

# Vers un méta-lexique pour le français : architecture, acquisition, utilisation

**Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie et Pierre Boullier**

INRIA Rocquencourt - Projet ATOLL

Domaine de Voluceau - B.P. 105

78152 Le Chesnay Cedex, France

benoit.sagot@inria.fr, lionel.clement@lefff.net

## Résumé

Nous présentons dans cet article une nouvelle ressource lexicale pour le français, bientôt librement disponible en tant que deuxième version du *Lefff* (*Lexique des Formes Fléchies du Français*). Il s'agit d'un lexique morphologique et syntaxique à large couverture, dont l'architecture repose sur une structure d'héritage de propriétés, ce qui le rend plus compact et plus aisément maintenable. Cela permet également une description des entrées lexicales indépendante des formalismes dans lequel il est utilisé. Pour ces deux raisons, nous utilisons le terme *méta-lexique*. Nous décrivons son architecture, différentes approches automatiques ou semi-automatiques pour acquérir, corriger et/ou compléter un tel lexique, ainsi que la manière dont il a été utilisé en lien avec une LFG et une TAG pour construire deux analyseurs du français à large couverture.

## 1 Introduction

La couverture et la précision d'une chaîne d'analyse du langage naturel ne dépend pas uniquement de la couverture et de la précision de la grammaire qui est utilisée. D'autres composants, parmi lesquels la chaîne de pré-traitement et le constructeur d'analyseurs, jouent un rôle majeur. Cependant, par son rôle central à tous les niveaux de la chaîne, le lexique a une importance capitale.

Cependant, la structuration et le développement d'un lexique, ainsi que sa mise en relation avec l'analyseur et donc la grammaire, sont des tâches difficiles. D'une part, un lexique à

large couverture comporte un nombre considérable d'entrées, qui se mesure en centaines de milliers. D'autre part, pour chacune de ces entrées, un grand nombre d'informations différentes sont nécessaires pour disposer d'une description qui satisfasse l'ensemble des besoins des autres composants (morphologie, syntaxe, ...).

Lors de la construction de notre lexique morphologique et syntaxique du français, qui comporte plus de 400 000 formes fléchies pour plus de 600 000 entrées, différentes techniques ont dû être mises en œuvre pour s'adapter à cette complexité. Nous présentons donc ici les différentes idées dont nous avons tiré parti pour concevoir l'architecture de notre lexique, et pour le valider, le compléter et le corriger. Nous montrons enfin comment il a été utilisé par différents analyseurs reposant sur différents formalismes.

## 2 Architecture et volume de données

### 2.1 Architecture

Nous présentons dans la figure 2 l'architecture du lexique, qui comporte deux phases. Tout au long de la chaîne, les données sont réparties en fichiers spécifiques à chaque partie du discours<sup>1</sup>.

La première phase est *morphologique*. Un fichier de couples lemme – classe flexionnelle et un fichier décrivant la flexion desdites classes servent d'entrée à un conjugueur, qui produit un fichier de formes fléchies. Ce fichier est complété par un fichier de formes fléchies entrées manuellement, pour gérer les variantes, abréviations, et autres phénomènes marginaux. Le fichier de formes fléchies associe à chaque entrée lexicale un lemme, une étiquette morphologique, et un indicateur morphosyntaxique. Ce dernier a pour rôle d'induire éventuellement des modifications

<sup>1</sup>Ce choix pourrait être revu pour tirer parti des mécanismes de morphologie dérivationnelle.

