Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 300 – 308

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Compiling texts for a specialized corpus in the biochemistry domain: theoretical and methodological aspects

Coral López Mateo, Françoise Olmo Cazevieille*

Universitat Politècnica de València, Camino de Vera s/n, Valencia 46022, Spain

Abstract

At the present time it is practically unthinkable to carry out a linguistic study without resorting to a corpus. In accordance with the type of study we wish to perform, we will compile a set of texts based on pre-established criteria (type of documents selected (on divulgation, research, notes, subject syllabus, etc.); authors responsible for contents, etc.) that will enable us to compile a good-quality and reliable linguistic study. Our work will include an account of the process of the compilation of specialized texts in the German language in the field of biochemistry and describe the creation of the corpus. The fact that most of the biochemistry texts are now published in English, and that biochemistry is now a multidisciplinary field, makes it difficult to compile texts and consequently complicates the design of the corpus itself. Basing our work on the conceptual structure of the domain under study, we will define our project and use a set of criteria that will guarantee a textual corpus that will be representative of the selected sub-field and that will subsequently facilitate the extraction of specialist terminology (Cabré, 1999; Adelstein, 2004).

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: biochemistry, conceptual structure; specialized textual corpus; selection criteria.

1. Introduction

Compiling a corpus is now common practice in linguistic research. Hunston (2002) identifies up to eight different types of corpus: specialized, general, comparable, parallel, learner, pedagogical, monitor, synchronic and diachronic. A specialized corpus is used to study a particular type of language, regardless of the level of specialization, but with

* Coral López Mateo. Tel.: +0-34-963877000 – ext. 75344 - clopezm@idm.upv.es
Françoise Olmo Cazevieille. Tel.: +0-34-963877000 – ext.75312 – folmo@idm.upv.es

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.448

a number of specific pre-established criteria as a guide to selecting the types of texts to be included in the corpus. The general aim of this work is to compile a representative specialized textual corpus in the field of biochemistry, i.e. one that includes the range of concepts commonly used in this area and that is valid for the extraction corpus from which the candidate terms will be taken.

This paper is divided into three very different basic parts. The first describes the structure of the field tree, which will provide us with the knowledge and indicate the boundaries of the study. Secondly, a description of the corpus will be given, and finally the method used to compile the texts of the corpus will be explained.

2. Constructing the conceptual structure of the biochemistry domain

In order to analyze the specialist language in a particular field, it is first essential to gather information relating to it. According to Cabré (1999) one can resort to monographic and general works on the subject, such as: manuals, monographs and articles, to specialists in the field and other documentary sources, such as encyclopaedias, hierarchical classifications and thesaurus.

This type of reference work provides assistance in drawing up the conceptual structure (field tree) of the subject and establishing the criteria to be used in the design of the corpus. In order to acquire information on the field of biochemistry, we mainly resorted to manuals and monographs, partly because hierarchical classifications and thesauri (UNESCO) usually offer quite a good general view of the discipline. After completing the field tree, it was given to a specialist in the subject to revise and validate.

Creating a conceptual structure of the discipline is essential because it forms the basis of the research and if the aim is to obtain reliable and representative results one must start from a solid and firm base. But, we must ask: what exactly is a conceptual structure or field tree*?* According to the German Institute for Standardization *(Deutsches Institut für Normung,* 1999), a conceptual structure is nothing other than a number of concepts related to each other or that have become interrelated and that represent a coherent set. Budín (1998) widens this definition by insisting not only on the relationships between its concepts but also on its ability to order knowledge. Sager (1993: 45) adds that "it is necessary to place a concept within the knowledge structure by which it is limited and defined, give it a denomination so that we can clearly refer to it and to define it as an act of clarification, confirmation or fixation of an element of knowledge". Felber & Picht (1984) compare the system of concepts (conceptual structure) with a building in which the concepts are the bricks and their interrelationships are the mortar. They also point out that for the building to take shape it must be designed in accordance with the following parameters: the objective, the field of knowledge with its specific methods and characteristics and the criteria used to systemize its concepts.

To sum up, we can say that a conceptual structure shows an ordered plan of a specific field of knowledge by means of the relationships between its concepts fixed in accordance with pre-established criteria and which allow the terminologist firstly to compare concepts and secondly to name and define them with the aim of clarifying, confirming or fixing them.

