La collecte et l'utilisation des données en morphologie

Nabil Hathout, Fiammetta Namer, Marc Plénat et Ludovic Tanguy

1. Introduction

L'évolution de la nature et de la quantité des données utilisée en morphologie et plus généralement en linguistique est directement liée au développement et aux progrès de l'informatique. Le développement des capacités de stockage a en effet permis la constitution de très gros corpus. Parallèlement, l'augmentation des capacités de calcul a rendu possible le traitement de grandes quantités de textes dans un temps acceptable. Si les premiers corpus en français de textes informatisés étaient essentiellement composés d'œuvres littéraires comme la base Frantext, les chercheurs ont par la suite disposé d'archives électroniques de journaux. Plus récemment, l'avènement du Web a fourni aux linguistes un corpus d'une taille jamais encore atteinte. En particulier, Internet permet la diffusion de corpus gratuits de plus en plus nombreux, présentant toute une palette typologique :

- dictionnaires, encyclopédies, lexiques, thésaurus, bases de données, bases de connaissances ;
- littérature ;
- journaux ;
- textes scientifiques, pédagogiques, techniques ;
- blogs, listes de discussion, forums...

Tous ces corpus permettent d'étendre et de mieux contrôler la base empirique qui sert aux études de la langue, du lexique et de la morphologie. Ils sont à l'origine des évolutions que l'on observe dans le rapport aux données en linguistique, et en particulier en morphologie : l'introspection et la consultation de dictionnaires qui fournissaient jusqu'ici les exemples attestés illustrant une règle sont relayés par la masse de formes et d'unités lexicales disponibles dans les corpus. Ceux-ci en effet permettent d'avoir à disposition un nombre considérable de néologismes, renouvelés sans cesse avec le développement constant du fonds documentaire indexé sur le Web, et aussi variés que le sont les domaines scientifico-techniques et les niveaux de langue dans lesquels ils sont créés. Mais bien entendu, cette créativité lexicale à portée de main a pour effet principal de remettre en question la justesse de certains principes, de modifier les contraintes pesant sur les procédés de construction lexicale, de faire passer certaines exceptions au rang de régularité…

2. Collecter des données sur le Web

Traditionnellement, les morphologues travaillaient sur des ensembles d'attestations compilées à partir de dépouillements manuels d'œuvres lexicographiques ou littéraires.
L'informatisation des dictionnaires et des textes rend aujourd'hui possible l'automatisation de cette collecte. Quelques secondes suffisent pour trouver l'ensemble des adjectifs en *-able* traités dans le *T.L.F.i* (*Trésor de la Langue Française informatisé*), et il ne faut que quelques minutes pour extraire les formes en *-ette* d'une dizaine d'années d'archives du *Monde*.

À côté de ces corpus « traditionnels », les linguistes utilisent de plus en plus couramment le Web pour trouver des attestations de lexèmes construits, de tournures syntaxiques… L'exploration du Web peut se faire manuellement en interrogeant un moteur de recherche comme Google, Yahoo... Il s'agit alors de découvrir des occurrences de mots possibles que l'on prédit en s'appuyant sur l'intuition. On peut par exemple trouver de cette façon plusieurs adjectifs comme *anti-inflammable*, *anti-explosible*...[1] La collecte d'attestations sur le Web peut également être effectuée au moyen d'outils comme WaliM (Namer, 2002) ou Webaffix (Tanguy & Hathout, 2002 ; Hathout & Tanguy, 2002). Ces deux outils sont très proches par leurs objectifs et par leur fonctionnement général. Tous les deux génèrent des requêtes par programme, les soumettent à un moteur de recherche puis effectuent différents nettoyages sur les résultats ramenés.

2.1. Le Web comme corpus

Le corpus sur lequel opèrent WaliM et Webaffix est donc le Web, et non un corpus classique. Comme le note G. Grefenstette (1999), nombre de linguistes peuvent, à juste titre, se montrer réticents dans l'emploi du Web comme source d'attestations, étant donné l'impossibilité technique de caractériser les pages sur le plan du domaine, du genre, du statut de l'auteur, de la validité du contenu... Il n'en reste pas moins que le Web constitue la masse textuelle la plus importante accessible pour une recherche linguistique.

Des projets comme WebCorp[2] mettent la technologie de base d'un concordancier à l'échelle du Web, en se fiant aux moteurs de recherche génériques. Ces moteurs sont actuellement le seul véritable moyen d'accès aux pages. (Des systèmes de parcours et d'indexation spécifique sont en cours de développement, même s'ils ne pourront de toute façon pas prétendre à l'exhaustivité d'un Google et autre Yahoo.) WaliM comme Webaffix n'échappent pas à cette règle, et opèrent comme tous les autres systèmes sur la partie du Web (dont la proportion est d'ailleurs inconnue) accessible par ces moteurs, et donc sur un sous-ensemble variable avec le temps, et sans critère de sélection identifiable.

Dans le cadre de la morphologie, le but affiché des études sur corpus qui recourent au Web est de constituer des collections d'exemples. Il existe cependant un biais, dont il faut avoir conscience : la nature illusoire du Web comme un corpus de « langue générale ». Si la variété des domaines abordés, et donc des sous-langages de spécialité représentés peut paraître suffisante, nous n'avons pas d'idée claire de la représentation de chacun de ces domaines.

