Un lexique syntaxique des verbes du français : VfrLPL

Stéphane Rauzy, Philippe Blache

Un lexique syntaxique des verbes du français : VfrLPL

Stéphane Rauzy, Philippe Blache
Laboratoire Parole et Langage
CNRS & Université de Provence
stephane.rauzy@lpl.univ-aix.fr, pb@lpl.univ-aix.fr

Résumé Nous présentons un lexique syntaxique des verbes du français. La ressource contient 8800 entrées environ (soit 6700 verbes distincts), pour lesquels nous produisons les formes conjuguées, leurs formes phonétisées correspondantes ainsi qu'un indice sur leurs fréquences d'usage. Pour chacun des verbes est donné son auxiliaire, son caractère pronominal et les informations caractérisant sa transitivité. Durant la constitution de cette ressource, nous avons apporté un soin particulier à valider les entrées produites en croisant nos résultats avec d'autres ressources de référence. Nous mettons à la disposition de la communauté une version préliminaire du lexique, la ressource électronique VfrLPL1.0.xml, pour laquelle les fréquences d'usage n'ont pas été recalculées. La ressource est librement distribuée par le Centre de Ressources pour la Description de l'Oral (site du CRDO : http ://crdo.fr). Ce travail s'inscrit dans un programme mené au Laboratoire Parole et Langage depuis quelques années, visant au développement et à la maintenance d'une ressource lexicale fiable et couvrante pour le français.

Abstract We present a syntactical lexicon of french verbs. The resource is rich of almost 8800 entries (6700 distinctive lemmas). For each verb is generated its set of conjugate forms, their phonetic counterpart, and the frequency of their usage. Syntactical informations such like the associated auxiliary, pronominal characterics and transitive nature is also given for each verb. The validation of the resource has been performed by crosschecking our data with others reference lexicons. We distribute a preliminar version of the electronic resource, VfrLPL1.0.xml, for which the frequencies have not been recalculated. The resource is freely distributed by the Centre de Ressources pour la Description de l'Oral (CRDO website : http ://crdo.fr). This work is part of a larger program started few years ago at the Laboratoire Parole et Langage which aims to develop and maintain a reliable large coverage lexical resource for french language.

Mots-clefs : Lexique, verbes du français, formes conjuguées, formes phonétisées, information syntaxique

Keywords: Lexicon, french verbs, conjugate forms, phonetic forms, syntactical information

1 Introduction

Le développement de lexiques syntaxiques occupe désormais en France une position particulière, notamment grâce au projet "LexSynt". Celui-ci a en effet permis de mettre en perspective les ressources existantes et de souligner les besoins pour les ressources futures. De plus, les lexiques sont désormais développés dans une perspective de libre distribution, ce qui facilite grandement la diffusion des informations et permet ainsi de progresser rapidement dans la spécification des informations devant et pouvant être contenues dans un lexique. Ce type de ressource a bien entendu un intérêt linguistique intrinsèque, mais constitue surtout un élément de base pour tout système de traitement automatique des langues. Le problème de la qualité des ressources dans le cadre d'une perspective de traitement automatique est alors une question centrale. Pour ce qui concerne les lexiques, cette question se décline de plusieurs façons :

– couverture : il est nécessaire pour obtenir un étiquetage morphosyntaxique fiable de minimiser le nombre de mots inconnus, donc de disposer d'un lexique aussi exhaustif que possible
– richesse et qualité des informations associées à chaque entrée
– validité des formes et des informations : il est nécessaire de vérifier d'une part les formes du lexique et d'autre part la validité des informations de chaque entrée

Parmi les projets de développement de lexique du français, on peut en souligner quelques uns reprenant à son compte tout ou partie de ces objectifs : Lefff (Sagot06), SynLex (Gardent06), Dicovalence (Eynde03), Morphalou (Romary04). Chacun constitue à un degré différent une contribution importante au problème. De notre côté, nous avons depuis plusieurs années également développé une ressource lexicale d'envergure, DicoLPL (VanRullen05), utilisée dans les outils développés par le LPL (étiqueteurs, analyseurs syntaxiques, outils d'aide à la communication, etc.). Cette ressource, comme il est d'usage dans ce type d'activité, est en constante évolution, et la validation de ses données progresse régulièrement. De plus, elle contient des informations souvent absentes d'autres projets, comme la phonétisation des formes ou l'indication de leur fréquence.

