Repérage de créations lexicales sur le Web francophone

Franck Sajous & Ludovic Tanguy
ERSS/CNRS & Université de Toulouse-Le Mirail

Résumé : *Nous présentons ici un ensemble de travaux autour de l'acquisition de créations lexicales en français utilisant le Web comme source de données. Ces travaux se découpent en deux périodes : la première se base sur l'utilisation d'un moteur de recherche généraliste, et est désormais obsolète au vu des possibilités d'interrogation actuellement proposées par les moteurs disponibles. La seconde met en place un logiciel dédié au parcours du Web francophone, baptisé* Trifouillette.

Introduction

Le Web, en tant que masse colossale de données textuelles numérisées est un lieu privilégié pour l'observation des créations lexicales. Puisque ces dernières apparaissent en permanence et sont rares, la masse du Web, son évolution exponentielle et sa mise à jour continue en font le seul choix possible pour une étude à large échelle. À cela se rajoute la spontanéité des productions que l'on y trouve, par opposition aux gros corpus numérisés classiques (banques textuelles ou corpus journalistiques). Bien entendu, ces avantages sont largement contrebalancés par les problèmes spécifiques que rencontre toute étude linguistique basée sur le Web, liés à l'absence de contrôle du contenu et de la forme, au manque d'information sur les documents manipulés et aux difficultés d'accès.

1 Créations lexicales : définition et utilisation

Les créations lexicales forment un sous-ensemble des néologismes pour lesquels la nouveauté repose uniquement sur la création d'un nouveau lexème (et non sur une évolution de sens ou une composition). Leur source principale est un processus de dérivation morphologique, principalement la suffixation. Quant à leurs motivations, on peut repérer comme principales catégories :

- des termes techniques : *aquamarquage*, *hémaglutination*, *immunofixation* ;
- des créations liées à des concepts récents : *pacser (se)*, *surencadrement*, *googler* ;
- des termes relevant de la langue populaire : *baisage*, *poilade* ;
- diverses créations transparentes dans leur interprétation : *pêchable*, *japonisation*, *europhobie*.

Leur détection se fait classiquement sur corpus en utilisant une liste de référence (généralement lexicographique), toute forme absente de la liste étant une création potentielle. Il est également possible de les repérer sur la base de leur fréquence : de telles créations sont généralement rares, voire des hapax.
Leur collecte intéresse particulièrement deux disciplines : la morphologie, qui trouve ainsi des données permettant d'affiner la description d'un processus dérivationnel ou d'un affixe particulier, mais aussi le traitement automatique du langage, qui peut ainsi enrichir les bases de données lexicales sur lesquelles nombre d'applications sont basées.

1.1 Utilisation en morphologie

Les travaux de description des affixes se basent traditionnellement sur des données issues de la lexicographie. Dans la plupart des cas ces données sont insuffisantes pour élaborer ou valider des modèles complets de leur fonctionnement. Des études récentes ont ainsi pu voir le jour grâce à la mise à disposition (via des méthodes automatiques de repérage sur le Web) de données nombreuses et récentes. Citons notamment les travaux de [Hathout et al 2004] sur le suffixe *-able*[1], qui ont permis de repérer des emplois qui ne peuvent être décrits par les principaux modèles en s'appuyant sur plus de 5000 attestations, et ceux de [Plénat et al 2002] sur le suffixe *-este*[2], qui au contraire se basent sur seulement une vingtaine de formes, alors qu'avant ces récoltes automatiques une seule avait pu être repérée. D'autres travaux enfin, comme [Dal et al 2004] étudient spécifiquement le phénomène de la concurrence suffixale, en se concentrant sur les cas où à partir d'un même verbe, plusieurs dérivés nominaux sont attestés, et dont un ou plusieurs sont des créations récentes non répertoriées dans les sources lexicographiques.

1.2 Utilisation en traitement automatique

Si un lexique nu est souvent de peu d'intérêt pour une application de TAL, un lexique morphologique est par contre largement utilisé, notamment par les outils d'analyse de corpus. La catégorisation des formes nouvelles apparaissant dans un corpus n'est généralement par un obstacle à une analyse de bas niveau (grâce à des procédés

[1] `http://www.univ-tlse2.fr/erss/ressources/morphologie/able-HPT.html`
[2] `http://www.univ-tlse2.fr/erss/membres/plenat/Deriveseneste.pdf`

de *word guessing*), mais certaines techniques plus complexes sont, elles, dépendantes de ressources enrichies et à large couverture. C'est dans ce cadre que les travaux présentés ici ont été utilisés pour enrichir la base Verbaction[3] qui comporte à l'heure actuelle 9000 couples *nom/verbe* tels que le nom est le nom d'action dérivé du verbe. Cette ressource est notamment utilisée par l'analyseur syntaxique Syntex [Bourigault et al 2005] pour l'héritage des structures argumentales acquises en corpus, ainsi qu'en recherche d'information pour l'expansion de requêtes.

