French collocations of cross-disciplinary scientific lexicon
WG1

Agnès Tutin, Thi Thu Hoai Tran, Olivier Kraif, Sylvain Hatier
Université de Grenoble Alpes, Laboratoire LIDILEM
e-mail: agnes.tutin@u-grenoble3.fr; Thi-Thu-Hoai.Tran@e.u-grenoble3.fr;olivier.kraif@u-grenoble3.fr; sylvain.hatier@u-grenoble3.fr>

This paper presents an ongoing project of lexical resources of French cross-disciplinary scientific lexicon in the framework of the Termith ANR project (http://www.atilf.fr/ressources/termith/), whose aim is to enhance the automatic indexing process of human science texts. This resource will be freely made available[1]. The cross-disciplinary scientific lexicon deals with words which cannot be used as terms but refer to methods, arguments, opinions and metatext, for example *hypothesis, that is why, surprisingly …* (for English, see Kosem 2010; Granger & Paquot 2010). In scientific texts, several kinds of cross-disciplinary scientific MWEs can be observed (Tutin 2014), among which we can mention:

- **Full phrasemes** (Mel'cuk 1998), i.e. non compositional MWEs, such as *prendre en compte* ('to take into account') or *point de vue* ('point of view').
- **Collocations**. We posit, in line with Explanatory and Combinatorial Lexicology (Mel'čuk, ibid), that these two different kinds of MWEs should be differentiated. As collocations are compositional, each component will receive a semantic tag and be stored within the database entry (e.g. s.v. *hypothèse* for *réfuter une hypothèse* ('to disprove a hypothesis') or *hypothèse valide* ('valid hypothesis'). Fully frozen MWE expressions (e.g. *mettre en évidence* 'highlight') will be stored as single units.
- **Semantico-rhetorical formulae**, which are typical (compositional) formulae in scientific genre associated with a specific rhetorical function, such as *as previously said*, *contrary to our expectations …*

In this paper, we focus on collocations of cross-disciplinary scientific lexicon. Collocation extraction is based on a varied set of 500 scientific articles in 10 disciplines of human sciences (Tran 2014). The 5 million word corpus has been parsed with the XIP syntactic parser and collocations have been extracted with the help of the Lexicoscope (Kraif & Diwersy 2014), a corpus tool for treebanks which includes a collocation extractor based on syntactic relations and association measures (see also Evert 2008; Seretan 2011). An element of collocations belongs to the cross-scientific scientific lexicon of single words, built by Hatier (2013) and colleagues.

For collocation extraction, the following criteria are used:

- One of the elements of the collocations belongs to the list of cross-disciplinary single words, that is, single words which are specific to the academic genre.
- Several statistical thresholds are used: frequency over 7 occurrences, log-likelihood ratio > 10.7 and a dispersion in at least 3 disciplines.

Several syntactic relations are used for collocations, using syntactic alternations such as passive:

- N prep N : *hypothèse de travail* (work hypothesis), *formulation d'une hypothèse* (formulation of hypothesis)

[1] Preliminary resources are already available on : http://scientext.msh-alpes.fr/scientext-site/spip.php?article39

- N Adj : *hypothèse valide* (valid hypothesis)
- N V : *les résultats confirment …* (results confirm…)
- V (prep) N : *faire une hypothèse* (to make a hypothesis…
- V ADV : *assumer pleinement* (fully assume)
- Adv Adj : *totalement absent* (totally absent, *significativement différent* (significantly different)

Extracted collocations are typical of the academic genre, but being cross-disciplinary, they also can be used to a less extent in other genres. Once extracted, the collocations are disambiguated and tagged with semantic labels. Syntactic distributional properties of collocations (determiners, position of adjectives …) will then be automatically extracted from our corpus. For example, for the collocation *faire-hypothèse* we observe the following syntactic properties:

Collocation		*Faire-hypothèse*	Examples
Determiners		l’ (91%), des (8%), une (0,5%),	
Syntactic alternations	Direct objet	92%	*Nous avons **fait** l’**hypothèse***
	Passive	2%	*Ainsi , aucune **hypothèse** n' est **faite** sur le caractère ...*
	Reduced passive	5%	*les conséquences des **hypothèses** générales **faites** sur l' illocutoire*
External subcategorization	Que-P	(with l’) : 58%	*On peut donc **faire** également **l'hypothèse que** c'est au cours de la Belle Époque que s'opère la transition*
	Prep-de	(with l’) : 12%	*On peut **faire l'hypothèse d'**une forte homogénéité des sujets ...*

These lexico-syntactic properties are very useful for implementation for automatic indexing, but also for language teaching for academic purposes.

References

Evert, Stefan (2008). “Corpora and collocations”, in: Anke Lüdeling/Merja Kytö (eds.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton/de Gruyter, 1212–1248.

Gledhill C. (2000), *Collocations in science writing*, Language in performance, 22, Tübingen, Gunter.

Granger, S., Paquot, M., (2010. The Louvain EAP Dictionary (LEAD) », *Proceedings of the XIV EURALEX International Congress* , Leeuwarden (The Netherlands), 6-10 July 2010, pp. 321-326.

Hatier, S. (2013), Extraction des mots simples du lexique scientifique transdisciplinaire dans les écrits de sciences humaines : une première expérimentation. Dans *Actes de Recital'2013* (p. 138–149). Les Sables d'Olonne, France.

Kosem, I. (2010). Designing a model for a corpus-driven dictionary of Academic English. PhD thesis. Aston University, Birmingham, UK.

Kraif, O., & Diwersy, S. (2014). Exploring combinatorial profiles using lexicograms on a parsed corpus: a case study in the lexical field of emotions. In P. Blumenthal, I. Novakova, & D. Siepmann (eds.), *Actes Du Colloque International Nouvelles Perspectives En Sémantique Lexicale Et En Organisation Du Discours*. Osnabrück, Allemagne: Peter Lang.

Mel’čuk, I. (1998). Collocations and Lexical Functions. In A. P. Cowie (ed), *Phraseology. Theory, Analysis and Applications,* 23-53. Oxford : Clarendon Press.

Seretan, Violeta. *Syntax-based collocation extraction*. Vol. 44. Springer Science & Business Media, 2011.

Tran, T. T. H. (2014). Les séquences lexicalisées à fonction discursive comme outil d’aide à l’écriture auprès des étudiants étrangers. Dans P. Blumenthal, I. Novakova, & D. Siepmann (éD.), *Actes Du Colloque International Nouvelles Perspectives En Sémantique Lexicale Et En Organisation Du Discours*. Osnabrück, Allemagne: Peter Lang.

Tutin, A. (2014). La phraséologie transdisciplinaire des écrits scientifiques : des collocations aux routines sémantico-rhétoriques. Dans A. Tutin & F. Grossmann (éD.), *L’écrit Scientifique : Du Lexique Au Discours. Autour De Scientext* (p. 27-44). Rennes: Presses Universitaires de Rennes.