Nous proposons ici de faire le point sur le développement de cette ressource lexicale en mettant en relief le type d'information qu'il contient. Nous nous attacherons plus particulièrement à décrire les techniques d'acquisition et de validation des informations verbales. Cette ressource pourra ainsi constituer une contribution à l'effort collectif pour différentes raisons et notamment sa couverture ainsi que la validation des informations. Le croisement des différentes ressources citées peut ainsi être envisagée, et permettre de produire de nouvelles ressources lexicales de qualité, quelque soit le mode de représentation d'information choisi.

2 Constitution de la ressource

Le lexique DicoLPL (VanRullen05) nous a servi de base pour constituer notre lexique syntaxique des verbes. A partir de cette ressource, nous avons dans un premier temps créé un modèle des conjugaisons des verbes du français (ie. un conjugueur). Dans un deuxième temps, nous avons complété les informations nécessaires à la constitution du lexique syntaxique en croisant nos données avec d'autres ressources existantes. Pour finir, nous avons défini des critères de sélection permettant d'établir la liste des verbes à inclure dans notre ressource.

2.1 Le conjugueur

Le lexique DicoLPL contient environ 7000 verbes défactorisés en 290 000 formes fléchis. Pour chaque lemme, nous avons regroupé ses formes fléchies et leurs catégories morphosyntaxiques associées, ainsi que la fréquence de chaque forme et sa phonétisation. Pour chacun des verbes, les formes ont été décomposées en un radical (commun à l'ensemble des formes) et un affixe terminal. La même opération a été réalisée sur les formes phonétisées.
Des tables de conjugaison ont été formées, regroupant les verbes possédant le même jeu d'affixes orthographiques et phonétiques. Nous avons obtenu environ 200 tables de conjugaison différentes. Une vérification manuelle de ces tables nous a permis de corriger nos données d'origine (erreurs de phonétisation, erreurs de graphie, mauvaise affectation de la catégorie morphosyntaxique, ...). Certaines tables ont été complétées en utilisant les conjugaisons validées du Bescherelle (Bescherelle06), notamment pour prendre en compte les modifications introduites par la réforme orthographique de 1990 (JO90). Après cette étape de validation, notre conjugueur est composé d'environ 150 tables de conjugaison.
Une des particularités de nos tables est de posséder une double clé d'entrée : orthographique et phonétique. Ainsi, certaines conjugaisons standards sont dédoublées dans notre base (eg. la conjugaison des verbes en `-ouer` est scindée en deux, selon que l'affixe phonétique se prononce `we` (en alphabet phonétique Sampa, comme `jouer`) ou se prononce `ue` (comme `clouer`).

Parmi les verbes extraits du dicoLPL, on a dénombré environ 2000 verbes qui ne possédaient pas une conjugaison complète (ie. des formes conjuguées manquantes). Pour la majorité d'entre eux, il s'agit d'une incomplétude des données d'origine qui peut être aisément corrigée en appliquant les tables de conjugaison. Pour d'autres en revanche – les verbes défectifs, certaines formes ne sont pas réalisées. Dans notre modèle, nous avons donc introduit des tables de défection qui rendent compte de l'absence de certaines formes dans la conjugaison des verbes défectifs. Notre conjugueur comporte 35 tables de défection, qui viennent se superposer aux tables de conjugaison (à la manière d'un masque ou d'un filtre) pour générer l'ensemble des formes fléchies valides pour un verbe donné.