2 Méthodes d'acquisition utilisant un moteur de recherche

La façon la plus directe (et la moins coûteuse en terme de développement) pour le repérage de ces nouvelles formes lexicales et d'utiliser la partie du Web indexée par un moteur de recherche généraliste, ainsi que le mode d'interrogation prévu par ce moteur. Deux méthodes ont pu être utilisées. La première, hypothético-déductive, est notamment celle proposée par l'outil Walim [Namer 2002], où des processus de dérivation sont automatiquement appliqués sur des bases connues, et les dérivations candidates ainsi obtenues sont recherchées telles qu'elles (e.g. gratiner → *?gratinage*, *?gratination*, *?gratinade*).
La seconde méthode, celle proposée par Webaffix [Tanguy et Hathout 2002] utilise quant à elle un principe inductif basé sur la possibilité qu'offr(ai)ent certains moteurs d'effectuer des recherches sur des formes sous-spécifiées en utilisant des jokers (du type "*age"). Cette approche nécessite bien entendu un ensemble de traitements spécifiques, notamment pour éliminer les diverses formes bruitées (noms propres, autres langues, erreurs typographiques, etc.). Cet outil a été utilisé de façon intensive et a permis un enrichissement massif des bases de données citées ci-dessus. Ces deux approches sont toutefois problématiques pour différentes raisons. La première ne permet simplement pas de repérer l'apparition de nouvelles bases, ces dernière devant être celles présentes dans des listes existantes et catégorisées (principalement lexicographiques). La seconde, qui échappe à cette critique, repose sur la disponibilité d'un moteur permettant des interrogations complexes. Ce fut le cas du moteur AltaVista jusqu'en 2003, avant son rachat par une société concurrente et la suppression de la possibilité d'une recherche par patrons. Dès lors, la méthode par induction n'est simplement plus possible et Webaffix n'est actuellement plus fonctionnel ! Ceci nous a conduit à développer une nouvelle approche, présentée ci-dessous.

3 Trifouillette, un outil de recherche automatique de créations lexicales sur le Web

Trifouillette est une application qui crée une base lexicale de toutes les formes lexicales nouvelles rencontrées en parcourant le Web, permettant de sélectionner celles dont le nombre d'occurrences est faible. Ses principales caractéristiques sont les suivantes :

- L'indépendance vis-à-vis des moteurs de recherche : elle est nécessaire, comme nous l'avons vu, pour pallier la faiblesse des modes d'interrogation proposés et pour des utilisations massives (les moteurs principaux limitant désormais les accès automatiques). Elle permet également de renouer avec une méthode inductive, et rend possible le repérage de dérivés dont la base est inconnue, comme *ViWiste* (passionné de voitures de la marque VolksWagen) ou *repositoire* (transfert de l'anglais *repository*). D'autre part, cette autonomie nous place à l'abri des changements brusques des moteurs et assure ainsi la pérennité de la solution développée (sauf changement majeur de l'architecture du Web).
- La recherche se veut non ciblée : lors du parcours du Web, toutes les formes lexicales nouvelles rencontrées sont stockées et non uniquement certaines formes correspondant à des critères prédéterminés.
- Il s'agit d'une architecture spécifique, plus légère que celle d'un moteur générique. Notamment, le stockage des pages (nécessaire pour un accès au contexte) est limité aux documents contenant une forme rare, et le mode d'interrogation de la base se concentre essentiellement sur les formes lexicales seules.

