WaliM : valider les unités morphologiques complexes par le Web

Fiametta Namer
UMR 7118 « ATILF » & Nancy Université
Fiammetta.namer@univ-nancy2.fr

Texte initialement paru dans *Silexicales* 3, pp. 142-150.

Internet a contribué, surtout depuis la fin du XXe siècle, au développement de l'approche dite extensive de la morphologie, qui consiste à appuyer les descriptions sur des corpus d'exemples aussi étendus que possible. Le morphologue a enfin la possibilité de confronter ses hypothèses à des masses de données dont la taille et la diversité en font des sources textuelles d'une richesse extraordinaire. Accéder facilement et rapidement à ces nouvelles ressources est une question à laquelle les informaticiens-linguistes ont apporté différentes réponses, parmi lesquelles WaliM, dont il est question dans cet article. Cet outil est toujours d'actualité en 2011 ; certes, il a subi depuis sa conception de nombreux changements, lui permettant de suivre l'évolution des moteurs de recherche, points d'entrée de la Toile. Cependant sa philosophie est restée la même : remplacer l'humain lors de l'interrogation du moteur Yahoo, au moyen de listes de requêtes de taille quelconque. Depuis la publication de l'article ci-dessous, WaliM n'a jamais cessé d'être utilisé. Son usage croissant est à l'image du succès grandissant de la morphologie dite extensive, et témoigne de d'affaiblissement graduel des réticences à l'égard de l'exploitation des données de la Toile.

1. Introduction

Un moyen de vérifier une hypothèse en morphologie consiste à tester l'attestation des résultats des règles de formation d'unités lexicales (désormais ULs) qui formulent cette hypothèse. Pour pouvoir s'effectuer à grande échelle, cette vérification nécessite alors un travail fastidieux de la part du linguiste, qui doit constituer une liste d'ULs construites reflétant ses intuitions et contrôler l'existence (ou l'absence) de chacune d'elles en les recherchant dans les corpus appropriés : lexiques, thésaurus, dictionnaires, bases de données, documents textuels … Une recherche automatisée sera plus efficace, et la fiabilité des résultats croît avec la taille des corpus consultés. Dans cet article, nous proposons un système de vérification automatique d'unités lexicales construites, baptisé WALIM (**W**eb et v**ali**dation en **M**orphologie), qui exploite l'immense réserve de ressources lexicales d'Internet. Sans vouloir prétendre fournir une réponse définitive aux intuitions linguistiques, WALIM participe néanmoins de façon non négligeable à leur confirmation (ou infirmation) en confrontant les listes d'ULs au fonds documentaire considérable, en expansion perpétuelle et extrêmement varié que constitue Internet. Dans un premier temps, nous discutons des avantages et des inconvénients d'Internet an tant qu'outil exploitable en linguistique de corpus (section 2). L'article se consacre ensuite à la présentation détaillée de WALIM (section 3), et à la description d'expériences récentes d'utilisation de ce système pour illustrer empiriquement les conditions d'application des procédés de construction de mots ou pour compiler les listes d'exceptions à ces procédés (section 4).

2. Validation empirique des hypothèses en morphologie

Le morphologue qui enquête sur les (ou l'évolution des) conditions d'application des procédés de formation des unités morphologiquement construites, a besoin de produire des exemples pour étayer ses hypothèses. Une partie de ces exemples sont parfois attestés, et proviennent le plus souvent de dictionnaires de la langue générale. Mais la plupart du temps, il s'agit d'innovations lexicales qui servent à tester la pertinence explicative de l'hypothèse défendue : ces néologismes sont recueillis à partir de documents souvent spécialisés. Plus la consultation des ressources documentaires et dictionnairiques renvoie un nombre important d'exemples différents apparaissant chacun avec une fréquence conséquente, plus l'hypothèse à laquelle ces exemples font écho est confortée. Une conséquence de cette observation est que la taille de la liste de néologismes obtenue est fonction de la quantité et de la variété des ressources lexicales disponibles.

Une autre façon de vérifier une hypothèse en morphologie, lorsque celle-ci préconise un ensemble de contraintes pesant sur la formation d'ULs, consiste à démontrer l'impossibilité de construire des exemples contrevenant aux règles proposées dans l'hypothèse. Pour ce faire, la quantité de documents disponibles est encore une fois une donnée cruciale : le statut de « mot impossible » sera d'autant plus légitime que la vérification aura pu se faire à grande échelle (cf. section 4.1). A contrario (cf. section 4.2. et 4.3), l'émergence de contrexemples dans ces mêmes conditions expérimentales donne l'opportunité de réexaminer la règle mise en défaut.

