Aide à la construction de lexiques morphosyntaxiques

Claude de Loupy
Sandra Gonçalves
Syllabs

Morphosyntactic lexica are a very important resource for natural language processing. Many exist; some are freely available for research. But many organisms still produce lexica, even for languages with available resources. In this paper, we present some techniques that can be leveraged to produce lexica more efficiently. Firstly, the format of the lexicon is important. We use a very simple format based on the association of a lemma and a flexion rule, avoiding dozens of entries for a single lemma. Secondly, the linguist must describe some basic elements: the tag list, the tool words and the flexion rules. Thirdly, a specific guesser makes the completion of the lexicon easier. We describe two ways of adding entries to the lexicon using a guesser which associates a lemma and a flexion rule to a word, or a flexion rule to a lemma.

1. Introduction

Les lexiques morphosyntaxiques sont une ressource fondamentale pour le traitement automatique des langues. De très nombreux lexiques ont été produits dans le domaine. Ils associent un mot avec une ou plusieurs catégorie/s grammaticale/s et un ou plusieurs lemme/s. Pourtant, la production de lexiques morphosyntaxiques a été peu étudiée et les méthodes aidant le linguiste dans sa démarche de création encore moins. On peut citer Grabar et Zweigenbaum (1999) et Nakov *et al*. (2003).

Dans cet article, nous décrivons une technique ayant pour but de faciliter le travail du linguiste pour la création d'une telle ressource. Nous présentons tout d'abord le lexique de Syllabs et expliquons les raisons du formalisme retenu par sa simplicité de manipulation. Nous décrivons ensuite les trois éléments de base d'un tel lexique qui doivent être développés en premier par le linguiste. Enfin, nous présentons deux méthodes d'aide au linguiste basée sur des guessers qui permettent d'associer à un mot son lemme et la règle de flexion associée ou une règle de flexion à un lemme.

2. Description de la ressource

Un lexique morphosyntaxique associe généralement à une forme une ou plusieurs catégorie/s grammaticale/s et un ou plusieurs lemme/s. La représentation la plus fréquente est du type suivant, issu du formalisme MulText (Ide et 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-
songe           songe         Ncms--
```

```
songes           songe         Ncmp--
statisticienne   statisticien  Ncfs--
statisticiennes  statisticien  Ncfp--
```

Figure 1: Lexique morphosyntaxique classique

Chaque ligne contient trois éléments: une forme fléchie, le lemme et les catégories morphosyntaxiques associées à la forme. Certains lexiques utilisent d'autres formats comme XML pour le lexique Morphalou (Romary *et al*. 2004) mais le principe de base est presque toujours d'associer à une forme une étiquette morphosyntaxique et un lemme. Or, ce format est difficile à gérer et sujet à de nombreuses erreurs. D'une part, la taille du lexique est très importante puisque, par exemple, un seul verbe français sera représenté par 51 formes fléchies. D'autre part, la vérification des entrées se fait forme par forme. Enfin, l'ajout d'entrées implique de gérer toutes les formes associées. Dans le meilleur des cas, le linguiste se fait alors aider par un fléchisseur mais l'opération reste assez lourde.

Nous avons donc décidé d'utiliser une forme de lexique beaucoup moins utilisée que l'on retrouve dans le DELAS de Gálvez (2003) et le Leff de Sagot (2006). Le concept de règles de flexions a été généralisé dans la définition même du lexique de Syllabs. Les entrées du lexique sont des lemmes associés à des règles de flexion. Par exemple:

```
abaisser      V1
brioche       N1
rendre        V9
songe         N2
statisticien  N13
```

Figure 2: Lexique morphosyntaxique "base de règles de flexions"

Où les `N1`, `N2`, `N13`, `V1` et `V9` font référence à des paradigmes de flexion associés à des descriptions morphosyntaxiques:

```
N1     0//Ncfs--|0/s/Ncfp--
N2     0//Ncms--|0/s/Ncmp--
N13    0//Ncms--|0/s/Ncmp--|0/ne/Ncfs--|0/nes/Ncfp--
V1     1//Vmip1s—|1/s/Vmip2s--|1//Vmip3s--|2/ons/Vmip1p--|...
V9     2/s/Vmip1s--|2/s/Vmip2s--|2//Vmip3s--|2/ons/Vmip1p--|...
```

Figure 3: Règles de flexion

La règle `V9` (qui s'applique à *rendre* par exemple), indique que pour générer la première forme, il faut supprimer les deux derniers caractères du lemme (*re*), ajouter le caractère *s* et on a alors le Verbe (`V`) non auxiliaire (`m`) de l'indicatif (`i`) présent (`p`) à la première (`1`) personne du singulier (`s`). Le format des catégories morphosyntaxiques est celui de MulText dont une forme très proche est utilisée dans Freeling (Carreras *et al*. 2004). Ainsi, les deux ressources ci-dessus permettent de générer un lexique morphosyntaxique de la même forme que celui de la Figure 1 sans en avoir les inconvénients. De plus, cette génération n'est absolument pas nécessaire en-dehors des outils d'analyse. Le linguiste manipule ainsi un lexique beaucoup plus simple à appréhender.

3. Phase de création des ressources de base

Trois ressources sont indispensables avant de commencer à alimenter un lexique morphosyntaxique: la liste des étiquettes particulières à la langue cible, la liste des mots outils et la liste des règles de flexion à utiliser.

3.1. *Liste des étiquettes*

Les étiquettes morphosyntaxiques indiquent la nature du mot qui est codé: son type grammatical (nom, verbe, etc.) son genre, son nombre, etc. Selon le niveau de finesse que l'on veut atteindre dans la description, ces étiquettes peuvent être assez complexes. La liste des étiquettes est

généralement différente d'une langue à une autre. Par exemple, les cas ne sont pas présents dans toutes les langues et ils sont différents selon la langue. Cela dépend aussi de l'application. On voudra peut-être repérer les mots composés dans les langues compositionnelles comme l'allemand.

Il existe de très nombreux formalismes de description morpho-syntaxiques plus ou moins détaillés. Nous avons choisi d'utiliser le formalisme MulText du fait de sa rigueur et parce qu'il a été adapté à plus d'une vingtaine de langues (bambara, bulgare, catalan, croate, tchèque, néerlandais, anglais, estonien, français, allemand, hongrois, italien, kikongo, lithuanien, occitan, russe, roumain, slovène, espagnol, serbe, suédois et swahili).

Des exemples d'étiquettes MulText sont donnés dans les Figure 1 et Figure 3. Notre lexique comporte 269 étiquettes différentes.