3.1 Architecture et fonctionnalités

Trifouillette est composée de trois modules : un robot ou *crawler*, un module de gestion des données et une interface utilisateur.
Le *crawler* se charge du parcours du Web et constitue une partie purement technique, mais non triviale, étant donné la déviance par rapport aux standards et normes relatifs au Web[4]. Nous avons déployé les techniques habituelles de ce genre d'outils, avec un parcours récursif des liens hypertextes à partir d'un ensemble de pages initiales. Les principaux types de liens sont traités (liens HTML classiques, redirections, scripts, etc.).
Un analyseur extrait des pages visitées ces liens, alimentant ainsi la base des pages à traiter ultérieurement, et les formes lexicales d'autre part. Celles-ci alimentent le module de gestion de données, composé d'une base lexicale et d'un ensemble de pages téléchargées. Ce module sait retrouver, pour une forme donnée, un ensemble de pages la contenant. L'utilisateur peut ainsi interroger la base en cherchant une forme ou un patron donné. Une liste de formes lui est alors proposée, ainsi que la possibilité d'accéder aux différents contextes et d'annoter les lexèmes.

[3] `http://www.univ-tlse2.fr/erss/ressources/verbaction/`

[4] Cette déviance est notamment encouragée par la permissivité des navigateurs qui prennent à leur charge la rectification de bon nombre d'erreurs. Charge aux auteurs de robots d'en faire de même.

3.2 Filtrage de l'information pertinente

C'est désormais un lieu commun que de mentionner qu'outre l'attrait que revêt le Web, à travers la masse de données qu'il recèle et la diversité de celle-ci, une part non négligeable de déchets (pour l'intérêt de nos recherches) est présente sur le réseau. Cela ne remet nullement en cause les bénéfices tirés de l'utilisation du Web dans nos travaux, mais, partant de ce constat, la majeure partie de nos efforts va se concentrer sur le filtrage de l'information pertinente. Pertinence signifie dans notre cas présence de véritables formes lexicales dans les documents, situées dans des contextes rédigés en langue française, exclusion étant faite de sigles, fautes de typographie ou d'orthographe, etc. Les différentes étapes de filtrage sont récapitulées fig. 1 et explicitées ci-après.

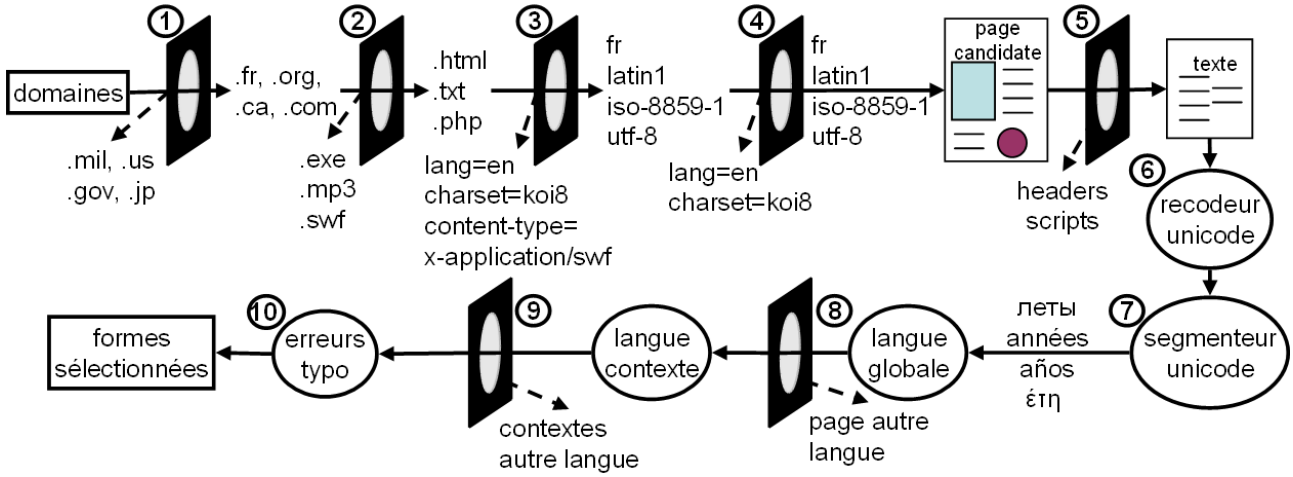

FIG. 1 – Chaîne de filtrage (flèches pleines : acceptation, flèches pointillées : rejet)

Caractéristiques globales des pages Au cours de la phase précédant l'analyse d'une page, plusieurs opportunités nous sont données de collecter des informations *a priori*, qui nous dispensent d'aller plus avant dans le traitement : sélection des domaines par domaine de premier niveau[5] (ou *top-level domain*, fig. 1, filtre 1), filtrage des liens par l'extension des noms de fichiers (supprimant ainsi les médias non textuels, fig. 1, filtre 2), sélection des pages selon les méta-données fournies, *i.e.* la langue, le type de contenu (texte ou autre) et le jeu de caractères utilisé. Ces méta-données, parfois fournies par les serveurs Web (en-têtes HTTP, fig. 1, filtre 3), rarement insérées dans les pages par leurs auteurs (méta-balises HTML, fig. 1, filtre 4), sont souvent absentes. Une page passant ces différents filtres va être analysée suivant les étapes ci-dessous.