2.2 Enrichissement de la ressource et sélection de la liste des verbes

Nous avons complété les informations sur les verbes extraits du DicoLPL en utilisant le conjugueur en ligne "Le Devoir Conjugal" (Baudoin01). Les informations retenues sont les suivantes :
– L'auxiliaire verbal : `avoir` ou `être`
– Le caractère pronominal du verbe : essentiellement pronominal, non pronominal ou dual (pronominal et non-pronominal)
– Le caractère impersonnel du verbe : personnel ou impersonnel
– La transitivité du verbe
Pour la centaine de verbes présents dans DicoLPL et absents du "Devoir Conjugal", nous avons collecté manuellement ces informations dans le Bescherelle (Bescherelle06) et dans le TLF (TLF04). Environ 200 verbes du DicoLPL n'ont pas été identifiés dans ces ressources et ont donc été écartés. Tous les verbes de la liste présentant une entrée multiple (pronominal et non-pronominal, auxiliaire `être` et auxiliaire `avoir`, personnel et impersonnel) ont été défactorisés. Après cette opération, la liste compte 8400 entrées environ.

Dans une deuxième phase, nous avons ajouté manuellement les verbes présents dans le Bescherelle et absents de notre liste (environ 350 entrées) et modifié le caractère pronominal d'environ 500 verbes qui présentent une double entrée (pronominal et non-pronominal) dans le Bescherelle. Cette liste intermédiaire contient 9300 entrées environ.

Nous retenons pour constituer notre liste finale de verbes les entrées qui sont présentes dans la liste intermédiaire et dans le TLF (TLF04). Ainsi, les verbes retenus appartiennent obligatoirement à l'intersection du Bescherelle et du TLF. Notre liste contient au final environ 8800 verbes défactorisés (soit 6700 lemmes distincts).

Le critère de sélection imposé, l'obligation pour le verbe d'être référencé dans au moins deux ressources standards distinctes (TLF et Bescherelle), disqualifie en pratique un nombre important de verbes (environ 500 entrées du Bescherelle et 2400 du TLF). Une inspection des verbes écartés révèle néanmoins que la majorité de ces entrées sont :
– Des graphies instables eg. `abatre`, `abbattre` pour `abattre` (TLF), `flacher` pour `flasher` (Bescherelle)
– Des néologismes peu fréquents eg. `abracadabrer`, `boulevarder`, ... (TLF), `tututer`, `gomorrhiser`, ... (Bescherelle)
– Des dérivations verbales peu fréquentes eg. `enrougir`, `hydroplaner`, ... (TLF), `retondre`, `détransposer`, ... (Bescherelle)
– Des verbes techniques ou des formes vieillies eg. `caracouler`, `introjecter`, ... (TLF), `anodiser`, `peausser`, ... (Bescherelle)

Nous avons pris le parti de ne pas incorporer ce type d'entrées dans la présente version de notre liste. Il est toutefois envisageable de reconsidérer ce choix pour les versions futures de la ressource.

3 Description de la ressource

Le lexique produit est présenté sous format XML. Chaque verbe de la liste est caractérisé par les informations suivantes :
– son lemme et sa phonétisation (attributs XML `lemma` et `phonemes`)
– son auxiliaire : `avoir` ou `être` (attribut XML `auxiliary="a"` ou `auxiliary="e"`)
– son caractère pronominal (voir section 3.1)
– son caractère défectif : personnel, impersonnel, autre (attribut XML `personnal="p"`, `personnal="i"` et respectivement `personnal="d"`)
– les informations sur la transitivité du verbe (voir section 3.2)
– une indication de sa fréquence d'usage (attribut XML `frequency`)
– l'indice de sa table de conjugaison (attribut XML `conjugation`)
– l'indice de sa table de défection (attribut XML `defection`)

L'élément XML représentant chaque verbe contient la liste des formes conjuguées valides du verbe (les éléments XML `form`), caractérisées par l'identifiant morphosyntaxique (attribut XML `msc`), la graphie (attribut XML `spelling`), la phonétisation de la forme (attribut XML `phonemes`) et une indication de sa fréquence (attribut XML `frequency`).