2.1. *Documents et dictionnaires*

L'importance de disposer de sources documentaires abondantes et variées pour vérifier les intuitions en morphologie est donc indéniable, que ce soit pour collecter le plus d'exemples illustratifs possible, ou au contraire pour confirmer l'absence de ceux-ci. Traditionnellement, toute forme de ressource écrite peut être exploitée : les dictionnaires de la langue générale ou étymologiques, les lexiques spécialisés dans un domaine, une époque ou un dialecte donné se veulent les garants du lexique attesté, et les journaux, les textes littéraires, les essais scientifiques etc. témoignent eux de l'innovation lexicale à un moment donné. Depuis quelques

années déjà, à la consultation de ces documents papier s'ajoute la possibilité d'effectuer des recherches dans des corpus électroniques, organisés le plus souvent sous forme de bases de données de taille considérable et munies d'interface facilitant la formulation des requêtes de l'utilisateur. Outre les dictionnaires et encyclopédies, ce nouveau format permet d'accéder rapidement et efficacement aux archives de journaux, aux œuvres complètes d'un romancier, etc., au moyen de CDROM ou par le biais d'interrogations sur Internet.

2.2. *Internet : le corpus universel*

2.2.1. Internet et linguistique de corpus

De plus en plus, Internet est plébiscité en tant que source de données et outils exploitables en linguistique. On y trouve en effet de nombreux services en lignes, parfois gratuits, incluant la consultation de dictionnaires ([TLFI][1]), la traduction dans de nombreuses langues (l'URL http://www.foreignword.com propose par exemple la traduction en ligne de 63 langues), l'interrogation de bases textuelles ([Frantext][1]). En outre, les linguistes travaillant sur corpus ont découvert récemment avec Internet une source immense, multilingue et gratuite, de corpus de textes écrits. Le contenu de ces textes, par ailleurs en renouvellement constant, présente tous les aspects typologiques possibles : le Web a stimulé de ce fait la recherche dans tous les domaines du traitement automatique des langues : extraction de corpus parallèles bilingues, acquisition lexicale, traduction automatique, classification sémantique de documents, études éthno- et sociolinguistiques, comme en témoignent les nombreuses références bibliographiques, parmi lesquelles on peut retenir (Resnik, 1999), (Brill, 2001), (Grefenstette, 1999), (Beaudouin et al., 2001). En ce qui concerne l'étude des unités morphologiques, Internet peut donc se voir comme un corpus formidable pour l'extraction d'exemples d'ULs construites.

Son utilisation comme source de corpus n'est certes pas sans comporter des problèmes pour son exploitation. En effet, le Web n'est pas fiable : un document en ligne peut contenir des fragments de textes appartenant à d'autres langues que le français, du code informatique, des fautes d'orthographe, enfin, le niveau de langue est non garanti. En outre, le Web n'est pas stable : en effet, le nombre et le type des pages indexées varient avec le moteur de recherche sollicité et avec le temps. Il est par conséquent impossible d'utiliser le Web pour étudier l'évolution l'un mot.

Néanmoins, les avantages que présente Internet le rendent très attractif en tant que ressource textuelle et lexicale : un très grand nombre de langues y est représenté, on y trouve tout type, tout format, de corpus écrits, dans tous les domaines de spécialité (sont consultables en ligne ou téléchargeables des dictionnaires, lexiques, thésaurus, encyclopédies, articles scientifiques, romans, cours, manuels d'utilisation, publicités, etc.). Internet reflète donc au mieux l'innovation lexicale, et offre des contextes d'utilisation les plus variés

2.2.2. Internet au service des morphologues

En tant que corpus « ouvert », et malgré les inconvénients qui viennent d'être mentionnés, Internet peut donc être vu par le morphologue comme un complément précieux aux dictionnaires et aux archives électroniques, quelle que soit leur taille. Dans cette perspective, il constitue un outil idéal en morphologie pour répondre à des questions comme : trouvera-t-on des néologismes, mots inconnus des dictionnaires de référence pour valider ou falsifier une règle ? Est-ce que la fréquence observée pour un type morphologique donné confirme ou infirme une prédiction quant à la productivité de ce type ? Est-il possible de faire des observations comparables dans plusieurs langues en ce qui concerne la formation des mots ?

[1] Cf. URL : http://atilf.atilf.fr/tlf.htm

Il existe fondamentalement deux types d'approches pour exploiter les ressources lexicales présentes dans les documents sur le Web. Le premier consiste à utiliser le Web en tant que base documentaire pour en extraire des ensembles d'unités lexicales appartenant à un type morphologique donné (de façon à répondre e.g. à la requête : « quels sont tous les noms en *-ette* ? »). De cette façon, le morphologue va pouvoir en étudier les contextes d'utilisation, se constituer des ressources lexicales, génériques ou spécialisées ((Jacquemin & Bush, 2000) se servent du Web pour le repérage des entités nommées), et éventuellement échafauder de nouvelles hypothèses de formation de mots. Cette démarche d'acquisition lexicale, proposée par Webaffix (Hathout & Tanguy, 2002) a l'avantage de produire un corpus important de néologismes ; cependant étant donnée l'hétérogénéité du Web et l'instabilité reconnue des résultats produits (cf. Selberg & Etzioni, 2000), une partie de cette production (i.e. les mots mal orthographiés ou appartenant à une langue étrangère) doit être filtrée. L'autre inconvénient de Webaffix est qu'il demande l'utilisation du caractère joker dans l'expression des requêtes, ce qui réduit les possibilités d'interrogation à quelques moteurs de recherche.

