# Mesure de l'alternance entre préfixes pour la génération en traduction automatique

Bruno Cartoni

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

**Résumé** La génération de néologismes construits pose des problèmes dans un système de traduction automatique, notamment au moment de la sélection du préfixe dans les formations préfixées, quand certains préfixes paraissent pouvoir alterner. Nous proposons une étude « extensive », qui vise à rechercher dans de larges ressources textuelles (l'Internet) des formes préfixées générées automatiquement, dans le but d'individualiser les paramètres qui favorisent l'un des préfixes ou qui, au contraire, permettent cette alternance. La volatilité de cette ressource textuelle nécessite certaines précautions dans la méthodologie de décompte des données extraites.

**Abstract** Generating constructed neologisms in a machine translation system is confronted to the issue of selecting the right affixes, especially when some affixes can be used alternately. We propose here an "extensive" study that looks into large textual data collections (web) for prefixed forms that have been automatically generated, in order to find out parameters that allow the use of both prefixes or, on the contrary, that prevent one or the other prefixation. The volatility of web resources requires methodological precautions, especially in data counting.

**Mots-clés :** morphologie, traduction automatique, génération, néologisme, études empiriques

**Keywords:** morphology, machine translation, generation, neologism, empirical studies

## Introduction

L'étape de génération des néologismes construits dans un contexte de traduction automatique est souvent confrontée à un certain nombre de difficultés, notamment lorsqu'il existe la possibilité de former, sur un même mot-base (e.g. *dimensionnel*), deux formes construites avec deux préfixes différents pour exprimer deux sens apparemment synonymes (*multidimensionnel/pluridimensionnel*). Dans cet article, nous proposons une méthodologie de morphologie « extensive » autour de ce phénomène d'alternance, encore trop peu étudié en morphologie, qui nous permet d'une part de le quantifier, et d'autre part de mettre au jour certaines caractéristiques autorisant ou empêchant cette alternance, pour pouvoir contraindre

les règles de génération. La ressource textuelle utilisée (l'Internet) fournit un nombre important de données, mais nécessite également la mise en place de certains affinages dans le traitement des résultats.

# 1 Traduction automatique et morphologie

Cette étude s'inscrit dans un projet de traduction automatique (TA) des mots construits (Cartoni 2005) pour lequel nous formalisons un certain nombre de règles de construction des lexèmes (RCL) bilingues, permettant de traduire automatiquement des néologismes construits absents des lexiques des systèmes de TA. Ces règles analysent tous les mots inconnus construits et individualisent le procédé de construction qui les sous-tend ; elles génèrent ensuite un équivalent de traduction dans la langue cible. Le système se concentre pour l'instant sur la préfixation en français et en italien. L'exemple ci-dessous montre une RCL bilingue de *réitérativité*, (ou mot italien X, préfixé en *ri*, peut être traduit par un mot français Y, préfixé en *re*).

**IT** : $RCL_{REITER}$ $[X]_V \rightarrow ri[X]_V \leftrightarrow$ **FR** : $RCL_{REITER}$ $[Y]_V \rightarrow re[Y]_V$

Ainsi, un mot italien construit (e.g. *ricostruire*) est analysé par la partie gauche de la règle (ici : *ri+ costruire*$_V$ ). La partie droite de la règle s'occupe ensuite de générer un équivalent de traduction en français (ici : *reconstruire*) en se basant sur la règle et sur la traduction de la base (costruire$_{IT}$=construire$_{FR}$). Ces RCL bilingues sont basées sur une étude contrastive des procédés de traduction, qui s'attache avant tout à mettre en parallèle des procédés en fonction du sens qu'ils permettent de construire.

## 1.1 Des règles de construction des lexèmes à exposants multiples ?

Cet appariement des RCL basé sur le sens construit implique que certaines règles mettent en jeu plusieurs affixes, sans qu'il soit possible, *a priori*, de contraindre l'emploi de l'un ou l'autre. Ce phénomène se retrouve notamment dans les préfixations de quantité, où l'on peut distinguer la règle de préfixation de « pluralité indéterminée » (qui met en jeu les préfixes français *pluri, multi, poly*, et les préfixes italiens *pluri, multi, poli*), et celle de « quantité unique » (mettant en jeu les préfixes *mono* et *uni* dans les deux langues), ainsi que dans les règles de préfixation évaluative, où l'on distingue notamment la valeur « grand » – FR : *méga, maxi*, IT : *mega, maxi* – et la valeur « bon » FR : *hyper, super*, IT : *iper, super*[1].

