French Cross-disciplinary Scientific Lexicon: Extraction and Linguistic Analysis

Sylvain Hatier, Magdalena Augustyn, Thi Thu Hoai Tran, Rui Yan, Agnès Tutin, Marie-Paule Jacques

LIDILEM – EA 609 – Université Grenoble Alpes

e-mail: {sylvain.hatier;magdalena.augustyn;thi-thu-hoai.tran;rui.yan;agnes.tutin;marie-paule.jacques}@univ-grenoble-alpes.fr

Abstract

This paper presents the work we carried out to extract and structure a specialized lexicon based on a corpus of French scientific articles in the fields of humanities and social sciences. The characteristics of the targeted lexicon may be summarized as follows: it is not domain-related inasmuch as it is shared by various disciplines; it serves to express the specific operations, naming the objects and exposing the results of research processes. In this view, the targeted lexicon studies here is genre-related. We designed this cross-disciplinary scientific lexicon (CSL) as a resource for several purposes: it may serve natural language processing, e.g. as a stoplist for automatic terms identification, as well as foreign language teaching. Indeed, students and scholars in the sciences need to acquire familiarity with the rhetoric of the research article, thus needing to master these words. We present here the two-stage creation of this lexicon: first, it was semi-automatically extracted from a corpus of 500 research articles spanning ten disciplines. Second, it was manually structured to reflect the semantics and rhetoric of science. This structure takes into account the lexico-syntactic properties of CSL nouns, adjectives, verbs and adverbs. The resource will be freely available for academic purposes.
Keywords: corpus linguistics; scientific writing; open lexical resources; natural language processing

1. Cross-disciplinary Scientific Lexicon: Definition and Purposes

In this paper, we present a lexical resource called “Cross-disciplinary Scientific Lexicon of French” (CSL). This lexicon is structured in semantic classes and subclasses and includes 1,768 entries with the following parts of speech: nouns, adjectives, verbs and adverbs. This online resource will be freely available. Multiword expressions including collocations, fixed expressions and rhetorical routines will be included.

The lexicon was automatically extracted from a large cross-disciplinary corpus in the Humanities and Social Sciences, tagged at the syntactic level with a dependency parser and compared to a large corpus of various French texts. As regards the semantic analysis, our semantic labelling is based on manual and automatic distributional analysis and syntagmatic relations.

This cross-disciplinary lexicon project has several purposes:

- As part of the TermITH[1]project, this lexicon is designed to improve the information extraction and especially automatic indexing using CSL to facilitate term identification (Jacquey et al., 2013)
- It will also be used for linguistic and epistemological studies, for example by identifying semantico-rhetorical routines incorporating CSL elements, for example when defining new terms (e.g. Jacques, 2011)
- It will also be very useful for learning and teaching activities, especially for learners of French as a Foreign Language. Several activities designed to help scientific writing and understanding have already been developed by our team (Hatier & Yan, 2015; Tran, 2014; Hartwell & Jacques, 2012).

Following Tutin (2007a), we define the CSL as a lexicon referring to the scientific procedures and the scientific discourse related to objects, essential in the argumentation and structuring of scientific discourse (Drouin, 2007; Paquot & Bestgen, 2009).

The CSL has three main properties:

- **Cross-disciplinarity**: It is genre-specific and not discipline-specific. For example, it does not include the terminology of linguistics or economics but words and expressions such as *hypothesis, that is why, analyse.*
- **Specificity**: It is specific to scientific writing in comparison to other genres, such as literary or newspapers ones.
- **Frequency**: It includes frequent words and expressions of scientific discourse.

Words and expressions belonging to CSL are underlined in the following extract from an anthropology article. We find single words such as *déplacement* ('transfer', 'shift'), *paradigme* ('paradigm'), *définition* ('definition') and multi-word expressions such as *point de vue* ('point of view') or *mettre en jeu* ('bring into play', 'involve').

Du point de vue anthropologique ce déplacement de paradigme met en jeu la définition même de l'identité[2]

('From the anthropological point of view, this paradigm shift involves the very definition of identity').

