From a Bilingual Transdisciplinary Scientific Lexicon to Bilingual Transdisciplinary Scientific Collocations

Patrick Drouin
Observatoire de linguistique Sens-Texte, Université de Montréal, Canada

With this paper, our main goal is to contribute to the existing research focusing on the study of the transdisciplinary lexicon in scientific writings. This lexicon includes abstract verbs and abstract nouns as well as a methodological lexicon that refers to the abstract lexicon used for the description of scientific activities and scientific reasoning. Our study involves two languages, French and English. The main goal of this paper is to test the idea that we can start from a raw bilingual scientific corpus and automatically build a list of bilingual transdisciplinary scientific collocations around the items from the transdisciplinary lexicon.

1. Introduction

Most linguistic studies dealing with the lexicon of scientific corpora are interested in subject area lexicon or terminology which leads to a general lack of description of the other types of lexical items contained in these corpora. The main exception to the previous statement is the work being done in the area of specialized language teaching like the studies of Coxhead (1998, 2000). In most cases, as pointed out by Tutin (2007), the lexicon itself is not what is being studied.

With this paper, our main goal is to contribute to the existing research focusing on the study of the transdisciplinary lexicon in scientific writings. Our study involves two languages, French and English. Prior work has been done on both languages in order to establish a transdisciplinary scientific lexicon. In most cases, researchers based their studies on earlier work aimed at the description of the general language lexicon. Usually, the goal of such work is to develop applications for second language teaching and learning.

For the French language, André Phal (1971) developed the *Vocabulaire général d'orientation scientifique*, also known as VGOS. This lexicon builds on what was done by Gougenheim *et al.* (1956) called *Français Fondamental*. For his research in the VGOS, Phal focused on what is known as hard sciences: mathematics, physics, chemistry and natural sciences. As far as the English language is concerned, Averil Coxhead (1998, 2000) has done impressive work on the *Academic Word List* (AWL) which contributes to the work done on the general language lexicon such as the *Basic English* by Ogden (1930) and the *General Service List* established by West (1953). In order to come up with the AWL, Coxhead (1998, 2000) used a corpus of texts in the domains of arts, commerce, law and sciences, thus taking into account subject areas outside of what is usually known as 'hard sciences'.

We consider that the lexicon used in scientific writings can be divided into three categories. The first one is the common basic lexicon, which includes function words such as determiners, auxiliary verbs and conjunctions, and content words of the general language. The second category is the transdisciplinary lexicon that includes abstract verbs such as *to think* or *to consider* and abstract nouns such as *idea*, *factor* and *relation*. It also includes a methodological lexicon that refers to the abstract lexicon used for the description of scientific activities and scientific reasoning. Examples of lexical items one would find in this category are *hypothesis*, *data* and *approach*. The last category of lexicon found in scientific writings is subject specific terminology, which refers to all concepts used in a particular domain.

Our study here is focused on the second category, which we called Transdisciplinary Scientific Lexicon (TSL), and its behavior in scientific writings. The main goal of this paper is to test the idea that we can start from a raw bilingual scientific corpus and automatically build a list of bilingual transdisciplinary scientific collocations around the lexical items from the second category described above (TSL).

2. From Scientific Corpora to Scientific Transdisciplinary Lexicon

2.1. Corpora description

In order to gather the transdisciplinary lexicon, we use freely available natural language processing tools and statistical techniques. Our identification method relies on two main ideas, the *distribution* of the lexical items in all scientific documents and the *specificity* of the TSL to scientific documents.

The techniques that we use for the current experiments rely on the comparison of two corpora: a transdisciplinary corpus (TC) and a reference corpus (RC). The first one is the corpus being mined for transdisciplinary vocabulary and it is a specialized corpus built from PhD theses and scientific papers. Thus, it is at the moment somehow genre specific. This corpus is an open corpus and we might include more documents of various genres as time passes.

Disciplines	Papers FR	Papers EN	Thesis FR	Thesis EN
Anthropology	233,699	226,470	254,956	259,264
Chemistry	213,239	206,616	191,034	224,647
Computer Science	207,445	210,649	247,612	238,250
Engineering	238,868	224,504	145,252	199,606
Geography	227,715	227,887	220,653	245,391
History	245,014	222,889	320,267	222,241
Law	234,784	238,867	374,830	242,857
Physics	214,546	215,145	197,867	196,559
Psychology	245,292	242,847	360,473	222,070
TOTAL	2,060,602	2,015,874	2,312,944	2,050,885

Table 1. Number of words per domain and per language in the TC

We rely on comparable transdisciplinary corpora totaling approximately 4 million words in both French and in English. All subject areas are evenly represented (about 200,000 words) and scientific papers and theses account for half of the corpora in both languages. Although the data is not exactly the same in English and French, we were able to come up with a balanced bilingual corpus. All documents included were published between 1997 and 2007. As for the work done by Coxhead (1998 and 2000), we do not restrict the subject areas selected to the usual boundaries of 'hard sciences'. The idea behind this decision is to provide a list of lexical items that would be more closely related to science as a textual genre than science as a list of domains.