3. WALIM

L'autre approche, que nous défendons dans cet article, consiste non pas à rapatrier des ressources lexicales depuis Internet, mais plutôt à se servir de la toile comme évaluateur d'hypothèses linguistiques, en exploitant la liste d'exemples du morphologue comme autant de requêtes automatisées, par l'entremise d'un ou plusieurs moteurs de recherche. À partir d'un ensemble d'unités lexicales linguistiquement motivées *a priori*, cette approche participe empiriquement à la confirmation, l'affinage voire à la remise en cause partielle d'hypothèses concernant le fonctionnement et les propriétés des unités morphologiques. Les listes d'UL utilisées comme requêtes peuvent être élaborées manuellement. Les résultats servent alors de guide pour répondre à des questions comme : l'hypothèse est-elle vérifiable à grande échelle ? Y a-t-il des exceptions ? Celles-ci font-elles apparaître de nouvelles contraintes ?

3.1. *Fonctionnement du système*

L'entrée de WALIM est une liste de requêtes qui sont soumises automatiquement à un moteur de recherche. Chaque requête, composée d'une unité lexicale unique, est l'un des termes construits reflétant l'hypothèse linguistique à valider. Cette liste peut-être constituée manuellement ou générée automatiquement. La seconde approche, qui suppose la maîtrise d'un langage de programmation approprié (e.g. Perl) est naturellement la plus adaptée quand la liste à soumettre est taille importante ou inconnue. Dans une partie des expériences relatées ci-dessous, le lexique source qui contient les bases potentielles des mots-requêtes construits est élaboré à partir du TLFNOME[2] qui constitue un lexique de référence de 98000 lemmes étiquetés catégoriellement. Dans une autre expérience, qui nécessite l'emploi de mots fléchis pour générer les unités morphologiques à tester, c'est le lexique de l'ABU[3] qui a été utilisé.

Le système proprement dit est un robot qui interroge un (ou plusieurs) moteur de recherche (MT) acceptant d'être utilisés de manière automatique[4]. Le MT que nous avons choisi d'utiliser est YAHOO. Ce choix s'explique par la couverture de ce MT : lorsque YAHOO ne retrouve pas de document correspondant à la requête, il interroge GOOGLE, ce qui offre une garantie étendue en termes de quantité de

[2] Le TLFNOME est le lexique composé des nomenclatures du TLF.

[3] L'ABU (Association des Bibliophiles Universels) diffuse au format électronique des textes libres en droits et des dictionnaires. Ces documents sont téléchargeables, ou interrogeables à distance. cf. URL : http://abu.cnam.fr/

[4] Par exemple, GOOGLE n'autorise la soumission de requêtes par robots que dans des conditions très précises.

documents indexés. Cependant, notre choix aurait été différent si les recherches de WALIM devaient se restreindre à un espace de recherche donné : ainsi, si l'on souhaite rechercher les contextes d'utilisation d'une UL dans le domaine biomédical francophone, l'un des moteur à interroger est CISMeF (cf. Darmoni et al., 2000), qui renvoie des articles scientifiques appartenant à cette thématique et libres de droits.

La valeur de la requête soumise à YAHOO, est chacune des unités morphologiques prises tour à tour. Pour respecter les règles déontologiques en vigueur sur Internet, l'écart minimal entre deux requêtes est de 16 secondes. L'utilisateur de WALIM s'attend à ce qu'une unité morphologiquement bien construite et non attesté dans les dictionnaires de la langue générale soit utilisé dans au moins un document en ligne, et que ce document soit récupérable via le moteur de recherche interrogé. Inversement, il voudrait qu'une unité dérogeant aux règles qui constituent son hypothèse de travail ne corresponde à aucun document accessible sur Internet.