Several studies focusing on this specific lexicon, especially for academic purposes, have already been developed for scientific English (Coxhead, 2002; Paquot, 2010) or for French, from textbooks (Phal & Beis, 1972). Pecman (2004) has studied the cross-disciplinary phraseology from a bilingual hard science corpus in the perspective of scientific writing. Drouin (2007) bases his study on a diverse French corpus of theses and a journalistic corpus to refine the description of this lexicon. To our knowledge, few studies have been conducted which specifically address the CSL in the Humanities and Social Sciences, in which procedures and methods differ from those of the experimental

[1] TermITH (Terminology and texts indexation in Human Sciences): ANR-12-CORD-0029 CONTINT. http://www.atilf.fr/ressources/termith (last accessed on 25 April 2016).

[2] Dubey Gérard, «Nouvelles techniques d'identification, nouveaux pouvoirs. Le cas de la biométrie», *Cahiers internationaux de sociologie* 2/2008 (n° 125), p. 263-279.

sciences, as reflected in the CSL. Furthermore, writing in the Humanities is an essential part of scientific activity and is often difficult to master for native and non-native students.

In the following sections, we present the method of semi-automatic extraction of CSL units. We then explain how we processed to enrich our lexicon with semantic and syntactic properties.

2. The Extraction of Cross-disciplinary Scientific Lexicon (CSL)

The CSL is extracted in two stages: an automatic extraction, and at the second stage, manual processing, involving several linguists.

2.1 Automatic identification criteria

Following Coxhead (2002), Drouin (2007), Paquot (2010), we first extracted an initial word-list based on statistical criteria from a 5-million word corpus of research articles.

The 5-million word research article corpus (RAC), from the Scientext[3] project, consists of 500 peer-reviewed research articles from ten disciplines of Humanities and Social Sciences. A diverse 120-million word reference corpus (RC) was also created, covering three genres: literature, newspapers and spoken language (including transcriptions and subtitles).

Extraction criteria were derived from the CSL linguistic properties mentioned above. Since this lexicon is specific to scientific writing, we expected that it would be more frequent in the RAC compared to RC and this overrepresentation of lexical units can be measured through several statistical tests (log-likelihood ratio, khi², log-odds, ratio).

We determined that the CSL units should meet the following criteria:

- **Global overrepresentation** (Drouin, 2007) in comparison with the diversified RC. The statistical measure adopted was also the simplest one: the ratio[4] (words are proportionally more frequent in RAC than in the reference corpus).
- **Overrepresentation** in at least four disciplines over ten to ensure that lexical units are actually cross-disciplinary.
- **Distribution**: RAC has been divided into 100 equal sections, a word should occur in at least 40/100 sections.
- **Occurrences as a single-word unit**: for example, whenever *point* was included in a multi-word expressions such as *point de vue* ('point of view'), it was not counted.

These criteria were applied to our RAC resulting in a list of 1,976 potential CSL items: 513 adjectives, 213 adverbs, 786 nouns and 464 verbs.

2.2 Manual processing

Although some lexical units positively met the criteria of specificity and distribution, they do not actually belong to CSL, notably lexical units referring mainly to the objects

[3] ANR Project "Corpora and research tools in humanities and social sciences" (2007-2010). Website: http://scientext.msh-alpes.fr. (last accessed October 15, 2015)

[4] We compared different measures, in a previous study (Hatier, 2013), and concluded that ratio, log-likelihood ratio and χ^2 produce equivalent results for CSL extraction.

of study in Humanities and Social Sciences, including *ouvrier* ('worker'), *école* ('school'), *ménage* ('household'), etc. Actually, there are two reasons for this noise in the extraction process. Firstly, these lexical units, as part of the papers' topics, are over-represented in the academic discourse in comparison with other genres, and therefore validate the specificity criteria. Secondly, because they refer to objects of study shared by various disciplines, they met the second criteria of distribution and consequently shared similar statistic properties with CSL. This lexicon is a characteristic of Humanities and Social Sciences corpora, and may not have been extracted if we had included articles in the field of natural sciences in our corpus. These two lexicons, lexicon of humanities and social sciences objects and CSL, are therefore complex to be automatically differentiated.

