ARTES: an online lexical database for research and teaching in specialized translation and communication

Mojca Pecman, Natalie Kübler
CLILLAC-ARP EA 3967
Université Paris Diderot – Paris 7
5 rue Thomas Mann
75205 Paris Cedex 13
PRES Paris Sorbonne Cité
E-mail: mpecman@eila.univ-paris-diderot.fr, nkubler@ eila.univ-paris-diderot.fr

Abstract

This paper presents the ARTES database (Aide à la Redaction de TExtes Scientifiques / Dictionary-assisted writing tool for scientific communication) and the underlying approaches to lexicography, terminology, languages for special purposes (LSPs), and lexical resources creation, behind the design of the database. This new type of lexical resource has been developed within the ARTES project, whose main objective is to explore the interaction between research and teaching in the areas of applied linguistics such as specialised translation and LSP communication. As a multilingual multidomain language resource targeting various LSP users – students, translators, experts, subject specialists, teachers, linguists –, the ARTES database offers a comprehensive approach to lexical resources: terminological, phraseological, domain-specific, domain-free, semasiological and ultimately onomasiological. The underlying research orientations are thus broad and allow to investigate various language mechanisms which operate on lexico-discursive level, and consequently to fine tune the database in order to take into account these various linguistic phenomena.

1. Introduction

The present paper is an overall study of the interactions between research and teaching in the domain of lexical resource creation that have been taking place within the ARTES project[1]. Launched in 2007 at Paris Diderot University - in the frame of the ESIDIS-ARTES scheme - the ARTES project was designed to bridge the gap between research and teaching in a number of related areas: terminology, phraseology, translation, LSPs, lexical resources, corpus linguistics, genre and discourse analysis. It enables us to tackle some core research problems related to the conceptual design of lexical resources which seek to integrate language phenomena currently observed through corpus analysis and on the linguistic levels of terminology, phraseology and discourse.

In 2010, a major tool was added to the project, an online database, opening up new avenues of research and facilitating the creation of lexical resources and development of dictionaries to meet the specific needs of speakers using LSPs: researchers, experts, translators, students, teachers. In addition to terminology, the database highlights phraseology, whether domain-specific or domain-free.

Although comparable to some extent to terminological databanks such as Termium[2], Grand Dictionnaire Terminologique[3] or Eurodicautom, now known as IATE[4], the ARTES database is closer to initiatives on lexical resources creation where teaching, research and database development are very closely related, such as DiCoInfo[5] database, and the latest DiCoEnviro[6] (L'Homme 2007) or WebTerm[7], a project of the Institute for Information Management in Cologne.

After these preliminary remarks, we shall first discuss the general approaches to specialised lexical resources adopted in ARTES database. The turning of the database into an online electronic dictionary by providing an interface for data access will be exemplified in the second part of the paper. In the last part, we evaluate the relevance of research conducted on terminology, phraseology and specialised discourse for a fine tuning of database architecture.

2. Creating lexical resources with ARTES database

The present section is an overview of the general scheme of the ARTES project and its goals, and of the ARTES database architecture designed to host LSP resources.

2.1 Presentation of ARTES project

ARTES is an ambitious and innovative project developed with the aim of bridging the gap between research and teaching in LSP translation and communication. It is carried out at the Paris Diderot University by a group of researchers working on LSP, Corpus Linguistics and Translation Studies: Kübler,

[1] ARTES project homepage: http://www.univ-paris-diderot.fr/artes

[2] The Government of Canada's terminology and linguistic databank: http://www.termiumplus.gc.ca

[3] Dictionary of the Office québécois de la langue française : http://www.granddictionnaire.com

[4] InterActive Terminology for Europe database: http://iate.europa.eu

[5] Dictionnaire fondamental de l'informatique et de l'internet : http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi

[6] Dictionnaire fondamental de l'environnement : http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi

[7] http://www.iim.fh-koeln.de/webterm/webtermsamm_e.htm

Pecman & Bordet 2011; Froeliger 2008; Humbley 2008; Kübler 2011; Kübler & Pecman forthcoming; Pecman 2005, 2008; Pecman *et al.* 2010; Volanschi et Kübler 2011. The target was to construct a model for collecting lexical resources in LSPs which would be flexible enough to allow developments in line with advances in research and changing learning needs. After testing several experimental models, we decided on SQL database technology with online applications for editing and retrieving data.

