Association automatique de lemmes et de paradigmes de flexion à un mot inconnu

Claude de Loupy (1,2), Michaël Bagur (1), Helena Blancafort (1,3)

(1) Syllabs – 15, rue Jean Baptiste Berlier, 75013 Paris
blancafort@syllabs.com; bagur@syllabs.com; loupy@syllabs.com
(2) MoDyCo – Université Paris 10, 200 Av. de la République, Nanterre
(3) Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona

Résumé

La maintenance et l'enrichissement des lexiques morphosyntaxiques sont souvent des tâches fastidieuses. Dans cet article nous présentons la mise en place d'une procédure de *guessing* de flexion afin d'aider les linguistes dans leur travail de lexicographes. Le *guesser* développé ne fait pas qu'évaluer l'étiquette morphosyntaxique comme c'est généralement le cas. Il propose pour un mot français inconnu, un ou plusieurs candidats-lemmes, ainsi que les paradigmes de flexion associés (formes fléchies et étiquettes morphosyntaxiques). Dans cet article, nous décrivons le modèle probabiliste utilisé ainsi que les résultats obtenus. La méthode utilisée permet de réduire considérablement le nombre de règles à valider, permettant ainsi un gain de temps important.

Abstract

Lexicon maintenance and lexicon enrichment is a labour-intensive task. In this paper, we present preliminary work on an inflectional guessing procedure for helping the linguist in lexicographic tasks. The guesser presented here does not only output morphosyntactic tags, but also suggests for an unknown French word one or more lemma candidates as well as their corresponding inflectional rules and morphosyntactic tags that the linguist has to validate. In this article, we present the probabilistic model we used as well as obtained results. The method allows a drastic reduction of the number of rules to validate.

Mots-clés : *guesser*, lexiques morphosyntaxiques, aide aux linguistes, induction des règles de flexion

Keywords: guesser, morphosyntactic lexica, aide to the linguist, induction of inflection rules

1 Introduction

Les ressources lexicales sont fondamentalement importantes en Traitement Automatique des Langues (TAL). Les lexiques morphosyntaxiques en particulier sont très utilisés par des traitements de relativement bas niveau qui sont donc la base de traitements ultérieurs[1]. Or, la gestion et la maintenance de tels lexiques est très consommatrice en temps et en énergie. Il n'est pas rare qu'un organisme travaillant en TAL ait à gérer plusieurs lexiques de ce type, que ce soit pour une même langue avec des spécialisations de ressources ou pour le traitement de plusieurs langues. Du fait de la créativité des langues, les lexiques doivent être en permanence mis à jour et il est important de penser des méthodes permettant de faciliter ce travail.

L'un des moyens les plus courants pour enrichir ces lexiques consiste à analyser des corpus, à récupérer les formes inconnues du lexique à enrichir et à présenter les plus courantes à un/e linguiste qui se charge alors d'introduire non seulement la forme en question mais également ses étiquettes morphosyntaxiques, son ou ses lemmes ainsi que toutes les formes fléchies du ou des lemmes en question. Cette tâche est généralement effectuée à la main ou en utilisant des outils spécifiques (fléchisseur) permettant de fléchir un lemme selon des paradigmes de flexion connus. Même dans ce dernier cas. L'ensemble de la tâche prend beaucoup de temps car il est nécessaire de décider quel est le lemme correspondant à une forme ainsi que le numéro ou la dénomination de la règle de flexion qui doit lui être associée.

L'objet de cet article est d'évaluer les performances de méthodes classiques utilisées dans les *guessers* de catégories grammaticales lorsqu'on les applique à l'association de couples (lemme, paradigme de flexion) à un mot inconnu. Un état de l'art est donné en section 2. Les fondements et les buts de l'expérience décrite sont ensuite exposés en section 3 afin d'expliquer pourquoi les méthodes précédentes ne sont pas satisfaisantes pour les conditions visées. La section 4 présente la méthode utilisée. Enfin, la section 5 présente l'évaluation des résultats.