Les études quantitatives sur le Web sont également sujettes à caution. Le manque de contrôle sur les documents, et l'absence de caractérisation globale de ce pseudo-corpus pose notamment des limites à la simple notion de fréquence. Bien qu'étant une information directement accessible par n'importe quel moteur de recherche, le nombre de documents renvoyés en réponse à une requête n'en est qu'une grossière approximation, et ce pour plusieurs raisons. Tout d'abord, l'unité du Web indexé est le document, et non pas l'occurrence lexicale : le nombre de documents gomme donc les répétitions d'une occurrence au sein des documents, et la notion d'hapax est elle-même rendue floue. Ensuite, le Web est le lieu par excellence de la reprise, de la citation et du plagiat. Des documents strictement identiques, mais accessibles à des adresses différentes seront considérés comme autant de réponses, augmentant artificiellement la fréquence d'un terme qui s'y trouverait. Le même phénomène s'applique à des citations partielles ou totales de documents.

[1] Ces adjectifs constituent des contre-exemples à l'hypothèse de (Fradin 1997, p. 100) selon laquelle les formes *antiA* où *A* est un adjectif en -*Vble* sont impossibles.
[2] http://www.webcorp.org.uk

Ainsi, la comparaison des fréquences d'unités lexicales sur le Web est à manier avec précautions. Si la distinction présence/absence (*i.e.* aucun document *vs. n* documents renvoyés) peut être prise en compte, tout comme les différences en ordre de grandeur (2 documents *vs.* 100 000 documents), les différences faiblement marquées ne sont a priori pas significatives.

La suite de la section est consacrée à Webaffix et WaliM, deux outils pour l'acquisition lexicale à partir du Web.

2.2. La boîte à outils Webaffix

Webaffix est à la fois une méthode et une boîte à outils d'enrichissement lexical à partir du Web, destinée à la création et la complétion semi-automatique de lexiques au moyen de formes dérivées morphologiquement. Il propose deux grandes fonctionnalités : la recherche de ces formes sur le Web et leur filtrage[3]. La boîte à outils contient trois composants :

1. un module de recherche par suffixe qui permet de découvrir sur le Web des formes qui correspondent à un motif[4] tel que la présence d'un suffixe graphémique donné ;
2. un composant de prédiction morphologique permettant de calculer les formes des lexèmes bases ou des lexèmes construits ;
3. un méta-moteur disposant de fonctionnalités dédiées à l'exploration lexicale du Web.

Webaffix propose deux modes de recherche de formes nouvelles. La première exploite un lexique existant ou une base de données morphologiques pour générer une liste de formes candidates au moyen du composant de prédiction morphologique puis utilise le méta-moteur pour rechercher sur le Web des attestations de ces formes. La seconde utilise le module de recherche par suffixe pour repérer sur le Web des formes qui correspondent à un schéma donné (par exemple celles qui finissent par *able*), sans aucune contrainte sur la base. Hathout *et al.* (2003) ont ainsi découvert de nombreuses formes en *-able* dont le radical n'apparaît ni parmi les verbes ni parmi les noms du *T.L.F.* et du *G.R.L.F.*

En plus des problèmes classiques en détection de formes nouvelles (noms propres, fautes d'orthographe, etc.) collectées par la recherche par suffixe, se pose la question du statut morphologique des résultats : s'agit-il bien de formes construites par le procédé morphologique auquel on s'intéresse ? Pour éliminer les formes qui ne sont pas dérivées, Webaffix propose deux méthodes de filtrage reposant sur la prédiction des formes des lexèmes bases. La plus simple consiste à rechercher sur le Web des attestations de ces formes. Dans ce cas, le critère pour le filtrage d'une forme comme *copolymérisable* sera l'attestation sur le Web d'une des formes du verbe *copolymériser*. Un filtrage plus strict est également disponible : rechercher des pages Web qui contiennent à la fois la forme candidate et l'une des formes de son lexème base.

2.3. Le système WaliM

WaliM est un système plus simple qui fait appel au moteur de recherche Yahoo pour trouver des attestations des mots-requêtes qui lui sont fournis en entrée. En sortie, ces candidats sont répartis en trois groupes : ceux qui n'ont aucune attestation sont potentiellement des mots construits « impossibles » ; ceux qui apparaissent dans un nombre suffisant de pages Web

[3] Par « filtrage », nous entendons l'élimination d'une partie des réponses erronées ramenées par le moteur de recherche.
[4] Ce module n'est plus opérationnel. Il faisait appel au moteur AltaVista, qui suite à son rachat par Yahoo, ne permet plus l'utilisation des caractères joker nécessaires à la recherche par suffixe.

sont enregistrés comme corrects ; ceux dont le nombre des attestations est inférieur à un seuil donné sont considérés comme moins fiables. Pour ces derniers, les résultats sont filtrés en supprimant les caractères non alphanumériques dans les mots de la première page ramenée puis en vérifiant que le mot requête est bien présent parmi ces derniers.

