Extraction automatique et préparation lexicographique d'unités polylexicales étendues

Sascha Diwersy, Olivier Kraif

Cette communication porte sur la conception et le développement du « lexico-graphe », un outil dédié à l'extraction automatique du « profil collocationnel » d'un pivot donné dans le cadre d'une colligation donnée (le pivot étant considéré dans une fonction syntaxique déterminée telle que sujet, objet, modifieur, etc.). Par *collocation*, nous subsumons l'ensemble des phénomènes régissant la combinatoire lexico-syntaxique d'un pivot donné, dans la perspective de Sinclair (2004). Dans le cadre du projet Emolex, nous avons mis en place une architecture permettant d'extraire et de comparer les profils combinatoires des pivots par l'extraction de ce que nommons les *lexicogrammes* (Heiden & Tournier, 1998) i.e. des matrices enregistrant, pour tous les cooccurrents syntaxiques d'un pivot, des mesures d'association statistiques (fréquence, loglike, t-score, etc.).

Bien que les *lexicogrammes* aient une structure essentiellement binaire, enregistrant des associations entre le pivot et ses collocatifs, il est possible d'extraire les lexicogrammes pour des *pivots complexes*, c'est-à-dire des pivots insérés dans un sous-arbre de dépendance comportant plusieurs autres unités (p.ex. *exprimer-(obj)->satisfaction*). Cette généralisation de notre modèle permet d'extraire de manière itérative, pour un pivot pris dans une colligation donnée, tout un ensemble *d'extensions collocationnelles* statistiquement significatives : ces extensions forment ce que nous appelons le *profil collocationnel*.

Pour cette étude nous avons réuni des corpus en 5 langues (allemand, français, anglais, espagnol, russe) comportant à la fois des œuvres littéraires contemporaines et des textes journalistiques (environ 100 millions de mots par langue pour le journalistique, et 20 millions pour le littéraire). Ces corpus ont été analysés syntaxiquement grâce à différents outils : XIP pour l'anglais (Aït-Mokhtar et al. 2001), Connexor pour l'allemand, le français et l'espagnol (Tapanainen & Järvinen 1997), DeSR pour le russe (Attardi et al. 2007), basé sur un modèle stochastique créé à partir du corpus arboré SyntagRus (Nivre et al., 2008).

Après avoir décrit l'architecture du système d'extraction, nous analysons les résultats obtenus à partir de l'exemple du pivot *admiration* pris en tant qu'objet direct. Dans une perspective lexicographique, nous proposons une structuration hiérarchisée des extensions lexicales de notre pivot, accompagnée d'exemples extraits du corpus. Pour *admiration*, nous obtenons par exemple 14 collocatifs verbaux (en fonction des seuils appliqués) donnant lieu à de nombreuses séquences, dont voici un échantillon:

- *forcer*
    - *forcer l'admiration*
        - *qui force l'admiration*
            - *précision qui force l'admiration*
- *vouer*
    - *vouer une admiration sans borne*
    - *vouer une profonde admiration*
    - ....
- *cacher*
    - *ne cache pas son admiration*
    - …

Pour chaque extension, nous donnons les informations suivantes :

- le sous-arbre syntaxique correspondant ;

- les exemples en contexte regroupés par réalisations morphosyntaxiques (celle-ci étant triées par ordre de fréquence) ;

- le paradigme étendu, dont fait partie le pivot, dans le contexte de l'extension collocationnelle : par exemple, à partir d'une expression telle que *ne pas cacher son admiration*, nous pouvons identifier le contexte collocationnel suivant : *ne pas cacher son + N*. Nous montrons comment un tel contexte permet d'identifier, de façon automatisée, des paradigmes assez riches, manifestant la structuration du champ sémantique étudié. Par exemple, à partir de *ne pas cacher son + N* , on trouve le paradigme : *satisfaction, inquiétude, déception, ambition, joie, sympathie, intention, amertume, scepticisme, préférence, colère, agacement, embarras, intérêt, pessimisme, volonté,* tandis que pour le contexte *exprimer son + N*, on trouve un paradigme assez différent, manifestant d'autres potentialités sémantiques (/situation positive ou négative/ + /affect centré sur le sujet/ vs /situation négative/ + /affect interpersonnel/) : *inquiétude, préoccupation, désaccord, solidarité, regret, soutien, émotion, indignation, souhait, gratitude, crainte, déception, doute, colère, mécontentement, désir*.

Dans la suite de nos travaux, nous prévoyons d'extraire les profils collocationnels de toutes les unités lexicales étudiées dans le cadre du projet Emolex (dans le champ sémantique des émotions), et d'établir des liens, de façon automatique, entre le résultat brut de ces extractions et les codages syntaxiques et sémantiques réalisés par notre équipe (Novakova et al. 2012).

Références

AÏT-MOKHTAR, S., CHANOD, J.-P., ROUX C. (2002) "Robustness beyond Shallowness: Incremental Deep Parsing", *Natural Language Engineering*, 8 :121-144.

ATTARDI, G., DELL'ORLETTA, F., SIMI, M., CHANEV, A., CIARAMITA, M. (2007) "Multilingual Dependency Parsing and Domain Adaptation using DeSR", In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, Prague.

HEIDEN S., TOURNIER M. (1998) Lexicométrie textuelle, sens et stratégie discursive, actes *I Simposio Internacional de Análisis del Discurso*, Madrid.

NOVAKOVA I., GOOSSENS V., MELNIKOVA I. (2012) Associations sémantiques et syntaxiques spécifiques. Sur l'exemple du lexique émotionnel des champs de surprise et de déception, *Actes du 3e Congrès Mondial de Linguistique Française*, Volume 1, pp. 1017-1029.

NIVRE, J., BOGUSLAVSKY, I. M., IOMDIN, L. L. (2008) "Parsing the SYNTAGRUS Treebank of Russian", *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008), Manchester, August 2008, p. 641–648.

SINCLAIR, JOHN MCH. (2004) *Trust the text : language, corpus and discourse*, London, Routledge.

TAPANAINEN, P., JÄRVINEN, T. (1997) "A non-projective dependency parser", In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, p. 64-74.

Résumé court

Nous présentons une méthode permettant l'extraction automatique des unités polylexicales caractérisant le profil collocationnel d'un pivot donné. A partir d'un corpus de textes journalistiques et littéraires d'environ 120 millions de mots, analysés syntaxiquement grâce au parseur Connexor, nous montrons comment constituer un réseau d'expressions hiérarchisées autour d'un pivot occupant une fonction syntaxique déterminé. Les unités extraites sont ensuite présentées de manière à faciliter leur exploitation lexicographique : affichage de l'arbre syntaxique, contextes d'occurrence, regroupement des expressions en fonction de leur réalisations morphosyntaxiques et indication du paradigme étendu auquel appartient le pivot dans le contexte de ses collocations. Nous chercherons enfin à relier ces extractions automatiques avec les codages syntaxiques et sémantiques du champ lexical des émotions réalisés par nos collègues du projet Emolex.