2 État de l'art

L'idée de fournir une aide supplémentaire aux linguistes dans cette étape d'enrichissement court depuis déjà un certain temps. Les premiers travaux en construction automatique de lexiques morphosyntaxiques n'utilisaient qu'un corpus comme source d'information. Ainsi, Jacquemin (1997) compare les terminaisons des mots de son corpus de manière à grouper des mots ayant un même stem puis regroupe les groupes de mots ayant la même suite de terminaisons. Goldsmith (1997 ; 2000) effectue également ce type de procédure en appliquant les méthodes de *Minimum Description Length* (Rissanen, 1989) et d'*Expectation Maximization* (Dempster et al., 1977)[2]. Schone & Jurasky (2000 ; 2001) utilisent la méthode du Latent Semantic Analysis (Deerwester et al., 1990).

---

[1] Nous n'entrerons pas ici dans un débat sur l'utilité ou non de ces lexiques en comparaison à d'autres moyens d'analyse plus sommaire comme les *stemmers* de type Porter (1980). Nous pensons que ces lexiques sont utiles et de nombreux autres travaux les utilisent, ce qui suffit à justifier un travail sur leur création.

[2] L'outil *Linguistica* résultant de ces travaux est téléchargeable à l'adresse suivante : http://linguistica.uchicago.edu/.

Dans tous les cas, les systèmes proposés produisent des listes de stems associés à des paradigmes de flexion permettant de générer les formes fléchies trouvées dans les corpus analysés. Le principal problème de ces méthodes est que les résultats générés sont très difficilement utilisables par des linguistes car ils n'ont que peu de liens avec les règles morphologiques auxquelles on s'attend. Ces travaux sont donc principalement utilisables dans un contexte totalement automatique avec toutes les erreurs que cela peut comporter.

D'autres expériences utilisent des ressources lexicales existantes ou des paradigmes de flexion déjà connus en plus des corpus à analyser. Nakov et al. (2003), travaillent sur l'allemand et commencent par extraire, de manière automatique, toutes les terminaisons possibles pour les mots présents dans un lexique. Ils extraient ensuite les mots présents dans un corpus qui sont inconnus dudit lexique, puis génèrent tous les stems possibles à partir des terminaisons pour chacun de ces mots. Ils utilisent enfin l'algorithme de *Maximum Likelihood Estimation* avec des paramètres préconisés par Mikheev (1997) afin de récupérer les règles de flexion les plus intéressantes (notions de qualité, de longueur et de fréquences).

Oliver & Tadiç (2004) présentent l'application sur le Croate de méthodes testées auparavant sur le russe (Oliver et al., 2003). Ils s'appuient sur un lexique morphosyntaxique existant dont ils extraient des paradigmes de flexion et sur un corpus à partir duquel ils complètent le lexique précédent. Le processus est découpé en 4 étapes : 1. Découpage des formes en un couple (*stem*, *ending*) à partir de toutes les terminaisons possibles puis regroupement de ces couples dans des paradigmes possibles issus du lexique existant. 2. Sélection pour chaque mot du paradigme permettant d'obtenir le plus de formes présentes dans le corpus. 3. Récupération des cas non décidables (même nombre d'entrées entre deux paradigmes possibles pour un mot). 4. Utilisation d'Internet pour résoudre les cas précédents (recherche de la présence sur Internet des différentes formes fléchies possibles à partir d'un mot et d'un paradigme) : les paradigmes ayant les formes les plus présentes sont validés.

Clément (2004) travaille sur l'extraction d'un lexique morphologique français à partir d'un corpus en s'appuyant sur des couples (lemme, forme) et en associant à un lemme une probabilité d'autant plus forte que beaucoup de formes peuvent lui être associées[3]. Ces travaux ont également été effectués sur le slovaque (Sagot, 2005) et le polonais (Sagot, 2007). Cette méthode, aussi puissante soit-elle et bien que permettant de produire rapidement des lexiques très volumineux, présente l'inconvénient de générer des entrées incomplètes. Ainsi, le Lefff contient des verbes dont toutes les formes ne sont pas présentes mais seulement celles qui ont été repérées dans le corpus. Il s'agit donc d'une méthode permettant de créer un lexique dédié à un corpus et non un lexique de langue.

