20

A Pragmatic Model of Text Classification for the Compilation of Special-Purpose Corpora

Chelo Vargas

1. Introduction

An important source of evidence for a terminologist comes from the different types of texts produced within a community of discourse or "epistemological community" (Alcaraz, 2000). Thus, the articles of specialised or professional journals, manuals, patents and so on, written by the specialists of a given area turn out to be an invaluable textual material —the corpus—, in which the behaviour of the terms can be observed and analyzed in a contextualized, effective and real way. I understand a corpus as the natural habitat where one comes across these lexical units.

The University of Alicante's research group *El Inglés Profesional y Académico*, (Academic and Professional English) is interested, among other topics, in the research and creation of bilingual terminological applications, mainly dictionaries and databases, aimed at the translator of specialised texts. Such applications must show, therefore, the real use of terms. For that matter, they include the equivalent in the target language, the real context in which those terms appear, the most common terminological collocations, definitions, and specific notes for their usage, to name a few. These linguistic data, essential in the process of translating, are mined from the corpora created for each specific subject field under consideration.

There exist a considerable number of articles that deal with the design and textual classification of a general corpus (*cf*., among others, Atkins *et al.*, 1992; Biber, 1993; Biber et al., 1998; Engwall, 1994; Leech, 1991, 1992b; Sinclair, 1987; Summers, 1993)[1]. However, not so much effort has been put in describing the necessary aspects involved in the design of a specialised corpus whose purpose is to create a terminological application. There are, of course, exceptions worth mentioning: the papers of the University of Surrey in Great Britain (Ahmad, 1993, Ahmad *et al*., 2001); those from the University of Ottawa in Canada (Bowker, 1996;

Meyer and Macintosh, 1996), and those from the Aarhus School of Business in Denmark (Bergenholtz and Tarp, 1995). In addition, there is a piece of reference work on this field by Pearson (1998), who more recently (2002) published together with Bowker a practical guide to use this type of corpus.

The objective of this article is to present a pragmatic model for text classification that can represent the different communicative situations, and consequently, the different levels of text specialization. The main purpose of the model is to serve as a useful tool to recover and select initially the written texts that will form part of a special-purpose corpus. With this objective in mind, in the first part of this article, I will revise briefly the concepts and criteria that theoretically frame this work. In the second part, I will provide the criteria and methods adopted to conceive the model, which will be developed in a sequential manner. This second part will conclude with an integral graphical view of the model. Finally, in the conclusion, I will supply some comments about the model.

2. Reference framework

2.1. Specialised discourse

From its beginning, linguistic research has been carried out under different perspectives. Initially, linguists deal with language considered as a system made up of signs and rules that govern the behaviour of such signs to eventually analyse the “language in action” (Alcaraz, 1990), that is, language as discourse or text. From this last conception of language, we take into consideration a series of extralinguistic factors that affect considerably the different linguistic productions in which any communicative act materialises.

I will use the term *discourse* in the sense given by Pragmatics, in other words, to refer to language in action, from a communicative point of view. It will encompass any kind of written or oral interaction. Similarly, I have decided to follow the trend that uses the term *specialised discourse*. Basically, this option responds to two reasons. Firstly, it seems that this term illustrates the type of discourse in which I will centre this study: texts which are produced and circulated within a specialised field; more specifically, an industrial sector: the natural stone. Secondly, it has to do with the descriptor *specialised* which I considerate a good choice, since it contemplates the notion of *continuum*. This notion becomes a fundamental aspect in order to approach texts with different degrees of specialisation. Although it can be argued that there are convergent or divergent features in the different texts that make us classify them within a certain type, function, etc., this segmentation is due to strictly functional and methodological reasons. Furthermore, *text* and *discourse* are used synonymously to refer to oral or written productions. Text is used to refer to a unit, whether oral or written, that includes not only a whole variety of characteristics belonging to its linguistic dimension, but also others concerning its communicative dimension.

Specialised texts occur during an act of either academic or professional communication and their main function is to express and transmit specialised knowledge of a varied level. They exhibit some pragmatic, linguistic and cultural characteristics relevant to the traditions of a given discursive community, which confer on them certain specificity within the number of texts produced in any language.

The concept of *fachsprahliche Texlinguistik* or *Fachtextlinguistik* (specialised text linguistics) comes from the Germanic textual linguistics, and was first introduced by Hoffmann, followed by Gläser and Kalverkämper (Schröder, 1991: 12). It emerged as a result of the orientation that Linguistics took on, not only towards textual linguistics, but mainly towards the research of specialised languages. In relation to the latter area, Göpferich (2000: 227) states that research in specialised languages has been centred in specialised texts, and, particularly, in text genres or types centred in the last fifteen years. Without a doubt, the efforts attempted for making progress in the construction of text typologies are relatively recent in the field of Linguistics.

The different approaches used to define the notion of specialised text, considering either linguistic or extralinguistic aspects have provided us with different text classifications. Incidentally, despite the fact that texts can be catalogued and labelled quite easily, there is not a commonly accepted text typology. Alcaraz (2000: 133-134) summarises the concept of *text type* within the context of Professional and Academic English in the following way:

> el conjunto de textos, escritos u orales, del mundo profesional y académico, que se ajustan a una serie de convenciones formales y estilísticas, entre las que sobresalen las siguientes: (a) una misma función comunicativa; (b) un esquema organizativo similar, llamado macroestructura; (c) como desarrollo de la macroestructura anterior, una modalidad discursiva semejante (narración, exposición, descripción, etc.) y unas técnicas discursivas equiparables (definición, clasificación, ejemplificación, etc.), […], que sirven de guía para que el receptor del mensaje espere una determinada experiencia discursiva; (d) un nivel léxico-sintáctico análogo, formado por unidades y rasgos funcionales y formales equivalentes (por ejemplo imperativos, pasivas, sintagmas nominales largos, etc., en los artículos de investigación científica); y (e) unas convenciones sociopragmáticas comunes, esto es, una utilización por profesionales y académicos en contextos socio-culturales similares.

This author (*ibid*.: 133) points out that in English speaking countries, the term *genre* is preferred to *text type* to classify typologically the oral and written texts produced in professional and academic spheres. The analysis of genre has been promoted from several research fields, such as Pedagogy, Linguistics, and Pragmatics and the Teaching and Learning of Foreign Languages, and predominantly by the trend of the Professional and Academic English with the work by Swales (1990) and Bhatia (1993).