Several earlier developments paved the way for acquiring knowledge necessary for designing this online database, namely BasTet[8] designed in 2006 by Claudie Juilliard, Terminom1[9] developed first in 2004 by Kübler and Juilliard and redeveloped in 2007 by Kübler and Pecman, and the LangYeast[10] combinatory dictionary developed by Volanschi (2008).

The architecture of the ARTES database is inspired by BasTet which was first developed using Microsoft Access database management system. In ARTES, new functions were added based on a better understanding of General Scientific Language (GSL) and the processing of domain-free phraseology (Pecman 2004, 2007, 2008). The transfer of technology from Access DBMS to an online SQL database was entrusted to (e)Kudji company. The new online database is currently under development but the main stages of transfer were achieved by November 2010.

Although not an XML database, The ARTES database was developed in agreement with TBX and TMF standards for terminological databases. The structure of a terminological entry, the specific data categories and the relational disposition of data adopted in ARTES scheme are very close to meta model of ISO 16642 standards and guidelines for creating terminological data collections. Nevertheless, in the present state of the tool, the terminological data collection (TDC) scheme recommended by ISO 16642 for providing information on concepts of specific subject fields is not yet integrated. In the future developments of ARTES DB, we intend though to provide a preliminary conceptual level analysis where concepts would act as pivots ensuring linguistic transfer between different languages, and consequently an access to data through ontologies.

2.2 General architecture of ARTES database

The ARTES database is a relational database designed to contain lexical information spread over a number of tables. Some aspects of the ARTES database architecture were already discussed in Pecman *et al.* (2010), and Kübler & Pecman (forthcoming). There are some forty tables in the ARTES database, which can be divided into three subgroups according to the type of information stored in the tables: data tables, labelling tables, and relational tables.

Data tables are central tables in which all key information is stored. There are six data tables in ARTES containing respectively sources, terms, contexts, definitions, specific collocations, generic collocations, and notes. The table for sources serves to record bibliographical references of textual or oral sources used for referencing definitions, contexts or notes. In the table for terms, all terms are recorded no matter what language or domain. The specifications on the language and domain are given through descriptors provided by labelling tables. Context tables serve to record contexts taken from various sources and which serve as examples for illustrating the usage of terms or collocational phenomena. The tables for definitions contain all the definitions of the terms recorded in the database. For each term, all relevant existing definitions are provided and, if necessary, a new definition is drafted. The table of specific collocations is designed to record the most frequent collocations associated with terminology. In this way, two tables were designed to separate specific collocations related to terms from generic collocations related to discourse functions. The table of generic collocations is designed to list frequent word combinations used in a variety of LSP domains and which are related to discourse functions, the latter being recorded in a label-type table (see hereafter). Finally, the table for notes contains various observations on resources stored in data tables (e.g. an additional commentary on the meaning of the term as a complement to its definition, or an observation on the choice of terms indicated as synonyms).

Labelling tables are the tables with pre-defined values which describe and classify the resources stored in the data tables. The predefined values act as labels or descriptors. They allow us to adopt a descriptive approach to language data. Some labelling tables contain closed-class type values, such as the table of grammatical categories linked to the table of terms. Other labelling tables are open-class tables and can be modified or completed according to the results of research conducted in relation with language resource creation. For instance, the table of discourse functions which offers some eighty classes for categorizing generic collocations according to their general meaning or function in LSP discourses, is an open-class table.

The relational tables are necessary for establishing various links between data (between equivalent terms across languages, between equivalent pairs of collocations, between synonyms within a language, between a hyperonym and its hyponyms, and so on).

It should be mentioned that in the relational database, all tables are eventually linked together in one scheme which forms the architecture of the database: terms are linked to definitions, contexts and specific collocations, which are in turn linked to sources and notes. Furthermore, terms can be linked to other terms to indicate language equivalences or synonym pairs or sets,

[8] http://wall.eila.univ-paris-diderot.fr/bastet (restricted access)

[9] http://terminom1.eila.univ-paris-diderot.fr (restricted access)

[10] http://ytat2.ijm.univ-paris-diderot.fr/LangYeast

and similarly collocations can be linked to other collocations to indicate equivalent pairs or to form semantic synonymous sets, and so on.