In order to evaluate the specificity of lexical units for our specialized corpus, we use a second corpus as a point of reference (the reference corpus). The French reference corpus was built from 30 million words taken from articles published in 2002 in the newspaper *Le Monde*. As far as English is concerned, we used parts of the British National Corpus (BNC). In order to come up with a corpus comparable to the one we used in French, we divided the BNC into genres using David Lee's classification (2001). Lee has divided the BNC into 46 genres and 8 super genres. For our experiment, we decided to consider only the texts that belonged to super genres *broadsheet national newspapers, regional and local newspapers* and *non-academic prose (non-fiction)* as they allowed us to gather a reference corpus that was quite similar to *Le Monde* both in size and in genres. We are aware that using a genre specific corpus as a point

of reference might not be the best scenario; but the limited availability of balanced French corpus similar to the BNC is a limiting factor that researchers in the field of natural language processing must face on a daily basis. We believe that using comparable corpora for both languages is a sound decision from a methodological point of view, as we want to obtain comparable results.

2.2. Corpora processing

The BNC documents were stripped from any tagging in order to start from raw text. Once again, the idea behind this first processing step is to apply the same methodology and the same tools for both languages. All further preprocessing of documents was performed using freely available tools.

The first step in preparing the documents is handled by *TreeTagger* (Schmid 2004), a part-of-speech (POS) tagger. The tool goes through the document and assigns a part-of-speech and a lemma to words. In order to be able to establish the real frequency of words contained in our corpus, we decided to simplify the tagging done by *TreeTagger* by discarding the actual form found in the corpora and keeping solely the lemma and the POS tag. Using such a simplification allows us to establish frequencies on lemmas instead of inflected forms. For example, we can now compute a single frequency for the verb *être_VER* instead of its various forms: *suis_VER*, *sont_VER*, *sommes_VER*, etc. Such a step is crucial for a statistical processing like the one described in the next paragraphs.

2.3. Corpus specificity

Corpus specificity is evaluated using a measure proposed by Lafon (1980) and called *calcul des spécificités*. It makes possible the comparison of the frequency of a word in a corpus (here our transdisciplinary corpora) to the frequency of the same word in another corpus (our reference corpora).

Corpus	**RC**	**TC**	**Total**
Frequency of word	*a*	*b*	*a+b*
Frequency of other words	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*N=a+b+c+d*

Table 2. Contingency table used to describe frequencies in corpora

The actual calculation is performed using the following formula (Lafon 1980):

$$\log P(X=b) = \log (a+b)! + \log (N-(a+b))! + \log (b+d)! + \log (N-(b+d))! - \log N! - \log b! - \log ((a+b)-b)! - \log ((b+d)-b)! - \log (N-(a+b)-(b+d)+b)!$$

This technique identifies three types of words based on their frequency according to a standard normal distribution: positive, negative or neutral specificities. This distribution is evaluated based on the observations made in the reference corpus and is used to compute a theoretical frequency. Simply put, the specificity scores represents the distance between the theoretical frequency and the frequency found in the TC.

The positive specificities have a frequency which is higher that could be expected in the transdisciplinary corpus. The negative specificities have a frequency which is lower that expected while items from the last group have a frequency in normal range. In our study we are solely interested in positive specificities. The rationale behind this decision is that we believe a significantly high frequency in the theses corpora means that a word is highly characteristic of such a corpus.

2.4. Corpus distribution

The second criteria used for the TSL extraction is that retained words must be distributed throughout the corpus. There are at least two ways to look at distribution when dealing with a corpus like the one we used for our experiment. The first approach would be to create sub-corpora based on subject area. From there, we look at the relative frequency (instead of the absolute frequency) of words in these sub-corpora, which might have different sizes. The other method would consist of making sure that the sub-corpora have the same size (computed in words) and then comparing the frequencies in the different parts of the corpus. In this way, we can simply compare the raw frequency of words across the corpus without taking into account the size of the sub-corpora. For our experiment, since we want the results to be completely independent of subject areas and since we want to keep the methodology simple, we decided to go forward with the first method. The statistical test used can also take into account the variation in size between sub-corpora.

