroniques et analyse automatique de textes. Masson, Paris.