We agree with Aguilar (2001: 22) on the fact that the graphic representation of a field tree should provide a general view of the field and thus make it comprehensible by complying with the following basic principles:

- *Univocity:* The representation should clearly and unequivocally reflect the different relationships and the classification criteria.
- *Ease of understanding*: The requirements should be designed for a specific group of end users. A system designed for teaching purposes should be different from one for specialists. In all cases, the users' knowledge level should not be overestimated.
- *Transparency*: To make conceptual relationships understandable, it is advisable to break up complex conceptual systems into a number of simpler systems.
- *Possible expansion*: The system should be organized in such a way that modifications can be made to it without

the need for complete re-structuring.

In order to obtain the structure of the biochemical field of knowledge, we based our efforts on the definition proposed by Mathews et al. (2006). According to these authors biochemistry is defined as a science that studies living beings at the molecular level by means of physical, chemical and biological methods and techniques. It is thus an interdisciplinary, experimental and investigative science that interacts with other disciplines such as organic chemistry, biophysics, medicine, nutrition, microbiology, cellular physiology, and genetics, among others. It is also a specific discipline with its own identity, distinguished by its emphasis on the structures and reactions of biomolecules, for its explanation of metabolic routes and their control, and by the principle that vital processes can be understood by means of chemical laws.

The representation of our field tree is in the form of a numerical list or classification. It is a polyhierarchical system divided into four sub-systems (see Table.1): molecular structures of living creatures (1), metabolic reactions (2), instrumental methods and techniques (3), and applications (4). These, in turn can be subdivided into another series of sub-fields as shown below.

Table 1. Conceptual structure of biochemistry

1.	**Molecular structures of living beings**	**3.**	**Instrumental methods and techniques**
1.1.	Biomolecules	3.1.	Chromatography
1.1.1.	Inorganic	3.2.	Electrophoresis
1.1.1.1.	Water	3.3.	Dialysis and ultrafiltration techniques
1.1.1.2.	Mineral salts	3.4.	Spectroscopy
1.1.2.	Organic	3.5.	Radioactive isotopes
1.1.2.1.	Carbohydrates	3.6.	Autoradiography
1.1.2.2.	Lipids	3.7.	Mass spectrometry
1.1.2.3.	Protides (nitrogenous compounds)	3.8.	Electronic microscopy
1.1.2.4.	Nucleic acids	3.9.	Radioimmunoassay
1.2.	The cell	3.10.	X-ray crystallography
1.2.1.	Animal	3.11.	Fluorometry
1.2.2.	Bacterial	3.12.	Immunoprecipitation
1.2.3.	Vegetable	**4.**	**Applications**
2.	**Metabolic reaction**	4.1.	Medicine and chemical therapies
2.1.	Enzymes	4.2.	Immunology
2.1.1.	Coenzymes	4.3.	Genetic engineering and cloning
2.2.	Metabolism	4.4.	Nutrition
2.2.1.	Metabolism of carbohydrates	4.5.	Clinical chemistry
2.2.2.	Metabolism of lipids	4.6.	Pharmacology
2.2.3.	Metabolism of protides	4.7.	Toxicology
2.2.4.	Metabolism of nucleic acids	4.8.	Nanotechnology
2.2.5.	Photosynthesis	4.9.	Ecology
		4.10.	Agriculture

Since we were dealing with a very wide field and had limited human resources, we had to set the volume of work to suit our capacity and therefore reduced the study domain. The decision to keep human biochemistry and rule out vegetable biochemistry was taken for two pragmatic reasons. The first was due to the large amount of existing research in this area, especially in medicine, therapeutic treatments, genetic engineering and cloning, which implies a correspondingly large number of publications and therefore of candidate texts for the corpus. The second was that

we give lectures in Chemical Engineering and Biomedical Engineering at the School of Industrial Engineering of the Polytechnic University of Valencia. We therefore considered it to be more interesting and appropriate to focus on human biochemistry for the sake of the academic profile of our students. We therefore discarded (see Table. 1) vegetable cells (1.2) from the first partial system, photosynthesis (2.2.5) from the second; the third was left intact, and excluded ecological applications (4.9) and agriculture (4.10) from the fourth.

In the following section we describe the design of our corpus and the criteria used to select the texts for inclusion in it.

3. Design of the specialized human biochemistry textual corpus

3.1. Representativeness and balance

First, we will review the meaning of "corpus" and then the criteria to be used in its design to make it representative and thus able to provide good-quality and reliable results to researchers. Santalla del Río (2005: 45-46) defines a corpus as follows: "A corpus is a set of texts in natural and unrestricted language stored in a homogeneous digital format, selected and ordered by explicit criteria, to be used as a model of a given state or language level in studies or applications more or less related to linguistic analysis."