Because of this noise, we then proceeded with a manual validation by expert linguists who assessed CSL membership. Their judgement was based on concordances from various disciplines and lexico-syntactic information extracted from the corpus syntactically analysed with XIP (Aït-Mokhtar, Chanod, & Roux, 2002). These elements (for example the most statistically significant noun subjects for a given verb) are displayed in order to illustrate the verb meaning.

In fact, the analysis of the word meaning cannot be performed out of context. Thus, concordances and statistically significant collocates are intended to facilitate word recontextualisation for expert linguists. Further, this phase allowed us to identify specific word uses, such as passive and pronominal verbal constructions. For example, the verb *situer* ('to locate') almost exclusively occurs in our corpus in the passive and pronominal constructions, as shown in the following examples:

Les valeurs obtenues se situent bien en dessous des valeurs généralement admises
('Reported values are far below generally allowed values').

In this experiment, expert linguists allocated, for each word, one of the three following labels: CSL use, General Abstract Lexicon (GAL) use and other use. The GAL (*année*, ('year') *changer* ('change'), *fin* ('end)') includes general abstract words, present in genres other than academic writing but over-represented in our corpus. Contrary to CSL, it does not refer to scientific reasoning or scientific activities.

We then calculated inter-rater agreement for this task, whose scores are displayed in the table below.

	Cross Disciplinary Lexicon use	**General Abstract Lexicon use**	**Other use**
Nouns	0.45 (83 %)	0.251 (64 %)	0.428 (72 %)
Verbs	0.234 (63 %)	0.081 (55 %)	0.337 (82 %)
Adjectives	0.755 (90 %)	0.615 (83 %)	0.746 (88 %)

Table 2: Inter-rater agreement

The first score is the Fleiss' kappa, the second, in round brackets, is the pairwise average. We observed that agreements for CSL use and GAL use were the lowest scores. Indeed, differentiation between those two lexicons is complex. Actually, GAL, as a non-terminological and frequency lexicon, has the same utility as CSL for term extraction

techniques and pedagogical applications. We then opted for integrating words labelled as CSL or GAL by experts into the CSL category.
At the end of the process, our CSL word list contained 1,311 items: 274 adjectives, 202 adverbs, 493 nouns and 342 verbs.

3. Linguistic Analysis of Cross-disciplinary Scientific Lexicon

Since a list of bare words was insufficient for the intended applications, we performed a semantic and syntactic analysis based on a syntactically analysed corpus and Natural Language Processing-based distributional analysis.

3.1 Structure of the lexicon

The CSL resource is based on word meaning as several lexical entries are polysemous (see below). For example, the word *développement* ('development') can refer to a section in a text or to a process. Each word meaning is described (e.g. *développement*-1) in Table 1 according to the following fields: lemma, part of speech, semantic class, semantic subclass, and definition/gloss. Examples and frequency information are also provided in the resource.

Meaning Identifier	Lemma	Part of speech	Class	Subclass	Definition / gloss
développement-1	*développement* ('development')	noun	{*communication*} ('communication')	{*document*} ('document')	*Exposé* ('report')
développement-2	*développement* ('development')	noun	{*processus_évolutif*} ('progressive process')	{*augmentation* ('increase')	*Croissance* ('growth')
développer-2	*développer* ('to develop')	verb	{*processus_évolutif*} ('progressive process')	{*augmentation* ('increase')	*Donner de l'extension* 'to expand'
strict	*strict* ('strict')	adjecti ve	{*modalité*} ('modality')	{*restriction*} ('restriction')	*Limité* ('limited')
strictement	*strictement* ('strictly')	adverb	{*modalité*} ('modality')	{*restriction*} ('restriction')	*Rigoureusement* ('strictly')

Table 3: Sample of CSL entries

This semantic classification allows one to the limit of the lemma level to reach a conceptual level of abstraction, necessary to consider more complex processing. For example, some semantic classes and subclasses are found across multiple parts of speech: the noun *développement-2* and the verb *développer* belong to the same semantic class {progressive process}; the noun *strict* and the adverb *strictement* also share the same semantic class and subclasses.

3.2 Meaning identification

The first semantic processing consisted in identifying cross-disciplinary meanings actually present in the Research Article Corpus and this for each lexical unit of the CSL,

since some CSL elements can be polysemous, as seen above with *développement* ('part of text' or 'increase').