D'un point de vue de pure formalisation morphologique, et dans une approche lexématique – notamment à la suite de travaux de (Fradin 2003) –, ces « groupes » de préfixes sont considérés comme les *exposants* d'une même règle de préfixation. Ainsi, nous formalisons des *RCL bilingues à exposants multiples*, qui décrivent les différentes possibilités de formation des mots en langues source et cible. La RCL de pluralité indéterminée pour la préfixation nominale prend alors la forme de :

**IT** *:* $RCL_{PLUR\_INDET}$ $[X]_N \rightarrow$ *multi*|*pluri*|*poli*$[X]_N \leftrightarrow$ **FR** : $RCL_{PLUR\_INDET}$ $[Y]_N \rightarrow$ *multi*|*pluri*|*poly* $[Y]_N$

Ces RCL à exposants multiples concernent évidemment les groupes de préfixes partageant une même valeur sémantique, et doivent être spécifiées en fonction des différentes catégories

[1] Ces groupes de préfixes inclut de nombreux préfixes, mais pour limiter l'étude, nous nous concentrons sur les préfixes les plus fréquents.

d'input et d'output. En français, par exemple, (Amiot 2005) rappelle que les trois préfixes de pluralité indéterminée renvoient tous à une idée de pluralité, et entrent dans la formation de noms (essentiellement sur base nominale) et d'adjectifs (sur base nominale ou adjectivale). Ils peuvent alterner avec d'autres formants exprimant d'autres idées de dénombrement (*multicellulaire/unicellulaire*) mais également entre eux (*multicellulaire/pluricellulaire*). Les préfixes de quantité unique (*mono, uni*) semblent présenter les mêmes caractéristiques en termes d'alternance, en italien comme en français. Enfin, les préfixes « évaluatifs » s'appliquent à des bases nominales et adjectivales pour former respectivement des noms ou des adjectifs (*bien→mégabien*$_{FR}$*, président → superprésident*$_{FR}$*, riforma→ iperriforma*$_{IT}$ ). Parmi ces préfixes, l'alternance semble inhérente à leur classe particulière, qui est souvent décrite comme extrêmement variable et volatile (Montermini à paraître).

## 1.2 Des préfixes interchangeables ?

Les études morphologiques décrivent certes certaines contraintes favorisant l'emploi de tel ou tel préfixe, comme l'origine latine ou grecque de la base et du préfixe, la catégorie de la base (les adjectifs préférant *multi* et *pluri*, les noms *poly*), ou encore le domaine (*poly* étant réservé à certains vocabulaires techniques). Mais malgré ces contraintes (lorsqu'elles existent), l'alternance entre préfixes paraît être un phénomène fréquent, qui mérite d'être étudié quantitativement et qualitativement.

Dans notre projet de système de traduction des néologismes construits, et tout particulièrement dans la partie qui traite de la génération en langue cible, nous avons besoin de connaître les conditions qui permettent cette **interchangeabilité** ou qui, au contraire, poussent à la **sélection** de l'une ou l'autre forme. L'étude de l'alternance entre préfixes doit donc permettre de découvrir des contraintes ou des tendances (de fréquence ou d'usage) permettant la sélection de l'une ou l'autre forme, ou au contraire, permettant de maintenir l'interchangeabilité et générer les différentes possibilités de traduction. Pour mesurer l'interchangeabilité de certains préfixes, les indices de productivité désormais classiques (Baayen et Lieber 1991) ne suffisent pas, car bien souvent, les préfixes sont l'un comme l'autre tout aussi productifs à l'intérieur du même groupe. Et même si ces indices peuvent *in fine* être la solution, ils ne nous épargnent pas de nous interroger au préalable sur l'alternance entre préfixes. Nous proposons donc dans la suite une méthode permettant d'appréhender l'alternance entre préfixes sous des angles différents.