3. Designing language resources for LSP communication and translation

In the ARTES project, the LSP communication and translation are tackled from the learner and professional perspective.

3.1 Taking into account teaching needs in LSP communication and translation

The ARTES database was designed to cater for teaching and learning needs in specialised translation at the department of Applied Languages of Paris Diderot University. The students in Master Studies in Specialised Translation and Language Engineering[11] are introduced in the theories, methods and applications of terminology, lexical resource creation, and corpus linguistics, with emphasis on corpus linguistic tools and information retrieval. A combination of these courses allows the students to develop skills and acquire knowledge crucial for achieving a high quality translation of LSP texts. The final result of the interaction of these various disciplines is presented in a form of Master's dissertation. The ARTES database is designed to allow students to participate in the project by creating LSP resources in relation with the text they translate. In turn, the database offers useful functions for teachers to help them follow students' work in progress and evaluate the resources compiled by students. Special effort was made to design the editing and management interface, commonly called the back-office, to anticipate these user situations.

Consequently, a very important feature of the ARTES DB project is that data is compiled mainly, but not exclusively, by LSP and translation learners. The overall methodology used to ensure the quality of data collected consists in three key procedures. The first one is the method itself followed by learners which is based on thorough comparable corpus analysis and an exchange with domain experts, which leads to a design of a domain ontology. The acquired knowledge on the domain, combined with the knowledge on terminology processing, allows the learners to select and process terms and relevant linguistic information adequately. The second procedure consists in the reviewing and validating resources by domain experts. An onging collaboration with experts in Earth and Planetary Sciences form a STEP department[12] and Institut de Physique du Globe de Paris (IPGP)[13] of Paris Diderot University, allows us to apply this procedure efficiently to a number of disciplines. The third and crucial stage in building language resources is the overall normalisation, correction and validation of resources by terminologists, which should be devised in the near future. We hope thus that the overall methodology will yield satisfactory results.

Students in Master's Studies are also invited to question the theoretical and methodological premises on which the description of language data in ARTES is based by testing them against “real life” translation problems.

The data recorded in the database can be retrieved via an online application specifically designed to take into account various LSP communication contexts.

3.2 Designing an online electronic dictionary for LSP users

The ARTES project had led to the design of an online terminological and phraseological database for storing and managing structure-rich information with the possibility for multiple criteria and multiple-level query. The database is searchable through a database application accessible online[14]. The access to data was devised with special care to targeted users: translators, teachers, students, domain experts and linguists. The intention behind the design of the interface for data access was to explore the possibilities for providing a dictionary-assisted writing tool for scientific communication. In the choice of the name for this application, the priority was given to target users which are largest in number: science students and experts who need to write articles or other text types in their second language within the scope of their discipline. It turns out that these are the most numerous users among students and researchers of Paris Diderot university which is a large multidisciplinary university hosting departments and research centres in Humanities, Sciences and Medicine.

The ARTES database has two interfaces: one for editing and management purposes and one for retrieving information and displaying it in the form of an LSP dictionary. The latter interface has been designed to display data recorded in the database functionally, following Leroyer's approach to functional lexicography according to which *development of a dictionary is determined by users needs and made to serve communication and knowledge-oriented functions in particular user situations* (Leroyer 2007: 110). The data disposition in the ARTES dictionary takes into account different users and user situations targeted by the tool - learners of terminology and translation studies, translators, learners in scientific fields, scientists, and linguists. Translators and translation learners need to find relevant information for translating concepts and phrases which they do not necessarily completely understand. On the other hand, scientists and science students need to find information that will help them formulate their ideas in the second language. Thus two opposing situations can be distinguished, leading to the necessity to make

[11] Master professionnel ILTS (Industrie de la langue et traduction spécialisée) : http://www.eila.univ-paris-diderot.fr/formations-pro/masterpro/ilts/index

[12] http://step.ipgp.fr

[13] http://www.ipgp.fr

[14] https://artes.eila.univ-paris-diderot.fr

ARTES both an encoding and a decoding dictionary. There is yet another function of the ARTES dictionary which enables the linguists and the teachers to navigate through data by specifying criteria for data selection. This function refers to a setting where a language specialist needs to retrieve data in order to construct useful material for teaching or research purposes. As shown in the Figure 1, there are three major accesses to data: through terminology, through phraseology, and by multiple criteria query.