3.2. *Format des résultats*

Les résultats de WALIM sont filtrés et reformulés de manière à en faciliter l'interprétation. C'est ainsi que l'utilisateur dispose de plusieurs options dont le déclenchement, à l'appel de WALIM, conditionne le nombre et le format des réponses. La recherche se limite ainsi, sur demande, aux URLs francophones ; quant à la réponse, elle peut renvoyer l'adresse des premières URLs indexées par le mot-requête, indiquer le nombre total de réponses trouvées, ou encore afficher le contexte d'utilisation du mot-requête apparaissant sur la page de réponses fournie par YAHOO. Si aucune option n'est sélectionnée par l'utilisateur, le type de résultat dépend, pour chaque UL testée, du nombre de réponses trouvées par le moteur. Il faut savoir que le moteur est susceptible de ramener un nombre parfois surprenant de réponses positives correspondant à la répétition d'une même erreur de typographie. Par exemple, en cherchant en 2002 les noms déverbaux (cf. section 4.2) on obtient 443 documents mentionnant *campage* (au lieu de *campagne*), 42 citant *culeur* (pour *couleur*) ; souvent les verbes au présent 3ème personne du pluriel sont orthographiés X**m**ent au lieu de Xent (d'où la récupération de : *appellement*, *amènement*, *appliquement*…). Une autre source fréquente d'erreurs vient de l'assimilation par YAHOO du tiret (et en général, de toute marque de ponctuation) à un séparateur de mots : 730 réponses apparemment positives à la requête *cadrement* ramènent en fait des documents contenant le terme *en-cadrement*.

Afin d'opérer une préselection des résultats, WALIM confronte à un seuil, arbitrairement fixé à 800[5], le nombre de réponses obtenues pour un mot-requête donné. Pour chaque mot-requête, tout nombre de réponses soit nul, soit supérieur au seuil fixé est simplement renvoyé en valeur. Les autres valeurs, inférieures au seuil et donc assimilables à des réponses moins fiables, entraînent la soumission renouvelée de la requête. Cette fois, WALIM procède à une détection de caractères non-alphanumériques : toute séquence qui ne s'identifie pas graphiquement avec le mot-requête (à l'exception de la casse) est éliminée (cette technique de vérification est utilisée également dans la reconnaissance de l'hyphénation dans les noms composés, cf. section 4.3). Les séquences restantes sont affichées avec le contexte d'utilisation apparaissant dans la réponse fournie par YAHOO.

Le résultat fourni à l'utilisateur consiste en sa liste d'unités morphologiques, où chaque UL est associée à une valeur interprétable : (1) un nombre supérieur au seuil donne à voir l'unité comme attestée hors dictionnaire ; (2) avec une valeur nulle, il est raisonnable de penser que le mot n'est pas utilisé : si le but est de démontrer qu'il s'agit d'un mot construit « impossible », alors l'expérience doit être renouvelée dans le temps et au moyen d'autres moteurs de recherche ; (3) enfin, les contextes d'utilisation du mot-requête associés aux valeurs inférieures au seuil permettent à l'utilisateur de juger de la validité de l'UL testée.

[5] Ce seuil est un paramètre que l'utilisateur peut modifier à l'appel du programme.

1. Applications

Cette dernière section est consacrée à la description de trois séries d'expériences. Dans chaque cas, le protocole est identique : la liste des unités morphologiquement construites à vérifier est constituée, soit manuellement, soit automatiquement. Dans le second cas, on veille à exclure de la liste les UL présentes dans le dictionnaire de référence[6]. Le moteur de recherche utilisé est YAHOO (interrogé entre 2000 et 2003), et en général le programme est exploité en mode par défaut (sans option, en dehors de la recherche de pages francophones). Quand on cherche à vérifier qu'une liste d'ULs contrevient à une règle et est donc « mal formée », on s'attend à des réponses nulles de la part de WALIM, qui est exécuté à plusieurs reprises, à différents mois d'intervalle.

4.1. *Formation de mots à base morphologiquement complexe*

La première série d'expériences explore, par l'utilisation de listes d'exemples *a priori* mal formées, les conditions de construction lexicale sur des bases morphologiquement complexes, c'est à dire dans des conditions morphosémantiques stables, puisqu'une partie des caractéristiques sémantiques de la base sont rendues prédictibles par le procédé de construction de mot qui en est à l'origine.

La première utilisation de WALIM, décrite dans (Dal & Namer, 2000), a pour objectif l'évaluation de l'interaction entre complexité d'une unité lexicale et construction de mots nouveaux. L'observation de quelques exemples conduit à penser que plus une unité lexicale est le résultat de constructions successives, moins son statut de base potentielle dans une opération morphologique est envisageable. Une liste de noms de propriété en *-ité* est générée automatiquement à partir des 441 adjectifs du TLFNOME de structure $[[X_{N/A}$ *-iser*$]_V$ *-able*$]_A$, qui de la sorte constituent des bases linguistiquement compatibles avec *-ité*. WALIM permet alors de constater non seulement que, bien que linguistiquement bien formés, très peu (10 %) de ces néologismes construits s'observent sur Internet (e.g. *scolarisabilité, synchronisabilité*), mais aussi que cette proportion ne varie ni qualitativement ni quantitativement avec le temps (la même expérience a été réitérée 2 ans de suite) ; enfin, ce résultat, passible d'une explication psycholinguistique (cf. entre autres, Baayen, 1996 ; Krott, Schreuder & Baayen, 1999), semble indépendant des langues puisqu'il s'observe en Italien dans des conditions expérimentales analogues (même moteur de recherche, procédés morphologiques équivalents : *-izzare, -ità, -(a)bile*, lexique de 429 unités, 17% de néologismes utilisés dans les documents en ligne, e.g. *sponsorizzabilità, storicizzabilità*).