# 2 Méthodes pour mesurer l'alternance

Dans cet article, nous présentons une série d'études en corpus permettant (1) d'évaluer l'existence de cette alternance et (2) de quantifier le degré d'alternance entre préfixes, en mesurant l'écart des fréquences entre les paires de formes possibles (p. ex. *mégaproduction/maxiproduction*). Ces fréquences comparées pour des paires de formes possibles sont obtenues en confrontant des paires de mots français générés automatiquement par nos règles de traduction automatique à partir de néologismes italiens avec la large ressource textuelle que constitue l'Internet. Cette recherche « extensive » (Hathout, Plénat et al. 2003) vise à acquérir, en corpus, le plus grand nombre possible d'attestations du procédé étudié en vue de faire apparaître des régularités nouvelles, ou, dans notre cas, des contraintes à appliquer sur la partie « génération » des RCL bilingues.

Comme toute méthode empirique, et particulièrement celles qui se basent sur la ressource textuelle que constitue l'Internet, les évidences d'une alternance ne doivent pas mener à une

conclusion trop abrupte, sans prendre des précautions d'analyse des données. En effet, une alternance constatée en corpus peut être le fait de deux phénomènes. Soit les préfixes sont complètement *interchangeables*, et les deux formes préfixées sont possibles et fréquentes dans les mêmes proportions, soit l'alternance relève d'un « accident », l'une des deux formes étant beaucoup moins fréquente que l'autre. Ainsi, une analyse plus fine des fréquences trouvées doit être menée, pour tenir compte de ces différences également d'un point de vue plus qualitatif.

Nous décrivons dans la suite la procédure d'acquisition des néologismes construits en italien et la génération des paires de traduction potentielles, ainsi que l'outil exploité pour effectuer les requêtes sur le web et récupérer les fréquences d'apparition des différentes formes testées.

## 2.1 Corpus et outils utilisés

Nous avons tout d'abord extrait d'un corpus journalistique italien (*La Repubblica*, 380 millions d'occurrences (Baroni, Bernardini *et al.* 2004)) les mots commençant par les préfixes de notre étude. Nous avons ensuite affiné cette extraction en ne gardant que les mots absents d'un lexique de référence de l'italien (Mmorph, 739 000 formes (Bouillon, Lehmann *et al.* 1998)) et en ne conservant que les mots dont la base est présente dans ce même lexique de référence. Pour chaque néologisme italien ainsi extrait, nous avons traduit semi-automatiquement leur base en français (*multidimensionale*$_{IT}$ → *dimensionnel*$_{FR}$), puis nous avons généré automatiquement les différentes possibilités de construction avec les préfixes possibles du français. Ainsi, pour l'exemple donné ci-dessus, nous avons généré automatiquement les mots français : *multidimensionnel, pluridimensionnel, polydimensionnel.*

Nous obtenons ainsi une liste de paires (ou de triplets) de mots possibles pour chacun des formes italiennes d'origine (611 formes préfixées en *maxi*, 714 formes en *mega*, 2155 formes en *super*, 731 formes en *iper*, 210 formes en *mono*, 263 formes en *multi*, 274 formes en *pluri*, et 94 formes en *poli*). Il convient en outre de préciser que nous n'avons procédé à aucune lemmatisation des données, étant donné que la flexion peut avoir son importance en néologie.

Pour évaluer empiriquement l'existence ou la fréquence de ces mots possibles, nous avons utilisé le robot Golf (Thomas 2008) qui permet de lancer une succession de requêtes sur le moteur de recherche Google, en spécifiant certaines options (langue, format du fichier, emplacement du mot recherché dans les pages web, etc.). Ce robot donne en retour le nombre de pages trouvées pour chaque requête. Évidemment, les données provenant de l'Internet sont à prendre avec précaution étant donné la *volatilité* de ce corpus : peu de fiabilité dans la correction (orthographique et terminologique), peu de contrôle linguistique, importante présence de noms de marque ou de jeux de mots. Malgré cela, le corpus du web reste une source inestimable pour un travail sur la productivité néologique.

En sortie de l'outil Golf, nous obtenons pour chaque forme la fréquence d'apparition sur l'Internet[2]. Mais cet ensemble de fréquences peut être analysé de différentes manières. Comme le rappellent (Hathout, Namer et al. (à paraître). ), nous pouvons prendre en compte la distinction *absence/présence* des formes recherchées, les différents *ordres de grandeur* (2

[2] Cette fréquence est en réalité le nombre de « documents » trouvés qui contiennent au moins une occurrence de la forme recherché. Il peut y avoir plusieurs occurrences dans un document, phénomène dont les moteurs de recherche ne permettent pas de rendre compte.

occurrences vs 3 millions d'occurrences), ou les *différences faiblement marquées* (presque autant d'occurrences pour les deux formes testées). Cette dernière comparaison est la plus intéressante en cas de préfixe alternant, mais doit être mise en relation avec les autres types de comparaison.