The function "Terminology in context" allows the user to interrogate the database for terminology which is domain-dependant and to display useful information in relation with each term. The data is categorised according to different settings: a term from the point of view of its meaning, its usage or its translation. The following function, "Discourse phraseology", provides a template for navigating through phraseology which is domain-free, yet frequently used in LSP communications, and includes collocations, collocational frameworks, expressions and other types of phraseological units (for example: *to be described elsewhere by, to be in a poor agreement with, to provide evidence for, tremendous amount of, etc.*). The access to this sort of data is provided via semantico-discursive categories which were pre-identified thorough multi-domain corpus analysis of phraseological data (Pecman 2004, 2007). The last function, "Multiple query search", intended for linguists, allows the user to retrieve data by multiple criteria and thus construct useful material for teaching or research purposes.

4. Improving ARTES database through research in terminology, phraseology and discourse analysis

The ARTES project is being developed in close relation with research in a number of connected areas: terminology, phraseology, translation, LSPs, corpus linguistics, genre and discourse analysis. The results of studies conducted in these various fields are implemented in the database whose architecture reflects the advances in our knowledge on LSPs. In this way, the solutions adopted in the ARTES database are to large extent based on switching between theories, observed linguistic evidences, and target users' needs. In the other words, we proceed by examining various theoretical premises in translation and LSP oriented terminology management and then by applying or adapting them in agreement with the observed linguistic phenomena, all to serve adequately various LSP speakers' needs defined in the context of the ARTES project. The present section presents some of the major aspects of lexical resource creation with the ARTES database for improving language-related research and applications in areas such as information retrieval, terminology, phraseology or discourse analysis.