3. Morphologie extensive

L'apport réel des données collectées sur le Web et dans d'autres ressources ne peut plus aujourd'hui être mis en doute. L'augmentation du nombre des données est ainsi directement à l'origine de généralisations nouvelles, de la mise au jour de faits rares... Les données en grand nombre permettent également de confirmer, d'infirmer, de nuancer des intuitions incertaines, mais aussi de concevoir des dispositifs expérimentaux susceptibles de renouveler au moins partiellement l'assise empirique de la discipline. Plusieurs expériences et études fondées sur une « approche extensive » de la morphologie illustrent clairement ces apports. Par « approche exensive », nous entendons la collecte du plus grand nombre possible d'attestations du procédé étudié en vue de faire apparaître des régularités nouvelles.

Ces études font également apparaître que le conditionnement des phénomènes morphologiques est beaucoup plus fin et la gamme des emplois beaucoup plus étendue qu'on ne pourrait le croire. Les linguistes comme les lexicographes sont souvent victimes d'effets de fréquence et surestiment leurs capacités de jugement.

3.1. Généralisations nouvelles

Les données qui ont donné lieu au plus grand nombre d'observations nouvelles sont sans doute celles qui sont rassemblées dans la base de dérivés en *-esque* constituée à l'ERSS sous la responsabilité de N. Serna. Cette base contient actuellement environ 3 000 enregistrements, chiffre qu'il convient de comparer à la petite centaine de lexèmes répertoriés dans le *Robert électronique*. À l'origine, cette base était destinée à étudier le comportement des voyelles finales des bases à finale vocalique devant un suffixe à initiale vocalique. Il s'agissait de déterminer quels contextes favorisaient le maintien de l'hiatus, la troncation de la voyelle finale, sa semivocalisation, ou encore l'insertion d'une consonne épenthétique. Il est toutefois assez vite apparu que les phénomènes de sandhi internes imputables aux contacts de voyelles étaient très loin d'être les seuls et que les accidents phonologiques obéissaient pour l'essentiel à un petit nombre de contraintes, comme par exemple les contraintes de taille ou les contraintes dissimilatives (*cf.* Plénat, ce volume ; Lignon & Plénat, ce volume).

Dérivés en *esque* sur base en *eur*. Un exemple illustre bien les progrès qu'autorise l'accumulation de données. Plénat (1997 : note 4), qui disposait d'un corpus d'environ 800 formes, observe que les bases en *-eur* perdent régulièrement leur rime finale devant *esque*. Cette hypothèse était extrêmement fragile, dans la mesure où les seuls bons exemples d'effacement de cette rime concernaient un trisyllabe (*tirailleur*, qui donne *tiraillesque*) et deux tétrasyllabes (*consommateur* > *consommatesque*, *déprédateur* > *déprédatesque*), mais un autre trisyllabe en était exempt (*enjôleur* > *enjôleresque*). Trois ans plus tard, la base s'étant étoffée de 400 ou 500 formes supplémentaires, Plénat (2000) montre, qu'en fait, les rimes constituées d'une voyelle antérieure moyenne (comme /ɛ/, /e/, /œ/ ou /ø/) et d'une consonne fixe se maintiennent en général lorsque la base fait deux ou trois syllabes (*Orwell* > *orwellesque*, *Le Pen* > *lepénesque*, *jongleur* > *jongleresque*, *coccinelle* > *coccinellesque*, *Aleikhem* > *aleikhemesque*, *enchanteur* > *enchanteresque*) ; la chute de la rime ne semble prendre effet qu'à partir de quatre syllabes, et ce d'une façon variable, puisque quelques exemples, comme *polichinellesque* ou *pantagruélesque* montrent qu'elle n'a pas toujours lieu.

L'augmentation du nombre des dérivés a ainsi remis en cause l'hypothèse initiale de M. Plénat : la chute de la rime finale de la base dans *tiraillesque* est donc un cas complètement isolé.

Conditionnements de la suffixation en *esque*. Les données actuelles de la base des dérivés en -*esque* font apparaître que :

- les bases de quatre syllabes peuvent perdre leur rime finale ou ne pas la perdre (*Polichinelle* > *polichinellesque* ou *polichinesque*, *Harry Potter* > *harrypotteresque* ou *harrypottesque*, *vétérinaire* > *vétérinaresque* ou *vétérinesque*, *ordinateur* > *ordinatoresque* ou *ordinatesque* ;
- les bases savantes à voyelles postérieures sont assez systématiquement utilisées, quand elles sont disponibles (*moteur* > *motoresque*, *professeur* > *professoresque*, *notaire* > *notaresque*).
- les bases de moins de quatre syllabes conservent normalement leur rime finale ;
- la rime finale des bases qui font deux syllabes ou plus tombe si la consonne finale est identique (le cas échéant au voisement près) à l'une des consonnes du suffixe (*Barthez* > *barthesque*, *Cherek* > *cheresque*) ;
- les bases trisyllabiques perdent (variablement) leur rime finale : *co**l**one**l*** > *colonesque*, *Be**n** Lade**n*** > *ben ladesque* (et *benladénesque*), *Tie**n**a**n**me**n*** > *tienanmesque*, *Wa**r**hamme**r*** > *Warhammesque* (et *Warhammeresque*), *In**t**erne**t*** > *internesque* (et *internetesque*). Mais il s'agit dans tous les cas de bases dont la consonne finale est représentée une autre fois, comme c'était déjà le cas dans *ti**r**ailleu**r*** (où l'on a deux /r/).