Since words can be highly specific to our transdisciplinary corpora and still be linked directly to one of the 9 subject areas (in other words, they can be terms), we want to make sure that words retained as potential TSL units are distributed in our TC. In order to be included in our list, a word both needs to appear in all domain specific sub-corpora and to have a high-specificity level. The following table presents the breakdown of the data selected by the specificity and the distribution criteria.

Parts of speech	English	French
Adjective	381	338
Adverbs	170	135
Nouns	551	611
Verbs	172	684
Total	1274	1768

Table 3. Breakdown of the data selected

Although some discrepancies are observed in the amount of data selected for both languages, the distribution of data, with the exception of the number of French verbs retained, is very similar. The discrepancy lies with the different tagging schemes used by the TreeTagger for English and French and our processing of the subsets produced by the tagger. The tagset used for French uses a finer grain description of the verbs and adds information about the tense to the lemma, which is not represented in the same way in the English tagset. We strongly believe that if we took this difference into account in a further processing of the corpora, this difference would be gone.

A sample output of TSL can be seen in Table 4. We included here the top 25 units for each language. It is interesting to notice that the items in bold are common to both list. This intersection is quite important, even if the list is very short. We have yet to complete the validation and the description of the lexical items retained in both language, but we believe that a fair amount of data is shared overall. All items are being looked at carefully in context and for each of them we provide one or more definition. The resulting lexicographical descriptions will be made available freely in XML format on the Web in the near future.

Rank	French	English
1	*type*	***model***
2	***modèle***	*analysis*
3	***fonction***	***function***
4	***phase***	***phase***
5	*objet*	***system***
6	*paramètre*	***structure***
7	***contexte***	*method*
8	*donnée*	*state*
9	*valeur*	*design*
10	*élément*	***interaction***
11	***structure***	***research***
12	*cas*	***surface***
13	*profil*	*order*
14	***effet***	*theory*
15	*siècle*	*process*
16	*section*	*component*
17	***interaction***	*type*
18	*forme*	***context***
19	***système***	*example*
20	*pointe*	*energy*
21	*figure*	*relation*
22	*ensemble*	*behavior*
23	***surface***	*domain*
24	***étude***	***effect***
25	*utilisation*	*approach*

Table 4. Top 25 lexical units selected

3. From Scientific Transdisciplinary Lexicon to Scientific Transdisciplinary Collocations

The list of TSL items described in the previous section is used as the starting point for collocation retrieval. For all items in the list, we extract from our corpora a list of statistically significant collocations. The extraction was performed using *Text-NSP* (Pedersen and Banerjee 2003), a set of tools dedicated to the extraction of n-grams (string entities of length *n*, e.g. character or word sequences) from a corpus. The package provides various means for the statistical analysis of n-gram occurrences. For the current study, we are solely interested in bigrams and the output of the tool will be limited to pairs of lexical items.

We extracted collocations based on a window of 3 words with a minimal frequency of 5 and a empirical threshold score of 6 using the log-likelihood test. This last test is one of the several measures proposed by NSP-Text to evaluate the strength of association between 2 words. In our study, we will focus on *VERB – NOUN* collocations identified for the nouns contained in the TSL for each language.

For each noun in our list, we select verbs that collocate significantly with the noun and build a list as shown in Table 5. The noun is considered to be the base of the collocation and the verbs, the collocates.

For the current experiment, all statistically significant collocations were kept and used since we want to automate the process of retrieving a set of Scientific Transdisciplinary Collocations automatically. For the actual lexicographical project under way, the list of collocations will be filtered out in order to distinguish between items like *perform – analysis* and *use – analysis* that have very different quality levels. The methodology also does not distinguish between contexts where the noun is subject of the verb and where it is used as an object of the verb.

