TALN 2005, Dourdan, 6–10 juin 2005

Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales

VanRullen T. , Blache P. , Portes C. , Rauzy S. , Maeyhieux J.-F. , Guénot M.-L. ,Balfourier J.-M. , Bellengier E.
Laboratoire Parole et Langage - CNRS - Université de Provence
29, Avenue Robert Schuman - 13100 Aix-en-Provence
{tristan,pb}@lpl.univ-aix.fr

Mots-clefs : dictionnaire, lexique, lexique noyau

Keywords: dictionary, lexicon, kernel lexicon

Résumé Nous présentons une plateforme de développement de lexique offrant une base lexicale accompagnée d'un certain nombre d'outils de maintenance et d'utilisation. Cette base, qui comporte aujourd'hui 440.000 formes du Français contemporain, est destinée à être diffusée et remise à jour régulièrement. Nous exposons d'abord les outils et les techniques employées pour sa constitution et son enrichissement, notamment la technique de calcul des fréquences lexicales par catégorie morphosyntaxique. Nous décrivons ensuite différentes approches pour constituer un sous-lexique de taille réduite, dont la particularité est de couvrir plus de 90% de l'usage. Un tel *lexique noyau* offre en outre la possibilité d'être réellement complété manuellement avec des informations sémantiques, de valence, pragmatiques etc.

Abstract We present a lexical development platform which comprises a lexical database of 440.000 lemmatized words of contemporary French, plus a set of maintenance tools. The lexical database is intended to be distributed and updated regularly. We present in this paper tools and techniques applied for the lexicon constitution and its enrichment, in particular the computation of lexical frequencies by morphosyntactic category. Then we describe various approaches to build an under-lexicon of reduced size, whose characteristic is to cover more than 90% of the use. Such a *kernel lexicon* makes it moreover possible to be really enriched by hand with semantic, valence, pragmatic information, etc.

1 Introduction

L'élaboration d'un lexique électronique peut sembler une tâche obsolète, de nombreux lexiques du français étant référencés. Cependant, force est de constater que cette affirmation doit être modulée. La première constatation est que seul un petit nombre d'entre eux est effectivement accessible. Il faut de ce point de vue souligner le rôle considérable joué par Bdlex (cf. [de Calmes98]) qui, dans le cadre des activités du GdR-PRC Communication Homme-Machine, a longtemps été le lexique le plus largement diffusé en contribuant ainsi puissamment à l'évolution du domaine en France. Le mode de diffusion constitue évidemment un aspect critique [1]. Un rapide survol des ressources lexicales libres d'accès pour le français permet d'en identifier deux :

- *Lexique* : il s'agit d'un lexique comportant 130.000 formes et comportant des informations morphosyntaxiques, phonologiques et des indications de fréquence (cf. [New01], `http://www.lexique.org/`).
- *ABU* : contient 300.000 formes avec indications morphosyntaxiques (cf. [ABU], `http://abu.cnam.fr/`).

On peut par ailleurs trouver quelques ressources verbales, par exemple :

- *Lefff* : il contient 200.000 formes verbales, avec les informations de base (temps, nombre, personne) (cf. [Clément04], `http://www.lefff.net/`);
- *Litote* : c'est une base contenant les formes conjuguées de 6.500 verbes. (`http://www.loria.fr/equipes/calligramme/litote/`)

Par ailleurs, il faut également signaler la démarche initiée par le Loria dans le cadre du projet *Morphalou* (cf. [Romary04], http`://loreley.loria.fr/morphalou/`). Ce projet fournira également à terme un lexique morphologique de 540.000 formes. Son intérêt tient d'une part au fait qu'il est collaboratif mais également qu'il s'inscrit dans le cadre du projet LMF (*Lexical Markup Framework*), proposant la normalisation du codage des informations linguistiques.