3.2. *Liste des mots outils*

Le simple fait d'introduire les mots outils dans un lexique permet souvent de couvrir plus de la moitié des occurrences des mots d'un corpus. Ils sont donc essentiels dans un lexique morphosyntaxique. Ils présentent aussi l'avantage d'être dans des classes fermées. Le comportement flexionnel des mots outils est cependant souvent particulier. L'une des premières tâches du linguiste doit donc être de s'attacher à les décrire de manière spécifique.

3.3. *Règles de flexion*

La dernière tâche de base est de créer la liste des règles de flexion (*cf*. Figure 3) ou du moins les règles de flexion les plus courantes. Le but n'est pas forcément de couvrir toutes les exceptions mais au contraire de rester dans la liste des règles très productives. Les exceptions pourront être gérées ultérieurement. Cette phase de création de règles de flexion peut être assez longue pour certaines langues très flexionnelles comme le finnois. Il est possible d'aider le linguiste en lui faisant valider des règles de flexion générées de manière automatique. Celles-ci peuvent être produites à partir d'un lexique existant même succinct. Le lexique Syllabs comporte actuellement environ 300 règles de flexion.

4. Ajout d'entrées au lexique

Une fois ces éléments de base créés, le linguiste doit compléter le lexique. Il faut donc disposer d'une liste de mots afin de pouvoir les ajouter. Deux solutions sont possibles. La première est basée sur un corpus et la seconde sur une liste de lemmes.

4.1. *Ajouts à partir d'un corpus*

Utiliser un corpus permet de travailler sur les mots les plus fréquents afin de les entrer en premier dans le lexique. On peut ainsi obtenir un lexique général assez rapidement car les 5.000 lemmes les plus fréquents permettent déjà d'avoir une très bonne couverture (97% de couverture statique obtenus sur un corpus d'une année du journal *Le Monde*). De plus, cela permet également de récupérer les mots inconnus spécifiques à un domaine en partant d'un corpus de textes spécialisés. La Figure 4 schématise la démarche suivie.

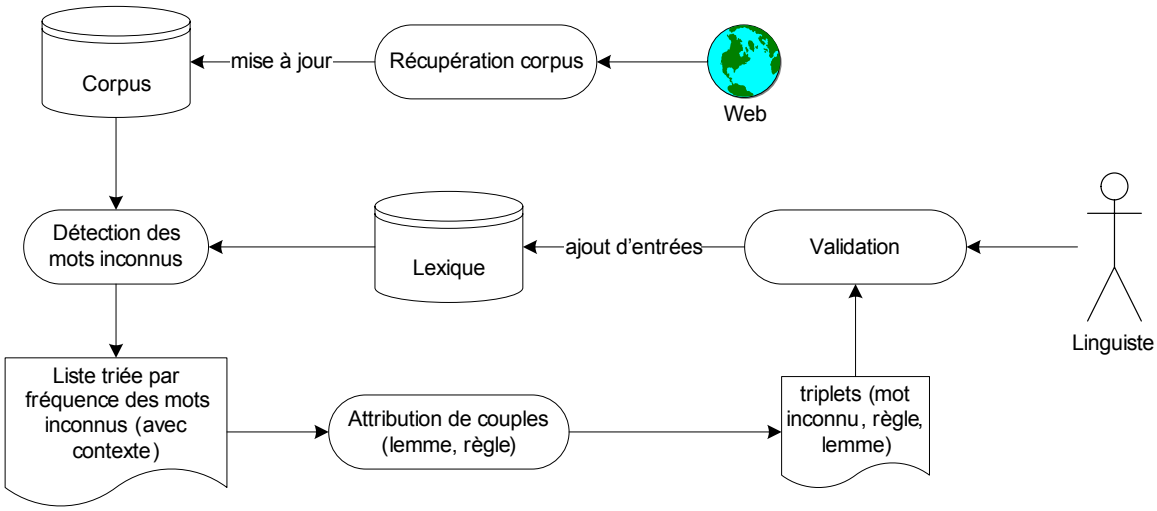

Figure 4 : Processus d'ajout d'entrées au lexique morphosyntaxique

La partie du haut permet de mettre à jour un corpus en utilisant par exemple des fils RSS. La détection des mots inconnus se fait par simple comparaison au lexique (après flexion bien sûr). L'attribution d'un couple (lemme, règle) à un mot inconnu se base sur un *guesser*.