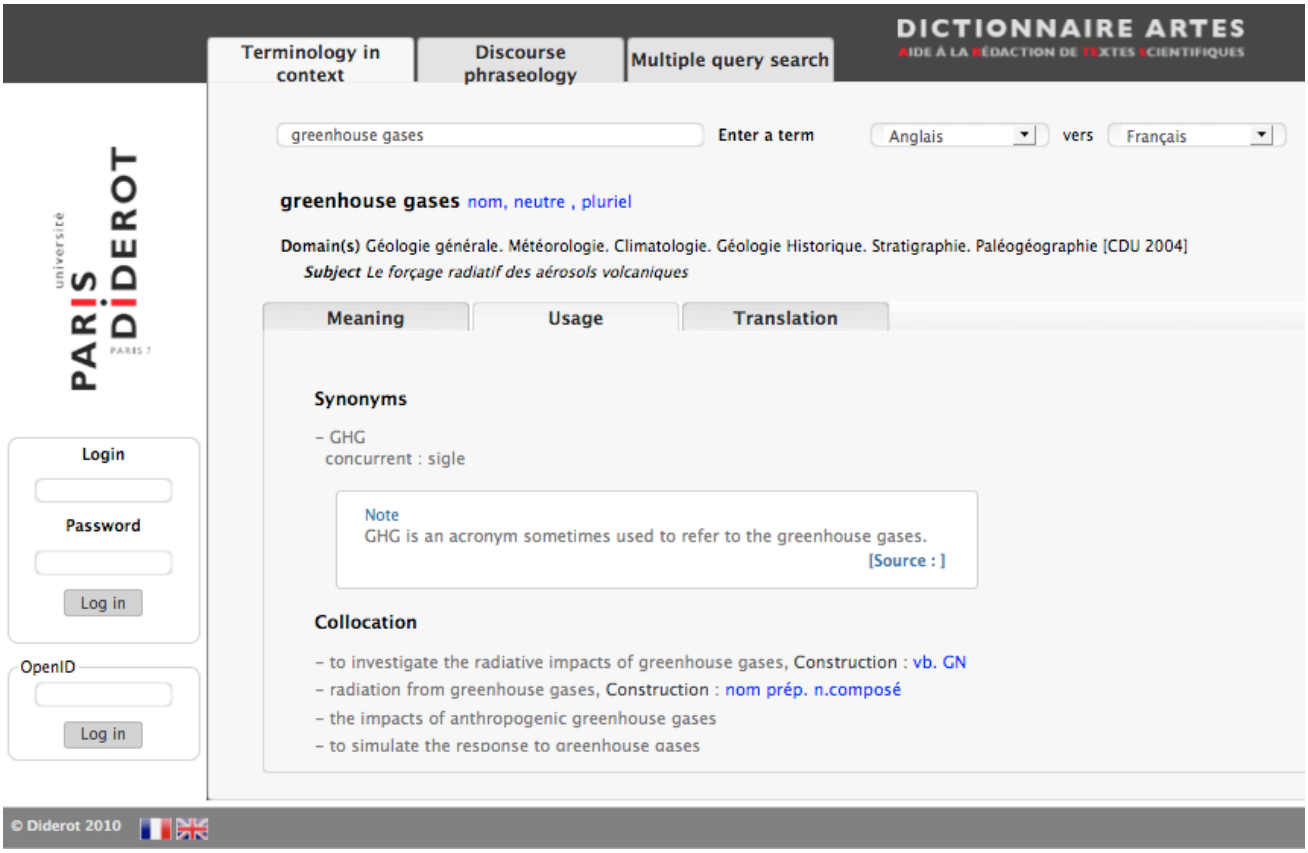

Figure 1: ARTES dictionary interface

4.1 Processing multidomain resources

Creation of lexical resources in a multidomain perspective raises the question of the organisation of lexical units according to different domains. It is well known that a term can have different meanings according to the different domains it may occur in. Assigning a term to a domain, or a series of domains, is often a complex task. Let us consider but one simple example: the word *water* which can be assigned to the domain of chemistry (where its definition could be "chemical compound consisting of two atoms of hydrogen and one atom of oxygen"), physics (where its definition could be "a liquid that changes its phase into ice at 0°C and into gas at 100°C"), or geology, climatology and meteorology (where its definition could be "the element of which seas, lakes, and rivers are composed, and which falls as rain and spouts from springs"), not to mention its role in the general language. The degree of precision of a domain specification is another difficulty we have to deal with. For example the term *fault*, defined as "a fracture in the Earth's crust that divides a geological area into two blocks which move relative to one another" can be assigned equally to the following domains: geology, seismology, plate tectonics, structural geology, geomorphology, endogenous geology, geophysics, and so on.

In the case of polysemy or in general, in an LSP database it is important to have an efficient system of descriptors for domain specifications. For the ARTES database, we have chosen to follow Universal Decimal Classification (UDC), which has systematic approach to classification, allows for exhaustiveness, and to choose a level of precision when specifying a domain. Consequently, in the ARTES database, a term can be easily assigned to one or more domains.

At this stage of the project, the resources are only being built and the domain coverage is not yet as large as the DB allows it. There are nevertheless some 23 000 terms, 25 000 collocations and 1 500 domains recorded in the DB. The collection of generic collocations is at initial stage and contains more then 100 entries.

4.2 Processing multilingual resources

So as not to be limited to a fixed number of languages when creating resources in LSPs, the architecture of the ARTES database was developed to allow for a multilingual approach. Each term can be assigned to one language specification. The table of language specifications contains some fifty languages. The pairs of equivalences can be established among any two terms or collocations of equivalent terms. For example, if *greenhouse gases* and *gaz à effet de serre* are defined as equivalents, it is then possible to align their respective collocations, e.g. *man-made green house gases* and *gaz à effet de serre anthropique, to reduce green house gases* and *réduire les gaz à effet de serre,* etc. It is thus possible to consider different types of units when working on the transfer of meaning from one language to another.

Nevertheless, establishing translational units is a very problematic matter, even in the exact sciences. In many cases the equivalences between terms are partial. We have thus added in the database a field for translation notes in order to indicate the contexts in which the equivalence is acceptable. For example, the concept of "rocks fabric" or "the fabric of a rock", in the domain of geology and mineralogy, is difficult to translate into French as it comprises the idea of rock's texture, composition and the disposition of its crystals. The French langue uses finally a loanword "fabrique" but which in common language is a false cognate meaning "factory". This difficulty is nevertheless particularly apparent in the domains where cultural differences are important. As the ARTES database is a multidomain language resource, the domains such as law, education or social sciences are also included. For example, in the domain of bankruptcy law the *distressed company* seems to be a suitable equivalent for *entreprise en difficulté*, but the cultural differences of law systems in English and French speaking countries, raise problems of translation, despite a European tendency for harmonisation, e.g. in US *distressed companies* are the matter of *bankruptcy courts* while in France *les entreprises en difficulté* are the matter of *tribunaux de commerce*. It would be though improper to say that *tribunaux de commerce* is the exact equivalent of *bankruptcy courts*.