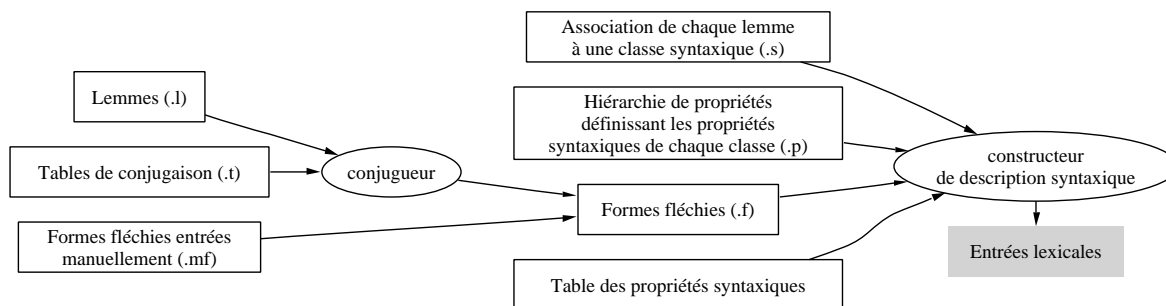


FIG. 1 – Architecture du lexique.

dans les informations syntaxiques dont héritera la forme. Ainsi, à un infinitif est associée un indicateur morphosyntaxique qui rend facultatif son sujet. Presque toutes les formes fléchies ont cependant un indicateur morphosyntaxique par défaut qui n'induit aucune modification.

Fichier des lemmes (morphologie) :
boire v74
Fichier des paradigmes flexionnels :
<TBL v74 boire>
boira "F3s" ;
boirai "F1s" ;
(...)
</TBL>
Fichier des lemmes (syntaxe) :
boire @verbe_standard
Fichier des patrons syntaxiques :
@verbe_standard {
< @verbe_transitif_direct
< @verbe_pronominal
}
@verbe_transitif_direct
< verbe
< passivable
< transitif_direct

TAB. 1 – Autour du lemme boire.

La seconde phase est *syntaxique*. Dans un premier fichier, chaque lemme est associé à un patron syntaxique, comme illustré dans la table ?? par le lemme boire. Dans un deuxième fichier, les patrons syntaxiques sont définis par héritage d'autres patrons et de propriétés syntaxiques atomiques (la disjonction est possible, cf. table ??). Dans un troisième fichier, les propriétés syntaxiques sont définies par une ou plusieurs opérations parmi les suivantes :

- apport à la structure syntaxique d'une *macro* syntaxique définie dans un fichier séparé par une représentation en structures de traits

avec partage de valeur,

- modification du *prédicat* associé à une entrée du lexique (utilisable par exemple en tant que couple attribut-valeur dans une TAG, *pred* en LFG, ou lemme dans d'autres théories...),
- attribution d'une *catégorie* à une entrée du lexique (utilisable par exemple en tant que terminal d'une TAG ou de la CFG sous-jacente à une LFG, en tant que catégorie dans d'autres théories...),
- attribution d'un poids autre que le poids par défaut, pour (dé)favoriser l'emploi d'une entrée lexicale pendant la désambiguïsation de la sortie d'un l'analyseur syntaxique.

Une telle structure hiérarchique permet de modifier aisément les informations associées à tous les verbes partageant une classe ou une propriété. Ceci facilite la correction d'erreurs et plus généralement la "maintenance" du lexique.

Enfin, les entrées lexicales sont regroupés dans le lexique morphosyntaxique final. Quelques entrées sont montrées dans la table ??.

## 2.2 Volume de données

Notre lexique comporte 404 366 formes fléchies distinctes représentant 600 909 entrées dont certaines sont factorisées (la première et la troisième personne du singulier du présent de l'indicatif d'un verbe du premier groupe sont regroupées en une seule entrée).

La répartition par lemmes donne entre autres à 10 024 adjectifs, 2 127 adverbes, 37 183 noms communs, 52 938 noms propres, 137 préfixes et suffixes<sup>2</sup>, 212 prépositions, 6788 verbes.