Un *guesser* (ou *devin*) est généralement utilisé pour deviner la catégorie grammaticale d'un mot inconnu (Chanod et Tapanainen 1995, Schmid 1995, Mikheev 1997, Brants 2000, Cucerzan et Yarowsky 2000, Vasilakopoulos 2003). Nous avons, ici, utilisé le guesser pour déterminer automatiquement tout d'abord la catégorie grammaticale, puis le lemme et la règle associés. Ce guesser est entièrement probabiliste et est basé sur les 5 dernières lettres des mots. Les probabilités d'association d'une terminaison avec une règle sont apprises à partir d'un lexique existant et appliquées sur les mots inconnus. On pourra trouver plus de détails dans l'article de Loupy *et al*. (2008).

Les performances obtenues sont relativement intéressantes. Le rappel et la précision peuvent être réglés en fonction des besoins. Pour un rappel de 52,6%, nous obtenons une précision de 85,8% avec, en moyenne, 0,7 propositions par mot. Ces résultats nous ont semblés utilisables et permettent d'avoir assez rapidement des résultats intéressants bien que le travail de filtrage soit loin d'être négligeable.

L'intérêt de ce guesser est de pouvoir traiter rapidement un grand nombre d'entrées. Le problème est qu'il a été entraîné sur un lexique déjà conséquent (50.000 lemmes) et ne peut donc être utilisé pour commencer le travail de création de lexique. Une autre méthode est donc à utiliser dans ce cas.

4.2. *Ajout à partir d'une liste de lemmes*

Afin de débuter le lexique, nous avons mis au point une méthode utilisant un guesser à base de règles qui s'applique sur une liste de lemmes. Il est plus facile de se baser sur une liste de lemmes que sur des mots fléchis et il est assez facile de trouver de telles listes.

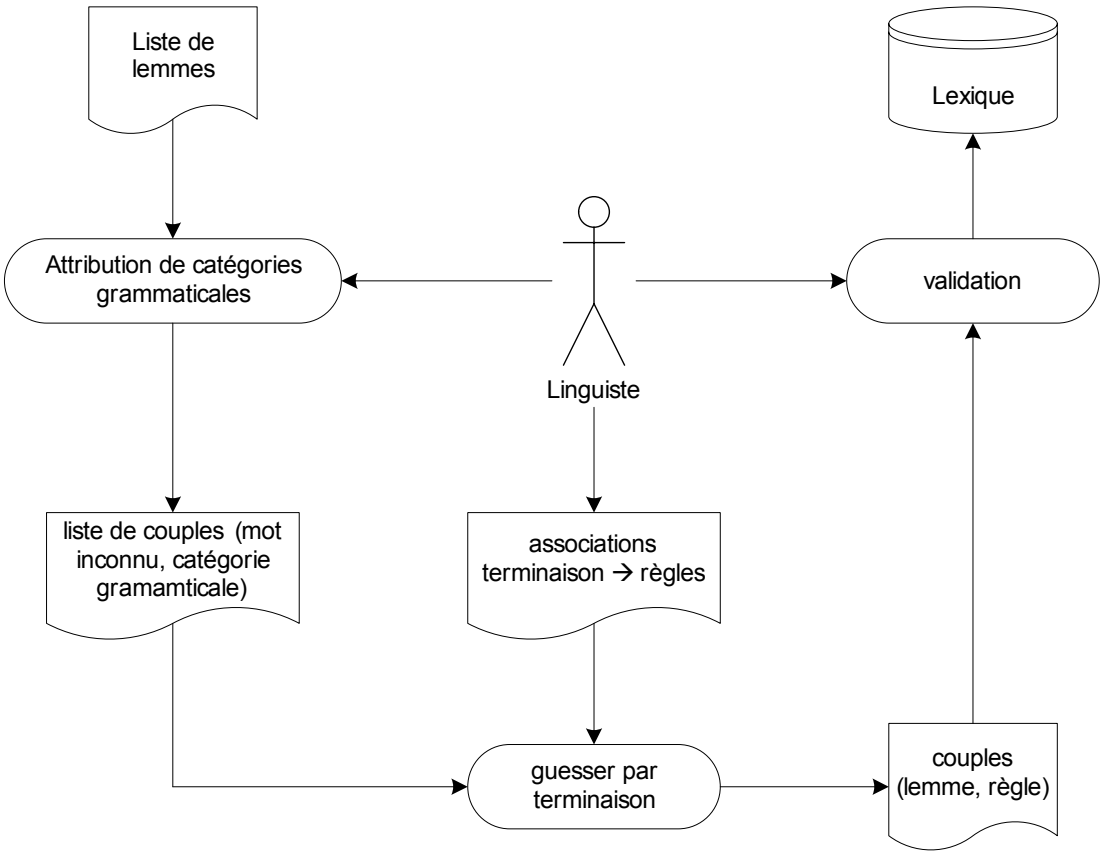

Figure 5: Processus d'ajout d'entrées à partir d'une liste de lemmes

La Figure 5 indique le processus de la méthode. Le linguiste commence par classer les lemmes par groupes de flexions. Cette catégorisation peut être plus ou moins poussée et l'on peu même s'en passer. Le tout est de trouver un équilibre entre le temps passé dans cette première intervention du linguiste (*Attribution de catégories grammaticales*) et dans la deuxième intervention (*Validation*). Plus la catégorisation sera précise, plus le temps passé sur la première phase sera long mais plus les résultats seront précis et donc le temps de validation sera plus court. Le guesser utilise des règles d'association prenant en compte le groupe flexionnel et des règles d'association (terminaison, règles de flexion). Nous avons mené des expériences avec une première catégorisation très précise.