Codages et segmentation en mots La première étape consiste à identifier le contenu textuel de la page (en-têtes, scripts : fig. 1, filtre 5) et à segmenter celui-ci en mots. Une attention particulière est portée au codage des caractères, un problème classique est celui du texte « *les problèmes soulevés par l'étude du français* » qui, codé dans le jeu de caractères *UTF8* dans une page déclarée en *latin1* serait analysée comme « *les problÃ¨mes soulevÃ©s par l'Ã©tude du franÃ§ais* » et entraînerait le repérage des formes {*les, probl, mes, soulev, s, par, l, tude, fran, ais*}. Un ensemble de tests et de conversions permet de représenter le contenu textuel en unicode (fig. 1, étape 6).

À partir de là, la segmentation en mots se fait aisément suivant les caractères délimiteurs classiques : espaces et signes de ponctuation (fig. 1, étape 7).

Détection de la langue L'étape suivante est la détection automatique de la langue (les informations globales n'étant pas fiables et les documents pouvant être bilingues). Ceci est nécessaire d'une part pour ne pas traiter les langues hors de notre champ d'étude et également pour ne pas traiter certaines pages rédigées en français, néanmoins indésirables (textes non rédigés tels les annuaires de personnels d'entreprise, résultats d'examens, liste de produits, nomenclatures diverses, etc.)

Nous projetons à cette fin sur le résultat de la segmentation en mots un lexique du français ainsi que des listes de mots grammaticaux du français et de langues proches comme l'italien et l'espagnol. Notre critère pour décider qu'une page est "globalement en français" est une combinaisons des facteurs suivants :

- proportion importante de formes connues du français (appartenant au lexique de référence) ;
- proportion importante de mots grammaticaux du français ;
- proportion importante de formes différentes ;
- proportion faible de mots grammaticaux d'autres langues.

Cette opération détermine la prise en compte d'une page au niveau global (fig. 1, sélecteur et filtre 8). Nous la réitérons au voisinage de chaque forme candidate afin d'ignorer les passages rédigés en langue étrangère (citations, pages bilingues, etc. : fig. 1, sélecteur et filtre 9).

Cette stratégie empêche l'insertion dans la base lexicale d'un nombre important de formes étrangères mais ne résout pas entièrement le problème : non seulement certaines pages, rédigées par exemple en espagnol, atteignent des scores qui surpassent ceux de pages françaises pertinentes, mais de plus, il nous est difficile de traiter de cette manière

[5] En supprimant les .mil, .gov, .de, .jp, .kp, etc. dont on suppose que le nombre de pages en français est anecdotique.

l'ancien français, l'occitan, le provençal, la catalan, etc. Les mots grammaticaux de ces langues sont souvent les mêmes et leur lexique n'est pas suffisamment disjoint de celui du français moderne.

Formes bruitées À ce stade, deux principales sources de bruit apparaissent majoritairement au niveau des formes retenues.

La première est liée à la présence ou absence d'espaces dans les documents. Leur origine en est souvent un *copier-coller* malencontreux d'un éditeur de texte vers un éditeur HTML ("*morphologie*" devient "*mor phologie*", "*la linguistique*" devient "*lalinguistique*"). Ce phénomène entraîne bien entendu la détection de formes non-pertinentes.

La seconde est constituée des fautes d'orthographes. La gamme d'illustrations de ce phénomène est large, de la coquille située au sein d'un texte globalement correct, aux forums inondés de fautes, d'abréviations, de *cyberlangue*, etc. Certains auteurs ne font que peu de cas de la correction, mais d'autres, beaucoup mieux intentionnés causent tout autant de difficultés à notre entreprise, comme les conseils de bon usage : « *on ne dit pas culturé mais cultivé !* », « *ne pas prononcer aréoport, mais aéroport* », « *on n'écrit pas acueil, mais accueil* », etc.

