# Un devin de microstructures pour importer ou normaliser des ressources lexicales

Mathieu Mangeot & Valérie Bellynck
Laboratoire GETALP-LIG
Bâtiment IMAG CS 40700
38058 GRENOBLE CEDEX 9, FRANCE
{mathieu.mangeot,valerie.bellynck}@imag.fr

Au sein des études lexicales, le dictionnaire est indissociable du corpus. Au moins trois modes d'interaction peuvent être envisagés. Tout d'abord, l'extraction d'unités lexicales d'un corpus nécessite presque toujours un dictionnaire de la langue. Ensuite, les dictionnaires modernes se construisent avec des vocables attestés dans des corpus et enfin, le dictionnaire lui-même peut servir de corpus. En particulier, les dictionnaires bilingues avec des exemples d'usage traduits fournissent d'intéressants corpus bilingues alignés.

L'intérêt des bases lexicales s'étend aux non-informaticiens qui produisent et utilisent des ressources lexicales. Ils voudraient profiter de ces technologies sans être confrontés aux concepts informatiques sous-jacents. Dans la suite de cet article, nous décrivons une chaîne de traitement automatisée permettant in fine de disposer de ressources lexicales en ligne pour les consulter et les éditer collaborativement.

Les premières étapes du traitement sont réalisées par iPolex, un entrepôt de données lexicales permettant d'importer et de stocker des ressources lexicales telles quelles. Lors de l'import d'une ressource, l'utilisateur est invité à renseigner les métadonnées décrivant cette ressource : nom, date de création, auteur, droits d'exploitation, macrostructure (structure des volumes et langues traitées). La microstructure des articles de chaque volume est ensuite devinée pour l'intégration des ressources dans Jibiki, une plateforme de manipulation générique en ligne de ressources lexicales (Mangeot & Chalvin, 2006; Zhang et al. 2014).

Il est possible d'importer dans Jibiki n'importe quel type de ressource lexicale (lexique, dictionnaire, base terminologique, monolingue, bilingue, multilingue, etc.) au format XML. Pour conserver la structure de chaque ressource, celles-ci sont manipulées à l'aide de pointeurs dans leur microstructure.
Chaque partie d'information clairement identifiée correspond à un pointeur spécifique appelé pointeur CDM pour Common Dictionary Markup (Mangeot, 2002). Nous avons choisi la norme XPath pour implémenter ces pointeurs. Par exemple, dans une ressource, le mot-vedette sera noté /volume/entry/hw/text(), la catégorie grammaticale sera notée /volume/entry/gram-block/pos/text(), etc.
Les pointeurs sont utilisées lors de l'import pour indexer chaque partie d'information décrite. Ils sont ensuite utilisés lors de la consultation pour effectuer des recherches multicritères. Par exemple, rechercher les articles dont le mot-vedette commence par « b », dont la catégorie grammaticale est un nom et dont le domaine est « botanique ».

L'import d'une ressource dans Jibiki nécessite donc de décrire ces pointeurs. Ce processus prend beaucoup de temps, est source d'erreurs et nécessite l'appropriation de concepts informatiques.
C'est pourquoi nous avons travaillé à la mise au point d'un programme permettant de deviner la valeur des pointeurs d'une ressource à partir d'une analyse poussée de la structure des articles et d'une série de règles empiriques.
Le programme est constitué de deux phases principales. La première consiste à calculer une signature lexicale du fichier à analyser. Pour chaque balise et chaque attribut XML du fichier, on compte le nombre d'occurrences. Le contenu de chaque attribut ainsi que le contenu textuel de chaque balise est ensuite analysé. On compte le nombre de valeurs différentes. Si celui-ci est inférieur à 50, chaque valeur est affichée avec son nombre d'occurrences. On compte également le nombre de valeurs classées par ordre alphabétique, le nombre moyen de caractères et de mots.

La signature lexicale peut également être utilisée pour effectuer des corrections sur une ressource. Par exemple, si une balise apparaît fréquemment, que son contenu est constitué d'une petite liste de valeurs qui se répète et que parmi ces valeurs, une valeur n'apparait qu'une fois alors que les autres valeurs

apparaissent plus fréquemment, il est possible que celle-ci soit une erreur (Mangeot & Enguehard, 2013). De même, il est possible de détecter des erreurs de structuration dans l'article. Par exemple, si un bloc d'exemples se trouve à deux endroits différents dans la structure avec une fréquence très élevée dans un cas et peu élevée dans l'autre.