L'autre série d'expériences sur les opérations morphologiques s'appliquant aux bases complexes a pour but de vérifier l'inadéquation sémantique entre types morphologiques catégoriellement compatibles.

Dans la première expérience de cette série, la liste d'ULs intuitivement mal-formées contient des verbes exprimant un changement d'état, construits sur des bases adjectivales en *-eux*[7]. Ces bases désignent des propriétés majoritairement vues comme **inhérentes** aux référents du nom recteur[8]. En revanche, la propriété transmise au référent du patient du verbe de changement d'état doit être **acquérable**. L'incompatibilité entre 'acquérabilité' et 'inhérence', soulignée dans les

[6] Par défaut, celui-ci est constitué par l'union des nomenclatures du TLF et du [RE].

[7] Quand ils sont construits sur base adjectivale, les verbes de changement d'état s'obtiennent en français par suffixation (*-iser, -ifier*) préfixation (*en-, a-*) ou conversion. La différence entre les instructions sémantiques de ces procédés a fait l'objet de nombreuses recherches. Cf. entre autres (Plag 1999), à propos de *-ize*, ou les travaux de Coralie Roger, cf. (Roger 2002) et (Roger 2003).

[8] Considérer la différence entre *plaine venteuse*, où le vent est constitutif de la plaine, et où la description est vraie même en l'absence de vent et *plaine ventée*, qui n'est vrai que lorsque le vent y souffle ; cf. (Corbin 1997) et (Aliquot 1996) pour une description détaillée de *-eux*.

études théorique, et expliquant l'absence de verbes en « Xosifier, enXoser, etc… », est ici mise en évidence par une absence totale de réponse aux (3365) verbes générés automatiquement à partir du référentiel (*verbosifier, aventos(e/i)r, vicios(e/i)r,* …) et utilisés comme requêtes par WALIM.

La seconde expérience, relatée dans (Namer et Dal, 2000) et (Namer 2001), a consisté à tester les verbes en *-iser* formés à partir d'adjectifs eux-mêmes construits en *-able* ; afin d'étudier la proximité de leur systèmes morphologiques, les vérifications de WALIM ont encore une fois porté sur le français et l'italien. Contrairement à l'expérience précédente, on a affaire ici à deux types morphologiques parfois cooccurrents dans les deux langues (*navigabiliser* ou *impermeabilizzare*) ; cependant la rareté des verbes en *-abiliser*, (il n'en existe en français que 14 sur les 700 verbes en *-iser* des dictionnaires de la langue générale) et le statut incertain de « verbe construit en synchronie sur base adjectivale » pour la plupart d'entre eux (cf. *amabiliser, culpabiliser*) vont dans le sens de l'existence d'un blocage sémantique entre ces deux types morphologiques, les adjectifs en *-able* (*-abile*) décrivant des propriétés latentes, donc endogènes au référent de leur nom recteur, et par conséquent encore une fois incompatibles avec la propriété exigée en position de base par les verbes de changement d'état en général et construits en *-iser* (*-izzare*) en particulier. Cette intuition est confirmée par WALIM, qui ne trouve aucun document correspondant aux quelque 700 verbes en *-abiliser/-abilizzare* générés automatiquement à partir de notre référentiel, la soumission des requêtes correpondantes ayant été répétée deux fois, à un an d'écart.

4.2. *Affiner les conditions d'application des règles*

Outre un outil de vérification, et parce qu'il atteste l'existence en corpus de mots inconnus des dictionnaires de langue, WALIM offre la possibilité d'affiner les conditions d'application des règles, par un recensement élargi des exceptions à celles-ci. Dans ce qui suit, les listes ne sont pas *a priori* mal formées, contrairement à 4.1. Les expériences relatées portent sur les nominalisations verbales en *-eur, -age, -ment* et *-tion*.