The concepts of genre and text type are nevertheless controversial and are debated continually in the vast linguistic literature about these subjects. Therefore,

we can find authors admitting the equivalence between these two notions (Alcaraz, 2000 and Stubbs, 1996), and others who defend that genre and text type are two different issues (Gläser, 1995: 141-2; Isemberg, 1987: 101). In relation to this matter, Trosborg (2000: IX) points out that whereas in German speaking countries the opposition between *Texsorten* (class of text, genre) and *Texttyp* (text type) is somehow clear-cut, it is not so for other linguistic communities. The former, according to those who are in favour of a distinction, refers to classifications not related to any specific linguistic theory and intuitively made by native speakers. The latter, in contrast, assigns a category related to a theory for the scientific classification of texts. For our purposes in this paper, I will adopt the definition of genre given, among others, by Alcaraz (2000: 133) and Stubbs (1996); therefore, the terms *genre* and *text type* will be used without distinction.

Ciapuscio (1994) carries out a critical review of text typologies made in the 70s and 80s up to 1994. This author favours a typology that proposes the approach based on the same line of cognitive textual analysis followed by Beaugrande and Dressler (1981), who describe the different levels of text classification. This classification takes into account the mental processes that guide the producer of a text to select the concepts and procedures that will result in a text. I believe that these complex approximations in which cognitive, linguistic and communicative features coexist can provide a much more exhaustive description of the different phenomena under consideration. Furthermore, I propose that this multilevel approach is more in tune with our concept of text and its direct relationship with the processes of linguistic comprehension and production.

2.2. Corpus linguistics (CL)

The research and the scientific development whose subject matter is specialised languages and, within these, terminology, take place nowadays in a framework sustained on the bases of theoretical and methodological principles which underline the importance of the use of linguistic corpora. A corpus is conceived as "a more powerful methodology from the point of view of the scientific method, as it is open to objective verification of results" (McEnery and Wilson, 1996: 13). Therefore, terminology and, more precisely, its practical activity or terminography benefit from the information a corpus may deliver to develop terminological applications. Discovering new or obsolete words, analysing new meanings assigned to a single lexical unit, detecting terminological collocations or combinations, gathering actual use instances, definitions and so forth can be quite readily performed if we work with corpora.

In the last few years, an increasing number of corpora, either of general or of specific purpose has been developed. Together with the growth of corpora, computer tools to analyse and exploit them have been improved. Nowadays, relevant terminological information, such as terms, combinations, contexts of use, definitions, etc. is semi-automatically extracted.

The compilation of a corpus has not been alien to lexicography or to language

teaching. Far from it, not only was it a common practice (*cf.* Francis, 1992: 17-22) but it has traditionally been the way of gathering data to describe linguistic issues. Currently, the notions of what a corpus is and how it must be created and exploited have undergone revolutionary changes since the late 70's because of personal computers. PCs became so popular that a greater number of researchers became involved in the study of natural language processing. The use of corpora together with the ever increasing sophisticated computer tools began in the 80's and since then it has contributed decisively in reviving and strengthening linguistic research based on corpus:

> The resurgence of corpus linguistics can be measured in terms of the increasing power of computers and of the exponentially increasing size of corpora, viewed simplistically as large bodies of computer-readable text (Leech, 1991: 9-10).

Moreover, the application of computers to operations such as compiling and building large corpora has been a key factor to overcome practical and theoretical objections. Nowadays, as a result of this connection, the term *corpus* contains intrinsically the characteristic of "machine readable" (McEnery and Wilson, 1996: 23)

Electronic corpora, thus, have become a privileged and unavoidable means to explore the structure of a language, to test linguistic hypotheses, to become aware of the different meanings a word can have in different contexts, and to develop new insights in natural language processing. Along with the popularization of the use of PCs and with the fact that the corpus acquires as an inherent characteristic its electronic form, there exist other aspects that were favourable for the *renaissance* of this applied discipline. I am referring, in the first instance, to the greater availability of a technological infrastructure in private and institutional hands (McEnery and Wilson, 1996), allowing access to online corpora (Biber *et al*., 1998: IX). Secondly, an outstanding development takes place in specific computer software –some of them commercialised– for corpus analysis (basically part-of-speech taggers and concor-dancers), and in Optical Character Recognition technology, which makes possible the release of a corpus compilation "from the logjam of manual input" (Leech, 1991: 10). Thirdly, it is worth mentioning that in the 80's CL began to adopt an eclectic orientation instead of following the extreme parameters set by American structuralists and generativists. The said eclecticism defends the coexistence of quantitative and qualitative methodologies. Thus, the opposition between "armchair linguists" and "corpus linguists" (Fillmore, 1992) happen to be dichotomies no longer held. It is, however, a must to strike a balance between these opposing poles to find mutual benefit: "the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body" (*ibid*.: 35). Lately, for the sake of reaching the highest degree of scientific rigour, a more decisive attitude favouring the combination and complementation of these two stances has been observed (Biber *et al.*, 1998; Fillmore, 1992; Stubbs, 1996, among others).

One of the hotly debated issues among corpus linguistics theoreticians is that of the status the CL must hold. The following question may arise when corpus linguistics is dealt with: are we implying that it is just a methodology or an entire new discipline? Undoubtedly, the discussion between linguists arguing that CL is an emerging paradigm and those who see it as a mere methodology is an unresolved matter. Stubbs (1996) believes firmly that CL has a theoretical status. This linguist argues that the research based on a corpus of large samples of texts means a new way of understanding the relationship between data and theory; thus, the way in which theory can be based on an existing corpus of the natural language becomes evident. In his opinion, theory can emerge inductively from data, putting forward a sounder foundation to computational corpus linguistics. Kennedy (1998: 79), on the other hand, understands that corpus linguistics should better be seen as a source of evidence for any linguistic theory that accepts the nature of empirical data.

Other authors such as McEnery and Wilson (1996), and similarly, Leech (1992a) do not consider that CL should be seen as an independent branch as syntax, semantics or socio-linguistics because these disciplines "concentrate on describing/explaining some aspect of language use" (McEnery and Wilson, 1996: 2). For these authors, CL does not have a specific object of study; and, in Leech's words (1992a: 105), it does not refer to an area of study, but works as a methodology that can be used in almost all the linguistic research fields without being an independent discipline.

Methodology, then, acquires the sense of "a set of tools and techniques" which other linguistic areas can make use of. Such set of tools and techniques can be applied to different disciplines without altering their theoretical tenants. According to this, it would be possible to have a type of corpus-based syntax *versus* conventional syntax, a corpus-based lexicography *versus* traditional lexicography, or a corpus-based terminography *versus* conventional terminography, among other examples. In these new disciplines the data, the theoretical premises and the results would be the same; the only difference would be in the instrument or techniques used. In other words, the difference would be on the procedure followed in order to solve the linguistic problem under consideration.