In order to list CSL element meanings in RAC, we referred to the *Dictionnaire Électronique des Mots* (DEM) (Dubois & Dubois-Charlier, 2010) and *Les Verbes Français* (LVF) (Dubois & Dubois-Charlier, 1997) created under the same principle of lexico-syntactic classes and which provided the semantic elements (definition, domain, synonyms) together with syntactic elements for each lexical entry. To our knowledge, these are the only freely available and large-scale resources for French. We identified meanings in RAC using concordances displayed by the Lexicoscope (Kraif & Diwersy, 2012), which is a query tool similar to Sketch Engine (Kilgarriff *et al.*, 2014).

For a word meaning to be considered as cross-disciplinary, we used two thresholds: a minimum of 20 occurrences and a minimum of 5/10 disciplines. If necessary, we manually added a new meaning in our resource when absent from the DEM and the LVF. As a result, our resource includes 1,768 CSL entries, 318 adjectives, 215 adverbs, 537 nouns and 698 verbs, amounting to a polysemy rate of 1.3 meaning per word.

3.3 Creation of semantic classes

The classification relies on a two-level typology, organized in semantic classes and subclasses. These classes are elaborated on the basis of lexico-syntactic properties, similar to Flaux and Van De Velde (2000) and Dubois and Dubois-Charlier (1997) studies, following the observation and analysis of combinatorial profiles [5] of CSL elements in RAC. We present below a lexicogram sample, which is a set of lexico-syntactic relations defining a word profile.

Lexicogramme | Graphiques

Show 25 entries — Search:

l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log.likel
article_NOUN	ce_DET	DETERM	735	10606	38321	7556564	10	2538,6444	1
article_NOUN	dans_PREP	PREPOBJ	383	10606	39680	7556564	10	835,3611	2
article_NOUN	This_NOUN	~NMOD	75	10606	478	7556564	10	571,5860	3
article_NOUN	publier_VERB	~U3_DEEPOBJ NMOD ~OBJ ~VMOD ~SUBJ ~DEEPOBJ ~U3_DE_VMOD	94	10606	2307	7556564	8	456,1285	4
article_NOUN	présent_ADJ	U3_ADJMOD	56	10606	1456	7556564	10	265,1835	5
article_NOUN	intituler_VERB	~U3_DEEPOBJ NMOD	33	10606	404	7556564	6	206,1274	6
article_NOUN	proposer_VERB	~SUBJ ~VMOD ~U3_A_VMOD ~OBJ	70	10606	8236	7556564	10	135,9957	7

Figure 1: Lexicogram sample for the noun *article*

[5] Combinatorial profiles, obtained through the Lexicoscope, are, according to Blumenthal (2008), «*l'image que donne du comportement d'un mot de base l'ensemble de ses collocatifs*» ('the reflection of a word behaviour given by all of its collocatives')

Lexicograms allowed us to observe significant collocates for a set of words, for example the verb *publier* ('publish') is very frequent with *article* ('article') as a direct object. In doing so, we identified CSL words that share similar collocates and therefore semantic properties. These collocates were then incorporated into lexico-syntactic tests used to check semantic class membership.

The DEM and the LVF dictionaries were also used to ensure the semantic consistency of the classification. We assumed that elements belonging to a same class/subclass:

- are co-hyponyms and meet the class definition, e.g.*: {communication/document}* subclass elements are defined as follows: 'N is a document in an act of communication'
- share statistically significant lexico-syntactic properties, *e.g.:* {*communication/document*} subclass elements – *article* ('article'), *texte* ('text'), *document* ('document') – frequently co-occur with demonstrative pronoun *ce* ('this')
- positively respond to lexico-syntactic tests derived from those lexico-syntactic properties, *e.g.:* {*communication/document*} subclass elements validate the following test: *Nous présentons dans ce N* ('We present in this N').

In order to define classes and subclasses, we drew upon several studies including Tutin (2007b), GermaNet (Hamp, Feldweg, *et al*, 1997) and Tran (2014).

Semantic processing for the four categories were performed concomitantly and collaboratively so as to ensure cross-category homogeneity. Thus, some classes such as {*modalité/certitude*} ('modality/certainty') cover all of studied categories, while other classes, such as {*discursif*} [6] are category-specific. Throughout its elaboration, the resource was collectively evaluated.