Ces différentes conditions peuvent être synthétisées comme suit : Pour qu'une rime en voyelle antérieure moyenne + consonne fixe tombe devant -*esque*, il faut (1) que la voyelle soit identique ou quasi identique à celle du suffixe et (2) que l'une au moins des trois conditions suivantes soit satisfaite :

- l'identité ou quasi-identité de la consonne avec une de celles du suffixe et une taille d'au moins deux syllabes (cf. *barthesque*) ;
- l'identité de la consonne avec une des consonnes internes de la base et une taille d'au moins trois syllabes (cf. *ben ladesque*) ;
- une taille soit au moins de quatre syllabes (cf. *vétérinesque*).

Ainsi, aux forces conservatrices qui tendent à préserver l'intégrité de la base et du suffixe s'opposent deux forces délétères qui tendent à atténuer l'éventuel caractère marqué de la forme résultante : les tensions dissimilatives et la tendance à l'accourcissement des formes trop longues. Cette conclusion restait hors de portée tant que les généralisations qui la fondent n'étaient pas extractibles de données numériquement trop faibles et ces généralisations elles-mêmes ne se précisent que peu à peu, au fur et à mesure que croît la quantité des données.

Plasticité sémantique des dérivés en *able*. L'augmentation de la quantité de données soumises à observation est également déterminante quand on s'intéresse aux dimensions catégorielle et sémantique de la description morphologique. Sur ce point, l'étude de la dérivation en -*able* de Hathout, Plénat & Tanguy (2003) donne d'assez bonnes indications sur le type de progrès que l'on peut attendre de l'accumulation de données nouvelles. Hathout *et al.* (2003) ont constitué pour cette étude une base de près de 5 000 adjectifs en utilisant successivement les deux méthodes de collecte de Webaffix et, pour certains dérivés, ont analysé systématiquement les emplois présents sur le Web.

Les dérivés en *-able* sont considérés d'ordinaire comme des déverbaux de sens « passif » (autrement dit, leur nom recteur correspondrait à un objet direct du verbe de base ou à un patient, selon que l'on considère la syntaxe ou la structure argumentale). Les récoltes montrent que, si c'est souvent le cas, le nom recteur peut aussi représenter un grand nombre d'autres types de participants au procès. Cette plasticité sémantique peut être illustrée en passant en revue les noms recteurs d'un dérivé comme *pêchable*. Sont pêchables en premier lieu les poissons, mollusques et autres batraciens. Mais le sont aussi certains lieux : les berges, ponts, digues... où peuvent se poster les pêcheurs et les rivières, étangs, courants... où circulent leurs proies. Les saisons, les jours et les circonstances atmosphériques peuvent eux aussi être ou non pêchables, selon que la pêche est ouverte ou non, suivant qu'il fait beau ou mauvais... On trouve aussi des contextes où *pêchable* est prédiqué du matériel de pêche (mouches ou nylon, par exemple). Enfin, ce ne sont pas seulement les participants du procès, mais aussi les propriétés de ces participants qui peuvent être qualifiées de pêchables ou d'impêchables : la taille des poissons ou la profondeur d'un cours d'eau peut être dite impêchable. Seuls les pêcheurs, en tant qu'agents, ne peuvent apparemment pas être pêchables. La collecte d'un nombre suffisamment important d'adjectifs présentant des emplois « circonstanciels » rend intelligibles des adjectifs comme *abordable* jusqu'ici considérés comme des exceptions. Le recours au Web et aux corpus donne ainsi accès à de grandes quantités de contextes pour chaque mot attesté qui permettent de décrire les procédés morphologiques avec une précision beaucoup plus grande et mieux prendre en compte la diversité de leurs sens observables.

L'approche extensive de la morphologie permet ainsi de réaliser des avancés notables, mais ces dernières ont un coût non négligeable : une base doit en effet être constituée pour chaque procédé étudié, ce qui demande un temps considérable. Si la collecte est désormais très rapide, la validation des données collectées exige en effet un long travail philologique. L'utilisation de corpus et du Web introduit ainsi de nouvelles pratiques en morphologie.

3.2. Faits rares

Les données massives constituent des ressources supplémentaires qui sont à la disposition du linguiste. Elles peuvent compléter utilement l'intuition mais ne peuvent ni la remplacer ni la suppléer. Il arrive ainsi fréquemment que, faute d'attestations, la validité d'intuitions fines sur la grammaticalité d'une forme ou d'un emploi ne puisse pas être établie. Les grands corpus, le Web en particulier, peuvent alors venir au secours du jugement de grammaticalité et permettent souvent de trouver des données que parfois une vie entière de lectures n'aurait pas permis de dénicher.

Une remarque de F. Yvon (s. d.) illustrer clairement ce type d'utilisation du Web. Dans un passage où il s'intéresse à la grammaticalité en matière de morphologie, Yvon signale comme problématique l'adverbe *charmamment*, pourtant construit sur *charmant* comme *galamment* l'est sur *galant*. (On sait qu'anciennement, l'adjectif *galant* était épicène et respectait donc la règle d'accord avec le nom féminin *mente* ; il est depuis devenu de règle que les adjectifs en *-ant* et *-ent* donnent des adverbes en *-amment* et *-emment* ; les grammaires ne citent sur ce point que trois exceptions : *lentement*, *présentement* et *véhémentement*.) Nous avons le sentiment qu'Yvon n'est pas assez catégorique et que *charmamment* est totalement agrammatical : pour nous, la seule forme concevable est *charmantement*. Mais nous avons trop intérêt à ce qu'il en soit ainsi pour être sûrs de notre jugement : s'il s'avérait que les adjectifs se terminant par /mɑ̃(t)/ donnent des adverbes en /-mɑ̃təmɑ̃(t)/ et non en /-mamɑ̃(t)/, comme permet de le croire aussi *véhémentement,* cela pourrait être un nouvel exemple de dissimilation. Nous avons donc interrogé le Web une nouvelle fois à l'aide de Webaffix. Il ressort de l'examen des données que celles-ci ne contiennent aucun bon exemple d'adverbe se