Leaving aside the issues we have briefly reviewed, in this paper CL -in its current trend- is conceived as a powerful methodology that provides useful and valid tools, which help to accomplish descriptive levels with more certainty than in the past.

From CL, researchers have developed various methods, tools and techniques to organise a representative sample of texts in a specific language and, subsequently, analyse this group of texts for lexicographic purposes (Sinclair, 1987). Thus, lexicographic corpora are based and arranged according to a particular text typology. However, there is not a universally accepted approach to carry out the task of classifying them by genres or text types. Each corpus project devises its own method of classification (Pearson, 1998: 53). In any case, text category is assigned according to different external parameters and, generally,

readily describable. Genres are also easily recognised as groupings of texts within a speech community, because they are "characterized by a set of communicative purpose(s) identified and mutually understood by members of the professional or academic community in which it regularly occurs» (Bhatia, 1993: 13). This is also the viewpoint of Swales (1990: 26) when he claims that "a discourse community utilizes and hence possesses one or more genres in the communicative furtherance of its aims".

Let's briefly summarise the text categorization used in the three most important lexicographic corpora: The Lancaster-Oslo/Bergen (LOB) corpus, the COBUILD corpus and The Longman/Lancaster Corpus (LLC).

The LOB[2] corpus is the equivalent to the Brown University Corpus of American English, known as the Brown Corpus, but is concerned with British English. It has been built using the same criteria as the Brown in that it contains: a) a million words; b) a uniform and approximate division by genres; c) 500 text samples; d) 2000 words per sample; e) published written sources. Although some of the criteria may be obsolete nowadays for the available computer resources, they are still recommended, particularly the one that states that a corpus should contain at least one million words. The LOB, in line with the Brown, classifies the texts in two great subdivisions: informative and imaginative. To the informative group belong journal articles and specialised texts while fiction, magazine articles are included in the latter subgroup. The COBUILD (Collins Birmingham University International Language Database) corpus does not establish the metacategories (informative or imaginative) mentioned above. It rather sorts the texts out according to the media used, i.e., newspapers, brochures, books, magazines and letter correspondence. New subcategories develop from the first five ones. The LLC, on the other hand, establishes the same metacategories as the Brown but subdivides them in ten superfields of which natural & pure science, social science, world affairs are three examples. What distinguishes the two corpora, however, is that the former is 'topic driven' whilst the latter is 'genre driven' (Summers, 1993: 192).

The conclusion that can be drawn from the design of these general corpora for lexicographic purposes is that in order to achieve the ends a specific corpus has been designed for, a great variety of text types, topics, knowledge areas and so on must be taken into account.

2.3. Corpus-based terminography

Corpora validity acquires greater and greater soundness in the identification and characterization of lexical units which according to specific contexts and situation activate their specialised value due to the fact that "it is the use of a lexical unit in a fixed expressive and situational context that provides it with the status of term" (Cabré 1999: 124). From here it derives the importance that should be attributed to communicative parameters when constructing a special-purpose corpus; thus we should consider the diversity of texts characteristic of an academic or professional field, arising from a given situation and having their peculiar functions.

The quality of a terminological project is directly related to the quality of the texts that a corpus is made up of (Bowker, 1996: 42). Text quality and representativeness, in this context, is much more relevant than text quantity (Meyer & Mackintosh, 1996: 268). By quality, we mean authenticity, one of the four features pinpointed by Sinclair (1997: 7): "all the material is gathered from the genuine communications of people going about their normal business". Representativeness, on the other hand, is a much debated issue among linguists. Apparently, they have not come with a definite conclusion as far as setting clear norms to build a corpus in the most representative manner (Clear 1992, 1997; Kennedy, 1998). Tognini-Bonelli (2001: 57) summarises this controversial issue as follows:

> a corpus [...] should be representative of a certain population and […] the statements derived from the analysis of the corpus will be largely applicable to a larger sample or to the language as a whole.

Thus, in its essence, a corpus, no matter which type described, must be representative of a certain population of linguistic events –the texts– which take place in a given language. The problem posed by the notion of representativeness is that a corpus is a finite sample out of an infinite population and, consequently, it is almost impossible to account for all the linguistic events because, according to Leech (2002: 4): "there are always new ones coming along that we haven't seen yet".

The small imbalances in quality and representativeness in general corpora are eventually overcome precisely by the great number of data they contain (100 million words in the case of the British National Corpus). However, it is difficult to create a specialised corpus of extensive dimensions for various reasons. Once acknowledged that there are no rules that account for the size a specialised corpus must have, it is the job of the compiler to decide how to deal with it depending on a number of variables, such as the needs and purpose of the project, availability of texts and time of completion, to name a few. Moreover, the lack of computer resources may also impose some restrictions on the desired design and size of the corpus (Engwall, 1994: 51). At any rate, the more quality and representativeness is observed in the texts composing the corpus, the more reliable it will be.

I have already mentioned that the design of a general corpus contemplates a wide range of areas, of topics, of types of texts with the purpose of building a representative sample of the language in question. As with general-purpose language, scientific and technical discourse is far from being homogeneous:

> Un simple análisis de la comunicación especializada real en situaciones profesionales de distinto signo muestra una multiplicidad importante de registros, en los que, sin abandonar el carácter especializado del conocimiento y su transmisión, se ponen de manifiesto una serie de características que coinciden con las que muestran otras unidades utilizadas en otros tipos de situación comunicativa (Cabré, 1999: 118).

As is the case with the general corpora (LOB, COBUILD, Brown, LLC), the specialised corpora must take into account all the different communicative situations that arise from the type of discourse they represent, keeping in mind also that an appropriate identification of terms depends on the rigorous classification of the texts that are going to be selected. In my opinion, a specialised corpus should, by definition, reflect its main communicative settings.

Nevertheless, a corpus is limited in that it can only represent some subsets of the language and not its total set. It is practically impossible for a corpus to include all and each one of the type of texts that can be produced in all and each one of the possible communicative situations in a given language or in one of its varieties. In order to compile a corpus, it is advisable then to select explicitly the linguistic uses that are going to be our focus of study.

A corpus that is balanced in terms of text types will satisfy the diverse conceptual, pragmatic and/or linguistic needs any terminographer may require. At a conceptual level, a corpus that reflects -through the conception and adoption of a pragmatic text typology (as is our case)- the different degrees of discourse specialisation is bound to become useful to the terminographer as he/she may immerse himself/herself in a completely new domain that is completely foreign to him/her to acquire relevant knowledge more easily. Thus, a corpus created in this way will enable the terminographer to come to terms with this new reality progressively. From texts addressed to semi-specialists or laymen, a terminographer will become aware of somewhat basic information in the form of definitions, explanations and synonyms that will shed light on the new field to gradually move into further complexity to be experienced in those texts addressed to professionals, where he/she will be faced with more complex technical information. At a pragmatic level, the terminographer, guided by frequency rates and term context use, will decide what is convenient or not to include in the terminological application that is under construction, but always depending on the prototypical user of such application (*cf.* Gómez & Vargas, 2004). Finally, at a linguistic level, Bowker (1995: 13) states that the inclusion of texts with different levels of specialisation provide the terminographer with a more complete image of the terminological diversity found in a given area of knowledge or activity.