Base	Collocates	Base	Collocates
modèle		*analysis*	
	adapter		*become*
	modèle		*conduct*
	adapter		*define*
	alimenter		*determine*
	apporter		*embed*
	appuyer		*employ*
	baser		*find*
	choisir		*follow*
	construire		*ground*
	...		*include*
	définir		*indicate*
	développer		*perform*
	élaborer		*present*
	éloigner		*provide*
	ériger		*require*
	...		*reveal*
	inspirer		*show*
	intégrer		*strip*
	...		*suggest*
	reprendre		*use*
	reproduire		
	représenter		
	servir		
	simplifier		
	suivre		
	sélectionner		
	tester		
	utiliser		
	valider		

Table 5. Sample output of collocates for *modèle* and *analysis*

4. From Scientific Transdisciplinary Collocations to Bilingual Transdisciplinary Scientific Collocations

4.1. Generating a bilingual lexicon

Establishing equivalence automatically from language is a challenging task; some could even say that it is an impossible task, and we tend to agree. In order to reach the goal we have set for ourselves with this study, we have two ways of approaching the problems: establish equivalence using 1) an existing bilingual resource or 2) using some brute force technique.

Since good and complete bilingual resources are hard to find (Fontenelle 1997) and we do not have access to one in electronic format, we explored a simple solution based on the idea of cognates (Simard *et al.* 1993) used on a regular basis to align bilingual corpora. We implemented an algorithm based on the work of Oliver (1993) that compares two input strings and gives in output a percentage of similarity.

The table illustrates examples of nouns in English and French that had a similarity of more than 90% for the TSL in both languages. Since we used a high similarity threshold, most of the pairs proposed by the algorithm are valid. Of course, using a lower threshold or comparing short strings will lead way to errors. In such a case, using a bilingual dictionary would be a better solution. Mining freely available dictionaries such as Wiktionary[1] or lexical resources

[1] http://en.wiktionary.org

such as WordNet and WOLF (WordNet Libre du Français)[2] could also provide a quick source for a bilingual dictionary.

EN	FR	Score
phase	*phase*	100.00
system	*system*	100.00
type	*type*	100.00
comparison	*comparaison*	95.24
transfer	*transfert*	94.12
principle	*principe*	94.12
experimentation	*expérimentation*	93.33
change	*échange*	92.31
construction	*reconstruction*	92.31
composition	*décomposition*	91.67
presentation	*présentation*	91.67
organization	*organisation*	91.67
effect	*effet*	90.91
phrase	*phase*	90.91
information	*formation*	90.00

Table 6. List of strings in potential cognates in both languages

4.2. Generating a bilingual graph to visualize the results

Starting with these pairs, we established the equivalence between the base of the collocations in both languages and linked the data. By using this technique, we were able to generate graphs that can allow us to visually represent the collocations in both languages. The data contained in Table 6 is converted to a GraphML[3] format that can be read by a number of graph visualization tools. Figure 1 and 2 were generated using yEd[4]. The following gives an example of a graph generated from the list of bigrams. The example is given using the DOT language[5], as it is simpler than GraphML. The first line of the graph is use to establish equivalence between the bases.

```
digraph graph_analysis {
        analysis -> analyse;
      analysis -> perform;
      analysis -> complete;
        analysis -> include;
        analysis -> conducts;
      analyse -> compléter;
      analyse -> poursuivre;
}
```