Tout d'abord, nous avons voulu vérifier l'inaptitude des verbes inaccusatifs à servir de bases pour la nominalisation en *-eur*, à partir de la liste des verbes répertoriés dans (Legendre, 1989) et suivant (Fradin et Kerleroux, 2003a et 2003b ; Kerleroux, 2004). La liste des 100 verbes dits inaccusatifs par Legendre servent de bases aux noms suffixés par *-eur* générés automatiquement. Les réponses rapportées par WALIM ainsi que les contextes d'utilisation des noms ont permis de mettre en évidence les faits suivants : 81% des noms-requêtes vérifient effectivement l'hypothèse suggérée : soit ils sont non-attestés (c'est le cas de **adveneur*), soit leur base n'est pas inaccusative : ainsi, l'existence des noms d'instruments *ralentisseur* ou *cuiseur* s'explique par la sélection par *-eur* de l'entrée transitive et donc non inaccusative de *ralentir* et *cuire*. Cependant, 19% des noms de la liste ramènent des réponses positives par WALIM, alors que leur base ne peut être qu'inaccusative. Une partie d'entre eux a une interprétation causative, formulable par « ce/celui qui fait V NP0 ». Ainsi, un *naisseur* caractérise un certain type d'éleveur alors qu'un *écloseur* est une sorte d'incubateur. Les autres ne peuvent en aucun cas être perçus comme des causatifs : *évanouisseur, persisteur, tousseur …* Suite à cette expérience, nous avons tenté de reformuler l'hypothèse de Legendre, en testant à grande échelle les conditions de création des néologismes en *-eur* à partir de 6000 unités verbes du référentiel : une fois mis en lumière l'ensemble des noms agentifs attestés dans les documents en ligne, pourra-t-on affirmer que les verbes restants sont inaccusatifs ? Et sinon, quelles autres conditions régissent cette formation ? La recherche a conduit jusqu'ici à l'obtention d'un millier de noms, désignant majoritairement des anciens métiers (*entêteur*), ou encore relatifs à des domaines d'activités spécialisés (*abondeur* est un terme juridique), ou à des technologies récentes (*abraseur* est un instrument).

Toujours partant des 6000 verbes du référentiel, nous avons constitué automatiquement la liste de tous les noms déverbaux d'activité potentiels en *-age, -ment* et *-tion*, en tenant compte, le cas échéant, de l'alternance possible entre base savante et base populaire. Par exemple, sur *éteindre*, on construit *éteignage, éteignement, éteign(a/i)tion,* mais aussi *extingage, extinguement, extinction*.

WALIM interroge YAHOO avec la liste de ces quelque 14500 déverbaux, et complète ainsi empiriquement les hypothèses formulées entre autre dans (Kelling, 2003a) et (Kelling, 2003b) à propos de la caractérisation des unités verbales sur lesquelles s'appliquent les procédés de construction de déverbaux en *-age*, *-ment, -tion*. Parmi les premiers résultats, on remarque que les bases savantes ne sont jamais sélectionnées par *-age* et *-ment*. D'autres résultats sont consignés dans le tableau ci-dessous.

Référentiel : 6000 verbes	**Vage**	**Vment**	**Vtion**
Pourcentage des $[Vsuf]_N$ inventés à partir des 6000 verbes	9,8% (*effarouchage*)	5,02% (*embossement*)	1,1% (*atrophiation*)
Nb de verbes sans $[Vsuf]_N$	10 (*manger*)		
Verbes bases de tous les $[Vsuf]_N$	44 (*appropriage, appropriement, appropriation argentage, argentement, argentation*)		
% Vage & Vment seulement	16,5% (*arasage, arasement*)		
% Vage & Vtion seulement	2,7 (*calibrage, calibration*)		
% Vment & Vtion seulement		2,7 (*consolidement, consolidation*)	

On observe (ligne 1) que c'est *-age* qui produit le plus de déverbaux hors dictionnaires utilisés dans les documents ; ces néologismes désignent souvent des techniques et domaines d'activités correspondant aux instruments en *-eur* mentionnés ci-dessus. À l'inverse, *-ment* est peu utilisé. Curieusement, le procédé le moins productif, comparativement aux données du référentiel, est le suffixe *-tion*. Les chiffres fournis par WALIM mettent par ailleurs en évidence des faits intéressants, qui peuvent influer sur l'élaboration de nouvelles hypothèses pour la formation des déverbaux de procès. Tout d'abord, pour certains verbes, il n'y a pas de déverbaux de procès ; bien sur, le déverbal peut être produit par un autre procédé que ceux testés ici (*voler* $\rightarrow_{\text{CONVERSION}}$ *vol*, *assassiner* $\rightarrow_{\text{-AT}}$ *assassinat*), mais l'existence de lacunes lexicales est indéniable, comme l'attestent l'absence de nominalisations pour *manger* ou *boire*. Ensuite, les doublons répertoriés (en *age/ment, age/tion*, etc.) doivent à terme contribuer à affiner la typologie des verbes intervenant dans les nominalisations : les verbes inaccusatifs, resp. transitifs, se retrouvent-ils dans ces doublons ? Les verbes des doublons partagent-ils le même aspect ? Correspondent-ils à la même classe sémantique ? Etc. Ces nouveaux indices permettront de préciser la définition du sens construit par les différents procédés.