terminant par *-mamment* ou *-memment*. *Charmamment* n'est attesté qu'une fois, dans une liste de discussion où le locuteur reconnaît peu après qu'il aurait dû écrire *charmantement*. Cette dernière forme en revanche apparaît dans une dizaine de bons exemples, et l'on trouve aussi un exemple de *aimantement*[5]. Une recherche rapide dans Frantext nous a appris ensuite qu'Albert Cohen a lui aussi utilisé *charmantement* dans *Mangeclous*. La question n'est pas tranchée, mais on doit maintenant considérer sérieusement l'hypothèse que, dans le cas des adverbes en *-ment*, la tendance à éviter les répétitions aboutit aux choix d'un thème inattendu.

3.3. Vers une morphologie expérimentale

Les observations qui fondent les travaux évoqués ci-dessus peuvent être reproduites, soit sur les mêmes données si elles ont été mises à la disposition du public, soit sur des données comparables, dans la mesure où leur mode d'obtention a été décrit. Et les prédictions peuvent être testées systématiquement sur des données nouvelles.

Prenons un exemple supplémentaire. Dal & Namer (2004) ont monté une expérience sur l'« échangisme entre bases » devant le suffixe *-ité,* en partant du constat que, bien que ce suffixe exige le plus souvent une base adjectivale, on trouve assez souvent dans cette position des toponymes là où l'on attendrait des gentilés : *ivoirité*, par exemple, est beaucoup plus fréquent que *ivoirianité*, et *portugalité* supplante entièrement, semble-t-il, *portugaisité*. Afin de déterminer la répartition de ces deux types de bases, les auteures ont réuni une centaine de noms de pays et de régions avec leurs variantes (ex. *Chine~sin-* ; *Danemark~dan-*) ainsi que les adjectifs ethniques correspondants, dérivés ou non (ex. *Hongrie*, *hongrois*, *magyar*). Elles ont ensuite construit sur ces formes tous les dérivés en *-ité* phonologiquement vraisemblables, puis, à l'aide de WaliM, déterminé et quantifié la présence sur le Web de chacune de ces formes-candidates.

Les résultats, frappants, opposent deux classes de suffixes : quand l'adjectif ethnique comporte les suffixes *-ain*, *-ien* ou *-éen*, c'est sa variante « savante » qui, en général, est retenue comme base du dérivé en *-ité*, de préférence au nom de pays (*marocan-ité*, *italian-ité*, *coréan-ité*) ; quand, en revanche, le gentilé comporte les suffixes *-ois* ou *-ais*, c'est ou bien un adjectif supplétif (*magyar-ité*) ou bien le toponyme (ou encore une variante présuffixale de celui-ci) qui sert systématiquement de base (*franc-ité*, *dan-ité*, *sin-ité*). Dans les autres cas de figure, les auteures ne discernent aucune régularité. Malgré ce qui vient d'être dit, il existe une classe de gentilés en *-ien* (ou *-éen*) qui, selon Dal & Namer (2004) ne pourraient pas servir, du moins tels quels, de base à une suffixation en *-ité* : ceux dans lesquels le suffixe est précédé de /n/ : *Lusitanie* donne *lusitanité* et non *lusitanianité*, et l'on a de même, selon les auteures, *estonité*, *iranité*, *jordanité*, *mauritanité*, *méditerranéité*, *palestinité*, *ukrainité*.

Ce cas de dissimilation préventive nous a paru assez intéressant pour que nous vérifiions, quelque mois après l'expérience initiale, la stabilité de ce résultat sur le Web. Tous les exemples fournis par Dal & Namer (2004) ont été confirmés (à ceci près que nous avons rencontré quelques attestations de *palestinianité*). Nous avons en outre trouvé *arménité*, *étasunité*, *ghanéité*, *guinéité*, *macédonité*, *slovénité*, au lieu de *albanianité*, *arménianité*, etc. Enfin, pour déterminer si la crainte d'une répétition de /n/ pouvait aboutir à la disparition du suffixe adjectival devant d'autres suffixes nominaux, nous avons mené une recherche du même type avec le suffixe *-itude*, et nous avons trouvé *arménitude*, *bosnitude*, *calédonitude*, *estonitude*, *macédonitude*, *slovénitude*, *ukrainitude* (et même *bourguignitude* !), sans jamais rencontrer les *arménianitude*, *bosnianitude*, ou *bourguignonitude* attendus.

[5] Nous ne pouvons nous faire de mal car nous sommes beaux et vous savez pourquoi....
Aimantement vôtre,
http://www.20six.fr/breizhie/archive/2004/11/

Le recours à l'intuition est lui aussi une sorte d'expérience. Mais c'est, en quelque sorte, une expérience privée et souvent incertaine, qui n'est pas remise en cause par des intuitions ultérieures ou extérieures. La répétibilité et la netteté d'expériences comme celle qui vient d'être décrite emporte la conviction. On peut préférer *palestinianité* à *palestinité*, mais on ne peut mettre en doute l'existence de la dissimilation chez un grand nombre de locuteurs.

