Automatic extraction of transdisciplinary scientific collocations in General Scientific Language

Patrick Drouin
Sophie Reid Triantafyllos

Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
Montréal, Canada
patrick.drouin@umontreal.ca

Recent advances in corpus linguistics and the availability of large specialized corpora in electronic format have allowed researchers interested in LSP to explore new techniques (linguistic, statistical, etc.) for handling their primary source of information. Corpus-based studies lead us to identify recurrent lexico-grammatical patterns that relate to what has been described in linguistics as "collocations". In this paper, we observe the characteristics of General Scientific Languague (GSL), focusing solely on the automatic extraction of transdisciplinary scientific collocations in this type of discourse.

One of the main postulates behind this paper concerns the concept of GSL itself. We believe, like other researchers (Phal 1971, Williams 1999, Pecman 2004), in the *a priori* existence of a common language of science. The definition put forward by Pecman (2004:148) builds on the concept of discourse community as proposed by Swales (1990): "*Pratique langagière spécifique à une communauté de discours composée de chercheurs en sciences exactes dont les objectifs communicatifs poursuivis émanent des préoccupations partagées par des scientifiques à travers le monde et indépendamment de leurs spécificités disciplinaires.*" Although we agree with this definition in general, we disagree that the corpus should limit itself to "exact sciences" and that it should include scientific productions. In this sense, our perception of GLS is somehow closer to the academic language of Coxhead (2002).

Lexical studies of GSL are not uncommon or new, early work has been done for French by Phal (1971) and lately, Coxhead (2002) worked on the lexicon of English. But the main focus of our study is collocations, which we define a linguistic expression made up of at least two lexical units: the *base* of the collocation which is "freely" chosen and the *collocate* which is chosen in a (partially) arbitrary way to express a given meaning and/or a grammatical structure contingent upon the choice of the base (Kahane & Polguère 2001:8).

The study of collocations in specialized corpora has gained momentum and multiple studies have been published in the recent years (Gréciano 1997, Williams 1999, Gledhill 2000, Luzon Marco 2000, Tutin *forthcoming*, Pecman 2004, L'Homme & Bertrand 2000). In the current study, we will focus on the automatic retrieval of VERB-NOUN collocations (*reject*/*refute hypothesis*) in GSL (Tutin *forthcoming*) and thus we will not address their description from a lexicographical (Tutin 2005) or terminological (L'Homme & Bertrand 2000) perspective or their use for second language production (Pecman 2004) although we believe our study could contribute to such work.

Most approaches to collocation extraction include a measure of association strength between two lexical items done using various statistical measures. From this point of view, our study does not stand out as we will use such measures to identify word pairs to be further analyzed. In order to measure the association strength between elements of a word pair, we use the *Ngram Statistics Package* (NSP) published by Ted Pedersen (Pedersen & Banerjee 2003).

We believe that a sole measure of association strength is not enough and that there is a need to evaluate to which degree the collocation retrieved is characteristic of GSL. One way of doing such measurement is to compare the behavior (from a frequency point of view) of a collocation in a GSL

corpus and in a non-GSL corpus. A statistical measure, called *specificity*, proposed by Lafon (1980), lends itself quite well to such a comparison.

Such a technique implies the use of two corpora. For this paper, we will use a 30 million words news corpus built from the newspaper *Le Monde* and a 4 million word GSL corpus put together from Ph. D. thesis taken from the Web. So as to maximize frequency measurements, both corpora will be tagged and lemmatized using a part-of-speech tagger (*TreeTagger*, Schmid 1994).

As we aimed to retrieve collocations that are used in GSL as a whole (independently of subject matter), we will also pay close attention to the importance of frequency distribution. Raw frequency distribution will be used as well as tf-idf (term frequency–inverse document frequency) which is a weight measurement often used in information retrieval and text mining. It is a statistical measure used to evaluate the importance of a word for sub-corpus in a corpus. The importance increases proportionally to the number of times a word appears in the document but it is offset by the frequency of the word in the corpus.

In our talk, we will describe our view of GSL, discuss the steps involved in compiling our corpora, present a quick overview of the tools and measures used and give the results of the evaluation of the collocations retrieved using the suggested techniques.

Coxhead, A. (2002). "The Academic Word List: A Corpus-based Word List for Academic Purposes, dans Language and Computers", *Proceedings of the Fourth International Conference on Teaching and Language Corpora*, p. 73-89.

Gledhill, C. (2000). "The discourse function of collocation in research article introductions", *English for Specific Purposes*, 19, p. 115-135.

Gréciano, G. (1997). « Collocations rythmologiques », *Meta*, (42)1, p. 33-44

Kahane, S. and A. Polguère (2001). "Formal foundation of lexical functions », *Actes du colloque COLLOCATION: Computational Extraction, Analysis and Exploitation*, p. 8-15.

L'Homme, M.C. and C. Bertrand (2000). "Specialized Lexical Combinations: Should they be Described as Collocations or in Terms of Selectional Restrictions", *Proceedings. Ninth Euralex International Congress*, p. 497-506

Lafon, P. (1980). « Sur la variabilité de la fréquence des formes dans un corpus », *MOTS*, 1, p. 128-165.

Luzón Marco, M. J. (2000). "Collocational frameworks in medical research papers: a genre-based study", *English for Specific Purposes*, 19(1), p. 63-86 .

PECMAN, Mojca (2004). « Exploitation de la phraséologie scientifique pour les besoins de l'apprentissage des langues », *Journée d'étude de l'ATALA, Traitement Automatique des Langues et Apprentissage des Langues*, p. 145-154.

Pedersen T. et S. Banerjee (2003). "The Design, Implementation, and Use of the Ngram Statistics Package", *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 12 p.

Phal A. (1971). *Vocabulaire général d'orientation scientifique (V.G.O.S.)*, Didier, Paris, 128 p.

Schmid, H. (1994) "Probabilistic part-of-speech tagging using decision trees", *Proceedings of the International Conference on New Methods in Language Processing*, p. 44-49.

Swales, J. M. (1990) *Genre Analysis: English in academic and research settings*, Cambridge University Press, Cambridge, 272 p.

Tutin, A. (*forthcoming*). *Modélisation linguistique et annotation des collocations : application au lexique transdisciplinaire des écrits scientifiques*, 18 p.

Tutin Agnès (2005), "Annotating Lexical Functions in Corpora: Showing Collocations in Context", *Proceedings of Second International Conference on Meaning-Text Theory*, 7 p.

Williams, G. C. (1999). *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*, PhD thesis, Université de Nantes, 337 p.