4. Corpus-based terminography

4.1. Using electronic text corpora in terminology work

As described in Section 3, traditional terminology argues for an onomasiological approach to gathering terminological data. Not only have objections to this method been raised, but Temmerman also claims that

"even though – in practice – terminographers have always started from understanding as they had to rely on textual material for their terminological analysis, one of the principles of traditional Terminology required them to (artificially) pretend that they were starting from concepts" (Temmerman 2000: 230; original emphasis).

In any case, terminographers have always had to compile extensive collections of texts, e.g. books, papers and other types of evidence for the language used, in order to acquire knowledge about the subject domain, to familiarise themselves with the conceptual system as well as to identify and extract the key terms (Cabré 1999: 117). While this approach is still being followed, especially in the initial stages of a terminology project, terminographers are increasingly taking advantage of computers to facilitate, accelerate and improve their work.

Cabré points out that computer science has provided resources and tools that are of use in practically all stages of terminological work (Cabré 1999: 164). As well as enabling terminologists to store terminological data electronically, computers allow the creation of an electronic type of terminological dictionary, which is usually referred to as a term bank. Furthermore, terminologists may compile electronic text corpora which they can then explore in order to gather and extract terminological information. The latter approach may be referred to as corpus-based terminology and is described in detail by Gamper and Stock, who define it as "a working method which explores a collection of domain-specific language material (corpus) to investigate terminological issues" (Gamper and Stock 1998/1999: 149).

Ahmad and Rogers have identified three main tasks for which electronic text corpora may be used to assist terminologists, namely to capture, validate and elaborate data (Ahmad and Rogers 2001: 740). Accordingly, corpora support the terminologist throughout a terminology project, both in the early stages when the key issues are to identify term candidates (i.e. to capture data) and to provide evidence for and about term candidates (i.e. to validate data), as well as in the core stages when the main tasks are to compile definitions and to select contextual examples (i.e. to elaborate data).
The fact that the use of corpora in terminology work is now generally accepted is partly a result of the increasing availability of electronic texts, enabling the terminologist to access large amounts of language data in order to collect the words and phrases belonging to a particular subject field. Making use of current language data also implies that "the prescriptive view of terminology work has given way to a more descriptive approach" (Maia 2002: n.pag.), an idea also discussed by Teubert, who defines descriptive terminology as "the kind of terminology work which is based upon current language use" (2003: 103). According to Teubert, the focus has shifted from hard terminology, with terms representing static concepts precisely described, to soft terminology without "binding definitions but contextually constrained attempts at definitions of temporary validity" (Teubert 2003: 104). He argues that the change has been brought about by the methodology of corpus linguistics and emphasises the importance of the internet as a "virtual corpus" (Teubert 2003: 103), which enables terminologists to extract terms from dynamic domain-specific special corpora.

The use of corpora in terminology has also been enhanced by recent developments in terminology.15 According to Sager,

"the increasing tendency to analyse terminology in its communicative, i.e. linguistic context, leads to a number of new theoretical assumptions and also to new methods of compilation and representation" (Sager 1990: 58).

This view is supported by sociocognitive terminology, which, in contrast to traditional terminology, accounts for the communicative aspect of language and therefore argues for the study of terminology in real language, such as texts written by domain specialists (Temmerman 2000: 16-17). Authentic texts provide information on how the world and the lexical items that are used to communicate about the world are understood (Temmerman 2000: 40).

### 4.2. Advantages of corpus-based terminography

While machine-readable corpora have been accepted in lexicography and language for general-purpose work for some time, their use and popularity in terminography or language for special-purpose work have been lagging behind. Arguing for their use, Bowker (1996: 30-31) points out three main

---

15 See Section 3.2.3
advantages of corpora in terminology, an approach which she refers to as the corpus-based approach to terminography or simply corpus-based terminography.

Firstly, machine-readable corpora enable terminologists to increase both the speed and the scope of their research. Not only can larger quantities of data be processed more rapidly, thereby exposing terminologists to a larger number of conceptual descriptions, but corpora also allow them to leave out the sections of a text that are terminologically irrelevant and to focus on those parts which are of interest from a terminological point of view (Bowker 1996: 31-32). The latter parts may be referred to as knowledge-rich contexts, containing "at least one item of domain knowledge that could be useful for conceptual analysis" (Meyer 2001: 281).

Secondly, in contrast to conventional term banks, which contain hardly any examples of terms in context, corpora present a variety of contexts and/or more extensive contexts which not only provide valuable supplementary information but also help to understand and use terms more effectively (Bowker 1996: 32-33).