Il reste donc un travail important pour parvenir à un lexique de qualité. Pour cela, une base lexicale doit avant tout être nettoyée de façon à proposer une couverture adéquate du français. Il ne sert à rien de constituer une ressource de 400 ou 500.000 formes si la plupart d'entre elles ne sont pas attestées. Le second aspect concerne le type d'informations contenu dans le lexique. Il est en effet nécessaire qu'un lexique contienne pour une même entrée autant d'informations que possible concernant ses propriétés morphologiques, syntaxiques, bien entendu, mais également sémantiques, phonétiques ou phonologiques. La forme phonétisée de l'entrée, la syllabation ou la fréquence sont par exemple autant d'informations précieuses pour la description.

Nous décrivons dans cet article la base lexicale développée au LPL. Cette base, construite autour d'un lexique morphologique, présente la particularité d'être couvrante, de contenir des information variées et d'avoir été validée sur corpus. Cette base est associée à une véritable *plateforme* de développement lexical, munie de divers outils de maintenance et d'accès. Après une présentation des principales caractéristiques de cette plateforme, nous en proposons une évaluation se fondant sur différents corpus. Nous décrivons de plus l'exploitation de cette base dans la perspective d'une étude lexicale du français contemporain.

[1] Nous nous associons de ce point de vue à la démarche aujourd'hui proposée par le projet Morphalou et nos ressources seront distribuées dans ce cadre

2 Le lexique complet

Le lexique que nous avons mis au point a fait l'objet de beaucoup d'études et de travaux d'amélioration. Nous aboutissons actuellement à un lexique défactorisé de plus de 444.000 entrées correspondant à environ 320.000 formes orthographiques différentes. Ce lexique est associé à un ensemble d'outils permettant sa maintenance, sa sécurisation et son interrogation. Ce projet est la base nécessaire à des applications du TALN qui auront besoin d'une ressource fiable, c'est pourquoi l'accent a été mis sur la maintenabilité de la ressource.

Les entrées du lexique *DicoLPL* sont basées sur des ressources libres et une acquisition semi-automatique. Comme le montre la figure 1, nous avons au départ recensé et incorporé des lexiques libres, tels *ABU* ou *Lexique.org*. Le formatage de notre lexique a nécessité un travail de transformation, de catégorisation, de phonétisation etc., afin de faire correspondre les entrées acquises. L'étape importante que constitue le calcul des fréquences lexicales est abordé dans la prochaine section.

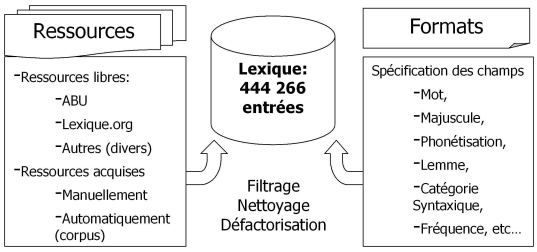

Figure 1: Conception du dictionnaire

Forme	Phonetisation	Frequence	Categorie	Lemme
de_traviole	d@_tRavjOl@	9	Rgp	de_traviole
dealer	dil@R	188	Ncms–	dealer
dealers	dil@R	368	Ncmp–	dealer
déambula	dea∼byla	5	Vmis3s–	déambuler
déambulai	dea∼bylE	0	Vmis1s–	déambuler
déambulaient	dea∼bylE	23	Vmii3p–	déambuler
déambulais	dea∼bylE	1	Vmii1s–	déambuler
déambulais	dea∼bylE	0	Vmii2s–	déambuler
déambulait	dea∼bylE	17	Vmii3s–	déambuler

Figure 2: Extrait du lexique

Le format du lexique, son codage, et son stockage ont été pensés afin d'accélérer son chargement dans les applications qui le requièrent. Ce lexique est en effet actuellement embarqué dans des applications de communication sur des machines ayant de petites capacités. D'autre part, il s'agit de permettre avec le même stockage un développement et des modifications manuels. C'est pourquoi un format ASCII, structuré en CSV tabulé classique a été choisi, plutôt qu'un standard XML ou qu'une forme binaire de type *base de données*. Ce choix a répondu à nos attentes et permet une transformation rapide dans d'autres formats tels que le XML répondant aux normes ISO (TC37/SC4) utilisées par le projet MORPHALOU par exemple.