### 2.2 Plusieurs mesures possibles

Différents types d'informations peuvent être tirés de la collecte de ces fréquences. Premièrement, l'absence de pages retournées donne un indice d'impossibilité de construction, même si cette absence ne dit rien dans l'absolu sur la possibilité de construction d'un lexème, et qu'elle peut provenir de problèmes informatiques du système ou du réseau, ou d'une divergence de construction entre l'italien et le français. Cette absence/présence nous permet tout de même d'évaluer l'ampleur de cette interchangeabilité (cf. section 3.1). Deuxièmement, en cas d'alternance entre deux formes d'une même paire, l'écart entre les nombres de pages retournées permet d'affiner l'analyse, d'une part, en comparant cet écart pour chaque paire (section 3.2 ci-après) et, d'autre part, en focalisant l'analyse qualitative sur certaines paires, quand les écarts de fréquence entre les deux formes sont très larges ou au contraire très restreints (section 3.3). Un écart important des fréquences entre deux formes permettra de discriminer l'un des deux préfixes, et un écart restreint montrera en revanche une importante interchangeabilité des préfixes du même groupe.

## 3 Résultats

### 3.1 Ampleur de l'alternance des préfixes

En comparant le nombre de paires de formes pour lesquelles Golf ne rendait de réponse que pour une seule forme préfixée (comme *monopoutre* (796 occ.)/*unipoutre* (0 occ.)) avec les paires de formes pour lesquels des réponses étaient retournées pour deux formes préfixées (comme *monosectoriel* (70 occ. )/*unisectoriel* (44 occ.)), nous pouvons estimer les proportions dans lesquelles il y a une alternance ou non, ce que résume la figure 1 ci-dessous. Dans cette figure, le premier préfixe donné correspond au cognat[3] de la forme italienne à partir de laquelle le mot a été généré. Le second est le préfixe « alternant ».

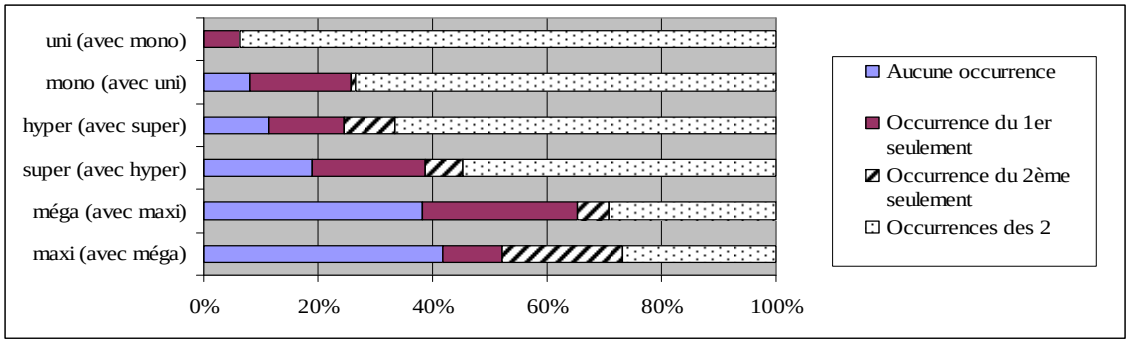


Figure 1 : Occurrences retournées pour chaque paire alternante

[3] C'est-à-dire les préfixes qui sont formellement très similaires entre les deux langues.

Trois types de commentaires peuvent être faits sur ce type de résultat. Premièrement, concernant l'absence d'occurrence, il est difficile de tirer des conclusions, tant les phénomènes d'absence dans des ressources aussi peu contrôlées que l'Internet ne représentent pas des données fiables. D'un point de vue traductionnel, il faut néanmoins remarquer l'absence importante de traductions (*aucune occurrence*, dans la figure 1) quelles que soient les paires en *méga*/*maxi*. Le deuxième type de commentaire porte sur les cas où il n'y a pas d'alternance entre deux formes possibles, mais où un seul mot préfixé est retrouvé (*occurrences du 1*$^{er}$ ou *du 2*$^{ème}$ *seulement*, dans la figure ci-dessus). Dans ce cas de figure, c'est avant tout la forme « cognat » de l'italien qui est retrouvée en français, avec une exception notable des formes italiennes en *maxi* qui sont majoritairement « traduites » en *méga* (près de 30 %, contre seulement 10 % en *maxi*). Enfin, le dernier commentaire porte sur le nombre important de cas où les deux formes ont été retournées (*occurrences des 2*, dans la figure 1). Ces cas interchangeables représentent plus de 70 % pour les alternances *mono/uni* dans les deux sens, et tout de même plus d'un quart pour les préfixés en *méga/maxi*.