Zanchetta & Baroni (2005) se base également sur un corpus pour extraire un lexique morphosyntaxique de l'italien[4]. Pour cela, ils commencent par utiliser TreeTagger (Schmid, 1994) afin d'obtenir la catégorie grammaticale et le lemme d'une forme donnée. Les lemmes ainsi obtenus ont ensuite été fléchis à l'aide de règles de flexions de l'italien. Nous noterons dans ce cas que le lexique généré contient l'ensemble des flexions d'un lemme donné. En revanche, la méthode d'application des règles est très manuelle puisque les auteurs ont

[3] Ces travaux ont permis la création d'un lexique morphosyntaxique du français, le Lefff, disponible gratuitement à l'adresse suivante : http://www.labri.fr/perso/clement/lefff.

[4] Ces travaux ont permis la production du lexique Morph-it!, disponible à l'adresse http://dev.sslmit.unibo.it/linguistics/morph-it.php.

appliqué des règles d'analyse des formes à fléchir avec de nombreuses corrections manuelles de manière à obtenir un résultat satisfaisant.

Par ailleurs, plusieurs expériences ont été effectuées dans un contexte lié plutôt à la morphologie dérivationnelle qui nous intéresse moins ici comme par exemple Dal & Namer (2000), Namer (1999), Hathout (2005), Hathout & Tanguy (2005).

3 Fondements et buts de l'expérience

Les travaux présentés ici se place dans un contexte plus large de construction de chaîne de traitement permettant d'aider les linguistes dans leur tâche de codage de ressources linguistiques (Loupy & Gonçalves, 2008). L'un des points de cette chaîne concerne l'inclusion de mots inconnus dans un lexique morphosyntaxique afin d'en augmenter sa couverture, soit sur un corpus spécialisé, soit sur la langue générale. Ces ajouts doivent être proposés aux linguistes selon un format correspondant à celui utilisé pour coder le lexique.

Habituellement, les lexiques morphosyntaxiques sont constitués de triplet (*forme*, *lemme*, *catégorie*). La figure suivante montre un exemple typique issu du lexique français MulText (Ide and Véronis, 1994).

```
abaisse            abaisser        Vmip3s-
abaissons          abaisser        Vmip1p-
brioche            brioche         Ncfs--
brioches           brioche         Ncfp--
rends              rendre          Vmip1s-
rends              rendre          Vmip2s-
rend               rendre          Vmip3s-
rendons            rendre          Vmip1p-
statisticienne     statisticien    Ncfs--
statisticiennes    statisticien    Ncfp—
```

Figure 1 : Structure habituelle d'un lexique morphosyntaxique

Ce type de format est difficile à manipuler et rend les lexiques difficiles à maintenir, à enrichir et à contrôler pour éviter des erreurs. Chaque forme fléchie est elle-même une entrée du lexique et est liée à son lemme et à l'ensemble de ses étiquettes. Pour des langues à flexion riche, cela conduit à un très grand nombre d'entrées pouvant aller jusqu'à plusieurs millions dans certaines langues comme le russe.

Afin d'avoir un meilleur contrôle de nos ressources lexicales, notre lexique SylLex est bâti sur une structure de type (*lemme*, *paradigme*) dans lequel les paradigmes décrivent l'ensemble des opérations de flexion à associer au lemme afin de générer ses flexions. Un format similaire est utilisé pour coder le DELAS (Gálvez, 2003). La figure suivante donne un exemple de ce format.

```
abaisser        V1
brioche         N1
rendre          V9
statisticien    N13
```

Figure 2 : Structure de SylLex

Les règles sont définies de façon à décrire les opérations de construction de flexions à partir des règles comme indiqué dans la figure suivante[5].

```
V1 0/a/Vmif3s--|0/ont/Vmif3p--|0/ai/Vmif1s--|0/ons/Vmif1p--|0/as/Vmif2s--
|0/ez/Vmif2p--|2/ait/Vmii3s--|...
```

Figure 3 : Description des paradigmes

Les paradigmes et le lexique peuvent être accédés via une interface rendant plus simple la manipulation des informations. Cette architecture de lexique est beaucoup plus facile à manipuler et à appréhender pour les linguistes.