On the other hand, when we have a series of synonymous units in a source and target language, they can all be considered as equivalent. Establishing multiple equivalent pairs is then necessary. The following examples taken from trans-discipline phraseology: *the present section concentrates on, our concern here is with, we shall concentrate here on* can be all considered as possible translations for *dans cette partie nous abordons, nous allons maintenant aborder, nous nous intéressons ici à*. In order to facilitate the processing of multiple equivalences across languages, we are currently modifying the ARTES database architecture in order to integrate synsets which can be defined within a language before aligning them across languages.

4.3 Processing terminological variation

Handling terminological variation when creating language resources is another complex matter. This issue relates to the phenomenon of synonymy, which in the ARTES project is tackled from a broad perspective and termed "concurrence", referring to a situation of competition between terms. In many cases the synonymy between terms is partial. We have thus added a series of descriptors allowing to determine the degree or the type of "concurrence" between terms: acronym, extended version of a term, reduced version of the term, partial synonymy, and so on. For example, in geology, *Moho* is a reduced version of a term for *Mohorovicic discontinuity*, *VLP* is an acronym for *very-long period*, *ice flow* and *ice creep* could be considered as partial synonyms, or near isonyms. An additional note on concurrent pairs explains the degree of superposition of

the meaning and the usage of concurrent terms.
Some more sophisticated phenomena of terminological variation are currently under study with a view to improving the method of processing terminology in the ARTES database, namely the case of complex nominal groups which appear as new terms and give rise to relatively high degree of variation within discourse: e.g. *naturally ventilated buildings* vs. *buildings ventilated naturally, buildings ventilated by natural means, buildings ventilated by natural convection*.
The current procedure for processing terminological variation contains several stages. The first one consists in accessing the nature of the variation through corpus and discourse analysis. Variation, specifically nominal variation, in LSP is generally considered as an indicator of neology. Nevertheless, in some instances, variation can play specific rhetoric or expressive effect. The second step consists in determining, again through corpus and discourse analysis, which variant is dominant, and which variants are the alternative ways of expressing the same concept. The dominant variant is encoded in the ARTES DB as a main entry, while all the relevant variations of the entry are recorded as its concurrents. The type of relation between the main term and each variant is precised, and a note is added to provide linguistic information on the usage of each variant, for instance explain the specific nature of a variant or its context, or circumstances, of use. As the ARTES DB is a relational DB, it is possible through the user's interface to search one of variants and to access the article of the main term.

4.4 Working toward a conceptual organisation of resources

The idea behind ARTES dictionary is to bypass classical alphabetic access to data by revealing multiple relations between data, some of which are particularly useful for understanding lexicon structure.
Generic, partitive, functional, instrumental, analogical... relations can be established between terms in order to highlight the conceptual organisation of lexicon of a particular domain. Retrieving data from the ARTES database in order to display lexical resources graphically is one of the perspectives we intend to develop in the near future.
By the same token, semantic preference and prosody relations, as defined by Sinclair (1987), Louw (1993) can be established between terms determining semantically cognate terms or terms sharing the same connotation. Few authors have studied these phenomena in LSPs, among them Tribble (2000), Hunston (2007) and Louw & Chateau (2010). Semantic preference and prosody have been studied extensively by the members of ARTES team (Kübler & Pecman forthcoming, Castagnoli *et al.* forthcoming) with a view to improving even further the linguistic information encoded in the ARTES database. These phenomena can indeed help us to enhance our knowledge of lexicon structure in terms of meaning and connotation.

One of the many ambitious approaches to data offered in ARTES is also the onomasiological access to collocations which are common to a variety of scientific discourses, as a help tool for drafting scientific texts (Pecman 2007, Pecman *et al.* 2010). This discourse phraseology has enriched studies on GSL (General Scientific Language) (Pecman 2004, 2007) which looks at ready-made patterns commonly employed by researchers and experts regardless of their discipline. In ARTES we have proposed 14 main classes and some 80 sub-classes for categorizing GSL phraseology in types, according to their meaning and function in LSP discourse.
Only very recently, the more comprehensive studies of a similar type of language resources, namely academic phraseology, have been carried out (cf. Durrant and Mathews-Aydınlı 2011; Simpson-Vlach and Ellis 2010). For example, Simpson-Vlach and Ellis (2010) propose an Academic Formula List (AFL) of most frequent lexical bundles found in academic communication, which are sorted according to major discourse-pragmatic functions. Nevertheless, if we compare the AFL with the collocations used in GSL, we find that the formulas used in academic setting are significantly different from those used in scientific setting. Moreover the methodologies for processing collocational phenomena for creating reusable lexical resources are still underexplored. In much the same way, the studies and resources on expert, rather then learner, trans-discipline phraseology are still lacking.