By *natural and unrestricted language*, this author understands that the texts should be original and their communicational situation should be contextualized; by *homogeneous digital format* we understand all the texts should be in the same format so that all can be finally used as a single text; and *selected and ordered by explicit criteria* we understand to mean that it should not be a mere collection of texts picked at random from any source, but rather a set of texts compiled in accordance with specific criteria determined by the purpose for which it was created, generally linguistic.

One of the essential aspects to be borne in mind in making a corpus is its representativeness. According to McEnery et al (2006) and other authors this is one of the features that distinguish a corpus from an archive, i.e. a set of unrelated texts. A corpus should use natural language, or a variety of natural language. It is of course impossible to include all types and varieties of language, so that only samples of the language or of its variety can be collected, and for these samples to be representative it is essential to fix very clear and precise text selection criteria based on the objective of the study to be carried out. This aspect is of the greatest importance if we are to obtain good and reliable results in our research. Another aspect related to representativeness is the balance among the samples, which in most cases is understood to mean the inclusion of all types of text in the language or variety in question. However, there are many different opinions on how to achieve a balance in selecting texts; Lemnitz & Zinsmeister (2006) point out that for this purpose it must be based on both external and internal criteria; the former should refer to the texts to be included in the corpus and the latter to the inclusion of a wide range of linguistic phenomenon. Atkins & Clear (1992: 6) define a *balanced corpus* as "a corpus so finely tuned that it offers a manageably small scale model of the linguistic material which the corpus builders wish to study". And add that achieving such a balance depends largely on intuition, and that only after the process has been completed can it be seen whether or not a balance has been achieved by the feedback from users. Hunston (2002) emphasizes that it is no easy task to create a balanced corpus, either because the texts included are all of different lengths or because it is impossible to include all the existing texts in equal proportions in a given language or in a variety of that language. In the former case, Hunston recommends including all the texts, however long they may be, and in the latter a corpus should be representative and balanced according to the purpose for which it was made.

After reviewing the opinions of different authors on representativeness and balance, we can state that a corpus should be representative according to the objective for which it was compiled and to its design, based of course on external and/or internal text selection criteria.

3.2. Text selection criteria

Ours is a specialized corpus in the field of human biochemistry and will include primary texts, i.e. written by and for specialists. According to Sinclair (1996) there are two types of general criteria: external and internal. The former refer to the social and cultural context and consider aspects such as date of publication and the origin, aim and form of publication. The latter criteria refer to purely linguistic aspects, such as word distribution and grammatical and lexical issues, etc. McEnery et al. (2006) and other authors recommend corpus compilers not to use these criteria because they condition the result of the analyses and the corpus would thus be biased by its own design. However, as the present tendency is to use both types of criteria we decided to follow this path.

Other specific criteria that should be taken into account include, for example, *quantity*, which was and still is a highly debated subject in which no clear conclusions have been reached. It has often been associated with *representativeness*. Sinclair's (1996) original opinion was that the larger the corpus the better, although five years later he rectified and stated that size really had nothing to do with the quality of a corpus. On the other hand, Hunston (2002) maintains that over-large corpora can be difficult to handle and smaller ones can be more efficient to work with. Meyer & Mackintosh (1996) argue that a specialized corpus can be much smaller than a general-purpose work, which must include samples of all the language. For Pearson (1998) length is not important if the previously established criteria are followed, but McEnery et al. (2006) point out that in the lexicology the number of texts or words is important and the corpus should be large, but in its terminography; the most interesting thing is that the corpus should offer an extensive terminological density, rather than a great number of texts or words. It is recommended to compile complete texts to avoid leaving out conceptual information of interest to the terminographer.

Quality is another important criterion to be considered when designing a corpus. Texts should be either up-to-date or of recent publication in order to be representative of the current state of knowledge in the field under study. It is preferable to use reliable texts from a recognized authority. Pearson (1998) recommends using already published texts since this implies that they have undergone a previous review and are therefore more likely to be reliable. The *language* should also be taken into consideration: Is it a monolingual or multilingual corpus? In the latter case: Do we want to design a comparable parallel corpus? How many languages do we wish to study? Texts should preferably be in the original version, i.e. translations will not be included so as to ensure they reflect the original terminology used in the specialized field. In addition, we should specifically establish the *level of the language*, which will depend on the authors (social group, age, etc), the subjects dealt with and the form (written-spoken, colloquial, learned, or specialist language).