4. Sémantique des composés néoclassiques

Les différentes études présentées ci-dessus portent toutes sur la morphologie dans la « langue générale ». Ainsi, aucune restriction particulière n'est imposée sur les domaines ou les genres des documents dans lesquelles les exemples sont collectés. Cette absence de contraintes sur les sources facilite la constitution de bases morphologiques de grande taille comme celles des dérivés en *-esque* ou *-able*. Mais l'augmentation du nombre des données est tout aussi fructueuse pour l'étude de procédés morphologiques dans des corpus mieux contrôlés. Les travaux de F. Namer sur la composition dite néoclassique illustre parfaitement cette approche. En étudiant de corpus spécialisés de grande taille, elle a mis en lumière des propriétés sémantiques et structurelles de lexèmes construits par cette composition, que les études théoriques antérieures n'ont pas pu découvrir, faute d'accès à ces données.

Ces corpus ont été réunis dans le cadre du projet UMLF[6], à partir de documents électroniques de toutes sortes : littérature scientifique, comptes-rendus hospitaliers anonymisés, lexiques, thésaurus, etc. Ils totalisent environ 12 millions d'occurrences. Leur vocabulaire comporte quelque 209 000 lexèmes, dont une grande partie est morphologiquement complexe.

4.1. Hypothèses théoriques

L'analyse sémantique de la composition morphologique, quel que soit son type (populaire *versus* néoclassique, selon le rapport catégoriel des éléments qui interviennent) doit prendre en compte le rapport de sens entre l'input (les éléments) et l'output (le composé), comme dans tout procédé de formation de lexèmes. Mais, de par la nature même de ce procédé, cette analyse dépend également du rapport de sens observable entre les éléments eux-mêmes.

Deux familles de sens sont identifiables en fonction du rapport entre les éléments du composé. On parle de type « additif », quand le sens du composé est la conjonction du sens de ses parties (l'adjectif *buccodentaire* fait référence à la fois aux dents et à la bouche), en opposition au type « déterminant-déterminé » caractérisant les noms ou adjectifs composés où l'apport sémantique des deux parties du composé n'est pas de même nature. En composition populaire, le composant déterminant, ou sémantiquement recteur, est placé à gauche (*jupe-portefeuille, allume-cigare*) ; en composition néoclassique il est placé à droite (*thalassothérapie*, *arachnophobe*). Le type déterminant-déterminé est lui-même subdivisé en deux sous-classes, selon que le composé entretient, ou pas, une relation d'hyponymie avec l'un de ses composants. Dans le premier cas, le composé appartient au même domaine conceptuel que le composant sémantiquement recteur (une jupe-portefeuille est une sorte de jupe, une thalassothérapie, une thérapie particulière). Dans le second cas, les deux termes font référence à des entités appartenant à des domaines conceptuels distincts (un allume-cigare n'est pas un type de procès, mais fait référence à une sorte de briquet, arachnophobe ne désigne pas un état psychologique (celui de 'craindre'), mais la propriété de ce(lui) qui craint les araignées). Le sens d'un composant de type déterminant-déterminé, en d'autres termes, se définit à partir du

[6] Le projet ACI UMLF « Lexique médical francophone unifié », a été financé par le MNERT de 2002-2004, et piloté par P. Zweigenbaum (INSERM), (*cf.* Zweigenbaum *et al.* 2005).

type de relation qu'entretient son composant recteur (souvent appelé « tête », *cf.* (Lieber, 1983 ; Williams, 1981)) avec son composant régi.

Comme le laissent entendre les exemples ci-dessus, le calcul du sens en composition obéit aux mêmes hypothèses théoriques selon que le lexème construit est issu d'une opération de composition populaire (*moissonneuse-batteuse, jupe-portefeuille, allume-cigare*) ou néoclassique (*buccodentaire, thalassothérapie, arachnophobe*).

L'étude du sens des composés telle que la présentent, par exemple Corbin (2004), Fradin (2000), Iacobini (2003), Villoing (2003) ou Warren (1990) sous-entend deux aspects. Tout d'abord, un composé est construit de façon binaire, et tout composé comportant plus de deux éléments relève d'un mécanisme itératif de composition. Ainsi, un *électrocardiogramme* se définit par rapport à *cardiogramme*, c'est-à-dire un moyen technique (l'électricité) pour l'obtention de ce tracé (des battements) du cœur. L'autre présupposé, qui constitue une généralisation du premier, réside dans la vision compositionnelle du sens des lexèmes construits : selon cette perspective, partagée dans les études théoriques récentes en morphologie fondées sur la régularité du lexique depuis (Aronoff, 1976), l'interaction entre affixation et composition est interprétable comme une succession d'opérations, dont l'ordre d'application est motivé linguistiquement. Le sens du lexème construit est alors dépendant de cette composition d'opérations. Ainsi, le sens de *anticéphalalgique* (‘contre le mal de tête’) est construit par le biais du préfixe *anti-* (Fradin, 1997), sur le résultat de la composition, endocentrique, de *céphal-* et *-algie*, *-ique* fonctionnant comme marqueur de classe ou intégrateur paradigmatique.