3. Towards a pragmatic model for LSP-text classification: criteria and method

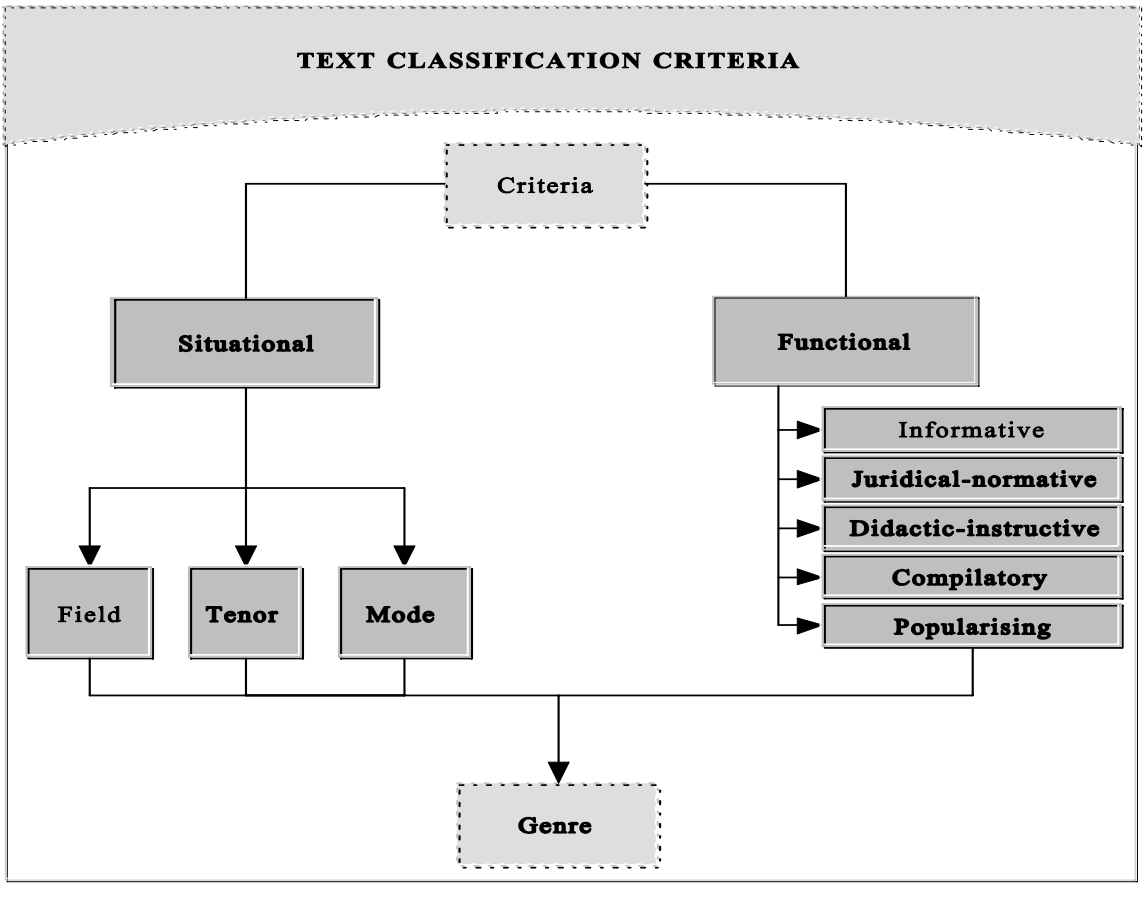
The decision of adopting a pragmatical and multidimensional approach (like the one proposed by Ciapuscio & Kugel, 2002; Göpferich, 1995; Gläser, 1995) has been taken with the purpose of elaborating a typology that accounts for the fundamental characteristics of the texts produced in a specific field. The typology is useful for the textual acquisition and construction of a corpus made up of a series of textual samples representative of the industrial sector of the natural stone.

On the first stage of the corpus compilation, Biber (1993) notes the importance of intending to cover all the situational parameters of the linguistic variation, that is

to say, the external criteria; given that these parameters are identifiable before a corpus is constructed. Moreover, he states that "there is no a priory way to identify linguistically defined types [of texts]" (*ibid*.: 245). In accordance with Biber, I consider the external characteristics of the texts[3], e.g., the extralinguistic context in which they generate and operate. How these external criteria can be identified is by means of a close analysis of the text communicative function, of the participants involved and of the communicative setting itself; in short, an analysis of all the socio-cultural criteria and categories involved in texts (Sinclair, 2003: 170).

At present, we could argue that in general the tendency is to use external criteria not only for the classification of texts, but also for the design of corpora (Pearson, 1998: 55). In fact, there are good reasons for that, as Atkins *et al*. (1992: 5) point out. Firstly, because external criteria can be determined without reading the text we intend to choose, making sure in this way that our selection is more objective, in that the compiler is not doing a previous linguistic judgement. Secondly, it is inevitable that the initial selection of the texts is based on external criteria, since it is only after we gather and analyse the text that we can find a series of linguistic features specific to each text which will contribute to its characterisation at an inner level.

I will adopt the pragmatic approach because I start from the assumption that the linguistic or internal characteristics of a genre depend on its communicative purpose. Therefore, the pragmatic or communicative typology will allow us to reduce the great variety of genres to a limited number of text categories and subcategories. Moreover, this approach, as Ciapuscio (1994, 2003) understands, is currently accepted as the best possible way for the development of meaningful typologies from a theoretical perspective because the various levels of knowledge



are analysed.

Pragmatic principles, then, govern our typology model. With the purpose of establishing common and opposing patterns that will allow us to distinguish and classify specialised technical texts more accurately, I have elaborated a model of specific features taking into account specialised bibliography (Ciapuscio and Kugel, 2002; Göpferich, 1995; Gläser, 1995; Halliday and Hasan, 1985). The typology is organised according to a multilevel dimension, taking into consideration two general analysis criteria: a) situational; and b) functional. These identify the relationship among the participants involved in a specialised communicative setting and that among the text categories established by communicative functions. Figure 1 shows the specific features resulting from the criteria mentioned above. These features will be described in detail afterwards.

