Section 16 - Projets en cours ; ressources et outils nouveaux

LexROM : un réseau lexical pour des familles de mots dans des langues romanes

Nuria GALA
Equipe TALEP LIF CNRS, Aix Marseille Université
nuria.gala@lif.univ-mrs.fr

Verginica BARBU MITITELU
Romanian Academy Research Institute for Artificial Intelligence
vergi@racai.ro

Mots clés

Lexique, morphologie, sémantique, traitement automatique des langues romanes, linguistique contrastive

Résumé

Le panorama des ressources lexicales s'est diversifié considérablement dans les dernières décennies grâce aux progrès technologiques. Dans le domaine de la (e-)lexicographie les avancées plus significatives concernent autant les contenus (par exemple, l'ajout d'informations sur des co-occurrences grâce à l'accès à de grands corpus, l'adaptation des informations à des utilisateurs et à des situations d'utilisation diverses, etc.) que les moyens d'accès aux contenus (recherches par mot clé ou avancées, etc.). Par ailleurs, le domaine du TAL (Traitement Automatique des Langues) a intensément collaboré à la création de ressources lexicales nouvelles, nécessaires au développement et à la validation de systèmes d'analyse morphologique, syntaxique et sémantique (bases de données lexicales explicites, réseaux lexicaux, etc.) (Gala et Lafourcade 2011). Cependant, les synergies entre le TAL et la lexicographie n'ont pas été suffisamment exploitées jusqu'ici, notamment en ce qui concerne la structuration des données -des listes à des réseaux de mots (Heid 2009)- et la création et la diffusion de ressources multilingues.

Ainsi, on ne peut que s'étonner de l'absence de ressources lexicales pour des langues typologiquement proches[1]. Si des ressources multilingues existent, par exemple EuroWordNet (Vossen 1998) ou Babelnet (Navigli et Ponzetto 2010), elles ont été principalement conçues pour mettre en avant les équivalences sémantiques entre les termes. A notre connaissance, aucune ressource ne permet de repérer des équivalences ou des divergences de forme dans des paradigmes de familles de mots (bien que la factorisation d'éléments communs ait été considérée, depuis longtemps, comme un des intérêts principaux pour justifier l'enseignement du latin et du grec ou des langues proches). L'inclusion de relations dérivationnelles dans divers wordnets devrait permettre la comparaison de familles de mots dans diverses langues, or à ce jour, parmi les langues romanes, seul le wordnet roumain présente ce type de relations (Barbu Mititelu 2012).

Dans cette communication, nous présenterons une ressource -en cours de construction- fondée sur la notion de famille de mots interlangue. La structuration du lexique des langues sous forme de familles, déjà décrit dans la littérature notamment par Bybee (1985), nous a semblé une propriété intéressante à explorer d'un point de vue multilingue dans le cadre de quelques langues romanes (à ce stade, français, roumain, catalan et espagnol). Notre objectif à terme est, ainsi, de proposer une ressource offrant des possibilités élargies pour des études contrastives, c'est-à-dire, des études portant sur des similarités ou des divergences (trous lexicaux) dans les paradigmes, autant d'un point de vue morphologique (base commune ou non (cf. figure 1), affixes similaires ou non), sémantique (même sens ou non, évolution du sens vers un sens figuré ou non, etc.) que quantitatif (taille de la famille similaire ou divergences importantes, cf. figure 2).

Par ailleurs, exploiter les ressemblances formelles au niveau interlangue présente un intérêt aussi bien dans des applications humaines -apprentissage du vocabulaire ou aide à l'écriture dans une langue étrangère- qu'en TAL -désambiguïsation sémantique, traduction[2]. Enfin, le développement et l'enrichissement d'un réseau multilingue comme LexROM soulève de nombreuses questions scientifiques et méthodologiques, certaines

1 De nombreux travaux existent, néanmoins, en ce qui concerne les corpus multilingues, utilisés en linguistique contrastive, en traduction, etc. (Granger 2003 ; 2010).

2 Une initiative de désambiguïsation lexicale pour les langues romanes a eu lieu il y a une douzaine d'années en TAL (Romanseval) mais bien qu'elle ait constitué un point de référence sur les systèmes de désambiguïsation sémantique, elle n'a pas donnée lieu à des ressources accessibles et librement utilisables).

déjà abordées dans Gala (2011), auxquelles nous apporterons des nouveaux éléments de réponse dans le cadre de notre communication.

Catégorie	*FR*	*CA*	*ES*	*RO*
A/N	petiot	petitó	pequeñito	micuț
N	**petit**	**petit**	**pequeño**	**mic**
Adv	petitement	petitament	pequeñamente	-
Nf	petitesse	petitesa	pequeñez	micime
Nf	rapetissement	-	-	micșorare
V	rapetisser	enpetitir	enpequeñecer	micșora

Figure 1. Famille morphologique 'petit', maintient d'une base commune en FR, CA et ES.

Catégorie	*FR*	*CA*	*ES*	*RO*
N	**chaise**	**cadira**	**silla**	**scaun**
N	chaisier	cadiraire	sillero	scăunar
N	-	cadiral	sillón	-
N	-	cadiram	sillería	-
N	-	cadirat	-	-
N	-	cadiratge	-	-
V	-	cadirejar	-	-
Nm	chaisier	cadirer	silletero	scăunar
N	-	cadireta	silleta	scăunel
Nm	-	-	sillín	-
N	-	-	sillar	-

Figure 2. Famille morphologique avec divergences quant à la taille des paradigmes (en espagnol, une vingtaine d'entrées avec la même base 'silla').

Références bibliographiques

[Barbu Mititelu 2012] V. Barbu Mititelu (2012) Adding Morpho-semantic Relations to the Romanian Wordnet. *Proceedings of LREC'12,* Istanbul, Turkey, 2596-2601.

[Bybee 1985] J. L. Bybee (1985) Morphology. A study of the relation between meaning and form. *Typological studies in Language 9*. Amsterdam: Benjamins.

[Gala 2011] N. Gala (2011) Developing a lexicon of word families for closely-related languages. Actes de *ESSLLI International Workshop on Lexical Resources (WoLeR).* Ljubljana, Slovenia, août 2011.

[Gala & Lafourcade 2011] N. Gala et M. Lafourcade (2011) NLP lexicons: innovative constructions and usages for machines and humans. *Electronic lexicography in the 21st century: new applications for new users (eLEX-2011)*. Bled, Slovenia, novembre 2011.

[Granger 2003] S. Granger (2003). The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies. Corpus-based Approaches to Contrastive Linguistics and Translation Studies. S. Granger, J. Lerot & S. Petch-Tyson (eds). Amsterdam/New York: Rodopi, 17-29.

[Granger 2010] S. Granger (2010) Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. Journal of Shanghai Jiaotong University.

[Heid 2009] U. Heid (2009) Aspects of lexical description for electronic dictionaries. Electronic lexicography in the 21st century (eLEX-2009), Louvain, Belgium.

[Navigli & Ponzetto 2010] R. Navigli et S. Ponzetto (2010) BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11-16 July 2010, pp. 216-225.

[Vossen 1998] P. Vossen (1998) EuroWordNet: A Multi-lingual Database with Lexical Semantic Networks. Kluwer, Dordrecht, The Netherlands.