When compiling the text extraction sources we searched for chemistry, biochemistry and molecular biology journals published in German for texts appropriate for our purpose and opted to use only the *Angewandte Chemie* journal as it contains only primary texts in that language. Other journals were also considered and rejected for not being specialized enough to meet our requirements as regards language level and communicational situation. The *Angewandte Chemie* belongs to the *Gesellschaft Deutscher Chemiker (GDCh)* and is published by Wiley-VCH publishers. It is a highly reputed journal that uses peer review and publishes original research dealing in the whole chemistry field. Its impact factor in 2013 was 11,336 (according to Journal Citation Reports, July 2014) and was 11th out of 178 journals in its class. It is published weekly (52 issues per year) in two editions: the German (*Angewandte Chemie)* and an international edition in English (*Angewandte Chemie International Edition).* Both editions have identical contents and are published on paper and online. However, it should be pointed out that most of the articles in the German edition are in English and it also contains articles translated into German, which were ruled out of the present research.

Table. 2. Our text selection criteria

OUR TEXT SELECTION CRITERIA	
Size	Complete texts of various lengths by different specialists: - Short reports (Zuschriften) from 3 to 7 pages (the most frequent case) - Articles (Aufsätze und Kurzaufsätze) from 15 to 30 pages (less frequent) - Novelties (Highlights) from 3 to 4 pages (less frequent) - Approximately 600 texts in total (450 papers, 84 articles and 73 novelties)
Time period	From 2010 to 2014 (inclusive) Weekly journal; 52 issues per year
Language	Monolingual: original texts in German; no translations
Language level	Authors: by specialists for specialists and students - Highly specialized texts - Written and published corpus (reviewed by peers)
Subjects	Research in human biochemistry
Form	Formal

3.3. Limits to selecting texts

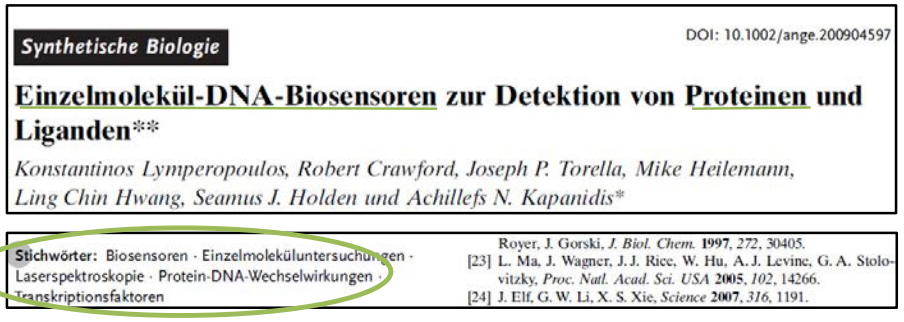
The compiled corpus is finite but open to expansion and modifications. We felt obliged to limit our selection for the following reasons: firstly, as mentioned above, due to the wide choice of available material we had to reduce this to a manageable size. Secondly, we met with problems in selecting specific texts in human biochemistry. The source journal includes articles on the whole range of chemistry topics, not only biochemistry. As the latter is a multi-disciplinary field, it is sometimes difficult to draw the line between biochemistry and associated fields, especially between biochemistry and organic chemistry. We finally opted to include the most representative chemical elements in biochemistry and excluded minority and less representative fields. Thirdly, we were also limited by the availability of texts in German, as most are now published in English. Finally, our criteria as regards communication and language level limited us to a single text extraction source which also included papers translated into German.

4. Compiling texts: method and examples

Besides the external criteria described above, we also included a single internal criterion, which was the *lexical*. We found the identification criterion of the specialized lexical units (SLU) by L'Homme (2004) very useful and applied it when selecting our texts. This author identifies SLUs in a field by studying the terms that surround it, and if these are from the field under study the SLU must also belong to it. Thus, if an SLU in a title or abstract or key words can be combined with others belonging to the human biochemistry domain, this indicates that the text belongs to this domain and can thus be considered as a candidate text. If there were doubts about this, then we read the entire text and consulted a specialist in the field to decide whether or not to include it in the corpus.

Below we give an example of our procedure in applying the SLU identification criteria and point out some of the difficulties involved.