4.2. Données récalcitrantes du domaine bio-médical

Si les données propres au domaine médical obéissent en général aux hypothèses établies pour la langue générale, deux séries au moins de composés néoclassiques sont rétifs aux principes de compositionnalité et de binarité énoncés ci-dessus. Ce constat conduit à se demander dans quelle mesure le composition néoclassique constitue un procédé ressortissant à la morphologie constructionnelle du français, d’autant plus que ces mêmes séries se rencontrent dans la plupart des langues européennes.

Il s'agit tout d'abord de deux sortes de noms dont le recteur, noté X, dénote une action médicale, et dont le sens impose une structure interne ternaire, que nous noterons YZX, plutôt que la décomposition binaire $Y+ZX$ attendue reflétant la relation hyponyme/hyperonyme. Dans le premier type de noms YZX, X régit deux constituants coordonnés et non ordonnés : c'est ce qui caractérise par exemple les paires synonymes : *urétrocystoplastie* et *cysto-urétroplastie :* « reconstruction chirurgicale (*plastie*) de la vessie (*cysto*) et de l'urètre (*urétr*) ». À ce premier type correspondent de nombreux noms, exprimant des activités d'observation (*-graphie, -métrie*) ou des interventions chirurgicales (*-pexie, -rrhaphie*). On note un second type de lien entre Y, Z et X assimilable à une relation prédicat-arguments, construisant des noms bâtis notamment autour de *-stomie* et décrivant au moyen des organes décrits par Z et Y les points de départ et d'arrivée de l'ouverture pratiquée chirurgicalement (*stomie*) : ainsi une *duodénoentérostomie* permet de faire communiquer le duodénum avec une partie de l'intestin grêle. Le fait que *entéroduodénostomie* est également attesté, avec le même sens, confirme que Z et Y sont non-ordonnés et appartiennent au même niveau d'analyse.

L'autre ensemble de termes obéissant à des règles différentes de celles établies pour la langue générale est constitué des noms à la fois composés et préfixés pouvant se représenter linéairement par *prefYXie* : X est l'élément recteur, Y l'élément régi ; en association avec le

marqueur de classe *-ie*, *pref* représente l'un des trois préfixes quantificateurs désignant l'absence (*a-*), l'excès (*hyper-*) ou l'insuffisance (*hypo-*) (*acéphalogastrie, hyperchondroplasie*). Les noms correspondant au modèle *prefYXie* décrivent une pathologie qui fait intervenir *Y* et *X*. On observe que le sens de ces noms est fonction de la portée du préfixe : celui-ci s'applique à *X* (*agastrémie* est l' « absence de sang=°ém dans l'estomac=°gastr »), à *Y* (*hypofibrinémie* est l' « insuffisance de fibrine dans le sang=°ém ») ou à la conjonction de *Y* et *X* (*acheiropodie* est l' « absence de mains=°cheir et de pieds=°pod ») selon la relation que ceux-ci entretiennent. Ainsi, si *X* ou *Y* est un constituant naturel de l'autre composant, la pathologie est due à la quantité anormale du constituant (les protéines dans *hyperprotéinémie,* le sang dans *anentérémie*). Si *X* est une activité physiologique, la raison de la pathologie est l'altération de cette activité, marquée par le préfixe (*hypochondroplasie* décrit le développement : °plas insuffisant : *hypo-* du cartilage : °chondr). Enfin, dans tous les autres cas, la pathologie est due à la quantité anormale de *X* et de *Y*. Ce dernier cas d'ailleurs ne s'observe à notre connaissance qu'avec le préfixe *a-*, et désigne des pathologies congénitales touchant l'embryon (*amyélencéphalie* est l'état caractérisé par l'absence de moelle épinière : °myel et de cerveau : °encéphal). Les principes théoriques de la langue générale, pour qui les procédés de construction s'ordonnent hiérarchiquement, sont inaptes à prendre en compte l'interaction observée en corpus entre la préfixation et les diverses relations qui peuvent relier *X* et *Y*.

La mise en évidence des données que nous venons de décrire a été rendue possible, à partir de ressources textuelles spécialisées massives, grâce à un ensemble de techniques applicables en séquence, et grâce à l'organisation finale des données sous la forme d'une base de données (Namer, à paraître ; Zweigenbaum *et al.* 2005). Ces données font clairement apparaître que les hypothèses théoriques en matière de composition savante ne prennent pas en compte la réalité du vocabulaire bio-médical. Celui-ci possède ses règles propres qui s'ajoutent (et parfois suppléent) à celles de la langue générale. Ajoutons enfin que, selon (Iacobini, 2004) il y a fort à parier que les nouvelles règles émergentes sont indubitablement transposables dans d'autres langues européennes, où s'observent des noms construits apparemment sur les mêmes modèles que les *YZX* (*uranostaphylorrhaphy*$_{EN}$, *Enterokolostomie*$_{DE}$) ou les *prefYXie* (*iperproteinemia*$_{IT}$, *anenteroneuria*$_{ES}$)[7].

Références

Aronoff M. (1976). *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press.

Bonami O., Boyé G. & Kerleroux F. (2006). « L'allomorphie radicale et la relation flexion-construction ». In Fradin B., Kerleroux F. & Plénat M. (éds) *Aperçus de morphologie du français*. pp. XX–XX. PUV.