Thirdly, a machine-readable corpus makes it easier to investigate syntactic and semantic information as well as linguistic patterns which are difficult to discover when scanning texts manually. For example, terminographers can look at concordances, also referred to as key words in context (KWIC), in order to reveal collocational information that may help to improve the use of terms immensely (Bowker 1996: 32). This aspect is of particular importance as it highlights the fundamental idea of working with corpora, best described in the words of the father of corpus linguistics, John Sinclair: "The ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before" (Sinclair 1991: 4). This idea is relevant for this study in two respects. As mentioned above, information on collocations shows how terms can be used. In addition, concordances enable the study of the relation between language and ideology. In his analyses of texts and text corpora, Stubbs shows that they can reveal patterns of language that institutions use to build up our linguistic, conceptual and ideological view of the world (Stubbs 1996: 59). According to Hunston (2002: 109), such patterns may convey messages implicitly, with the reader being neither intuitively nor consciously aware of them.
4.3. Building terminological corpora

Sinclair (2004a: 79) describes corpus building as a process comprising two stages, corpus design and implementation, which for practical reasons cannot be completely separated. While in the design stage corpus builders establish the principles on which the structure of the corpus will be based, the implementation stage requires them to put these principles into practice.

In the corpus design phase, the corpus builder has to determine the criteria on which the texts that form the corpus will be selected. These criteria include, but are not restricted to, the text mode (e.g. spoken or written language), the type of text (e.g. books, journals or reports), the domain of the text, the language (or languages) and the date of the text (Sinclair 2004b: 4).

The implementation stage, also referred to as the stage of text collection or capture, is concerned with the actual selection of the texts that are to be included in the corpus. The easiest and most common way of collecting corpus texts is the use of data that already exist in electronic format and can be found on websites or in internet archives (Baker 2006: 31). If existing electronic sources are not available, texts can be captured by scanning them or, as a last resort, by keying them in manually (Baker 2006: 34).

The following general design issues have been identified as being relevant for terminological corpora.

- The content of the corpus will naturally be determined by the specific purpose and the objective of the research. By raising a particular research question or, in a terminology project, by deciding on a specific domain to be investigated, the corpus builder establishes one of the key design criteria.

- The corpus content may be restricted by the availability of resources in a machine-readable form as well as the copyright of the materials (Bowker 1996: 40).

- Another key issue in corpus design is corpus size, i.e. the number of words the corpus is made up of. Baker points out that corpus size is basically dependent on the type of language under consideration (2006: 31). According to Ahmad and Rogers (2001: 735-736), special-language corpora may be smaller than general-language corpora, putting forward two main arguments to justify smaller corpora in terminology work. First, they state that "the specific purpose of the terminology that is being compiled may naturally restrict the text type [...] selected for the corpus" (2001: 736).
Second, as special-language texts usually show less lexical and grammatical variation than general-language texts, patterns of language may be seen more clearly in smaller language samples. Consequently, they believe terminological corpora start to become useful in the tens of thousands of words and consider a corpus size of approximately 100,000 words a "good starting point for corpus-based terminology management in a highly specialized discipline" (2001: 735). While other research is less explicit about corpus size, the idea of special-language corpora requiring significantly fewer words than general-language corpora is generally accepted (Baker 2006: 29; Bergenholtz and Tarp 1995: 95-96; Bowker 1996: 42; Meyer and Mackintosh 1996: 268). This idea is succinctly described by Varantola, who observes that the "my-corpus-is-bigger-than-yours" rhetoric has given way to a "my-corpus-is-smarter-than-yours" rhetoric (2002: 174).

- The length of the individual texts that form the corpus represents another basic issue in building a terminological corpus. Texts selected for a general-language corpus may be samples of a specified size (e.g. 2,000 words) taken from complete texts. Special-language corpora, however, must not have any restrictions on the length of individual texts but should consist only of complete texts. This characteristic is also referred to as text integrality (Meyer and Mackintosh 1996: 268). Using full text corpora, as this type of corpus is sometimes also called (Kennedy 1998: 21), is of vital importance, the more so as terms and conceptual descriptions may appear anywhere in a given text (Bowker 1996: 42-43). By selecting only text samples the terminologist risks missing valuable information.