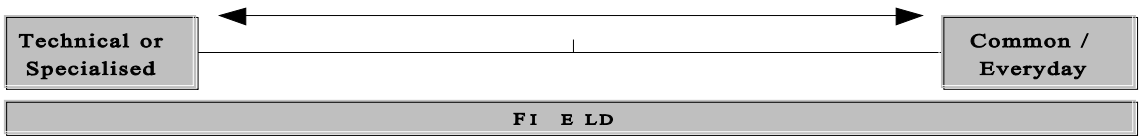
3.1. Situational criteria: elements of a communicative situation

Assuming every text has been engendered in a particular situational context, it is essential to establish what sort of features allows a better description of such context (Halliday, 1978).

In my research, I will consider three specific variables conventionally used to describe the context of situation. I am referring to field, tenor and mode of the discourse (Halliday and Hasan, 1985).

3.1.1. Field

In the process of text production and reception, Halliday & Hasan (1985: 12) develop, as it is well known, three register levels: that of field, of tenor and of mode. They define field as "what is happening, the nature of the social action that is taking place: what is it that the participants are engaged in, in which language figures as some special component?" Thus, field encompasses the social interactions happening in a wider process that includes language as part of it. In the field, a situation can be technical or everyday, or be some place in between these two poles. Eggins (1994: 73) states that the feature 'field' varies along a dimension of technicality. Therefore, it is possible to establish a first distinctive feature between general and specialised discourse, represented in figure 2.



What remains to be seen is whether or not the feature 'field' can be applied to industry, more particularly to natural stone, our subject matter. We apparently can if we refer to classification systems proposed by reliable sources such as a Thesaurus. In *Spines Thesaurus*[4], for example, under number 12, we read *Industry- Production and Distribution* and, within it *Extractive Industry*, which is the

field our study on natural stone can be placed in.

The concept ‘field’, thus, acquires the nuance of technical or specialised in contrast with general or everyday. For our purposes, the notion of ‘specialised’ will be a constant since all the written material have been generated in the natural stone domain. As I mentioned at the beginning of this paper, the descriptor ‘specialised’ is understood as a gradient, a *continuum*. Therefore, the texts will be more or less specialised depending on a variety of factors, such as topic, recipient, communicative goal, the specific situation, content abstraction level, to name some.

3.1.2. The tenor or participant relationship.

The feature ‘tenor’ is concerned with the people taking part in the process of communication, the nature of the participants as well as their status and roles. But precisely who are the participants of a specialised communication process?

Cabré (1991: 153-4) claims that only those participants who have a specific knowledge in a professional field acquired through learning can produce and intervene in the production-reception process of a specialised communication. In order to produce specialised knowledge; the specialist studies, analyses, interprets or conceptualises reality. This knowledge developed in different fields take form in the shape of oral and written texts, which are eventually circulated through diverse publication means (magazines, monographs and the like); through different formats (on paper, electronically, audio …) and also through different text types (doctoral dissertations, articles, reviews, brochures and so on).

The recipients, on the other hand, can vary. They range from experts to non-experts. Experts can be identified as the professionals whilst non-experts can be said to be laymen, the general public. A third group can be identified as being halfway in between: the semi-experts. They can be depicted as having certain in-depth knowledge about a certain subject matter.

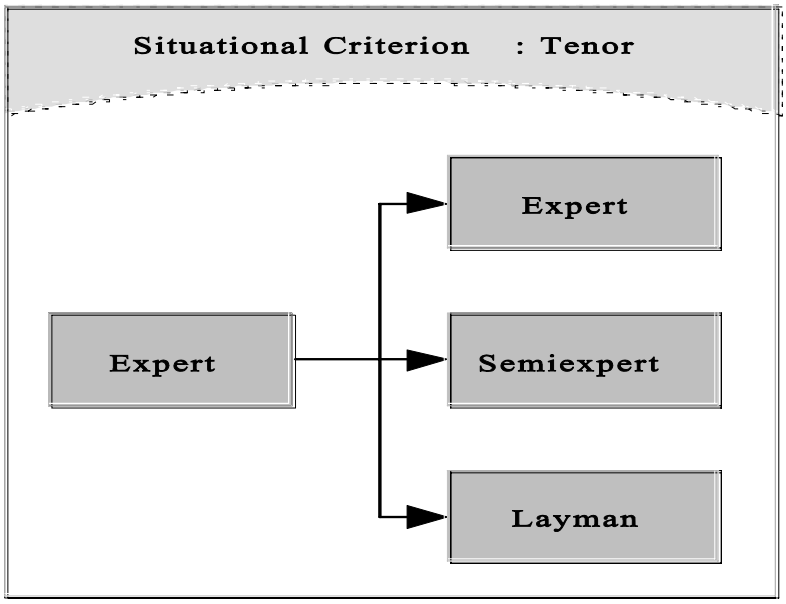
Thus, a clear restriction can be appreciated in the number of specialised knowledge producers as opposed to the general discourse producers. Furthermore, notwithstanding this restriction, one gets a wide range of variability that accounts for the degree of topic specialisation of the addressees. The professional communication will satisfy the varied audiences that respond to the different levels of expertise: the professionals of the same or of a related area, the students who major in that specific area or in a related one, layman, etc.

Next, I will establish the relationships among the participants of the communication in the selected area of this study, which is represented in figure 3.

3.1.3. Mode

The feature ‘mode’ refers to the way in which a text is originally produced, and will also add to it the function of describing the channel in which the text is produced (Sinclair & Ball, 1996: 9).

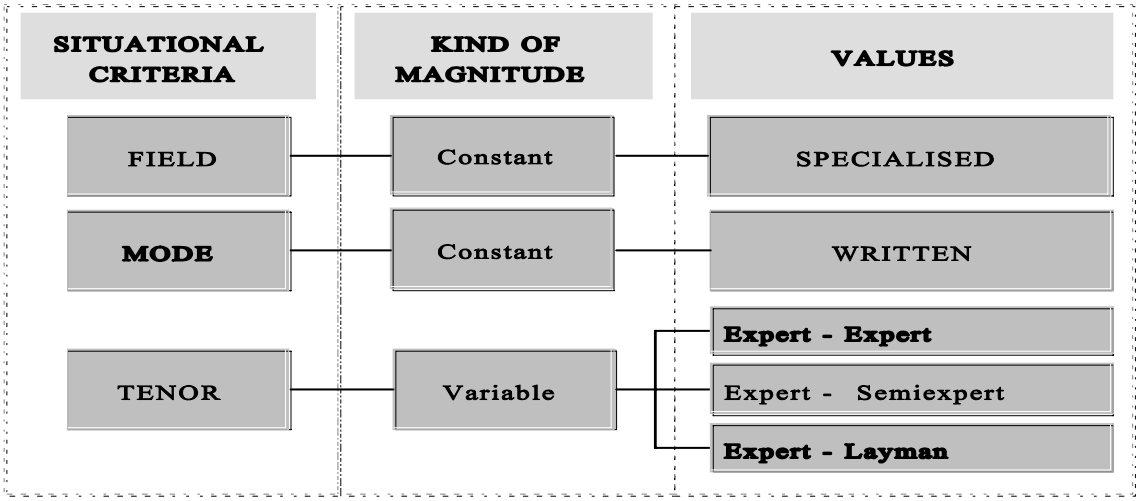
Two basic modes can be found: oral and written. However, a new mode has



been proposed in the document EAGLES (*ibid*): the electronic mode. It is added as a separate category because the authors consider that there are texts that can only be located on electronic media, such as Web pages, e-mails and the like. There are still other authors, such as Atkins *et al*. (1992), who identify as many as five different modes: written, written to be read, written to be spoken, spoken and spoken to be written.