4.5 Integration of complex lexical items such as collocations, collocational frameworks and prefabricated sentence builders

In line with advances in corpus linguistics, the ARTES resources are constructed giving priority to context for determining the meaning and the usage of terminological units. Terms are considered as main entries in the database, while collocations, collocational frameworks and ready-made sentence builders are handled as secondary entries. They behave as preferential contexts of use, which provide useful information on the combination profile of terms in LSP communicative situations.
Studies on collocations and translational problems from a corpus perspective (Kübler 2003, Pecman 2004, Volanschi 2008) have encouraged us to separate specific collocations (associated with terminology) from generic collocations (associated with discourse functions). The information on specific collocations avoids collocational blends when using highly scientific or technical terms in second language communication (e.g. the adjectival term *buoyant* used in a comparative form *to be more buoyant* corresponds in French to a nominal term modified by an adjective: *avoir une plus grande flottabilité*). Similarly, the information on generic collocations allows the user to go further in achieving native-like communicative skills. The generic collocations are often associated to lexical units which are domain non-specific (e.g. *aspect,*

approach, method, study, result, limit, question, problem, evaluate, describe, etc.) with which they enter into collocation (e.g. *to raise a question*, *promising results, experimental approach, etc*.), or they act as sentence builders (e.g. *the most complete account of this problem is found in..., our conclusions focus on aspects such as...*). Both collocations, specific and generic, are analysed in the ARTES database according to their syntactic structure (e.g. *to raise a question*: vb_noun, *experimental approach*: adj_noun) and offer a very useful information for LSP users, particularly when communicating in a second language or working in translation perspective.

5. Conclusion

The ARTES database is an innovative approach to creating lexical resources where database development and an in-depth linguistic analysis of language phenomena are closely interwoven. The originality of this tool lies in its comprehensive approach to language items of relevance in LSP translation and communication, encompassing terminological, phraseological and discoursal elements. The contrastive approach to languages and to scientific disciplines extends further the coverage of lexical resources stored in the ARTES database. This multiple approach makes of the ARTES database an interesting framework for conducting research on a variety of linguistic phenomena observable in relation with LSP. Furthermore, a growing variety of LSP users (translators, teachers, students, experts and linguists) motivated the design of applications that ensure entering and retrieving information from the database by taking into account different user situations. Although designed for a dictionary type use, the ARTES database offers many possibilities for extracting lexical resources and thus anticipate new situations of use: for instance linking the terminological and phraseological data stored in the ARTES database to an online concordancer would allow to display lexical items in larger contexts for distribution analysis. The future research will focus on exploring this new orientation of research. Finally, we think that research in terminology and phraseology from a lexical resources creation perspective can lead the way to a better understanding of language in terms of cognition, description and teaching.

6. Acknowledgements

The design of the ARTES database was supported by the funds provided by the Department for Applied Linguistics and Intercultural studies - UFR EILA – at Paris Diderot University. We would like to thank Pascal Cabaud, from the team System within EILA department, for his ongoing participation in the project.

7. References

Castagnoli, S., Ciobanu, D. Kunz, K., Kübler, N., Volanschi, A. (forthcoming). Designing a Learner Translator Corpus for Training Purposes. in Kübler N. (Ed.), *Practical approaches of theorical models for language corpora and language-related teaching.* Peter Lang: Bern.

Durrant, Ph., Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30(1), pp. 58--72.

Froeliger, N. (2008) Le facteur local comme levier d'une traductologie pragmatique. *Meta, le Journal des traducteurs*, 55(4).

Humbley, J. (2008). Les dictionnaires de néologismes, leur évolution depuis 1945 : une perspective européenne. In J.-F. Sablayrolles (Ed.), *Néologie et terminologie dans les dictionnaires*, Paris : Honoré Champion. Collection Lexica, Mots et dictionnaires, pp. 37--60.