```
nom féminin variable
nom féminin invariable
nom masculin variable
nom masculin al->aux
nom masculin invariable
nom masculin pluriel en "es" (import anglais)
nom masculin avec féminin
nom masculin avec féminin en er/euse (import anglais)
nom masculin et féminin variables
nom masculin et féminin invariable
adjectif variable en genre et nombre
```

Figure 6: catégories flexionnelles

Les résultats sont alors très bons puisque nous obtenons une précision de 96,6% et un rappel de 99,7% pour un nombre moyen de propositions par lemme de 1,03. Le nombre de propositions étant faible, la validation est très rapide. Le temps passé sur la catégorisation nous a semblé acceptable mais il nous faudra mener des expériences utilisateurs afin de mesurer le gain en

temps du linguiste selon qu'il utilise ou non cette méthode et selon la précision de la catégorisation préalable.

5. Conclusion et perspectives

Nous avons présenté un certain nombre de démarches pour la création de lexiques morphosyntaxiques ainsi qu'un formalisme de codage et des aides basées sur un guesser permettant au linguiste de coder plus vite. Les méthodes doivent maintenant être validées en mesurant le gain réel en termes de temps passé par le linguiste durant le processus de codage (cela dépendra aussi de l'efficacité de l'interface et des informations données au linguiste pour sa validation).

Par ailleurs, un certain nombre de points pourraient être améliorés. En particulier, les associations erronées (lemme, règle de flexion) devraient pouvoir être détectées dans un certain nombre de cas en vérifiant que toutes les flexions générées existent bien, par exemple sur le web.

Un travail devra également être fait sur les mots composés qui sont beaucoup plus difficiles à gérer du fait des difficultés d'application d'un guesser. Enfin, l'aspect multilingue devra être pris en compte et nous allons tester cette méthode sur des langues autres que le français et non latines pour lesquelles les apports pourraient être moins intéressants.

Remerciements

Nous remercions Helena Blancafort pour sa lecture et ses suggestions dont nous avons profité.

Références

Brants, T. (2000). "TnT—A Statistical Part-of-Speech Tagger". Dans *Proceedings of the 6th ANLP Conference*, Seattle, USA.

Carreras, X. *et al*. (2004). "FreeLing: An Open-Source Suite of Language Analyzers". Dans *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.

Chanod, J. P.; Tapanainen, P. (1995). "Creating a Tagset, Lexicon and Guesser for a French Tagger". Dans *Proceedings of the European Chapter of the ACL SIGDAT Workshop "From text to tags: Issues in Multilingual Language Analysis*". Dublin, Ireland. 51-57.

Cucerzan, S.; Yarowsky, D. (2000). "Language Independent, Minimally Supervised Induction of Lexical Probabilities". Dans *Proceedings of ACL-2000*. Hong Kong. 270-277.

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)*.

Grabar, N.; Zweigenbaum, P. (1999). "Acquisition automatique de connaissances morphologiques sur le vocabulaire médical". *Actes de TALN 1999*. Cargèse, France. 175-184.

Ide, N.; Véronis, J. (1994). "MULTEXT: Multilingual Text Tools and Corpora". Dans *Proceedings of the 15th International Conference on Computational Linguistics*. COLING'94, Kyoto, Japan. 588-92.

Loupy, C. de; Bagur, M.; Blancafort, H. (2008). *Guesing flexion rules; soumis à COLING 2008*.

Mikheev, A. (1997). "Automatic Rule Induction for Unknown-Word Guessing". Dans *Computational Linguistics* 23 (3). ACL. 405-423.

Nakov, P. *et al*. (2003). "Guessing Morphological Classes of Unknown German Nouns". Dans *Proceedings of Recent Advances in Natural Language Processing (RANLP'03)*. Borovetz, Bulgaria. 319-326.

Romary, L.; Salmon-Alt, S.; Francopoulo, G. (2004). "Standards going concrete: from LMF to Morphalou". Dans *Workshop on Electronic Dictionaries, COLING-04*.

Sagot, B. (2006). *Analyse automatique du français: lexiques, formalismes, analyseurs*. Thèse de doctorat en informatique. Paris: Université Paris VII.

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

Vasilakopoulos, A. (2003). "Improved Unknown Word Guessing by Decision Tree Induction for POS Tagging with TBL" Dans *Proceedings of CLUK*. Edinburgh.

Véronis, J.; Khouri, L. (1996). *MulText lexical specifications: application to French*. URL: *http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX.LangSpec.fr.html.*