In our project, we will only concentrate on the two main ones, oral and written, as we believe that focussing in all five will not be feasible, basically because not only will the study of the five modes will take us far too much time because it will be imperative to try to find out if a text has been created for an oral or written target in mind. Besides, we will be concerned only with written texts rather than with oral ones. Oral texts can be extremely cumbersome when being transcribed to be of much use in the actual building of a corpus given the time limitation imposed by the project. At any rate, text corpora turn out to be, without exception, written in the end.

Figure 4 summarises the above situational criteria and their distinctive features.



3.2. Functional criterion: communicative functions

Communicative functions are concerned with classifying text types or genres according to their communicative purpose. Sorting out texts, whether expositive, evaluative, informative, and the like, cannot be made without experiencing some degree of difficulty. Jackobson (1981) made us aware that any text is bound to contain a number of functions but eventually they portray a dominant one; the one that defines the main purpose of the text. More recently, Ciapuscio & Kugel (2001) support Jackobson's perception. They believe that the functions of expressing, contacting, informing and addressing are conceived in such a way that each one of them includes the previous one. For this reason, in the model used in this study I have adopted the functions established in Göperich (1995) and Gläser (1995). They have been grouped in five main functions: informative, juridical-normative, didactic-instructive, popularising and compilatory[5]. In accordance with the last of the authors mentioned above, functions are linked, on the one hand with the type of relationship established between the interlocutors (the tenor), and on the other with the resulting genre.

The functions and their corresponding genres acquire different meanings depending on the relationship established among the participants. In connection to expert-to-expert relationship, the following functions and genres have been taken into account:

- Informative texts: as their name implies, purport to transmit academic/professional knowledge with a professional audience in mind. The genres in which this function is represented —in the area considered here— are the articles of specialised journals, specialised commercial articles, project reports, technical reports, catalogues with technical specifications, yearbooks, etc.;

- *Juridical-normative texts:* their communicative function is to establish reference

standards. The information transmitted deals with norms that regulate products, services or test methods, or it can also be about the rights to exploit an invention or about recommendations destined to achieve uniformity. The genres included in this group are patents, work regulations and the test norms;

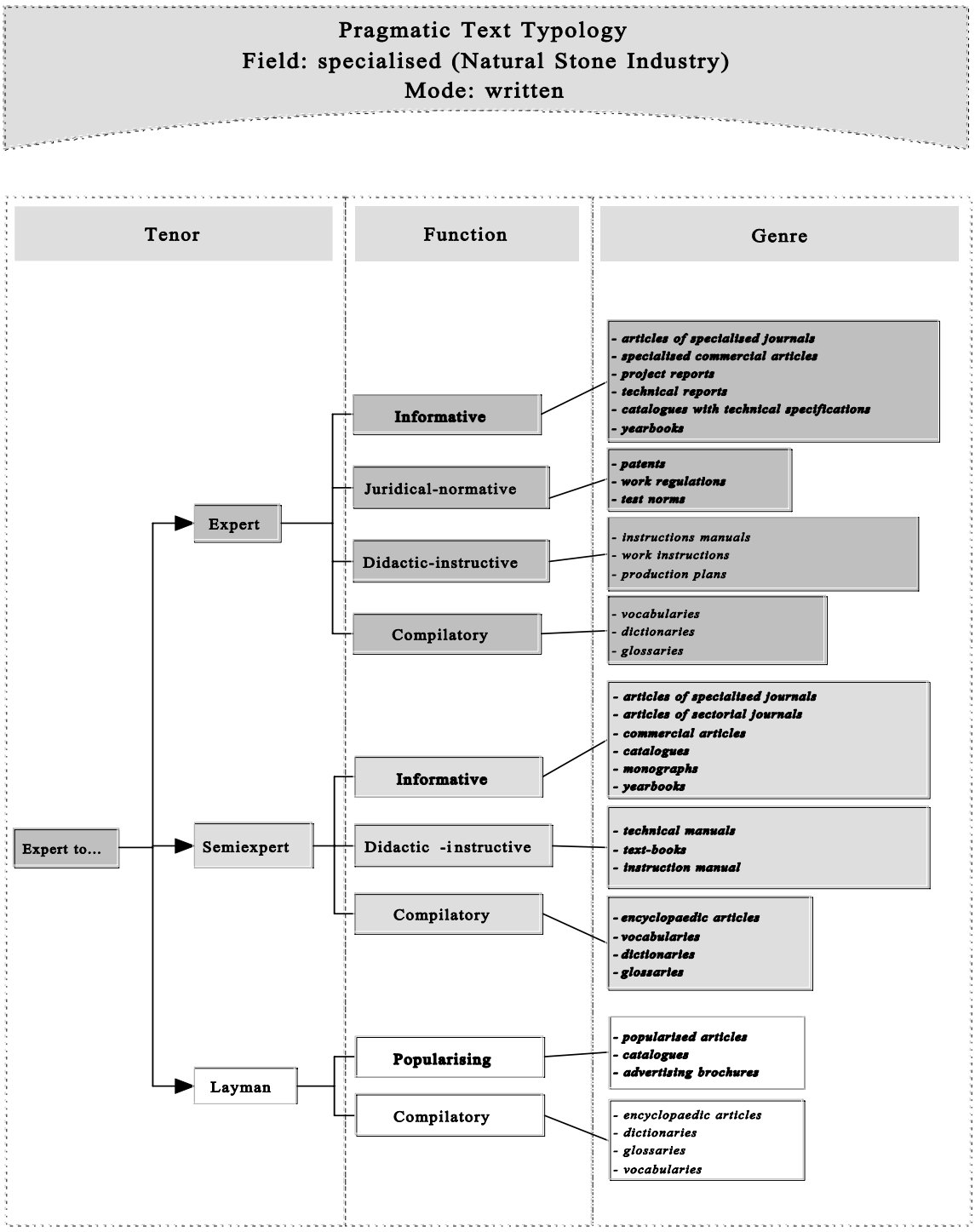
- *Didactic-instructive texts:* the purpose of the transmitted information is to illustrate a practical application. The genres taken into account here are the instruction manuals, work instructions and production plans;

- *Compilation texts*: they have a communicative function aimed at the unification and standardization of the used terminology in a given area. Examples of genres included in this classification are vocabularies, dictionaries and glossaries, especially those of a prescriptive nature.