Attitude face aux erreurs Les deux types de problèmes précédents peuvent naturellement être détectés et traités automatiquement à l'aide de stratégies de réparation et d'un lexique de référence. Il est ainsi facile de repérer que *tellesque* n'est pas un dérivé en *-esque* mais bien une collision de *telles que*. En revanche, la forme *pâquestes*, également rare et inconnue du lexique, peut être scindée sur le même principe en *pâques* et *tes*. Ce terme a été relevé dans le message d'un forum, daté du 28 mars (période de Pâques), dans le court passage « *bises pâquestes* ». Étant donnée la rareté des dérivés en *-este*, se priver de celui-ci serait un tort. Le même problème se pose pour les insertions et omissions de lettres : *hivernable* pourrait être corrigé en *hivernale*, de même qu'*entarter* pourrait être corrigé en *entartrer*, éliminant la collecte de ces deux formations pertinentes.
D'une manière générale, nous avons décidé de ne pas surcorriger et de stocker le maximum de formes, quitte à noyer l'information pertinente dans un bruit important. Le système se contente de proposer une indication sur l'estimation de la pertinence d'une forme (fig. 1, étape 10) et un système d'annotation manuel *a posteriori*.

Cas limites Conjointement aux difficultés exposées plus haut, il demeure des erreurs difficilement décelables et traitables, voire des erreurs volontaires et stylistiques. On peut citer de manière non exhaustive les forums de médecine où s'expriment des patients atteints de dyslexie et de dysorthographie, les institutions qui jouent avec la langue (Oulipo, Collège de Pataphysique), les exercices à trous (les trous pouvant représenter des parties de mots) et... les linguistes : « *paradigme d'un verbe inexistant [...] par exemple, on pourrait conjuguer* rutambrer *ou* brédir, *mais pas* jitre ».

3.3 Résultats et perspectives

L'exploitation de notre application en étant encore à ses débuts, il est trop tôt pour dresser un bilan définitif. Néanmoins, d'après nos premières observations, notre moteur visite 100 000 à 600 000 pages par jour, dont une partie n'est pas analysée ou rejetée après analyse. Cela représente 2 à 35 millions de mots traités par jour parmi lesquels 2 000 à 70 000 nouvelles entrées dans la base. Nous avons atteint au bout de la première semaine 3 millions de pages stockées et 580 000 formes lexicales. Cette première moisson a déjà apporté son lot de formes lexicales nouvelles. Une évaluation manuelle sur un échantillon, tiré aléatoirement, de 100 hapax et 100 entrées de fréquence inférieure à 100 montre qu'environ 15% des formes sont pertinentes. Ce score de précision est nettement plus élevé dès lors qu'on ne considère que les formes correspondant à un patron suffixal. Nous comptons améliorer ce résultat par un traitement plus complet des erreurs, déjà développé mais non mis en place au moment de l'évaluation.
Une interface utilisateur permet l'interrogation par expressions régulières, l'annotation collaborative manuelle de la pertinence des formes et la possibilité de définir ses requêtes avec prise en compte d'un anti-lexique. Enfin, un système d'alerte par e-mail permet de signaler les nouvelles trouvailles (pour un patron donné) au fur et à mesure du parcours du Web.
Ces fonctionnalités sont pour l'instant restreintes aux seuls membres de notre équipe de recherche, la phase de mise au point étant toujours en cours. Une version accessible à tous est prévue pour le courant de l'année. Enfin, nous sommes tout à fait ouverts à toute collaboration sur ce sujet, que ce soit pour ajouter des traitements ou détections spécifiques sur l'architecture existante, ou pour fusionner les principales tâches (notamment le crawl) avec des projets connexes de la communauté du TAL.

Bibliographie

Bourigault D., Fabre C., Frérot C., Jacques M.-P. et Ozdowska S. (2005) : *Syntex, analyseur syntaxique de corpus*, Actes de TALN'05, Dourdan.
Dal G., Lignon S., Namer F. et Tanguy L. (2004) : *Toîle contre dictionnaires : analyse morphologique en corps de noms déverbaux concurrents*, Colloque Noms Déverbaux, Lille.
Hathout N., Plénat M. et Tanguy, L. (2004) : *Enquête sur les dérivés en* -able. Cahiers de Grammaire, N°28.
Namer, F. (2002) : *WALIM : Valider les unités morphologiques complexes par le Web* Actes du forum Silexicales n°3.
Plénat M., Lignon S., Serna N. et Tanguy L.(2002) : *La conjecture de Pichon* Corpus, N°1, pp 105-150.
Tanguy L. et Hathout N. (2002) : *Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web* Actes de TALN'02, Nancy.