C'est donc dans ce contexte qu'il a été décidé de construire une chaîne de traitement des mots inconnus. Les méthodes présentées plus haut présentent toutes des inconvénients par rapport à cela. Celles qui n'utilisent que les corpus ne peuvent permettre de proposer des suggestions propres par rapport à un format faisant référence à des paradigmes de flexions établis. Les méthodes qui se basent sur une association (*forme*, *lemme*) ne permettent pas non plus de se projeter de manière fiable sur ces paradigmes. L'extraction automatique de paradigmes à partir d'un lexique de type (forme, lemme, tag) présente d'ailleurs toujours des erreurs ou des complexités inutiles du fait de la présence de mots ambigus. Ainsi, le verbe *payer*, du fait de ses variantes (*paye*/*paie*) génèrera un paradigme différent du paradigme canonique du verbe *aimer* alors que la prise en compte de variantes et donc d'un *stem* supplémentaire permet d'éliminer ce problème.

Seule la méthode de Zanchetta & Baroni pourrait convenir mais elle est beaucoup trop manuelle et demanderait un travail très important de constitution d'un fléchisseur automatique pour chaque nouvelle langue.

Nous avons donc mis en place une méthode permettant d'associer directement un couple (*lemme*, *paradigme*) à un mot inconnu. Cette méthode peut ensuite être couplée avec les processus présentés dans les autres travaux afin de donner plus de poids à des couples dont les formes générés sont présentes dans un corpus ou sur Internet.

Le lexique français sur lequel ont été faites les expériences présentées dans cet article contenait 60 000 couples (*lemme*, *paradigme*), il s'agit donc d'un petit lexique qui demande justement un important travail pour en augmenter la taille.

4 Méthode utilisée

Le travail présenté ici est basé sur des méthodes classiquement utilisées dans les *guessers* et adapté à la génération de couples (*lemme*, *paradigme*).

4.1 Les guessers

Il y a peu de publications dédiées aux *guessers*. La plupart du temps, les procédures de *guessing* sont décrites à l'intérieur de descriptions de *taggers* dès que l'on parle de mots

[5] On pourra noter que l'encodage des étiquettes est basé sur les spécifications de MulText (Ide & Véronis, 1994). Malgré quelques changements, ce format nous a paru le plus pratique et il a surtout l'avantage d'avoir été testé sur au mois 20 langues (Véronis & Khouri, 1995).

inconnus (Chanod & Tapanainen, 1995 ; Schmid, 1995). Mikheev (1997) considère que les guessers de catégorie (*Part Of Speech* ou *POS*) peuvent utiliser trois indices :

1. Les préfixes. Si *impossible* est un mot inconnu mais que *possible* est connu dans le lexique, il est probable que le POS du mot *impossible* soit le même que celui du mot *possible*. La précision de cet indice est élevée mais la couverture est en revanche très faible (respectivement 93,5% et 6,5% selon Mikheev).

2. Les suffixes. Dans beaucoup de langues, les suffixes indiquent des propriétés grammaticales (pluriel, temps, etc.). La précision de cet indice est encore plus élevée (96,8%) mais la couverture reste limitée (26,5%).

3. Les terminaisons (*endings*). Les terminaisons sont les dernières lettres des mots. Il ne s'agit pas de suffixes (ou alors elles le sont par hasard) car elles peuvent être plus longues ou plus courtes que les suffixes réels des mots dont elles sont extraites. Elles n'ont pas de signification grammaticale. Selon Mikheev, les terminaisons permettent d'obtenir des performances de l'ordre de 91,9% en précision et 78,2% en rappel.

Le *guesser* décrit ici ne fait appel qu'aux terminaisons mais nous comptons utiliser d'autres informations, les suffixes pouvant facilement être extraits de nos paradigmes.

4.2 Description du *guesser* utilisé

Habituellement, les *guessers* sont utilisés pour calculer le POS le plus probable pour un mot inconnu, c'est-à-dire $P(t|w)$ où t représente une étiquette morphosyntaxique et w le mot inconnu. Ici, nous devons évaluer la probabilité $P(l, R|w)$ où l_j représente un lemme et R un paradigme.