The relationship among the participants, on the other hand, may be that between an expert and a semi-expert in which case the functions and genres are:

- *Informative texts*: they perform the same function as above with the difference that the audience exhibits an average level of specialisation. This function can be found in genres such as specialised or sectorial journal articles, commercial articles, catalogues, monographs, yearbooks, etc.;

- *Didactic-instructive texts*: they perform quite the same function as for the above group, i.e. illustrate a point, a practical point. The genres included in this category are technical manuals, text-books, and the instruction manuals (for machines, materials, anchors, assembling, maintenance, etc.);



· *Compilation texts*: they offer a general view of knowledge in an easily accessible way. Examples of genres included in this group are encyclopaedic articles, vocabularies, dictionaries and glossaries.

If the communicative function contemplates the interaction between a professional and a layman, then two types of texts are expected:

- *Popularised texts*: their communicative function consists of introducing a topic from a general perspective. Examples of genres included are popularised articles, advertising brochures, catalogues, etc.;
- *Compilation texts*: their communicative function is to offer a general and more accessible view of a specific knowledge and facilitate its access. Examples are encyclopaedic articles, vocabularies, dictionaries and glossaries, among others.

Before we move any further, I would like to underline the fact that any attempt to segment and classify reality, whether or not of linguistic nature, is downright arbitrary and artificial because it responds to our needs for understanding what surrounds us. Having said that, I think that the model proposed here, as other models, presents a subjective division of a linguistic reality that responds mainly to methodological needs and criteria. Moreover, it is not possible to divide the whole range of texts produced in a specific area and place each one of them in a well-defined and demarcated compartment. In certain cases, to assign a text to one of these compartments stated in the model may be a complex task due to the particular characteristics that the mentioned text presents.

Textual typologies vary depending on the observer; therefore different professionals (translators, documentary makers, terminologists, etc.) establish different typologies depending on their needs. This aspect illustrates that any given textual classification has a pragmatical base (Gläser, 1995). In our case, the proposed model responds to the purpose for which it was designed, i.e., to create a corpus of written texts that represented the field of natural stone and that later on could be exploited in order to obtain the different terminological units used in this sector. For that, it was necessary to establish beforehand the hierarchical structure of our population (elements chosen for this study); which is in keeping with Biber's proposal (1993). In short, it is important to define among other things, the different text types present in a given population.

Figure 5 above shows the pragmatic text typology for the field under consideration, the natural stone.

4. Conclusion

The above model should be considered as a work tool useful for the initial selection and evaluation of written texts that will make up a specialised corpus for terminographical purposes. I believe it offers valid criteria for the formation of a corpus from which we are going to extract terminology because it anticipates relations of different levels of specialisation among the participants of the communication. This parameter allows to establish different aims of the discourse

and to predict textual results because the conjunction of the situational and functional criteria determines linguistic resources that an author will use in order to convey the knowledge and organisation of the text.

In a descriptive corpus-based terminography it is essential to establish various text types with different levels of specialisation because each type provides linguistic, pragmatical and conceptual information of a varied nature. Written texts by a specialist aimed at another specialist with the highest level of specialisation will be very different from the texts aimed at a lay reader in the same subject, placed in the lowest scale level because the terms and concepts used will not be the same. Thus, each level of specialised texts shows different phenomena linked to terms. I am referring here to the variation in denomination; that is to say, the use of synonyms and paraphrases, or explanations, definitions, etc.; in short, linguistic information that is very useful in descriptive terminography.

All in all, I consider that the introduced model helps speed up the compilation of a corpus. However, it has its limitations, as it reduces the varied range of discursive parameters needed to explain in depth the complex topics of study found in specialised texts. It will be necessary to continue research in order to go into a more exhaustive textual analysis matrix in greater depth, and prove, at the same time, the usefulness of the model in other projects of specialised corpus belonging to different professional and academic areas. This will allow us to make the necessary adjustments in order to enlarge its validity.

Notes

1. Apart from these references, it is possible to consult specific information on the design of different corpora, such as Brown, LOB, FLOB, FROWN, among others, on ICAME:

[http://khnt.hit.uib.no/icame/manuals/].

2. The Lancaster-Oslo/ Bergen, the so-called the LOB, is the result of the cooperation of Leech (Lancaster University), of Johansson (U. of Oslo) and of the Norwegian Computing Centre for the Humanities in Bergen. It contains one million words and the written samples date back to 1961. It deals with British English. Its compilation covers the years 1970-1978. For an exhaustive list of textual categories in this corpus you can visit the following web page:

[http://helmer.aksis.uib.no/icame/lob/lob-dir.htm#lob1].

3. When building up a corpus for linguistic analysis, many researchers (Atkins *et al.*, 1992; Sinclair *et al*., 1996; Pearson, 1998; Sinclair, 2003, among others) emphasise the importance of two types of linguistic criteria for the selection and classification of texts found in the corpus: the internal and the external ones.

4. From a structural point of view, a thesaurus is defined as controlled and dynamic vocabulary lists with semantic and generic relationship to be applicable to any domain. From a functional point of view, a thesaurus is a terminology control tool used to provide the natural language in a document with a precise language. It is also used to classify the different disciplines, which is precisely the use we have made of it. More information about Spines Thesaurus can be found at:

[http://pci204.cindoc.csic.es/tesauros/SpinTes/Spines.htm].

5. The texts sorted out under compilation type -vocabulary lists, glossaries, dictionaries, etc.- could very well belong to another group purporting another function. For example, they can be placed under normative or prescriptive texts when the situation is that they have been produced in an expert-to-expert environment because they may behave as a standardisation tool. They can also be placed under popularisation texts when addressed to laymen with the object of delivering complex notions in an easy way.

Works Cited

Ahmad, K. (1993): *A "Pragmatic" Approach for Representing Terminology*. University of Surrey, CS-93-13.

Ahmad, K. & M. Rogers, (2001): "Corpus Linguistics and Terminology Extraction". In Wright, S.E. & G. Budin, eds., *Handbook of Terminology Management*. Amsterdam/Philadelphia: John Benjamins, vol.2, 725-760.