In the example text selected (see Fig.1) no problems were found in identifying SLUs. In both the title and the key words there are explicit terms from the biochemistry field plus other concepts from the proposed field tree. The first system, *molecular structures of the living being*, includes *Proteinen*, and the second, *chemical reactions*, *Protein-DNA-Wechselwirkungen*, *Transkriptionsfaktoren* and the third, *instrumental methods and techniques*, includes *Laserspektroskopie*.



Synthetische Biologie

DOI: 10.1002/ange.200904597

Einzelmolekül-DNA-Biosensoren zur Detektion von Proteinen und Liganden**

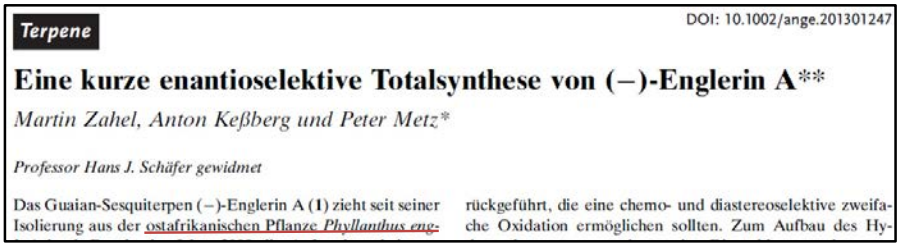
*Konstantinos Lymperopoulos, Robert Crawford, Joseph P. Torella, Mike Heilemann, Ling Chin Hwang, Seamus J. Holden und Achillefs N. Kapanidis**

Stichwörter: Biosensoren · Einzelmoleküluntersuchungen · Laserspektroskopie · Protein-DNA-Wechselwirkungen · Transkriptionsfaktoren

Royer, J. Gorski, *J. Biol. Chem.* **1997**, *272*, 30405.
[23] L. Ma, J. Wagner, J. J. Rice, W. Hu, A. J. Levine, G. A. Stolovitzky, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 14266.
[24] J. Elf, G. W. Li, X. S. Xie, *Science* **2007**, *316*, 1191.

Fig. 1: Example of lack of problems in identifying an SLU.

In the case of terpenes we had to carefully read the paper to determine whether they were produced by animals, micro-organisms or vegetables. Most belonged to the vegetable world.



Terpene

DOI: 10.1002/ange.201301247

Eine kurze enantioselektive Totalsynthese von (−)-Englerin A**

*Martin Zahel, Anton Keßberg und Peter Metz**

Professor Hans J. Schäfer gewidmet

Das Guaian-Sesquiterpen (−)-Englerin A (**1**) zieht seit seiner Isolierung aus der ostafrikanischen Pflanze *Phyllanthus eng-* rückgeführt, die eine chemo- und diastereoselektive zweifache Oxidation ermöglichen sollten. Zum Aufbau des Hy-

Fig. 2: Example of the need for careful reading of the text to identify SLUs: terpenes of vegetable origin.

Terpen-Biosynthese

DOI: 10.1002/ange.201209103

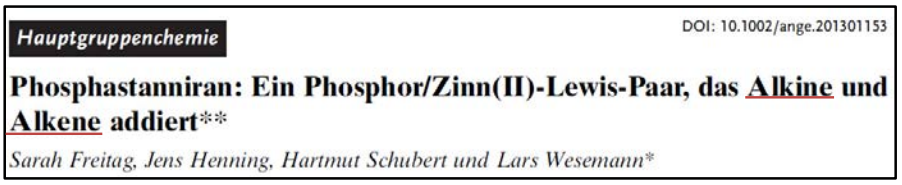
Schnelle chemische Charakterisierung bakterieller Terpen-Synthasen**

*Patrick Rabe und Jeroen S. Dickschat**

Fig.3 Example of the need for careful reading of the text to identify SLUs: terpenes of bacterial origin.

In the example in Figure 2, the text was rejected because the terpenes cited were of vegetable origin and were thus outside our field tree. The example test in Figure 3 was selected since the terpenes cited were from bacteria.

Although alkanes, alkenes and alkynes are naturally present in a variety of forms they are not biologically classified as essential materials. As explained above we limited the research field to predominant and basic biochemical molecules, so that any texts containing the terms or SLUs *alkanes, alkenes and alkynes* were discarded.