Pour la classe des préfixes de pluralité indéterminée, qui contient donc trois préfixes alternants, les résultats sont encore plus frappants, comme le montre la figure 2 ci-dessous :

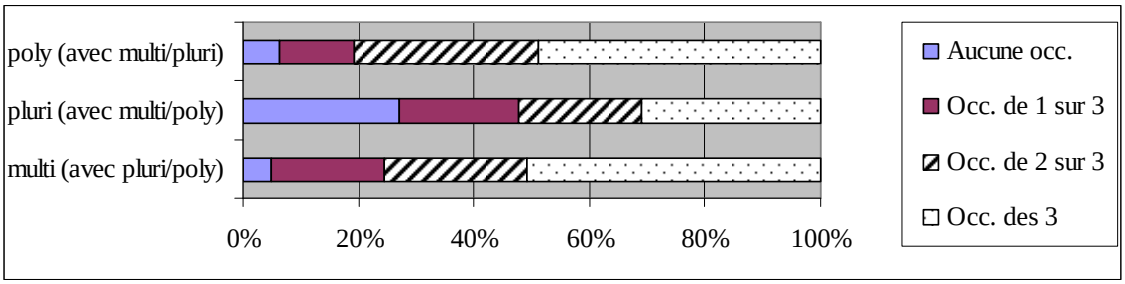


Figure 2 : Occurrences retournées pour chaque triplet alternant

Si l'on tient compte de l'alternance entre deux ou entre trois préfixes, nous constatons qu'elle est possible dans plus de 50 % des cas. Ces deux graphiques confirment donc qu'il existe un alternance/interchangeabilité potentielle entre les préfixes d'une même valeur sémantique. Mais à nouveau, la qualité des données extraites de l'Internet peut être sujette à caution, et des précautions d'analyse s'imposent. Dans la suite, nous proposons deux méthodes d'analyse permettant de regarder de plus près ces paires alternantes et d'affiner ces résultats.

## 3.2 Plusieurs types d'alternance : le rapport de fréquence

Les données extraites dans l'expérience précédente peuvent être analysées de différents points de vue. Nous proposons ici de classer et de comparer les différentes paires de mots alternants, pour pouvoir qualifier de manière plus précise cette alternance et voir s'il s'agit d'une interchangeabilité réelle, ou si la présence des deux formes relève plutôt de l'accident et serait due à la volatilité des données de l'Internet. Pour cette partie, nous travaillons sur les listes de mots alternants en français, quel que soit le mot italien desquels ils proviennent. Ainsi, pour des raisons de validité méthodologique et de simplicité de traitement, nous considérons dans cette section l'alternance dans sa dimension monolingue uniquement.

### *3.2.1 Calcul d'un score d'alternance pour chaque paire*

Pour pouvoir comparer les paires d'alternances et leur fréquence, il est intéressant de calculer, pour chaque forme un « score d'alternance », qui est obtenu en divisant le nombre d'occurrences d'une forme préfixée avec un préfixe par le nombre total d'occurrences préfixées par les deux préfixes. Par exemple, pour la paire *multiconfessionnel/pluriconfessionnel*, le robot Golf a retourné respectivement 8310 et 1100 occurrences. Nous attribuons donc à cette paire le score de 0.883103082, obtenu par division du nombre d'occurrences de la première forme, par le nombre total d'occurrences des deux formes. Nous obtenons ainsi des données comparables pour chaque paire d'alternance, située sur une *échelle d'alternance* entre 0 et 1. Plus le rapport est proche de 0.5, plus les deux formes sont aussi fréquentes l'une que l'autre et plus nous avons affaire à une interchangeabilité très fortement marquée.