3.4 Experimenting automatic classification

The resource is not conceived as static. Therefore, in order to facilitate the maintenance while reducing its cost, we conducted two experiments involving automatic classification methods. Both methods were useful to enhance our semantic classification by basing it on lexico-syntactic similarities automatically computed from the analysed corpus.

The first experiment consisted in a semi-automatic classification based on prototype theory (Rosch, 1973). A set of semantic classes (e.g. {*communication support*}) together with their corresponding prototypes (e.g. *chapitre* ('chapter'), *figure* ('figure'), *travail* ('work')), i.e. the most representative class elements, were manually defined. The lexico-syntactic properties shared by the prototypes (e.g. subject of the verb *to show*) were computed to create a class profile. A measure taking into account the lexico-syntactic characteristics of any new word to add can then be computed to help identify the word class. This method mainly allowed us to highlight the relevant definitional

[6] The semantic class {discursive} elements, adverbs as *pourtant*, ('however') *potentiellement* ('potentially'), *ci-dessous* ('below'), have an argumentative function or are used to organize text structure.

features for each class.
The second experiment relied upon the formal concept analysis method named *Galois lattices* (Bendaoud, Toussaint, & Napoli, 2010). With this method, we obtained a representation of the class hierarchy and an identification of definitional properties through the class intent. For example, we observed that elements of the semantic class {*déterminant*} ({'determiner'}), such as *groupe* ('cluster'), *type* ('kind'), *ensemble* ('set') share the following lexico-syntactic properties: subject of the verbs *apparaître* ('appear') and *constituer* ('constitute', 'form'), object of the verb *considérer* ('consider').

Those two experiments helped us identify the typical properties of the semantic classes. For example, we extracted several lexico-syntactic features for the semantic class {*communication support*}. Lexical units of this class can be:

- the subject of the following verbs: *présenter* ('to present'), *proposer* ('to offer'), *décrire* ('to describe')
- in a syntactic relation with the following adjectives: *nouveau* ('new'), *majeur* ('major'), *premier* ('first'), *dernier* ('last')
- the object of the following verbs: *s'intituler* ('to entitle'), *consacrer* ('to devote'), *évoquer* ('to mention')
- in syntactic relation with demonstrative pronouns: *ce travail* ('this work'), *ce chapitre* ('that chapter').

In addition, these experiments showed how complex a consistent classification is because of polysemy which entails non-homogeneous clusters.
The example of *selon* illustrates this phenomenon. *Selon* can refer to two main meanings: *according to* and *based on*. By observing the set of its significant collocates, we can identify two syntagmatic disjointed sets:

- {*conception, auteur, définition*} ('concept', 'author', 'definition') co-occur with *selon* in its meaning "according to"
    - ex : ***Selon*** *cette conception, le premier pilier repose sur le régime général […]* ('According to this concept, the first pillar relies on the general system')
- {*lieu, valeur, état*} ('place, value, state') cooccur with *selon* in its meaning "based on"
    - ex : *Il est, en conséquence, équivalent de déterminer un salaire* ***selon*** *le lieu de travail…* ('In consequence, it is equivalent to determine a salary based on the work place…').

Automatic classification results are thus compromised by the polysemy of lexical units as well as by the polysemy of the features of those lexical units.

4. From Lexicon to Patterns

In addition to semantic properties, we added syntactic properties to our resource for a subset of CSL verbs. We collected and modelled lexico-syntactic patterns for a sample of semantically related verbs. We aimed at extracting specific structures assuming that

the verbs semantic proximity will also be revealed in similar syntactic properties. Inspired by Hanks' CPA model (2013), a systematic approach to the syntagmatic and semantic description of verbs, we associated verb meanings with patterns, which fulfil important rhetorical functions (expressing one's opinion, cause and effect, reviewing the literature, etc.) and thus can be useful for NLP applications. For most CSL verbs, patterns are rarely ambiguous and are limited in number (at most two patterns are identified for most verbs).