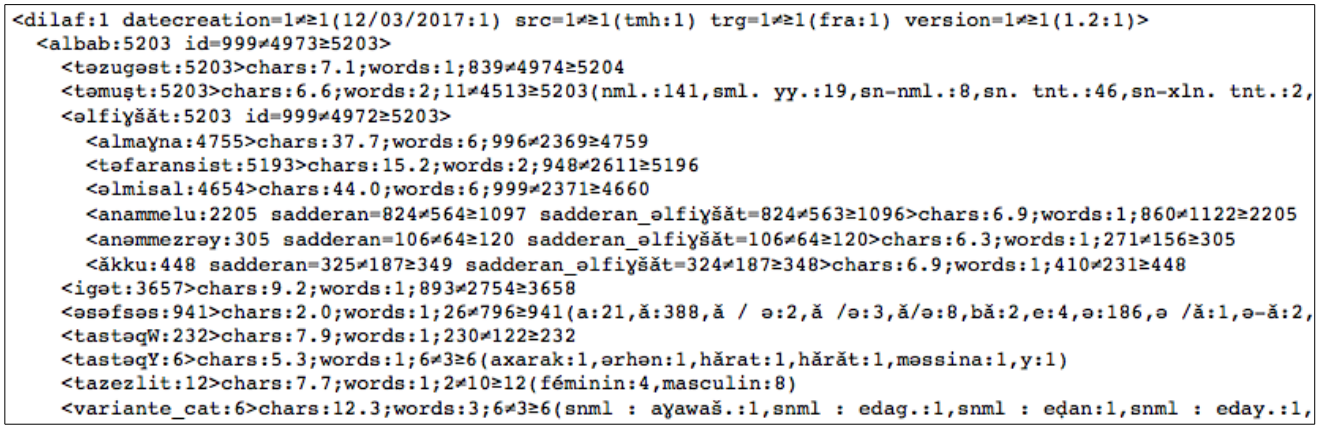

Figure 1 : signature lexicale du dictionnaire DiLAf tamajaq-français

Dans la deuxième phase, la signature lexicale est utilisée pour calculer une série d'hypothèses pour chaque pointeur CDM. Par exemple, si le contenu d'une balise est très souvent classé par ordre alphabétique, qu'il y a une majorité de valeurs différentes, qu'il y a en moyenne un seul mot, que cette balise apparaît au début de l'article, il y a une forte probabilité que ce soit le mot-vedette (élément *<təzugəst>* dans la figure 1). Autre exemple : si le contenu d'une balise est constitué d'une petite liste de valeurs qui se répète, qu'il y a en moyenne un ou deux mots, que cette balise apparaît après le mot-vedette, il y a de fortes probabilités que ce soit la catégorie grammaticale (élément *<təmuṣt>* dans la figure 1).

Une fois les hypothèses calculées, le devineur les affiche à l'utilisateur qui peut les corriger manuellement s'il constate des erreurs. L'utilisateur peut également ajouter de nouveaux pointeurs spécifiques à sa ressource s'il désire indexer une partie d'information qui n'est pas décrite par des pointeurs CDM existants. Si l'utilisateur ne maîtrise pas les subtilités de la norme XPath, il peut malgré tout importer sa ressource dans Jibiki à condition que les pointeurs CDM de l'article et du mot-vedette aient été correctement détectés.

L'outil a été utilisé avec succès par quelques personnes pour importer une cinquantaine de ressources. Il est disponible en source ouverte sur le répertoire Github suivant : https://github.com/mangeot/ipolex
Une image Docker est également disponible à cette adresse : https://hub.docker.com/r/mangeot/ipolex/
Celle-ci permet d'installer l'outil très simplement à l'aide d'une seule commande.

**Références**

**Mathieu Mangeot** (2002) *An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language.* Proc. of International Standards of Terminology and Language Resources Management, LREC 2002 workshop, Las Palmas, Islas Canarias, Spain, 28 May 2002, pp 37-44.

**Mathieu Mangeot & Antoine Chalvin** (2006) *Dictionary Building with the Jibiki Platform: the GDEF case.* Proc. of LREC 2006, Genoa, Italy, 23-25 May 2006, pp 1666-1669.

**Mathieu Mangeot & Chantal Enguehard** (2013) *Des dictionnaires éditoriaux aux représentations XML standardisées. Chapitre 8, livre "Ressources Lexicales : contenu, construction, utilisation, évaluation",* Eds. Nuria Gala & Michael Zock, Linguisticae Investigationes Supplementa, John Benjamins Publishing, Amsterdam, Pays-Bas, 24 p.

**Mathieu Mangeot & David Thevenin** (2004) *Online Generic Editing of Heterogeneous Dictionary Entries in Papillon Project.* Proc. of COLING 2004, ISSCO, Université de Genève, Switzerland, 23-27 August 2004, vol 2/2, pp 1029-1035.

**Ying Zhang, Mathieu Mangeot, Valérie Bellynck & Christian Boitet** (2014) *Jibiki-LINKS: a tool between traditional dictionaries and lexical networks for modelling lexical resources.* Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex) 2014 (Eds. Michael Zock, Reinhard Rapp, Chu-Ren Huang), Dublin, Ireland, 23 August 2014, 12 p.