## 3.3 Répartition sur les échelles d'alternance

Les rapports d'alternance ainsi obtenus sont ordonnés, ce qui permet d'obtenir une représentation de ces paires en fonction de la préférence pour l'un ou l'autre des préfixes, ou en fonction de leur interchangeabilité. Au milieu de l'échelle, nous avons individualisé une fenêtre d'analyse autour du rapport 0,5 (entre les rapports 0,4 et 0,6), dans laquelle se situent les paires d'alternance où la préférence pour l'un ou l'autre des préfixes n'est quasiment pas marquée et donc où les préfixes semblent interchangeables. Aux extrémités de l'échelle, les préférences sont plus nettement marquées[4]. Ainsi, les échelles que nous présentons dans la figure 3 ci-dessous montrent les proportions de paires où l'alternance est « parfaite » et celles où il existe une préférence plus marquée pour l'un ou l'autre des formes préfixées.

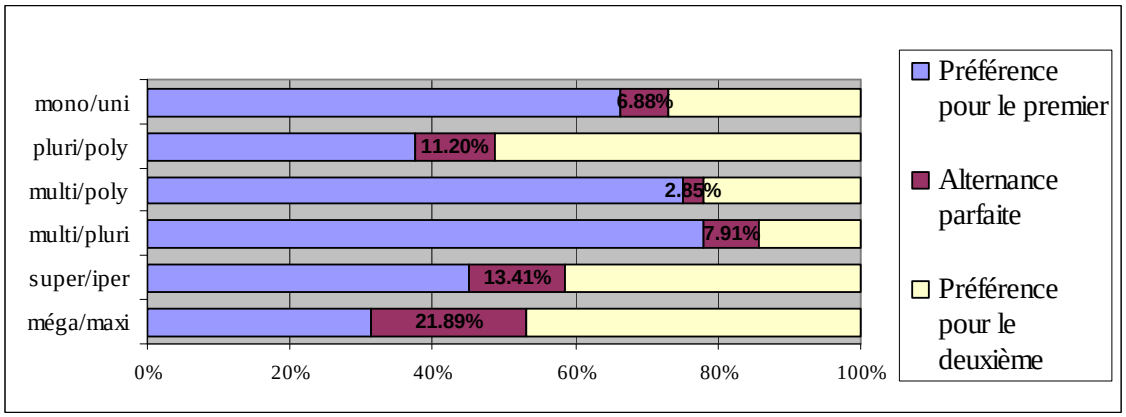


Figure 3 : Echelles d'alternance

De nombreux enseignements peuvent être tirés d'un tel graphique. Premièrement, l'alternance « parfaite », ou interchangeabilité, est la plus importante dans les deux paires de préfixes évaluatifs. Ce n'est pas étonnant, étant donné que ces préfixes sont très souvent décrits comme très instables, et n'apportant pas une « instruction » sémantique précise. La sélection de l'un ou l'autre des préfixes dépend donc de nombreux autres facteurs que de la seule règle

[4] Pour des questions de représentativité, les paires dont les deux formes n'étaient trouvées qu'une fois par Golf ont été exclues.

de préfixation (pression lexicale, mode, autres paramètres pragmatiques). Deuxièmement, dans les paires comportant *multi,* c'est ce dernier qui a la préférence, ce qui est sans doute le reflet de sa forte productivité en français. Cette même tendance se retrouve dans la paire *mono/uni,* où le premier apparaît comme prédominant. Troisièmement, la quasi-absence d'alternance « parfaite » entre *multi* et *poly* montre bien la spécificité de ce dernier préfixe. En revanche, une alternance semble possible entre *poly* et *pluri.*

## 3.4 Étude qualitative sur les paires d'alternance

En classant ces paires alternantes en fonction de leur écart de fréquence, nous obtenons un échantillon de paires de mots où les deux formes sont quasiment aussi fréquentes (c'est la fenêtre d'alternance parfaite du graphique ci-dessus). Il est donc particulièrement intéressant de regarder de plus près quel type de base permet la double préfixation, sans montrer une préférence pour l'une ou pour l'autre des formes. A l'inverse, les groupes les plus « extrêmes » contiennent des paires où l'existence d'une alternance entre deux formes relève davantage de l'accident, dû à l'extrême volatilité de la ressource textuelle utilisée. Par exemple, sur l'échelle *méga/maxi,* à l'extrémité de la préférence pour *méga,* nous avons trouvé la paire *mégawatt* (40500)/*maxiwatt* (41). Ces cas montrent une alternance entre une forme fortement lexicalisée, et une forme qui relève du jeu de mots, ou d'un nom propre. En les regardant de plus près, et en comparant systématiquement leur contenu, nous pouvons extraire les caractéristiques qui semblent favoriser la préfixation.