4.3. *Pister les néologismes construits : le cas des $[VN]_N$*

Pour finir, nous décrivons ici la dernière vérification, actuellement en cours, qui est menée au moyen de WALIM. À partir du lexique ABU des 267872 formes fléchies étiquetées du français sont générés l'ensemble des quelque 23 millions de couples (V, N) possibles, où V est à la 3ème personne du présent de l'indicatif, N est au singulier, et V et N ont moins de 7 caractères. Ces couples sont organisés en noms composés séparés par un trait d'union. L'expérience consiste à illustrer, confirmer, voire étendre, par une recherche à grande échelle, les hypothèses défendues dans (Villoing, 2003) quant aux contraintes régissant la formation des noms composés de type VN.

Les caractéristiques graphiques des mots composés impliquent que l'on prenne un certain nombre de précaution dans le rapatriement des résultats : en effet, le tiret présent entre les composants n'est pas indexé par YAHOO (pas plus que par tout autre MT), qui l'assimile à n'importe quel séparateur (espace, point, virgule …). En conséquence, nombre de couple verbe, nom qui constituent des candidats noms-composés plutôt improbables ramènent un nombre parfois non négligeable de résultats apparemment positifs, surtout quand s'ajoute l'ambiguïté catégorielle de certains noms : ainsi, WALIM nous annonce qu'« abaisse-ton » est retrouvé dans plus de 100 documents, alors que les séquences qui y figurent réellement correspondent à « abaisse ton ».

WALIM procède alors à une vérification systématique des résultats, en comparant à celle de la requête la chaîne de caractère indexée dans le document ramené. Cette comparaison n'est réalisée que dans la première page de réponses de YAHOO[9]. Si la comparaison réussit, l'environnement d'utilisation du nom composé candidat est renvoyé en valeur, pour que l'utilisateur décide de sa validité.

2. Conclusion

On peut envisager d'utiliser WALIM pour guider la validation de nombreuses autres hypothèses en morphologie : (1) explorer l'innovation lexicale constatée à petite échelle, et contredisant des études théoriques ; on cherchera e.g. à identifier le type des bases verbales non transitives et non inaccusatives (cf. Horn, 1980) des néologismes adjectivaux en *-able*, suivant l'exemple de *roulable, surfable, dansable* ; (2) définir des règles d'acquisition automatique de sens : par exemple, en disposant d'environnements phrastiques des noms de propriétés en *–ité*, cf. (Fradin & Kerleroux, 2003a), (Rainer, 1988) ou des verbes désadjectivaux de changement d'état servant d'input à WALIM, on se donnera des indices pour caractériser sémantiquement les adjectifs en position de base (sont-ils prédicatifs ou relationnels ? etc.) ; (3) justifier l'existence de procédés morphologiques (apparemment) concurrents : la (in)validation de doublons adjectivaux permettra de proposer des contraintes (sémantiques, phonologiques, etc.) pour expliquer les distributions complémentaires des bases par certains procédés (*-ain* vs *-ais*, *-al* vs *-ique*), ou pour, à l'inverse, préciser les variations de sens des dérivés obtenus à partir de la même base (*dormant / dormeur, vicieux / vicié, ferreux / ferrique, apposage / apposement / apposition*).

En conclusion, WALIM est un système qui présente certainement un certain nombre de faiblesses. En effet, son fonctionnement est dépendant de l'interface des MT interrogés, ce qui sous-entend que toute modification de celle-ci entraîne une mise à jour du programme ; d'autre part, l'utilisation d'Internet comme base lexicale interdit tout travail contrastif en diachronie : la rapidité avec laquelle des nouveaux documents sont ou ne sont plus indexés par les MT explique qu'on ne puisse jamais obtenir la même réponse à deux époques différentes. Enfin, la diversité des documents ramenés demande un filtrage soigneux des résultats, dont une partie est manuel : c'est pourquoi il faut voir WALIM comme un système d'aide à la décision, et Internet comme un réservoir lexical en perpétuelle évolution et couvrant tous les domaines lexicaux possibles.

En contrepartie de ses inconvénients, WALIM est un moyen simple et efficace mis à la disposition des linguistes non informaticiens pour examiner le lexique offert par Internet : il leur évite la répétition fastidieuse de requêtes manuelles. Son usage peut porter à la découverte de néologismes, à la reformulation de règles, à l'élaboration de nouvelles hypothèses. Il rend possible l'étude de plusieurs langues, de domaines lexicaux, de registres de langue particuliers. Finalement, il concourt à l'élaboration d'hypothèses en morphologie, mais aussi en psycholinguistique (par la production d'exemples pouvant nourrir une analyse des erreurs typographiques les plus fréquentes) ou en syntaxe. Le programme, ainsi que des échantillons de listes

[9] Par défaut, les réponses de YAHOO sont affichées 20 par 20.

d'exemples et un manuel d'utilisation sont librement accessibles à l'URL : http://www.univ-nancy2.fr/pers/namer/WALIM/.