Corbin D. (2004). « French (Indo-European: Romance) ». In Booij G., Lehmann C. et Mugdan J. (éds), *An International Handbook on Inflection and Word Formation*. Volume 1. New-York : Mouton - Walter de Gruyter.

Dal G. & Namer F. (2005). « L'exception infirme-t-elle la notion de règle ? Ou le lexique construit et la théorie de l'optimalité ». *Faits de langue*, **25**, .

Fradin B. (1997). « Esquisse d'une sémantique de la préfixation en *-anti* ». *Recherches linguistiques de Vincennes*, volume 26. pp. 87-112.

[7] Respectivement traduits par : *uranostaphylorraphie*, *entérocolostomie*, *hyperprotéinémie*, *anentéroneurie*.

FRADIN B. (2000). « Combining forms, blends and related phenomena ». In DOLESCHAL U et THORNTON A.-M. (éds.) *Extragrammatical and Marginal Morphology*. pp. 11–59. München : Lincom Europa**.**

GREFENSTETTE G. (1999). « The WWW as a Resource for Example-Based MT Tasks ». In *Proceedings of the ASLIB 'Translating and the Computer' Conference*, London. Invited Talk.

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

HATHOUT N. & TANGUY L. (2002). « Webaffix : finding and validating morphological links on the WWW ». In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pp. 1799–1804, Las Palmas de Gran Canaria : ELRA.

IACOBINI C. (2004). « Composizione con elementi neoclassici ». In GROSSMANN M. et RAINER F. (éds.) *La formazione delle parole in italiano*. pp. 69–96, Tübingen : Niemeyer.

JACQUEMIN C. & BUSH C. (2000). « Combining lexical and formatting clues for named entity acquisition from the Web ». In *Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pp. 181–189, Hong Kong.

LIEBER R. (1983). « Argument Linking and Compounds in English ». *Linguistic Inquiry*, **14**(2), 251–85.

LIGON S. et PLÉNAT M. (2006). « Échangisme suffixal et contraintes phonologiques ». In FRADIN B., KERLEROUX F. et PLÉNAT M. (éds) *Aperçus de morphologie du français*. pp. XX–XX. PUV.

NAMER F. (2002). « Valider les unités morphologiques par le Web ». In FRADIN B., DAL G., HATHOUT N., KERLEROUX F., PLÉNAT M. et ROCHÉ M. (éds) *Silexicales 3: les unités morphologiques*. pp. 142–150, Villeneuve d'Ascq, Presses Universitaires du Septentrion.

NAMER F. (2003a). « Productivité morphologique et complexité de la base : le système Morphête ». *Langue Française*, **140**, 79–101.

NAMER F. (à paraître). « Le modèle Lstat: ou comment se constituer une base de données morphologique à partir du Web ». *Revue Québécoise de Linguistique*. Montréal, Erudit**.** **32**(1).

PLÉNAT M. (1988). « Morphologie de adjectifs en *-able* », *Cahiers de grammaire* **13**, 101–132.

PLÉNAT M. (1997). « Analyse morpho-phonologique d'un corpus d'adjectifs en *-esque* ». *Journal of French Language Studies*, **7**, 163–179.

PLÉNAT M. (2000). « Quelques thèmes de recherche actuels en morphophonologie française ». *Cahiers de lexicologie*, **77**, 27–62.

PLÉNAT M. (2006). « Les contraintes de taille ». In FRADIN B., KERLEROUX F. et PLÉNAT M. (éds) *Aperçus de morphologie du français*. pp. XX–XX. PUV.

RESNIK P. (1999). « Mining the Web for bilingual text ». In *Proceedings of the 37th Meeting of ACL*, pp. 527–534, Maryland, USA.

TANGUY L. & HATHOUT N. (2002). « Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web ». In J.-M. PIERREL, éd., *Actes de la 9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, pp. 245–254, Nancy : ATALA.

VILLOING F. (2003). « Les bases des opérations de construction morphologiques : des unités sémantiquement spécifiées. Illustration à la lumière de la composition [VN]N/A en français ». In FRADIN B., DAL G., HATHOUT N., KERLEROUX F., PLÉNAT M. et ROCHÉ M. (éds) *Silexicales 3: les unités morphologiques*. pp. 213-219, Villeneuve d'Ascq, Presses Universitaires du Septentrion.

WARREN B. (1990). « The importance of combining forms ». In DRESSLER W. U., LUSCHÜTZKY H. C., PFEIFFER O. E et RENNISON J. R. (éds.) *Contemporary Morphology*. pp. 111-32, Berlin, New York, Mouton - Walter de Gruyter.

WILLIAMS E. (1981). « On the Notions 'Lexically Related' and 'Head of a Word' ». *Linguistic Inquiry*, **12**(2), 245-74.

YVON F. (s. d.). Introduction au Traitement automatique des langues naturelles. ENST. http://perso.enst.fr/~demoulin/taln.ps

ZWEIGENBAUM P., BAUD R., BURGUN A., NAMER F., JARROUSSE E., GRABAR N., RUCH P., LE DUFF F., FORGET J. F., DOUYERE M. et DARMONI S. (2005). « UMLF: a unified medical lexicon for French ». *International Journal of Medical Informatics*, 74(2-4), 119-124.