Furthermore, from a pedagogical point of view, verb constructions are fundamental to master scientific writing since "insufficient knowledge of verbs [is] a serious handicap for learners" (Granger & Paquot, 2009: 2). A comparison between the Scientext corpus, a Chinese learner corpus and a novice native writer corpus (Hatier & Yan, 2015), has demonstrated that these two groups significantly underused the verb patterns expressed in examples such as *considérons ces facteurs* ('let us consider these factors'), *cela s'explique par* ('this can be explained by'). We therefore decided to include these verb patterns in our resource.

First, we automatically extracted all syntactic relations involving the verbs for each of their occurrences in order to build subcategorization frames. Second, the frames were manually grouped into patterns, that is, lexico-syntactic configurations including semantic properties, especially the collocations.

The lexico-syntactic configurations along with the main collocations found in the predicate-argument structure are shown in Table 4 for the verb *expliquer* ('explain') (1635 occurrences), including its two meanings '*faire comprendre*' ('make someone understand') and '*constituer la raison de*' ('be the reason of').

Meaning	Semantic class/subclass	Pattern		
		SUBJ	Verb	OBJ/Complement
Faire comprendre ('make someone understand')	{analyse_information} / {interprétation} ('analysis_information') / ('interpretation')	*il, elle, on, ,nous, je* ('he, she, I/we')	expliquer ('explain')	complétive *que* / *pourquoi* / *comment* ('that/why/how clause')
		il, elle, on, nous, je ('he,she,/we/I')	expliquer ('explain')	difference ('difference')
Constituer la raison de ('be the reason of')	{relation} / {implication_cause} ('relation') / ('involvement/reason')	*ceci, cela, facteur, faiblesse, différence, raison* ('this, that, factor, weakness, difference, reason')	expliquer ('explain')	*différence, phénomène, choix, situation, résultat, problème* ('difference, phenomenon, choice, situation, result, problem')
		ceci, cela, faiblesse, phénomène, résultat ('this, that, weakness, phenomenon, result')	(s') expliquer (par) ('be explained by')	*fait, facteur, différence, effet* ('fact, factor, difference, effect')

Table 4: Example of lexico-syntactic patterns for the verb *expliquer* ('explain')

Following CPA, for example, the patterns sharing the meaning '*constituer la raison de*' can be represented as follow:

1. [[objet scientifique|relation|qualité=cause]] explique [[événement=conséquence]]
   implicature: [[objet scientifique|relation|qualité=cause]] constitue la raison principale de [[événement=conséquence]]
2. [[événement=conséquence]] s'explique par [[objet scientifique|relation|qualité=cause]]
   implicature: [[événement=conséquence]] peut être justifié par [[objet scientifique|relation|qualité=cause]][7]

Thus, the relevant collocations of a verb are expressed as a semantic type in double square brackets. The attribution of semantic types is based on our CSL noun classification and thus depends on corpus data. The implicature consists of a definition of the verb pattern.

A verb pattern consists of a set of syntactic relations with information about meaning, arguments, frequency and preferential collocate semantic classes. Class and subclass features are used to enrich patterns. Examples from corpora are also added in order to illustrate the pattern use. In prospect of developing pedagogical applications, we would like to exploit the use of these verb patterns for teachers of French as a foreign language so that they can help non-native learners in their academic writing.

Furthermore, the semantic properties, together with syntactic relations of verb pattern enabled us to consider automatic extraction of semantico-rhetorical routines (see Tutin & Kraif, 2016), which are valuable access to observe and study intertextual and thematic relations in text (Legallois, 2012).

5. Conclusion

The Cross-disciplinary Scientific Lexicon list was designed with the help of corpus linguistics techniques, using syntactically analysed corpora, including a scientific corpus and a general language corpus. The treebanks were used in order to facilitate the extraction of CSL and the semantic treatment, with distributional semantics techniques, but also to extract semi-automatically grammatical patterns, in a CPA-like approach. The semantic labelling of words was not fully automatized, due to the widespread polysemy of CSL, and expert linguists were requested throughout the semantic process.

The CSL is likely to be a useful resource for applied linguistics, especially for building pedagogical activities for French as a foreign language in the academic background, but also for native French students, in order to observe typical grammatical constructions or specific semantic fields. Its semantic and syntactic properties also enable NLP applications such as information retrieval, but also linguistic and epistemological studies.