Bibliographie

Aliquot S. (1996), *Référence Collective / Sens Collectif : La notion de collectif à travers les noms suffixés du lexique français*, thèse de doctorat, Université de LilleIII.

Baayen H. (1996), « Derivational Productivity and Text Typology », in *Journal of Quantitative Linguistics, Vol. 1*, N°1, pp. 16-34.

Beaudouin V., Fleury S., Habert B., Illiouz G., Liccope C. et Pasquier M. (2001), « TyPWeb : décrire la Toile pour mieux comprendre les parcours », in *CIUST'01*, Paris.

Brill E. (2001), **«** Empirical NLP: Does the web change everything ? **»**, in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and Computational Natural Language Learning (CoNLL-ACL'01)*, Toulouse.

Corbin, D. (1987), *Morphologie dérivationnelle et structuration du lexique*, 2 vols., Tübingen, M. Niemeyer Verlag ; 2ème ed. Villeneuve d'Ascq, PUL, 1991.

Dal G. & Namer F. (2000), « Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations », *T.A.L.* 41(2), pp. 423-446.

Darmoni S., Leroy, J.P., Thirion, B. et al. (2000), « CISMeF, a structured health resource guide », in *Methods Inf Med*, 39(1).

Fradin, B. & Kerleroux, F. (2003a), Quelles bases pour les procédés de la morphologie constructionnelle ? In *Silexicales 3 : les unités morphologiques*, eds. Bernard Fradin et al., 76-84. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Fradin B. & Kerleroux, F. (2003b), "Troubles with lexemes", *3d Mediterranean Morphology Meeting (MMM3) (selected papers)*, Barcelona, pp. 177-196.

Grefenstette G. (1999), "The WWW as a Resource for Example-Based MT Tasks",, in *Proceedings of the ASLIB 'Translating and the Computer' Conference*, London.

Hathout N. & Tanguy L. (2002), "Webbaffix : finding and validating morphological links on the WWW", in *Proceedings of LREC2002*.

Horn L. (1980), "Affixation and the inaccusative hypothesis", in *CLS* 16-1, pp. 134-146.

Jacquemin Ch. & Bush C. (2000), « Fouille du Web pour la collecte d'entités nommées », in *Actes de TALN'00*, EPFL, Lausanne.

Kerleroux F. (2004), » Sur quels objets portent les opérations morphologiques de construction ? » *Lexique* 16, pp. 85-124.

Kelling C. (2003a), "The role of agentivity for suffix selection", paper presented at *3d Mediterranean Morphology Meeting (MMM3) (selected papers)*, Barcelona, pp. 197-210.

Kelling C. (2003b), « Verbes psychologiques et nominalisations», In *Silexicales 3 : les unités morphologiques*, eds. Bernard Fradin et al., 92-99. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Krott A., Schreuder A. et Baayen H. (1999), **"**Complex words in complex words", in *Linguistics* 37-5, pp. 905-926.

Legendre G. (1989), "Unaccusativity in French", Lingual 79(2/3), pp. 95-164.

Namer F. (2001), « Génération automatique de néologismes bilingues morphologiquement construits en français et italien », *Atelier Traduction automatique et applications en grandeur réelle, TALN'01*, Tours.

Namer F. & Dal G. (2000). "GéDériF : Automatic Generation and Analysis of Morphologically Constructed Lexical Resources », in *LREC'2000*, Athènes.

Plag I. (1999), *Morphological Productivity. Structural Constraints in English Derivation*. Berlin/New York: Mouton de Gruyter.

Rainer F. (1988), "Towards a theory of blocking: the case of Italian and German quality nouns", in C. Booij & J. van Marle (eds) *Yearbook of Morphology*, pp. 155-185.

[RE] = *Le Grand Robert de la langue française,* CDROM, 1994.

Resnik Ph. (1999), "Mining the Web for Bilingual Text", in *Proccedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland.

Roger C. (2002), « 'Conversion totale' vs 'changement partie' : une hypothèse pour établir l'identité sémantique de, respectivement, *-ifi(er)* et *-is(er)* », communication présentée au Séminaire de Morphologie de l'université de Nanterre, 18 mai 2002.

Roger C. (2003), « De la pertinence de la notion de paradigme pour les procédés de construction des verbes de changement d'état », In *Silexicales 3 : les unités morphologiques*, eds. Bernard Fradin et al., 179-187. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Selberg E. & Etzioni O. (2000), "On the instability of the Web Search Engines", *Proceedings of the RIAO 2000*.

Villoing F. (2003), « Les bases des opérations de construction morphologiques : des unités sémantiquement spécifiées. Illustration à la lumière de la composition $[VN]_{N/A}$ du français », In *Silexicales 3 : les unités morphologiques*, eds. Bernard Fradin et al., 213-219. Villeneuve d'Ascq: Presses Universitaires du Septentrion