Hauptgruppenchemie

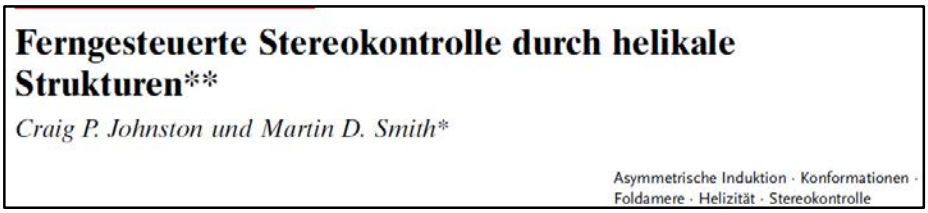
DOI: 10.1002/ange.201301153

Phosphastanniran: Ein Phosphor/Zinn(II)-Lewis-Paar, das Alkine und Alkene addiert**

*Sarah Freitag, Jens Henning, Hartmut Schubert und Lars Wesemann**

Fig. 4: Example of the difficulty of defining SLUs: alkanes, alkenes, alkynes.

In the next case (see Fig.5) the difficulty lay not in identifying SLUs but in the uncertainty of whether or not the paper had been translated. Although translations normally cite the name of the translator, we were surprised by the fact that the Highlights (novelties) were usually published in German even though some of the authors did not have German surnames nor did they belong to a German, Austrian or Swiss university. In these cases we did not include the texts in the corpus.


Ferngesteuerte Stereokontrolle durch helikale Strukturen**

*Craig P. Johnston und Martin D. Smith**

Asymmetrische Induktion · Konformationen · Foldamere · Helizität · Stereokontrolle

Fig. 5: Example of the difficulty of selection: possible translated text.

5. Conclusions

This paper describes the creation of a conceptual structure in biochemistry in which the fields and sub-fields under study were limited to human biochemistry. We also designed a textual corpus based on the objective of our research and on specific external and internal criteria to select texts that were representative of and faithfully reflected the specialized language used in the field of human biochemistry. To compile the texts we adapted the SLU identification criteria proposed by L'Homme for selecting texts. During the specialized text selection phase the difficulties encountered included: the shortage of texts written in the German language, which we got over by using a relatively long publication period of five years. Another problem was making a divide between human biochemistry and organic chemistry, which we solved by identifying SLUs and consulting experts in cases of doubt.

Finally, we would like to emphasize that due to the low availability of highly specialized texts, which is now becoming more frequent in specialist fields in German and languages other than English, there is an obvious need for terminological studies to collect, disseminate and share the terminology used in these languages, which are the basis of scientific language.

References

Adelstein, A. (2004). *Unidad léxica y valor especializado: estado de la cuestión y observaciones sobre su representación*. Institut Universitari de Lingüística Aplicada/Universitat Pompeu Fabra.

Aguilar, L. (2001). *Lexicología y terminología aplicadas a la traducción*: *Curso práctico de introducción*. Barcelona: Universitat Autònoma de Barcelona.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1-16.

Baxmann-Krafft, E. & Herzog, G. (1999). *Normen für Übersetzer und technische Autoren*. Berlin; Wien; Zürich: Beuth Verlag.

Budin, G. (1998). *Theorie und Praxis der übersetzungsbezogenen Terminologiearbeit*. WUV-Univ.-Verlag.

Cabré, M. T. (1999). *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Felber, H., & Picht, H. (1984). *Métodos de terminografía y principios de investigación terminológica*. Editorial CSIC-CSIC Press.

Hunston, S.(2002). *Corpora in applied linguistics*. Cambridge University Press.

Lemnitzer, L., & Zinsmeister, H. (2006). Korpuslinguistik: Eine Einführung. Gunter Narr Verlag.

L'Homme, M. C. (2004). *La terminologie: principes et techniques*. Pum.

Mathews, C. K., Van Holde, K. E., & Ahern, K. G. (2002). *Bioquímica. (3rd ed.)*. Madrid: Pearson Education.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies*. London: Routledge.

Meyer, I., & Mackintosh, K. (1996). The corpus from a terminographer's viewpoint. *International Journal of Corpus Linguistics, 1(2)*, 257-285.

Pearson, J. (1998). *Terms in context* (Vol. 1). Amsterdam: John Benjamins Publishing.

Sager, J. C. (1993). *Curso práctico sobre el procesamiento de la terminología*.[Traducción de L. Chumillas de A Practical Course in Terminology Processing (1990]. Madrid: Fundación Germán Ruipérez.

Santalla del Río, M. ª P.(2005). La elaboración de corpus lingüísticos. *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas, Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico*, 45-63.

Sinclair, J. (1996). Preliminary recommendations on corpus typology. *EAGLES Document TCWG-CTYP/P (available from http://www. ilc. pi. cnr. it/EAGLES/corpustyp/corpustyp. html).*