<sup>2</sup>En effet, nos outils permettent de décomposer des mots inconnus issus de morphologie compositionnelle tels que *anti-Bush-né* en *anti-* / *Bush* / *-né*.

passer	v	[pred="passer<(subj ssubj vsubj),(obj),(†-obj)>", cat=v, @W ] ;
passer	v	[pred="passer<(subj ssubj vsubj),acomp>", cat=v, @W , @AASubj] ;
passer	v	[pred="passer<(subj ssubj vsubj),pour-acomp>", cat=v, @W , @AAPourSubj] ;
passer	v	[pred="passerSe<(subj),(de-obj)>obj", cat=v,@pron, @W ] ;
passer	v	[pred="passerSe<(subj),de-vcomp>obj", cat=v,@pron, @W , @CtrlSubjDe] ;
petit_à_petit	500 adv	[pred="petit-à-petit", cat=adv ] ;

TAB. 2 – Quelques exemples d’entrées du lexique.

### 3 Aquisition, complétion et correction

La constitution d’un lexique est une tâche ardue, en raison à la fois du nombre d’entrées nécessaires pour former un lexique à large couverture et de la complexité des informations à associer à chaque entrée pour former un lexique de qualité.

L’architecture présentée dans la partie précédente permet une factorisation importante des informations. Cependant, ces informations doivent être malgré tout obtenues d’une manière ou d’une autre, puis complétées et/ou corrigées. Pour faciliter ces étapes, et mise à part l’adjonction purement manuelle ou la récupération de ressources lexicales libres de droits, nous avons utilisé différentes techniques.

#### 3.1 Acquisition automatique de lexique morphologique

Les entrées morphologiques des catégories syntaxiques à morphologie riche peuvent être acquises automatiquement. Une validation manuelle est nécessaire pour garantir la qualité du lexique obtenu, mais ceci prend un temps considérablement moins élevé que les techniques purement manuelles.

Nous avons décrit dans (Clément et al., 2004) une telle méthodologie, et son application aux formes verbales d’un corpus du *Monde Diplomatique* a débouché sur la publication en ligne du lexique morphologique *Lefff* dans sa première version ([www.lefff.net](http://www.lefff.net)). Pour résumer, la méthodologie repose sur l’idée que l’hypothèse de l’existence d’un lemme peut être faite si de plusieurs mots différents présents dans le corpus sont interprétés à moindre coût comme étant des variantes morphologiques dudit lemme. Dans notre système, ceci repose sur une description préalable des classes flexionnelles. Une telle technique permet d’approcher une couverture exhaustive du corpus dont on part. L’expérience montre que cela permet d’acquérir de nombreux mots absents des

ressources classiques, souvent des mots dérivés (préfixés) ou récents (techniques).

#### 3.2 Détection automatique de mots manquants

Nous avons également utilisé notre correcteur orthographique *SXSPELL*, décrit ailleurs, pour détecter dans un gros corpus les mots inconnus du lexique et leur proposer des corrections. En fonction de l’existence ou non pour un mot inconnu d’une correction à faible coût, et en fonction du nombre d’occurrences de ce mot, on peut constituer automatiquement une liste de formes fléchies inconnues du lexique et devant probablement y figurer. Cette technique est bien plus manuelle que la précédente, mais elle permet d’augmenter rapidement la couverture du lexique pour des catégories à la flexion moins riche que les verbes.

#### 3.3 Acquisition automatique de multi-mots manquants

Un des points faibles des lexiques est souvent le manque de couverture pour les multi-mots (expressions comme "pomme de terre" ou "France 2"). Nous avons utilisé à titre expérimental des techniques statistiques simples inspirées de travaux plus systématiques comme ceux de (Dias et al., 2001) pour acquérir automatiquement de telles expressions. Ici encore, une validation manuelle est indispensable.

#### 3.4 Détection automatique d’entrées syntaxiquement erronées

Utilisant diverses idées développées en parallèle par (van Noord, 2004), nous avons construit, à partir de l’analyse syntaxique de gros corpus, des tables mesurant un *taux de parsabilité* pour chaque mot. L’idée est qu’un mot qui apparaît souvent dans des phrases inanalysables a de bonnes chances d’être syntaxiquement incomplet

ou incorrect dans le lexique.<sup>3</sup> Il ne reste plus qu'à la corriger manuellement.