Hunston, S. (2007). Semantic prosody revisited. International *Journal of Corpus Linguistics*, 12(2), pp. 249-268.

Kübler, N. (2011). Working with different corpora in translation teaching. In A. Frankenberg-Garcia, L. Flowerdew, G. Aston (Eds.) *New Trends in Corpora and Language Learning.* London: Continuum.

Kübler, N. (2003). Corpora and LSP translation. In F. Zanettin, S. Bernardini & D. Stewart (Eds.), *Corpora in Translator Education*. Manchester: St Jerome Publishing, pp. 25--42.

Kübler, N. & Pecman, M. (forthcoming). The ARTES bilingual LSP dictionary: from collocation to higher order phraseology. In S. Granger & M. Paquot (Eds.) *Electronic lexicography*. Oxford : Oxford University Press.

Kübler, N.; Pecman, M., Bordet, G. (2011). La linguistique de corpus entretient-elle d'étroites relations avec la traduction pragmatique ?. In M. Van Campenhoudt, T. Lino, R. Costa (Dir.) *Passeurs de mots, passeur d'espoir: Lexicologie, terminologie et traduction face au défi de la diversité.* Actes des huitièmes journées de Lexicologie, Terminologie, traduction (LTT) 15-17 Oct. 2009 Lisbonne, 579--592.

L'Homme, M-C. (2007). De la lexicographie formelle pour la terminologie : projets terminographiques de l'Observatoire de linguistique Sens-Texte. in *Actes du colloque BDL-CA* (Bases de données lexicales : construction et applications), 23 avril 2007, OLST, Université de Montréal, pp. 29--40.

Leroyer, P. (2007). Bringing corporate dictionary design into accord with corporate image. From words to messages and back again. In Gottlieb, H. & J. E. Mogensen (Eds.) *Dictionary vision, research and practice: selected papers from the 12th International Symposium on Lexicography*, Copenhagen 2004. Terminology and Lexicography Research and Practice 10. Amsterdam/Philadelphia: John Benjamins, pp. 109--117.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157-176). Amsterdam: John Benjamins.

Louw, B. & Chateau, C. (2010). Semantic Prosody for

the 21st Century: Are Prosodies Smoothed in Academic Context? A Contextual Prosodic Theoretical Perspective. In S. Bolasco, I. Chiari, L. Giuliano (Eds): *Statistical Analysis of Textual data: Proceedings of the tenth JADT Conference*

Pecman, M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique*. Thèse de Doctorat en linguistique. Université de Nice-Sophia Antipolis.

Pecman, M. (2007). Approche onomasiologique de la langue scientifique générale. *Revue française de linguistique appliquée.* « Lexique des écrits scientifiques », 12(2), pp. 79--96.

Pecman, M. (2008). Compilation, formalisation and presentation of bilingual phraseology: problems and possible solutions. In S. Granger & F. Meunier (Eds.) *Phraseology in language learning and teaching*. Amsterdam/Philadelphia: John Benjamins, pp. 203 --222.

Pecman, M., Juilliard, C., Kübler, N., Volanschi, A. (2010). Processing collocations in a terminological database based on a cross-disciplinary study of scientific texts. *Cahiers du Cental.* Proceedings of eLex2009, Université catholique de Louvain, Louvain-la-Neuve, Belgique, pp. 249--262.

Simpson-Vlach, R & Ellis, N. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistic*, 31(4), pp. 487 --512.

Sinclair, J. (1987). *Looking up: An account of the* COBUILD *project in lexical computing and the development of the Collins COBUILD English language dictionary.* London/Glasgow: Collins.

Tribble, C. (2000). Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals. In L. Burnard & T. McEnery, (Eds.), *Rethinking language pedagogy from a corpus perspective:* Papers from the Third International Conference on Teaching and Language Corpora. New York: Peter Lang, pp. 74--90.

Volanschi, A. (2008). Étude et modélisation des phénomènes collocationnels : Implémentation dans un système d'aide à la rédaction en anglais scientifique, Thèse de doctorat en linguistique. Université Paris Diderot.

Volanschi, A. & Kübler, N. (2011). The impact of metaphorical framing on term creation in biology. *Terminology* 17(2).