L'identifiant morphosyntaxique est codé par un vecteur de traits dérivé des formats Multext et Grace) :
– trait 1 : `V` pour Verbe
– trait 2 : `a` pour l'auxiliaire `avoir`, `e` pour l'auxiliaire `être` et `m` pour les verbes principaux

– trait 3 : le mode de la conjugaison, `n` pour l'infinitif, `i` pour l'indicatif, `s` pour le subjonctif, `c` pour le conditionnel, `m` pour l'impératif et `p` pour les participes.
– trait 4 : le temps de la conjugaison, `p` pour le présent, `i` pour l'imparfait, `s` pour le passé simple et `f` pour le futur
– trait 5 : la personne de la conjugaison si il y a lieu, `1`, `2` ou `3`.
– trait 6 : le nombre de la conjugaison, `s` pour singulier et `p` pour pluriel.
– trait 7 : le genre de la conjugaison pour le participe passé, `f` pour le féminin et `m` pour le masculin

L'identifiant morphosyntaxique `msc="Vmsi3p-"` signifie ainsi qu'il s'agit de la forme d'un verbe principal à la troisième personne du pluriel du subjonctif imparfait.

3.1 Caractère pronominal du verbe

Nous avons adopté un système de notation à trois valeurs pour rendre compte du caractère pronominal des verbes :
– Les verbes essentiellement pronominaux, c'est-à-dire qui ne s'emploient qu'à la forme pronominale eg. `se repentir`. La valeur de l'attribut `pronominal` rendant compte du caractère pronominal du verbe est dans ce cas noté `o` (pour obligatoire).
– Les verbes qui n'acceptent pas la forme pronominale. Principalement, il s'agit des verbes intransitifs, qui faute de complément d'objet direct et indirect, n'acceptent ni l'emploi pronominal réfléchi ou réciproque, ni l'emploi pronominal passif. La valeur de l'attribut `pronominal` est dans ce cas noté `i` (pour interdit). Quelques verbes intransitifs acceptent toutefois un emploi pronominal subjectif eg. `se mourir`, nous avons créé deux entrées pour ce type de verbes.
– Les verbes qui sont référencés classiquement comme pronominaux et non-pronominaux eg. `laver`. Dans notre lexique, ces verbes occupent deux entrées dont l'attribut `pronominal` prend respectivement pour valeur `i` et `o`.
– Les verbes qui ne sont pas référencés comme pronominaux et non-pronominaux, mais qui sont néanmoins susceptibles d'accepter un emploi pronominal passif, eg. `se manger` dans `le riz se mange chaud`. Nous affectons à l'attribut `pronominal` de ces verbes la valeur `p` (pour possible).

3.2 Transitivité

Le codage que nous avons adopté pour rendre compte des propriétés de transitivité des verbes du lexique est présenté figure 2. L'attribut XML associé à la transivité est noté `transitivity`.

3.3 Défection

Certaines tables de défection ont pu être générées automatiquement. C'est le cas par exemple pour les verbes fonctionnant avec l'auxiliaire `avoir` et qui n'acceptent pas de complément d'objet direct (donc les verbes intransitifs et intransitifs directs d'après la figure 2). Ces verbes ne peuvent être passivés. De plus, le participe passé de ces verbes ne s'accordant pas, les formes

Transitivité	code	COD	COI
transitif direct	td	p	a
		p	p
transitif indirect	ti	a	p
		p	p
transitif	t-	p	a
		a	p
		p	p
intransitif	i-	a	a
intransitif direct	id	a	a
		a	p
intransitif indirect	ii	a	a
		p	a
transitif ou intransitif	--	a	a
		p	a
		a	p
		p	p

FIG. 1 – Les traits codant les informations sur la transitivité du verbe. Les propriétés de transitivité varient selon que le verbe requiert la présence (p) ou l'absence (a) d'un complément d'objet direct (COD) et d'un complément d'objet indirect (COI). La table de vérité ci-contre récapitule les classes de transitivité généralement rencontrées, eg. un verbe est transitif indirect si il requiert la présence d'un COI, la présence du COD étant optionnelle.

du participe passé au masculin pluriel, féminin singulier et féminin pluriel ne sont jamais réalisées. Une table de défection a été créée pour rendre compte de ce phénomène.

Il y a bien entendu une exception à cette règle, les verbes `obéir` et `désobéir`, qui bien qu'intransifs directs, acceptent néanmoins la forme passive eg. `Seront-elles obéies ?`. Ces deux verbes ne suivent donc pas la table de défection mentionnée précédemment.