### *3.4.1 Alternance dans la classe des préfixes de quantité*

En analysant la fenêtre d'alternance « parfaite » des paires en *multi/pluri* et des paires *pluri/poly,* nous constatons une présence importante d'adjectifs, comme[5] *multicommunaux*/*pluricommunaux* (47/69), *multicentenaire*/*pluricentenaire* (1220/1410), *plurispécialisé*/*polyspécialisé* (10/8), *plurifonctionnel*/*polyfonctionnel* (4710/3710) et quelques bases nominales, mais utilisées majoritairement comme adjectifs : *multipathologie*/*pluripathologie* (65/64), *plurifonction*/*polyfonction* (142/110) *multifonds*/*plurifonds* (712/577). La même prépondérance des adjectifs dans la fenêtre d'alternance parfaite se retrouve pour la classe des préfixes de quantité unique, par exemple *monoconfessionnel*/*uniconfessionnel* (49/66), *mononational*/*uninational* (281/255).

A l'inverse, dans les fenêtres « extrêmes », nous avons constaté un grand nombre de bases nominales là où la préférence était fortement marquée pour la préfixation en *multi* et en *pluri,* comme par exemple les paires *multirécidivistes*/*plurirécidivistes* (29 500/2), *multicouche*/*pluricouche* (109 000/3), *plurithérapie*/*polythérapie* (325/3460), *plurivaccination* /*polyvaccination* (3/39). Il semblerait donc que la base adjectivale soit plus propice à l'alternance dans le groupe des préfixes de quantité.

### *3.4.2 Alternance dans la classe des préfixes évaluatifs*

Sur l'échelle *super/hyper*, dans les trois fenêtres d'analyse (l'alternance parfaite et les deux extrêmes) nous avons constaté que le nombre de bases adjectivales était très différent. En

[5] Les nombres entre parenthèses indiquent la fréquence retournée par Golf pour les deux formes.

effet, pour la fenêtre où les formes en *super* sont les plus fréquentes, il n'y a que 10 % d'adjectifs, le reste étant composé de bases nominales, comme *superproduction*/*hyperproduction* (132000/13600), *supercatégorie*/*hypercatégorie* (256/19), *superrégions*/*hyperrégions* (81/6), *superperformance*/*hyperperformance* (2080/122). En revanche, parmi les paires « alternantes » et les paires privilégiant la forme en *hyper*, les bases adjectivales représentent plus de la moitié des effectifs.

Ce grand nombre de noms construits avec *super* est sans doute lié au fait que ce préfixe était à l'origine utilisé dans des règles de construction d'un sens « locatif » (que l'on ressent encore dans *superstructure*, par exemple), ce qui n'a jamais été le cas pour *hyper*.

Ainsi, *hyper* semble favoriser les adjectifs, avec lesquels il peut parfois entrer en alternance avec *super*. En revanche, devant une base nominale, c'est avant tout *super* qui devrait être favorisé. En revanche, l'échelle d'alternance *méga/maxi* ne nous apporte pas de tendance aussi affirmée que pour les préfixes précédents.

## 4 Quelles contraintes pour la génération automatique ?

De l'étude de ces fenêtres, nous avons mis à jour un certain nombre de contraintes, qui, si elles ne sont pas forcément « morphologiques », témoignent en tout cas d'une certaine préférence dans la formation des mots nouveaux. Ainsi, pour générer une forme nominale de pluralité indéterminée, nous favorisons l'emploi de *multi*, alors que pour générer des formes adjectivales de cette même classe, nous maintenons la possibilité de générer les trois formes possibles, comme le montrent les deux règles ci-dessous :

**IT** : $RCL_{PLUR\_INDET}$ $[X]_N \rightarrow$ *multi*|*pluri*|*poli* $[X]_N \rightarrow$ **FR** : $RCL_{PLUR\_INDET}$ $[Y]_N \rightarrow$ *multi* $[Y]_N$

