

Une plateforme d'analyse lexicale pour le développement du Français Fondamental

Philippe Blache, Marie-Laure Guénot & Cristel Portes
Laboratoire Parole et Langage - CNRS / Université de Provence
29 av. Robert Schuman 13621 Aix en Provence cedex
{pb;mlg;portes}@lpl.univ-aix.fr

L'élaboration du français fondamental a été un travail précurseur non seulement du point de vue de ses objectifs et ses applications, mais également par la méthode employée. Il s'est agi en effet d'interpréter une analyse statistique sur un corpus de français parlé, avant que la linguistique de corpus n'apparaisse avec ses outils. Nous proposons dans cet article une nouvelle élaboration du français fondamental s'appuyant sur l'exploitation d'une plateforme d'analyse lexicale utilisant les techniques et ressources de traitement automatique des langues développées au LPL. Cette étude offre bien entendu la possibilité d'une analyse diachronique, mais permet également d'envisager la possibilité de disposer d'un outil pouvant de façon très simple effectuer régulièrement une mise à jour des données d'origine. La plateforme utilisée exploite un lexique morphologique de 450.000 formes (soit environ 50.000 lemmes), contenant différents types d'information comme la forme phonétiques ou la fréquence. Les outils associés permettent d'automatiser l'analyse des corpus notamment pour ce qui concerne les tâches de segmentation des textes, d'étiquetage, de désambiguïsation, de fréquentage ainsi que pour l'évaluation des résultats.

Le lexique utilisé est donc très couvrant, mais une analyse sur corpus montre que 90% des formes rencontrés représentent environ 2% de notre lexique général. Nous avons donc créé un *lexique noyau* (ci-après *LN*) des 10.000 formes les plus fréquentes, et l'avons confronté à des corpus de genres divers pour vérifier son efficacité. Nous avons ainsi constaté que si 10.000 mots permettent effectivement d'étiqueter 90% de corpus écrit, en revanche 4.000 formes suffisent à étiqueter la même proportion de corpus strictement oral (sur la base d'un corpus de 500.000 mots). En effet, en comparant les résultats fournis par l'étiquetage de corpus écrits et de corpus oraux, on constate que la répartition des catégories ainsi que la variabilité au sein d'une même catégorie, dépendent de la nature du corpus :

- A l'oral, on trouve une quantité nettement supérieure de marqueurs de formulation (phatiques, connecteurs, etc.).
- A fréquence égale, les mots lexicaux (noms, verbes, adjectifs, adverbes) sont plus variés et moins fréquents à l'écrit, et plus fréquents mais moins variés à l'oral. En d'autres termes, on utilise autant de verbes à l'oral qu'à l'écrit, mais on a tendance à utiliser plus fréquemment les mêmes verbes à l'oral.
- Le reste des catégories (prépositions, déterminants, conjonctions,...) est réparti à peu près de la même façon entre écrit et oral.

Nous avons donc comparé ce *LN* avec la partie vocabulaire du *Français Fondamental* (ci-après *FF*), en particulier pour avoir une idée des changements diachroniques qui ont pu affecter le français oral en plus de soixante ans. Cette comparaison a nécessité une adaptation du *LN* : les entrées par *formes* de *LN* ont ainsi été synthétisées en entrées par *mots* (correspondant à peu près à la notion de lemme) comparables à celles du *FF*, et seules les 1.063 entrées les plus fréquentes ont été conservées (contre 10.000 formes dans le *LN* d'origine). A la suite de quoi nous avons effectué des tests de comparaison entre *LN* ainsi réduit et *FF* qui nous ont mené aux résultats suivants :

- *LN_réduit* contient plus de noms (52% contre 40%) et moins de verbes (11% contre 22%) que le *FF* : ceci est partiellement dû au fait qu'un certain nombre de radicaux qui sont présents sous leur forme verbale dans *FF* se retrouve sous leur forme nominale dans *LN* (notamment danser>danse, employer>emploi, intéresser>intérêt). Plus marginalement, *LN* comporte plus de mots grammaticaux (29% contre 27%) et moins d'adjectifs (7% contre 11%).
- Malgré ces différences, les rapports entre fréquence et catégories grammaticales dans *LN_réduit* est très comparable à celui du *FF* présenté dans le chapitre 2 de l'ouvrage de 1964 (Gougenheim, G. ; Rivenc, P. ; Michéa, R. & Sauvageot, A., 1964, *L'élaboration du Français Fondamental*, 1^{er} degré, Didier : 114-117) : les mots grammaticaux sont plus nombreux dans les hautes fréquences et de plus en plus rares au fur et à mesure que la fréquence décroît contrairement aux noms qui ont le comportement inverse. Ici comme là, les verbes sont plus réguliers dans leur répartition alors que les adjectifs se comportent à peu près comme les noms.
- Les deux lexiques partagent 654 entrées soit 62% de leur effectif. Dans cet ensemble commun, les mots grammaticaux sont mieux représentés que dans chacun des lexiques (36% au lieu de 27-29%). Ce résultat est cohérent avec la nature des mots grammaticaux, plus stables que les mots lexicaux.
- Enfin, entre 1954 et 2005, les mots nouveaux du lexique fondamental sont majoritairement des noms alors que les verbes sont les moins bien représentés.

Outre ses applications didactiques, le français fondamental a été une expérience remarquable d'analyse du français parlé en offrant une image précise à un instant donné de l'état de la langue. Nous proposons d'utiliser les techniques modernes du traitement automatique des langues pour élaborer un outil d'observation automatisant la plupart des tâches nécessaires à la mise à jour des données initiale offrant ainsi la possibilité d'une étude diachronique tout en fournissant des données quantifiées précises sur la base d'analyse de corpus très volumineux.