Apprentissage du lexique des langues romanes à l'aide d'une ressource lexicale fondée sur la notion de familles et séries de mots

Nuria Gala, Nabil Hathout, Véronique Rey

Apprentissage du lexique des langues romanes à l'aide d'une ressource lexicale fondée sur la notion de familles et séries de mots

N. Gala, LIF-CNRS, Marseille

N. Hathout, CLLE-ERSS, CNRS & UTM, Toulouse

V. Rey, SHADYC-EHESS, Marseille

Mots clés

Ressource lexicale, langues romanes, familles de mots, séries morphologiques, morphologie constructionnelle, analogie formelle.

Résumé

Les langues romanes ont été à l'origine de la dialectologie en France et elles sont attestées comme relevant d'une même langue mère. La trace de leur origine commune est clairement perçue, notamment, par la similarité des unités lexicales. Néanmoins, la similarité entre des langues proches n'a pas été jusqu'ici décrite dans une ressource lexicale accessible par le Web et destinée à l'apprentissage de ces langues.

La notion de similarité a cependant été étudiée avec des objectifs différents. Quelques travaux en traitement automatique des langues ont, par exemple, exploité la structure du vocabulaire pour l'obtention automatique d'équivalences dans des domaines spécialisés (Langlais et al. 2008), pour traduire des mots inconnus dans des systèmes de traduction automatique (Langlais et Patry, 2007) ou pour calculer la proximité entre langues (Lepage et al. 2009). En linguistique typologique, de nombreuses études comparatives ont cherché à rendre compte de la similarité du lexique entre différentes langues dans le but d'établir des familles de langues (par exemple, Greenberg, 2000 ; Ruhlen 1997).

En revanche, la similarité des familles (ensemble de mots qui partagent un même radical et l'essentiel de leurs propriétés sémantiques, paradigmes lexicaux ou *clusters* (Bybee, 1985)) et des séries morphologiques (ensemble de mots qui entrent dans une même série analogique constructionnelle, par exemple contrôlable, dérivable, lavable) entre langues proches, à notre connaissance, n'a pas été étudiée. Dans les études citées ci-dessus, les mots sont en effet comparés entre les différentes langues individuellement ; cependant, la similarité des familles et des séries de mots n'a pas été évaluée globalement au niveau des lexiques entiers. Une telle comparaison permet par exemple de pointer des divergences comme *cesto* dans l'exemple suivant : pain, panier, compagnon (fr), pa, panera, company (cat), pan, *cesto*, compañero (es). De surcroît, il n'y a pas de nos jours une ressource linguistique capable de rendre

visible les familles et des séries de mots au sein de langues proches et, par conséquent, de rendre l'apprentissage de telles langues plus facile.

Dans cette communication, nous nous proposons de présenter une ressource en cours de construction dont la principale caractéristique est le fait d'appréhender des ensembles de mots regroupés en familles et en série, non seulement dans une même langue, mais aussi dans des langues proches (français, catalan, espagnol, italien). La création de cette ressource morphologique constructionnelle repose sur deux groupes de travaux. Le premier, réalisé par Gala et Rey (2008) a permis de constituer un lexique de familles de mots du français et de les caractériser. Gala et al. (2009, 2010) se sont d'autre part intéressés à la caractérisation sémantique des familles de mots, ouvrant la voie à un accès lexical par les unités lexicales elles-mêmes mais aussi par les traits sémantiques qui leur sont associés. Le second, réalisé par Hathout (2008, 2009), développe un ensemble de méthodes visant à faire émerger la structure morphologique du lexique à partir des propriétés sémantiques et formelles des mots.

Cette ressource permettra de combler un manque flagrant. Il existe en effet un lexique morphologique multilingue pour les langues germaniques (anglais, allemand et néerlandais), la base CELEX (Baayen et al. 1995). La ressource en cours de réalisation pour les langues romanes disposera en plus de liens de traductions permettant des comparaisons fines entre les langues donnera à voir à la fois des continuités sémantiques pour certaines familles ou séries de mots et des ruptures dans d’autres cas.

Les atouts d'une telle ressource se trouvent autant au niveau théorique que pratique. D'un point de vue théorique, la manipulation du lexique inter-langues en fonction des familles et des séries de mots permettra d'étudier le lexique des langues romanes d'un point de vue phonologique, morphologique et sémantique jusqu'ici peu exploré. D'un point de vue applicatif, une telle base offrira des avantages au niveau de l'apprentissage du vocabulaire des langues romanes : les regroupements en familles et en séries offrent une façon originale d'appréhender le lexique des langues, autant pour des apprenants de langue maternelle ou secondaire que dans des cas de pathologies particulières (dyslexie etc.). Leurs comparaisons permettra d'identifier les difficultés induites par différents types de continuités et de divergences (faux amis dans les cas de similarités formelles, supplétions dans les cas de divergences formelles, etc.).

Références

R. H. Baayen, R. Piepenbrock & L. Gulikers, The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

Bybee, Joan L. 1985.*Morphology: A study of the relation between meaning and form. A*msterdam: John Benjamins.

Gala N., Rey V. et Zock M. (2010) A tool for linking stems and conceptual fragments to enhance word access. In the *seventh international conference on Language Resources and Evaluation (LREC)*. La Valetta, Malta, mai 2010.

Gala N. et Rey V. (2009) Acquiring semantics from structured corpora to enrich an existing lexicon. In *E-lexicography in the 21st century : new applications, new challenges*, Louvain-la-Neuve, octobre 2009.

Gala N. et Rey V. (2008) Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. Actes de TALN 08: *Traitement Automatique des Langues Naturelles*, Avignon, juin 2008.

Greenberg J. 2003, *Les langues indo-européennes*, Belin : Paris.

Hathout N. (2009) Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie. Habilitation à diriger des recherches. Universités de Toulouse II-Le Mirail.

Hathout N. (2008) Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the Coling workshop Textgraphs-3*, pp. 1-8, Manchester.

Langlais P., Yvon F., Zweigenbaum P. (2008) Analogical translation of medical words in different languages. Dans Bengt Nordström and Aarne Ranta, editors, LNAI 5221, *Proceedings of the 6th International Conférence GoTAL 2008, Advances in Natural Language Processing*, pages 284–295, Gothenburg, Suède..

Langlais P. et Patry A. (2007) Translating unknown words by analogical learning. Actes de la conférence *Empirical Methods in Natural Language Processing and Computationnal Natural Language Learning* (EMNLP-CoNLL), pages 877-886, Prague.

Lepage Y., Lardillieux A., Gosme J. (2009) Commonality across vocabulary structures as an estimate of the proximity between languages. *4th Language & Technology Conference* (LTC'09), Poznan, Pologne.

Ruhlen M. 2001, *L'origine des langues*, Belin : Paris.