En fait, étant donné que les paradigmes de flexion donnent toutes les formes qui sont associées à un lemme, il suffit de calculer la probabilité $P(t, R|w)$. Le fait de connaître le POS et la règle permet de retrouver très facilement le lemme tout en conservant l'information sur le mot analysé. Une fois le paradigme et l'étiquette trouvés, le lemme peut être déduit sans risque d'erreur.

Les statistiques sont apprises sur le lexique existant en utilisant un arbre classique sur les terminaisons. Les terminaisons considérées sont de longueur 1 à 5. Pour chaque mot inconnu, 5 terminaisons sont utilisées, représentées par les n ($1 \leq n \leq 5$) dernières lettres du mot inconnu. Pour chaque terminaison, un ou plusieurs lemmes candidats sont associés avec une fréquence calculée sur le lexique. Une terminaison peut être associée à plus d'un couple (*lemme*, *paradigme*).

L'approximation suivante est effectuée :

$$P(t, R|w) \approx \sum_{i=1}^{5} \rho_i * P(t, R|[w]_i)$$

où $[w]_i$ représente la terminaison de longueur *i* du mot *w*.

et ρ_i est un facteur de lissage. Elle est basée sur l'entropie selon la formule suivante :

$$\rho_i = \frac{1 - H_i}{\sum_{j=1}^{5}(1 - H_j)}$$

où H_i représente l'entropie des terminaisons de longueur *j* au sein du lexique (vis-à-vis de leur association à un couple (t, R)). On a

$$H_i = \frac{\sum_{n \in \mathcal{N}_i} h(n)}{Card(\mathcal{N}_i)}$$

$$h(n) = -\sum_{(t,R) \in \mathcal{T}_n} P(t,R) * \log\left(P(t,R)\right)$$

Où $\mathcal{N}_i$ représente l'ensemble des nœuds à la profondeur i (chaque nœud contient une seule terminaison $[w]_i$) et $\mathcal{T}_n$ est l'ensemble des paires (t, R) possibles au nœud n.

Lorsqu'un mot inconnu est rencontré, toutes les probabilités $P(t, R|w)$ sont calculées pour tout couple (t, R) étant associé à l'une des terminaison $[w]_i$ du mot w. Seuls les couples dont le score est supérieur à un certain seuil θ sont conservés.

En dernier lieu, une simple vérification de cohérence est effectuée. Le calcul des probabilités est effectué sur les terminaisons mais la mise en relation avec un paradigme R permet d'accéder au suffixe de la forme associée au tag t correspondant. Il suffit alors de vérifier que le suffixe du paradigme est conforme au mot inconnu. Si ce n'est pas le cas, le couple (t, R) est éliminé des possibles.

Une fois un couple (t, R) il est alors immédiat de récupérer un couple (l, R) donnant le lemme du mot inconnu. Ce couple est présenté aux linguistes avec l'ensemble des flexions possibles.

Comme précisé plus haut, il est tout à fait possible (et nous le ferons) dans cette dernière phase, de vérifier que les formes ainsi générées existent dans un corpus ou sur le web.

5 Évaluation

Seules les classes ouvertes (adjectif, adverbe, non, verbe) ont été traitées. L'évaluation a été effectuée en prenant aléatoirement 90% du lexique pour l'entraînement et 10% pour les tests. 10 permutations ont été effectuées afin d'éviter des problèmes locaux spécifiques. Les performances sont indiquées dans le tableau suivant. Les chiffres représentent les scores moyens sur les 10 passes et la première colonne indique le seuil θ utilisé pour la sélection des solutions.

θ	Précision	Rappel	Nombre de couples (lemme, paradigme) proposés
0	14.3%	91.5%	14.8
0.025	51.8%	90.5%	2.7
0.05	68.4%	84.5%	1.8
0.075	75.6%	73.7%	1.3
0.1	80.2%	66.9%	1
0.15	85.8%	52.6%	0.7
0.2	88.9%	40.2%	0.5

Table 1 : Performances du guesser

Ces résultats sont du même ordre que ceux trouvés par Oliver & Tadiç (2004) puisque pour une précision de 84,5, ils obtiennent un rappel de 38,4. Il est impossible de pousser la comparaison plus loin car les langues de travail sont différentes mais nos résultats semblent donc cohérents.