Alcaraz Varó, E. (1990): *Tres paradigmas de la investigación lingüística.* Alcoy: Marfil.

_____. (2000): *El inglés profesional y académico.* Madrid: Alianza Editorial.

Atkins, B.T.S., J. Clear, & N. Ostler (1992): "Corpus Design Criteria". *Literary and Linguistic Computing*, 7(1): 1-16.

Bergenholt, H. & S. Tarp (1995): *Manual of Specialised Lexicography: the Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: John Benjamins.

Bhatia, V. K. (1993): *Analysing Genre: study of its application to professional genres*. London: Cambridge University Press.

Biber, D. (1993): "Representativeness in Corpus Design". *Literary and Linguistic Computing*, 8(4): 243-257.

Biber, D., S. Conrad, & R. Reppen (1998): *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Bowker, L. (1996): "Towards a Corpus-Based Approach to Terminography". *Terminology*, 3(1): 27-52.

Bowker, L. & J. Pearson (2002): *Working with Specialized Language. A practical guide to using corpora*. London/New York: Routledge.

Cabré, M.T. (1999): *La terminología: representación y comunicación*. Barcelona: Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.

Ciapuscio, G. E. (1994): *Tipos textuales*. Buenos Aires: Eudeba.

_____. (2003): *Textos especializados y terminología*. Barcelona: Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.

Ciapuscio, G.E. & I. Kugel (2002): "Hacia una tipología del discurso especializado: aspectos teóricos y aplicados. In García Palacios, J. & M.T. Fuentes Morán, eds., *Texto, terminología y traducción*. Salamanca: Ediciones Almar, 37-73.

Clear, J. (1992): "Corpus Sampling". In Leitner, G., ed., *New Directions in English Language Corpora*. Berlin/New York: Mouton de Gruyter, 21-31.

De Beaugrande, R., & W. Dressler (1981): *Introducción a la Lingüística del Texto.* Barcelona: Ariel.

Eggins, S. (1994): *An Introduction to Systemic Functional Linguistics*. London: Pinter Publishers, Ltd.

Engwall, G. (1994): "Not Chance, but Choice: Criteria in Corpus Creation". In Atkins, B.T.S. & A. Zampolli, eds., *Computational Approaches to the Lexicon*. Oxford: Oxford

University Press, 49-82.
Fillmore, Ch. J. (1992): "'Corpus linguistics' or 'Computer-aided armchair linguistics'". In Svartvik, J., ed., *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*. Stockholm, August 1991. Berlin/NewYork: Mouton de Gruyter, 35-60.
Francis, W. N. (1992): "Language Corpora B.C.". In Svartvik, J., ed., *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*. Stockholm, August 1991, Berlin/New York: Mouton de Gruyter, 17-32.
Gläser, R. (1995): *Linguistic Features and Genre Profiles of Scientific English*. Frankfurt am Main: Peter Lang.
Gómez, A. & Ch. Vargas (2004): "Aspectos metodológicos para la elaboración de diccionarios especializados bilingües destinados al traductor". In *Actas del II Congreso "El español, lengua de traducción"*. Bruselas: ESLEtRA, 365-398. Available online: [http://www.toledo2004.net/html/contribuciones/gomez-vargas.htm].
Göpferich, S. (1995): "A Pragmatic Classification of LSP Texts in Science and Technology". *Target*, 7(2): 305-326.
_____. (2000): "Analysing LSP Genres (Text Types): From Perpetuation to Optimization in Text(-type) Linguistics". In Trosborg, A., ed., *Analysing Professional Genres*, Amsterdam/Philadelphia: John Benjamins, 227-247.
Halliday, M. A. K. (1978): *Language as a social semiotics: The social interpretation of language and meaning*. London: Arnold.
Halliday, M. A. K. & R. Hasan, (1985): *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective.* Oxford: Oxford University Press.
Isenberg, H. (1987): "Cuestiones fundamentales de tipología textual". In Bernárdez, E., comp., *Lingüística del texto*. Madrid: Arco/Libros, 95-129.
Jakobson, R. (1981): *Lingüística y poética*. Madrid: Cátedra.
Kennedy, G. D. (1998): *An Introduction to Corpus Linguistics*. London/New York: Longman.
Leech, G. (1991): "The state of the art in corpus linguistics". In Aijmer, K. & B. Altenberg, eds., *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman: 8-29. Available online: [http://angli02.kgw.tu-berlin.de/corpus/art.htm].
_____. (1992a): "Corpora and theories of linguistic performance". In Svartvik, J., ed., *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, 105-122.
_____. (1992b): "100 million words of English: the British National Corpus". *Language Research* 28(1): 1-13.
_____. (2002): "The Importance of Reference Corpora". In *Corpus lingüísticos. Presente y futuro*. Donostia: UZEI. Available online: [http://www.uzei.org/corpusajardunaldia/06_gleech.pdf].
McEnery, T. & A. Wilson (1996): *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Meyer, I. & K. Mackintosh (1996): "The Corpus from a Terminographer's Viewpoint". *International Journal of Corpus Linguistics* 1(2): 257-285.
Pearson, J. (1998): *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.
Schröder, H. (1991): "Linguistic and Text-theoretical Research on Languages for Special Purposes. A thematic and bibliographical guide". In Schröder, H., ed. *Subject-oriented Texts. Languages for Special Purposes and Text Theory*. Berlin/New York: Walter de Gruyter, 1-48.
Sinclair, J. (ed.) (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 ELT.

_____. (1996): *Preliminary recommendations on Corpus Typology*. EAG-TCWG-CTYP/P. May 1996, Pisa: EAGLES. Available online: [http://citeseer.ist.psu.edu/430619.html].

_____. (2003): "Corpora for lexicography". In Sterkenburg, P. ed., *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins, 167-178.

Sinclair, J. & J. Ball, (1996): *Preliminary Recommendations on Text Typology*. EAG-TCWG-TTYP/P, June 1996. Pisa: EAGLES. Available online: [http://citeseer.ist.psu.edu/437698.html].

Summers, D. (1993): "Longman/Lancaster English Language Corpus -Criteria and Design". *International Journal of Lexicography*, 6(3): 181-208.

Stubbs, M. (1996): *Text and Corpus Analysis. Computer-assisted Studies of Language and Culture*. Oxford/Cambridge (MA): Blackwell Publishers.

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

Tognini-Bonelli, E. (2001): *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.

Trosborg, A. (1997): *Text Typology and Translation*. Amsterdam/Philadelphia: John Benjamins.

_____ (ed.) (2000): *Analysing Professional Genres*. Amsterdam/Philadelphia: John Benjamins.