Notre lexique se structure sous une forme défactorisée (une ligne par quadruplet *[Mot, Phonetisation, Categorie, Lemme]* par opposition à d'autres lexiques pour lesquels une seule ligne est réservée pour chaque forme orthographique.

L'extrait de lexique donné dans la table 2 met en évidence les caractéristiques de son format. On y observe la défactorisation du mot *déambulais*.

Certaines colonnes ont été réservées pour un usage ultérieur; les mots acceptés dans ce lexique ne doivent pas être des affixes, mais toujours des mots (simples ou composés) du langage courant. Ainsi, les préfixes et suffixes tels que *anti, hecto, isme* ou *able* en sont rejetés.

Le codage des champs du lexique est lui aussi contraint: les fréquences correspondent au nombre d'occurences de chaque entrée mesurée sur les corpus d'apprentissage. Les valeurs sont des entiers et ne représentent pas des pourcentages. Les valeurs de traits des catégories de chaque entrée sont formatées selon un codage dérivé de Multext et de Grace. La forme phonétisée est exprimée à l'aide de l'alphabet standard Sampa, qui permet un codage phonétique en texte brut sans faire appel à des polices de caractères spécifiques.

3 Plateforme d'enrichissement du lexique

Le lexique *DicoLPL* est une ressource en évolution. Nous présentons ici quelques uns des outils qui permettent son enrichissement.

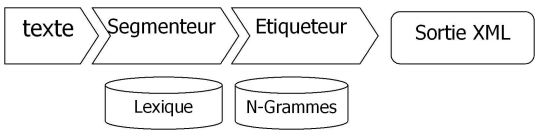

Figure 3: segmenteur et étiqueteur

Deux outils - un segmenteur et un étiqueteur- mettent en relation les mots d'un texte fourni en entrée avec les mots du lexique. La figure 3 illustre leur usage. Le segmenteur, basé sur des automates simples, effectue un découpage du texte en *tokens*. C'est à partir de ces informations que l'étiqueteur effectuera la désambiguisation en contexte des catégories à attribuer à chaque token.

La technique de désambiguisation que nous utilisons s'inspire des techniques stochastiques existantes. Nous avons cependant préféré développer notre propre étiqueteur afin de correspondre au mieux avec la précision des traits morphosyntaxiques que nous employons. Une première évaluation de l'étiqueteur sur le corpus du projet *Multitag* (cf. [Paroubek00]) a donné des résultats par catégorie variant de 60% à 99%. Le score moyen calculé sur le corpus de référence Multitag est de 95%.

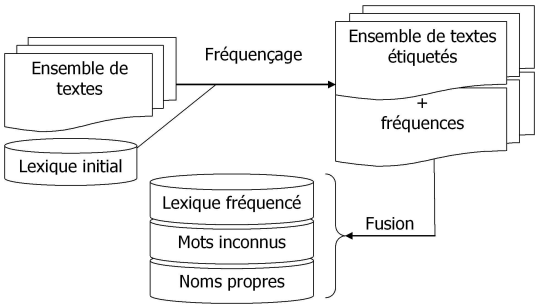

Figure 4: calcul des fréquences lexicales

Afin d'enrichir le lexique et de calculer au besoin les fréquences lexicales spécifiques à un ensemble de corpus, nous avons développé un outil de fréquençage. Comme l'indique la figure 4, cet outil fait appel aux résultats de l'étiquetage pour en déduire les fréquences des entrées du lexique, pour chaque couple *(mot, catégorie)*. A partir d'un lexique initial, étant donné un ensemble de textes, nous obtenons en sortie du fréquenceur un lexique des mots inconnus, un lexique des noms-propres et une nouvelle version du lexique initial, dont les champs *fréquence* sont mis à jour.

La version actuelle de *DicoLPL* dispose des fréquences acquises sur 153 millions de mots tirés du journal Le Monde, de ressources littéraires gratuites, de transcriptions de corpus oraux et des textes spécifiques (domaine médical, corpus de mails etc.).