Étant donné que notre but est d'aider les linguistes dans leur tâche, il n'est pas envisageable de proposer près de 15 couples à valider (tous les couples possibles sont fournis avec $\theta = 0$). De plus, plusieurs méthodes présentées dans l'état de l'art vérifient l'existence d'une forme fléchie générée au sein d'un corpus ou sur Internet. Si 15 paradigmes de flexion sont associés en moyenne à un mot inconnu, sachant qu'il y a en moyenne 25 flexions par paradigmes, cela fait 375 requêtes par mot inconnu en moyenne. Le temps nécessaire devient alors beaucoup trop important, si l'on veut générer des propositions rapidement. En particulier, pour des recherches sur Internet, 1000 mots inconnus demandent 375 000 requêtes, ce qui ne se fait pas en une nuit et risque de poser des soucis avec le moteur utilisé.

Nous pouvons constater que l'utilisation d'un seuil très faible permet à la fois de diminuer considérablement le nombre de propositions (plus de 5 fois moins), d'augmenter de manière très nette la précision (3,6 fois plus élevée) tout en ne diminuant quasiment pas le rappel (1 point de perte).

Le calcul d'une probabilité $P(l, R|w)$ permet donc d'améliorer de manière très nette la vitesse de codage, que ce soit par la présentation de moins de possibilités aux linguistes ou par la diminution des vérifications à effectuer. Par ailleurs, cette méthode permet d'obtenir des entrée de lexique qui sont complètes puisque générées à partir de paradigmes de flexion validés au préalable.

6 Conclusion et perspectives

Le système décrit dans cet article permet d'associer à un mot inconnu des couples (*lemme*, *paradigme de flexion*). Les performances ne sont pas très élevées si on les compare à celles qu'obtiennent les guessers sur les étiquettes morphosyntaxiques. Néanmoins, l'utilisation de cette méthode permet de proposer peu de choix à la validation humaine tout en gardant une très bonne couverture. Cela permet également de diminuer considérablement le nombre de formes à tester, que ce soit sur un corpus local ou sur le web.

Par ailleurs, de nombreuses pistes d'amélioration sont à envisager. Déjà, la vérification de l'existence des formes fléchies générées par les couples (l, R), telle que plusieurs travaux la pratiquent, devrait apporter une nette amélioration. De plus, les indices que représentent les préfixes, les suffixes grammaticaux et le contexte dans lesquels apparaissent les mots inconnus devraient également donner des résultats encore plus intéressants.

Enfin, les expériences présentées ici ont été menées en utilisant l'ensemble des règles de flexion. Or, certains paradigmes non productifs (verbe *être* par exemple) polluent les probabilités ci-dessus. Il convient donc de refaire ces expériences en évaluant la productivité des paradigmes utilisés. Dans le même ordre d'idée, la fréquence des paradigmes dans le lexique d'apprentissage n'a pas été utilisée. Or, par analogie avec d'autres phénomènes linguistiques, nous pouvons supposer que le simple fait d'associer aux inconnus les règles selon leur fréquence devrait permettre d'augmenter encore les performances.

Le traitement des formes polylexicales demandera une attention particulière. Cependant, le codage adopté pour ces formes dans notre lexique (règle de flexion à appliquer aux

composants si c'est le cas + connaissance de la tête du mot composé s'il y en a une) nous permet de faire un apprentissage assez simple en reprenant la méthode présentée ici. Nous testerons les résultats obtenus ainsi.

Lors de prochaines expériences, nous évaluerons les lexiques produits en termes de couverture et le temps que de tels outils peuvent faire gagner lors du codage d'informations morphosyntaxiques par des linguistes. Bien que cette évaluation puisse être biaisée par l'interface qui sera utilisée, les résultats en sont importants pour déterminer s'il est possible de créer des lexiques fiables et complets de manière rapide en conservant une validation manuelle dans la boucle.

Enfin, nous comptons également effectuer des comparaisons de performances de langue à langue dans le même esprit que de précédentes expériences de comparaison (Blancafort & Loupy, 2008).

Références

BLANCAFORT H., LOUPY C. DE, (2008). Comparing languages from vocabulary growth to inflection paradigms – A study run on parallel corpora and multilingual lexicons. Actes de SEPLN'2008. Madrid, Espagne.