### 3.5 Acquisition automatique d'informations syntaxiques ciblées

Dans certains cas spécifiques, il est possible d'acquérir automatiquement des informations syntaxiques. Nous n'avons mené d'expérience dans ce domaine que sur deux points. Le premier concerne les prépositions associées aux formes verbales : un comptage simple des couples verbes-préposition présents dans le corpus permet d'avoir une idée assez précise de la composante oblique de la sous-catégorisation d'un verbe.

Le second point concerne les verbes supports. En effet, des statistiques élémentaires sur les occurrences de noms communs suivant immédiatement une forme verbale permettent d'acquérir très rapidement une liste importante de couples verbe support – nom prédicatif.

D'autres méthodes sont envisageables, comme celles proposées lors de l'ARC RLT : il est désormais possible d'analyser de très gros corpus avec des grammaires permissives (affectant par exemple à tous les verbes des cadres de sous-catégorisation moins restrictifs, en profitant de la structure d'héritage du lexique), puis d'extraire par des techniques statistiques des informations de nature syntaxique. Toutefois, nous n'avons pas encore eu le temps de mener cela à bien.

## 4 Utilisation

Notre lexique peut être vu comme un *méta-lexique* car les informations qui le composent sont présentes d'une part comme résultat d'un mécanisme d'héritage de propriétés, et d'autre part sous une forme indépendante du formalisme. Ceci comprend :

- la forme fléchie,
- éventuellement un poids (cf. plus haut),
- une catégorie,
- un prédicat (comme dit plus haut, c'est le lemme de nombre de formalismes, le *pred* de LFG, ou un trait comme les autres en TAG),
- un cadre de sous-catégorisation, dans un format à la LFG mais immédiatement conver-

<sup>3</sup>Naturellement, dans certains cas, ceci est plutôt la trace d'une incomplétude de la grammaire à propos d'une construction dont le mot est spécifique.

tible en tout autre format,

- une liste de macros morphosyntaxiques dont l'expansion peut être définie par l'utilisateur en respectant son formalisme grammatical.

De fait, nous avons utilisé ce lexique lors de la campagne EASy d'évaluation des analyseurs syntaxiques, tout en présentant deux analyseurs totalement différents. Le premier est un analyseur reposant sur une TAG calculée à partir d'une métagrammaire, et utilisant l'environnement DyaLog de développement d'analyseurs (Villemonde de la Clergerie, 2002). Le second est un analyseur LFG reposant sur une grammaire développée manuellement et utilisant le système Syntax (technologies décrites par exemple dans (Boullier, 2003)). Les formalismes sont donc bien différents, mais ils ont pu utiliser le même lexique.

## 5 Conclusion

Nous avons présenté une version préliminaire d'un lexique morphologique et syntaxique du français à large couverture, ainsi que diverses techniques utilisées pour le renseigner, le compléter et le corriger. Nous avons signalé son utilisation dans divers analyseurs reposant sur divers formalismes ainsi que son architecture reposant fortement sur l'héritage de propriétés. A ce titre, il peut être considéré comme un méta-lexique. Enfin, nous prévoyons de le rendre accessible librement dans un proche avenir.

## References

- Pierre Boullier. 2003. Guided Earley parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT'03)*, pages 43–54, Nancy, France, April.
- Lionel Clément, Benoît Sagot, and Bernard Lang. 2004. Morphology Based Automatic Acquisition of Large-coverage Lexica. In *Proceedings of LREC'04*, pages 1841–1844, May.
- G. Dias, S. Guilloré, J.C. Bassano, and J.G.P. Lopes. 2001. Extraction automatique d'unités lexicales complexes : Un enjeu fondamental pour la recherche documentaire. *Traitement Automatique des Langues (T.A.L.)*, 41(2).
- Gertjan van Noord. 2004. Error mining for wide-coverage grammar engineering. In *Proc. of ACL 2004*, Barcelona, Spain.
- Éric Villemonde de la Clergerie. 2002. Construire des analyseurs avec DyALog. In *Proc. of TALN'02*, June.