D'autre part, la forme phonétisée des entrées est obtenue grâce à un phonétiseur inspiré du projet *Syntaix* (cf. [Di Cristo01]) pour la conception d'un système de synthèse vocale. D'autres champs (sémantique, valence verbale etc.) nécessitent toujours une validation manuelle.

L'évaluation des outils et du lexique est réalisée avec les techniques suivantes: Nous pouvons mesurer la **couverture** du lexique pour un corpus donné (calculer le quotient *nombre de mots reconnus / nombre total de mots*). La couverture actuelle du lexique représente 96% des corpus analysés (153 millions de mots). Lorsque nous souhaitons une information plus fine concernant l'étiquetage, il faut alors disposer d'un corpus de référence, pour lequel chaque mot est associé à une catégorie morphosyntaxique certifiée. Il est alors possible de mesurer les **scores de rappel et précision** pour chaque catégorie. C'est dans ce cadre que nous avons pu calculer un score de 95% sur le corpus de référence Multitag.

4 Un lexique noyau du français contemporain

Une fois le lexique constitué, il est nécessaire de vérifier sa couverture. Par ailleurs, l'analyse des corpus décrits plus haut permet de fournir des indications pour la constitution d'un dictionnaire minimal (ou *lexique noyau*) du français ayant une couverture maximale (un tel sous-lexique est toujours spécifique à un ensemble de corpus). Cette ressource est d'une grande importance pour le futur: Il n'est pas possible d'enrichir un grand lexique manuellement. Or, nombre d'informations ne peuvent aujourd'hui être acquises totalement automatiquement, notamment les informations sémantiques. Un lexique noyau permet d'identifier un nombre limité d'entrées lexicales qu'il est possible d'enrichir y compris manuellement. L'objectif est à terme de disposer d'une ressource lexicale très complète, comportant des informations syntaxiques, sémantiques, voire pragmatiques. Un lexique limité aux 10.000 formes les plus fréquentes couvre en moyenne 90% du français. Il s'avère donc intéressant de sélectionner un lexique noyau du Français contemporain avoisinant cette taille. La qualité de l'information concernant la fréquence de chacune des entrées du lexique complet permet de concevoir un lexique noyau (dorénavant LN) des mots les plus fréquents. C'est aussi l'occasion d'évaluer diachroniquement l'évolution du lexique de base du français depuis "Le Français Fondamental" (cf. [Gougenheim64] et [Blache05]). Nous avons sélectionné les formes pertinentes du LN grâce à une méthode simple utilisant une fréquence seuil (une autre méthode basée sur une réflexion à propos des types de catégories à conserver indépendamment de leur fréquence s'est révélée moins efficace et a été abandonnée). Ainsi, pour obtenir un dictionnaire de 10.000 formes (LN10) avons nous sélectionné les 10.000 entrées les plus fréquentes du lexique général DicoLPL, c'est-à-dire toutes les formes dont la fréquence est supérieure à 1091. Différentes versions de LN de taille croissante ont été produites suivant la même méthode : LN15 (fréquence>613, 15.017 formes), LN20 (fréquence>389,19.990 formes) et LN30 (fréquence>193, 30.018 formes) afin de comparer leurs couvertures et choisir le meilleur rendement taille/couverture.

DicoNoyau	Corpus écrit	Corpus oral
LN10 (f>1091)	88,63%	91,56%
LN15 (f>613)	91,07%	93,60%
LN20 (f>389)	92,50%	94,60%
LN30 (f>193)	94,08%	96,46%
DicoLPL	96,21%	99,02%

Figure 5: couvertures par taille de lexique et par type de corpus

Nous avons soumis les différentes versions du LN à un test de couverture sur deux types différents de corpus: un corpus écrit (580.000 mots extraits d'articles publiés dans le journal Le Monde) et un corpus oral (435.000 mots et regroupe le *Bristol Corpus*, un ensemble de 95 entretiens enregistrés et transcrits par Kate Beeching (1988-1990), ainsi que des corpus de parole recueillis au LPL).