[2] http://alpage.inria.fr/~sagot/wolf.html

[3] http://graphml.graphdrawing.org/

[4] http://www.yworks.com

[5] http://www.graphviz.org/doc/info/lang.html

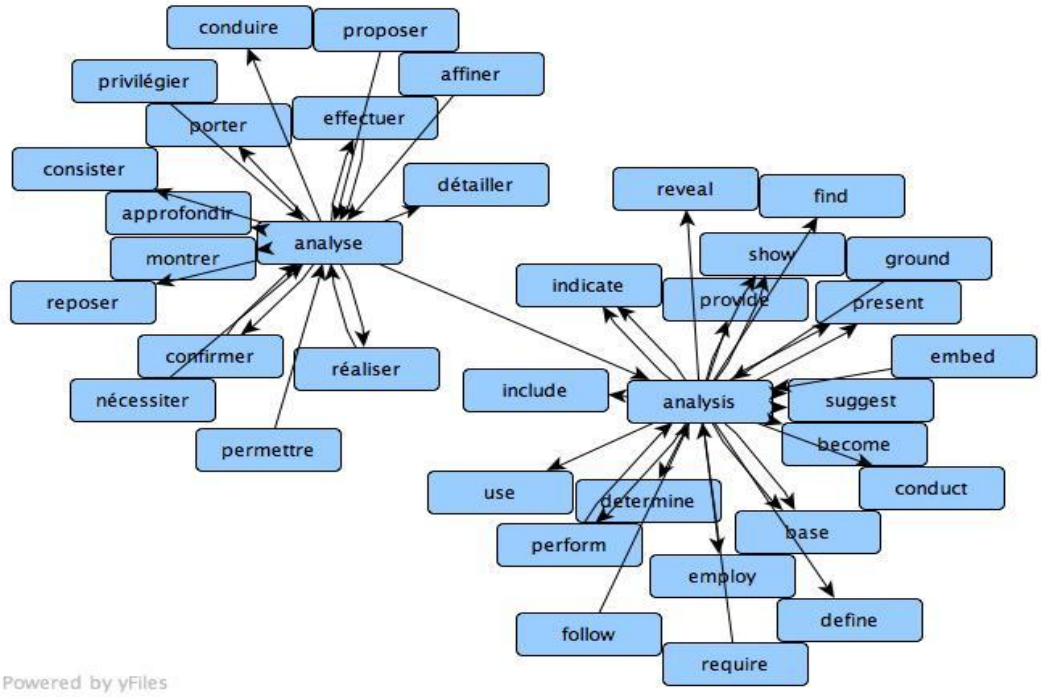

Figure 1. Graph representing the collocates of *analysis* and *analyse*

The first example, built automatically, shows that we can link the pair *analysis* – *analyse*. A experienced lexicographer can quickly see that having access to such graphs allows to then establish bilingual pairs such as *(analysis, suggest)* – *(analyse, proposer)* or *(analysis, conduct)* – *(analyse, effectuer)*. Of course, from this point, we could push the idea of automatically linking the data further and having access to a bilingual dictionary could have allowed us to link both elements of the pairs from the start (*proposer* – *suggest*, *effectuer* – *conduct*, *montrer* – *indicate*). On the other hand, at this point, we would much rather leave this decision process to the lexicographer.

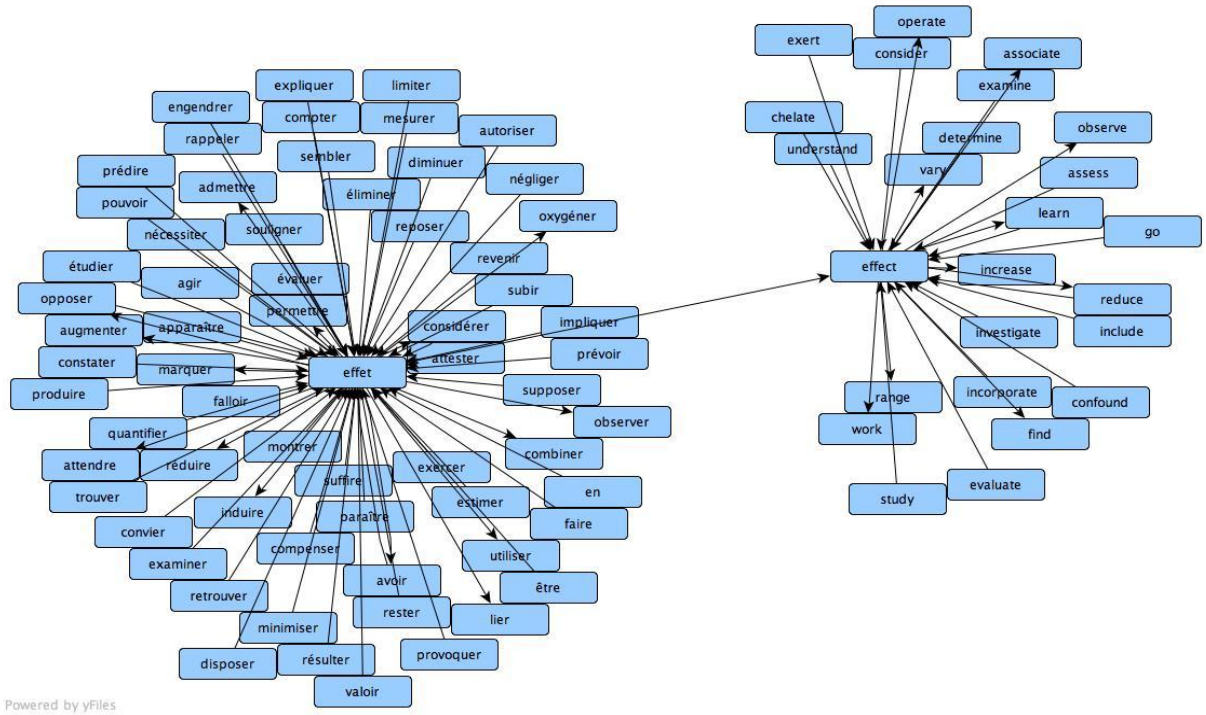

Figure 2. Graph representing the collocates of *effect* and *effet*

Although the figure build around the pair *effect – effet* is not as well balanced as the previous one, the same type of valuable information could be used to draft dictionary entries is present. It is quite obvious from this figure that both words behave quite differently in the corpora; various explanations could be found for such difference.