**IT** : $RCL_{PLUR\_INDET}$ $[X]_A \rightarrow$ *multi*|*pluri*|*poli* $[X]_A \rightarrow$ **FR** : $RCL_{PLUR\_INDET}$ $[Y]_A \rightarrow$ *multi/pluri/poly* $[Y]_A$

Pour les préfixes évaluatifs, nous avons mis à jour la forte tendance à la sélection de *super* pour les bases nominales, sans doute à cause de l'origine sémantique de ce préfixe. Il sera alors plus sûr que les bases nominales des préfixations évaluatives soient toujours construites avec *super* en français. Pour les bases adjectivales en revanche, les deux préfixes semblent possibles, et il convient alors de maintenir les deux possibilités :

**IT** : $RCL_{QUAL+}$ $[X]_N \rightarrow$ *super*|*iper* $[X]_N \rightarrow$ **FR** : $RCL_{QUAL+}$ $[Y]_N \rightarrow$ *super* $[Y]_N$

**IT** : $RCL_{QUAL+}$ $[X]_A \rightarrow$ *super*|*iper* $[X]_A \rightarrow$ **FR** : $RCL_{QUAL+}$ $[Y]_A \rightarrow$ *super/hyper* $[Y]_A$

Enfin, maintenant que nous avons évalué la possibilité de l'alternance, nous connaissons les cas où elle est réellement possible et où les deux formes peuvent être générées sans risque d'erreur. Nous pouvons à présent envisager soit de laisser pour l'étape de post-édition le choix entre les deux formes, soit d'utiliser des méthodes de sélection plus « simples » (basées sur des indices de productivité ou des principes de cognaticité) mais désormais sans prendre le risque de générer une mauvaise traduction.

## Conclusion et discussion

L'interchangeabilité potentielle entre affixes est un vrai défi pour la génération automatique de néologismes construits. L'étude extensive des constructions morphologiques sur de larges ressources textuelles nous a permis de rendre compte de l'ampleur de cette interchangeabilité et nous a permis d'individualiser certaines tendances. La connaissance de ces dernières permet de faire une « sélection » pour générer un néologisme construit. Ce genre d'étude permet

également d'améliorer les connaissances sur certaines règles de préfixation morphologique, et de confirmer certaines assertions théoriques.

En outre, la méthodologie proposée pour les besoins de notre projet de traduction automatique pourrait être réemployée pour d'autres études sur l'alternance. En effet, étant donné qu'elle est basée sur des mots « possibles » issus de mots construits dans une autre langue proche, elle permet de travailler sur des données beaucoup plus restreintes que si l'on envisageait tous les mots possiblement préfixables avec ces préfixes. De plus, si une telle méthode fournit des informations avant tout sur les fréquences d'utilisation, elle laisse également entrevoir certaines tendances de productivité.

## Références

AMIOT D. (2005). Plusieurs vs poly-, pluri- et multi- *La quantification côté déterminants et côté préfixes*. N. Flaux et D. Amiot, Artois Presse Université.

BAAYEN H. et LIEBER R (1991) Productivity and English derivation *Linguistics* 29(5), 801-43.

BARONI, M., BERNARDINI S, COMASTRI F, PICCIONI L., VOLPI A., ASTON G. MAZZOLENI M, (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. Actes de *LREC 2004* Lisbon. 1771-1774.

BOUILLON P., LEHMANN S, MANZI S., PETITPIERRE D. (1998). Développement de lexiques à grande échelle. Actes du *Colloque des journées LTT de TUNIS*, Tunis, 71-80

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

FRADIN B. (2003) *Nouvelles approches en morphologies*. Paris, Puf.

HATHOUT N., NAMER F., PLENAT M., TANGUY L. (à paraître). La collecte et l'utilisation des données en morphologie. *Aperçus de morphologie du français*. B. Fradin, F. Kerleroux et M. Plénat (éd.), Presses Universitaires de Vincennes.

HATHOUT N., PLENAT M, TANGUY L. (2003) Enquête sur les dérivés en -able *Cahiers de Grammaire* 28 49-90.

MONTERMINI, F. (à paraître). La préfixation évaluative. *Aperçus de morphologie du français*. K. F. Fradin B., Plénat M. Vincennes, Puv.

THOMAS C. (2008). *Google Online Lexical Frequencies User Manual (Version 0.9.0)* http://www.craigthomas.ca/docs/golf-0.9.0-manual.pdf, consulté 04.02.2008