CHANOD J.P., TAPANAINEN P., (1995). Creating a Tagset, Lexicon and Guesser for a French Tagger; Proceedings of the European Chapter of the ACL SIGDAT Workshop From text to tags : Issues in Multilingual Language Analysis, pp. 51-57, Dublin, Irelande.

CLÉMENT L., SAGOT B., LANG B., (2004). Morphology based automatic acquisition of large-coverage lexica. In Proceedings of LREC'04, Lisbonne, Portugal. pp. 1841-1844.

CUCERZAN S., YAROWSKY D., (2000). Language Independent, Minimally Supervised Induction of Lexical Probabilities. Proceedings of ACL-2000, Hong Kong, pp. 270-277.

DAL G., NAMER F., (2000). GéDériF: automatic generation and analysis of morphologically constructed lexical resources. In actes de Second International Conference on Language Resources and Evaluation, Athens, Grèce, pp. 1447-1454.

DEMPSTER A.P., LAIRD N. M., RUBIN D. B., (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B 39(1):1-38.

GALVEZ, C., (2006). El diccionario electrónico: un instrumento para la unificación de términos en la indización automática. Linguax: Revista de Lenguas Aplicadas (ISSN 1695-632X).

HATHOUT N., TANGUY L., (2005). Webaffix : une boîte à outils d'acquisition lexicale à partir du Web. In Revue Québéquoise de Linguistique. Volume 32, numéro 1.

HATHOUT N., (2005). Exploiter la structure analogique du lexique construit : une approche computationnelle. In Cahiers de Lexicologie. Volume 87, numéro 2.

IDE N., VÉRONIS J., (1994). MULTEXT: Multilingual Text Tools and Corpora. Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, Kyoto, Japon, 588-92.

LOUPY C. DE, GONÇALVES S., (2008). Aide à la construction de lexiques morphosyntaxiques. Actes de EURALEX 2008. Barcelone, Espagne.

MIKHEEV A., (1997). Automatic Rule Induction for Unknown-Word Guessing. In Computational Linguistics vol 23(3), ACL 1997. pp. 405-423.

NAKOV P., BONEV Y., ANGELOVA G., GIUS E., VON HAHN W., (2003). Guessing Morphological Classes of Unknown German Nouns. In Proceedings of Recent Advances in Natural Language Processing (RANLP'03). pp. 319-326. Borovetz, Bulgarie.

NAMER F., (1999). Le traitement automatique des mots dérivés : le cas des noms et adjectifs en -et(te). in D. Corbin, G. Dal, B. Fradin, B. Habert., F. Kerleroux, M. Plénat & M. Roché éds, La morphologie des dérivés évaluatifs, Silexicales 2, pp. 169-179. Villeneuve d'Ascq.

OLIVER A., TADIĆ M., (2004). Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In: Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC 2004), p. 1259–1262. Lisbonne, Portugal.

OLIVER A., CASTELLÓN I., MÀRQUEZ L., (2003). Use of internet for augmenting coverage in a lexical acquisition system from raw corpora. In: Proceedings of the RANLP 2003 International Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages (IESL 2003). Borovets, Bulgarie.

PORTER M., (1980). An algorithm for suffix stripping. in Program, n°14, 130-137.

RISSANEN J., (1989). Stochastic Complexity in Statistical Inquiry. World Scientific Publishing Co, Singapour.

SAGOT B., (2005). Automatic acquisition of a Slovak Lexicon from a Raw Corpus. In Lecture Notes in Artificial Intelligence 3658 (Springer-Verlag), Proceedings of TSD'05, Karlovy Vary, République Tchèque. pp. 156-163.

SAGOT B., (2007). Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In: Proceedings of LTC 2007, Poznań, Pologne.

SCHMID H., (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing. September 1994.

SCHMID H., (1995). Improvements in part-of-speech tagging with an application to German. in Proceedings of the ACL SIGDAT-Workshop, pp. 47-50.

VÉRONIS J., KHOURI L., (1995). Étiquetage grammatical multilingue: le projet Multext. Traitement Automatique des Langues, 36(1/2), 233-248.