Our collocations being extracted at the form level do not allow us to distinguish between multiple meanings of *effet* and *effect* that would lead us to different subsets of collocates. In an ideal world, NLP tools should build sets like {$effet_1$; $collocate_1$, $collocate_2$, $collocate_3$, ... } {$effet_2$; $collocate_1$, $collocate_5$, $collocate_7$, ... } where $effet_1$ and $effet_2$ point to two different meanings. From such results, one could see intersection between collocates of the different bases and also unique contributions in each subsets. Another explanation for the discrepancies between languages in figure 2 could be that both forms behave completely differently in the two languages. For example, *effet* is often found in the French corpus in the phrase *en effet*, a phenomenon that could very well increase the frequency of the form and allow is to collocate with different verbs.

5. Conclusion

We presented a technique that allows to built automatically a set of bilingual Transdisciplinary Scientific Collocations starting from raw scientific corpora of thesis and papers in English and French. In order to do so, we first identified a Transdisciplinary Scientific Lexicon (TLS) in each languages and extracted collocations around a subset of the TSL (limited to nouns).

Between these subsets in both languages, we automatically established equivalence between languages and generated graphs representing the collocations when linked from one language to another through the base of the collocations. Our initial results lead us to think that such graphs could be very useful for various purposes, the most obvious being specialized lexicography. It is our opinion that the process could benefit from using existing bilingual dictionaries or from direct human input to establish equivalence so as to improve the quality and the coverage of the results obtained.

References

Coxhead, A. (1998). *An academic word list*. Wellington: Victoria University of Wellington.

Coxhead, A. (2000). 'A New Academic Word List'. In *TESOL Quarterly* 34 (2). 213 - 238.

Drouin, Patrick (2003). 'Term extraction using non-technical corpora as a point of leverage'. In *Terminology* 9 (1). 99-117.

Fontenelle, T. (1997). 'Using a bilingual dictionary to create semantic networks'. In *International Journal of Lexicography* 10 (4). 275-303.

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

Gougenheim G., Michea R., Rivenc, P.; Sauvageot, A. (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier.

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

Lee, D. (2001). 'Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle'. In *Language Learning & Technology* 5 (3). 37-72.

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

Ogden, C. K. (1930). *Basic English: a general introduction with rules and grammar*. London: Kegan Paul, Trench, Trubner.

Oliver, J. (1993). 'Decision graphs - an extension of decision trees'. In *Proceedings of the 4th International Conference on Artificial Intelligence and Statistics*. 343-350.

Pecman, M. (2004). 'Exploitation de la phraséologie scientifique pour les besoins de l'apprentissage des langues',. In *Actes des Journée d'étude de l'ATALA, Traitement Automatique des Langues et Apprentissage des Langues*. 145-154.

Pedersen T.; Banerjee, S. (2003). 'The Design, Implementation, and Use of the Ngram Statistics Package' In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*.

Phal A. (1971). *Vocabulaire général d'orientation scientifique (V.G.O.S.) - Part du lexique commun dans l'expression scientifique*. Paris: Didier.

Rayson, P.; Garside, R. (2000). 'Comparing Copora Using Frequeny Profiling'. In *Proceedings of the Workshop on Comparing Corpora, 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000*). 1-6.

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

Simard, M., Foster, G.; Isabelle, P. (1993). 'Using cognates to align sentences in bilingual corpora'. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing*. 1071-1082.

Stuart K.; Botellea Trelis, A. (2006). 'Collocation and knowledge production in an academic discourse community'. In *Proceedings of the 5th International AELFE Conference*. 238-245

Tutin, A. (*to appear*). 'Modélisation linguistique et annotation des collocations : application au lexique transdisciplinaire des écrits scientifiques.

Tutin, A. (2007). 'Présentation du numéro Autour du lexique et de la phraséologie des écrits scientifiques'. In *Revue française de linguistique appliquée* 12 (2). 5-13.

Weir, G.; Anagnostou, N. (2007). 'Exploring Newspapers: A Case Study in Corpus Analysis'. In *Proceedings of ICTATLL 2007*.

West, M. (1953). *A General Service List of English Words*. London: Longman.