3.4 Phonétisation

Pour chaque forme fléchie, la forme phonétisée est donnée en alphabet standard Sampa. La phonétisation des formes de chaque conjugaison a été extraite en recoupant les informations présentes dans le DicoLPL. Certaines conjugaisons ont été corrigées manuellement et la phonétisation des lemmes des verbes absents de DicoLPL a été ajoutée.
Pour certaines formes conjuguées, la distinction entre deux entrées phonétiques ne se fait que sur la durée de production des phonèmes. Nous avons adopté la convention de doubler ces phonèmes pour lever l'ambiguïté, par exemple (distinction présent-futur) :

```
<form msc="Vmip1p-" spelling="courons" phonemes="kuRo~"/>
<form msc="Vmif1p-" spelling="courrons" phonemes="kuRRo~"/>
```

ou encore (distinction présent-imparfait) :

```
<form msc="Vmip2p-" spelling="envoyez" phonemes="a~vwaje"/>
<form msc="Vmii2p-" spelling="envoyiez" phonemes="a~vwajje"/>
```

4 Conclusion

Nous avons constitué un lexique syntaxique des verbes du français, riche de 8800 entrées environ (soit 6700 verbes distincts), pour lesquels nous générons à partir d'un conjugueur les formes fléchies du verbe et leurs formes phonétisées correspondantes. Un indice sur la fréquence d'usage de chaque forme est aussi proposé pour les verbes présents dans le lexique DicoLPL.

Pour chacun des verbes sont fournies les informations syntaxiques telles que l'auxiliaire à employer, le caractère pronominal du verbe, si il est impersonnel ou défectif, et les informations caractérisant sa transitivité.
La constitution de cette ressource a pris pour base le lexique DicoLPL, mais nous avons apporté un soin particulier à valider les entrées produites en croisant nos résultats avec d'autres ressources de référence (Bescherelle, TLF, ...).

Nous mettons à la disposition de la communauté une version préliminaire de ce lexique de verbes, la ressource électronique VfrLPL1.0.xml. La ressource est librement distribuée par le Centre de Ressources pour la Description de l'Oral (site du CRDO : http ://crdo.fr). Dans cette version préliminaire, les fréquences d'usage de chaque forme n'ont pas été recalculées sur corpus et proviennent des données du DicoLPL. La prise en compte des informations syntaxiques (transitivité, caractère pronominal, etc.) pour le recalcul des fréquences d'usage sur corpus est en cours.

Références

Baudoin M. (2001), "Le Devoir conjugal : de la conceptualisation à la diffusion", ALSIC, Vol. 4, numéro 1, juin 2001, pp. 91-102 (site web : http ://www.pomme.ualberta.ca/conjugateur/)

Bescherelle H. (2006), "La conjugaison pour tous", Edition Hatier, Paris, France

Clément L., Sagot B., Lang B. (2004) "Morphology based automatic acquisition of large-coverage lexica", in proceedings of LREC'04

Gardent C., Bruno G., Perrier G. & I. Falk (2006) "Extraction d'information de sous-catégorisation à partir des tables du LADL", in actes de TALN-06

Eynde K. & Mertens P. (2003) "La valence : l'approche pronominale et son application au lexique verbal", in Journal of French Language Studies 13, 63-104.

Rapport du conseil supérieur de la langue française (1990) "Rectification de l'orthographe", Journal Officiel de la République Française du 06/12/1990

Romary L., Salmon-Alt S., Francopoulo G. (2004). "Standards going concrete : from LMF to Morphalou" Workshop on Electronic Dictionaries, Coling 2004, Geneva, Switzerland.

Sagot B., Clément L., Villemonte de la Clergerie E., Boullier P. (2006) "The Lefff 2 syntactic lexicon for French : architecture, acquisition, use", in proceedings of LREC'06, Gênes

Imbs P. (2004), "Trésor de la langue française informatisé", Paris, France : CNRS éditions, 2004 (CNRS, Paris et ATILF, Nancy) (site web : http ://atilf.atilf.fr/tlf.htm)

VanRullen T. , Blache P. , Portes C. , Rauzy S. , Maeyhieux J.-F. , Guénot M.-L. ,Balfourier J.-M. , Bellengier E. (2005) "Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales", in actes de TALN, pp. 41-48 (Juin 2005 : Paris, France)