[7] [[scientific object|relation|property=reason]] explain [[event=consequence]]
implicature : [[scientific object|relation|property=reason]] can be the main reason of [[event=consequence]]
[[event=consequence]] can be explained by [[scientific object|relation|property=reason]]
implicature: [[event=consequence]] can be justified by [[scientific object|relation|property=reason]]

In the very near future, the CSL will be available in an online database[8], with many corpus-based examples. We are also currently expanding the resource with multiword expressions, including collocations, full phrasemes and routines.

References

Aït-Mokhtar, S., Chanod, J.-P., & Roux, C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering, 8*(2-3), 121–144.

Bendaoud, R., Toussaint, Y., & Napoli, A. (2010). L'analyse Formelle de Concepts au service de la construction et l'enrichissement d'une ontologie. *Revue des nouvelles technologies de l'information*, 133–164.

Coxhead, A. (2002). The academic word list: A corpus-based word list for academic purposes. *Language and Computers, 42*(1), 73–89.

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée, Vol. XII*(2), 45-64.

Dubois, J., & Dubois-Charlier, F. (1997). *Les verbes français*. Larousse.

Dubois, J., & Dubois-Charlier, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages, (3)*, 31–56.

Flaux, N., & Van de Velde, D. (2000). *Les noms en français: esquisse de classement*. Editions Ophrys.

Hamp, B., Feldweg, H., & others. (1997). Germanet-a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (p. 9–15). Citeseer.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Mit Press.

Hartwell, L. & Jacques, M. P. (2012). A corpus-informed text reconstruction resource for learning about the language of scientific abstracts. *Proceedings of the International EuroCALL Conference. CALL: using, learning, knowing*. University of Gothenburg (Sweden).

Hatier, S. (2013). Extraction des mots simples du lexique scientifique transdisciplinaire dans les écrits de sciences humaines : une première expérimentation. In *Actes de la 15e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2013)* (p. 138–149). Les Sables d'Olonne, France.

Hatier, S., & Yan, R. (2015). Comparaison de constructions verbales entre un corpus d'apprenants et un corpus d'articles de recherche. *8es Journées Internationales de Linguistique de Corpus (JLC2015)*, Orléans.

Jacques, M.-P. (2011). Nous appelons X cet Y : X est-il un terme émergent ? In K. Kageura & P. Zweigenbaum (éd.), *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence* (p. 31–37). Paris, France: INALCO.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*(1), 7–36.

[8] http://lidilem.u-grenoble3.fr/ressources/corpus-du-labo/

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

Kraif, O., & Diwersy, S. (2012). Le Lexicoscope: un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. In *19e conférence TALN*.

Legallois, D. (2012). La colligation: autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique?. *Corpus*, (11).

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. Bloomsbury Publishing.

Paquot, M., & Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. *Language and Computers, 68*(1), 247 269.

Pecman, M. (2004). *Phraséologie contrastive anglais-français: analyse et traitement en vue de l'aide à la rédaction scientifique* (Thèse doctorat). Université de Nice-Sophia Antipolis. UFR Lettres, arts et sciences humaines, France.

Phal, A., & Beis, L. (1972). *Vocabulaire général d'orientation scientifique, VGOS: part du lexique commun dans l'expression scientifique*. Crédif.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Tutin, A. (Éd.). (2007a). Lexique des écrits scientifiques. *Revue française de linguistique appliquée, XII*(2).

Tutin, A. (2007b). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. In *Actes de la 14ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*. Toulouse, France.

Tutin, A., & Kraif, O. (Forthcoming). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents. *Lidil* 53.

Tran, T-T-H. (2014). *Description de la phraséologie transdisciplinaire scientifique et réflexions didactiques pour l'enseignement à des étudiants non-natifs. Application aux marqueurs discursifs* (Thèse de doctorat). Université de Grenoble, Grenoble.

Acknowledgements

We thank the Région Rhône-Alpes for supporting our work by a grant. We would also like to thank Evelyne Jacquey and Laurence Kister for their great help throughout extraction and manual validation, and Laura Hartwell for her helpful review. This research was also supported by the French Agence Nationale de la Recherche under grant ANR-12-CORD-0029 (TermITH project).