Les résultats présentés dans la figure 5 appellent plusieurs commentaires: nous constatons d'abord que la couverture du lexique général DicoLPL (dernière ligne) n'est pas totale et qu'elle est meilleure pour le corpus oral que pour le corpus écrit, remarque qui vaut aussi pour les autres dictionnaires. Ceci s'explique selon nous par le fait que l'écrit utilise un vocabulaire beaucoup plus étendu et varié que l'oral. On constate aussi que les performances de couverture s'améliorent régulièrement au fur et à mesure que le LN contient plus de formes, ce qui est bien sûr attendu. Il faut néanmoins noter qu'il existe un saut qualitatif plus important entre LN10 et LN15 qu'entre LN15 et LN20 ou LN20 et LN30 alors même que l'écart de taille entre ces deux derniers est plus important. Le dictionnaire noyau de 15000 formes apparaît donc comme la version optimale pour obtenir la plus grande couverture avec un nombre réduit de formes.

5 Conclusion

La plateforme de développement de lexique décrite dans cet article répond à un certain nombre de besoins à la fois en termes de richesse d'informations, mais également de développements de lexiques spécialisés en produisant des fréquences spécifiques. Notre approche permet de rationaliser le choix des entrées sur lesquelles travailler en proposant la construction d'un lexique noyau élaboré sur la base d'une véritable analyse de la langue. L'enrichissement manuel de petits lexiques avec des informations sémantiques, pragmatiques etc. s'en trouve facilité. C'est pourquoi nous défendons la démarche qui consiste à concentrer les efforts sur un sous-lexique dont la couverture a été vérifiée sur corpus. D'autre part, un lexique de petite taille offre de nombreuses possibilités d'études sur l'usage avec notamment les *réseaux sémantiques*, les *petits mondes* etc. (voir à ce propos [Ferrer01]).

Le fait de disposer d'un grand lexique de formes n'en reste pas moins un atout, puisque c'est à partir d'une telle ressource que peuvent être extraits des sous-lexiques *ad hoc* couvrant des types de texte de domaines divers que le *fréquençage* permet d'isoler.

Enfin, la tâche de constituer une telle ressource est immense. Nous souhaitons la voir s'améliorer avec le temps, ce qui suppose sa diffusion, sa confrontation à l'usage et un retour de la communauté. La plateforme décrite ici comportant une série d'outils de maintenance, il est ainsi possible d'envisager une mise à jour régulière des informations. Au total, notre contribution viendrait s'inscire dans le mouvement de mise à disposition de ressources du français initié par les différents projets signalés plus haut.

Références

Association des Bibliophiles Universels, "ABU. Dictionnaire des mots communs", in La Bibliothèque Universelle, http://abu.cnam.fr/DICO/mots-communs.html. CNAM.

Blache P., M.-L. Guénot & C. Portes (2005), "Outils et ressources pour la mise à jour du Français Fondamental", in Proceedings of Français Fondamental: 50 ans de travaux et d'enjeux.

Clément L., B. Sagot & B. Lang (2004), "Morphology-Based Automatic Acquisition", in proceedings of LREC-04.

de Calmès M. & G. Pérennou (1998), "BDLEX : a Lexicon for Spoken and Written French", in proceedings of LREC-98

Di Cristo & P. Di Cristo (2001), "Syntaix : une approche métrique-autosegmentale de la prosodie", in revue TAL, 42:1

Ferrer R., Cancho I. & Sole R. (2001), "The small-world of human language", Proceedings of the Royal Society of London, B 268, 2261– 2266 url = "citeseer.ist.psu.edu/ferrer01small.html"

Gougenheim, G. ; Rivenc, P. ; Michéa, R. & Sauvageot, A. (1964), "L'élaboration du Français Fondamental", 1er degré, Didier : New B.

Pallier C., L. Ferrand & R.Matos (2001), "Une base de données lexicales du Français contemporain sur Internet : Lexique ", in L'Année Pschologique, 101

Paroubek P. & M. Rajman (2000), "MULTITAG, une ressource linguistique produit du paradigme d'évaluation", in Actes de la conférence TALN-2000

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