- The corpus builder has to decide between a closed (or static) corpus, which, once completed, does not change in size, and an open (or dynamic) corpus, which may be continually changing in size (Baker, Hardie and McEnery 2006: 64). Considering the speed with which terminology changes, a terminological corpus should be open, enabling the corpus builder to add new texts and delete obsolete documents when the need arises (Meyer and Mackintosh 1996: 269).

- Another important design issue for terminological corpora is the extent of domain coverage they aim at. Before starting to select documents for a corpus, the corpus builder must delimit the domain that the corpus is intended to represent. This task may prove extremely difficult, especially as domain boundaries are hardly ever unambiguous and well-defined (Meyer and Mackintosh 1996: 260 and 268). Once the boundaries of the domain have been clearly defined, the corpus builder has to compile texts with the
objective of creating a corpus that covers all the aspects of a domain, including subdomains and related domains, as adequately as possible (Meyer and Mackintosh 1996: 270). Bowker (1996: 45) refers to this idea as **conceptual balance**.

- Like other types of corpora, a terminological corpus is to be balanced and representative. The issue of creating balanced and representative corpora has been occupying corpus linguists for decades. While, according to Sinclair, representativeness and balance "are not precisely definable and attainable goals", they should still be seen as target notions, guiding the corpus builder in his task (Sinclair 2004b: 9). Although it may be easier to judge the balance of a special-language corpus than that of a general-language corpus (Ahmad and Rogers 2001: 734), the question of what actually determines the balance and representativeness of terminological corpora remains open (Bowker 1996: 44). Generally speaking, a corpus may be considered as balanced if it "contains texts from a wide range of different language genres and text domains" (Baker, Hardie and McEnery 2006: 18) or, according to another definition, if the proportions of different kinds of text it contains "correspond with informed and intuitive judgements" (Sinclair 2004b: 8). For terminological corpora, Bowker (1996: 44) recommends covering the terminology of the domain as exhaustively as possible and therefore using many different types of documents as sources of terminological information.

- Finally, terminologists also have to pay attention to the types of texts used as the basis for their work. As mentioned above, terminologists draw on different types of documents in order to obtain terminological information. According to Meyer and Mackintosh (1996: 270), the range of text types is highly dependent on the field of study since different domains produce different text types. They point out that terminologists also have to account for the variation in text type within a single domain, and in particular, the variation in the technicality of texts, including texts written by and/or for experts and which are therefore highly technical, as well as texts written by and/or for non-experts (Meyer and Mackintosh 1996: 270).

### 4.4. Analysing corpora

The analysis of the corpora compiled for the purpose of this study is assisted by concordance software. The WordSmith tool family represents an integrated
set of programs for looking at how words are used in texts and consists of three major instruments, viz. WordList, KeyWord and Concord.

According to Hunston (2002: 67), key words are a valuable starting point for analysing specialised corpora. The identification of key words, the words which may be considered key, requires the generation of a word list, which is basically a list of all the distinct words in a corpus showing the number of occurrences of every word with the possibility of sorting them by frequency or in alphabetic order. In a first step, therefore, a word list of the corpus under investigation is created using WordSmith's WordList.

The keyness of a word in a text or collection of texts may be characterised in terms of importance and "aboutness" (Scott 2007: 3-4), in the sense that it indicates that the word is important and shows what the text is about, respectively. Scott and Tribble define "keyness" as "what the text 'boils down to' [...], once we have steamed off the verbiage, the adornment, the blah blah blah" (Scott and Tribble 2006: 56). WordSmith calculates key words by comparing the frequency of each word in the word list of a smaller, more specialised corpus with the frequency of a larger, more general one and lists the key words for the former (Hunston 2002: 68). More precisely, WordSmith's KeyWord function is used to compare the word list with the word list of the British National Corpus (BNC), which is used as a reference corpus. For every word in the corpus under investigation, KeyWord contrasts the patterns of frequency and calculates a keyness score. It has been decided to set a maximum of 800 key words as this number is deemed to be sufficient for the analysis. Furthermore, it is advisable to work through the initial list of key words in order to remove noise as well as words which are clearly not relevant from a terminological point of view, viz. grammatical words (e.g. articles, conjunctions, prepositions) and words that are characteristic of legal texts (e.g. shall, article, paragraph).

---

16 The calculation of the keyness of a word involves computing its frequency in the smaller corpus, the number of words in the small word list, its frequency in the reference corpus and the number of running words in the reference corpus and then cross-tabulating these numbers. WordSmith uses several statistical tests, including the classic chi-square test of significance and Ted Dunning's Log Likelihood test (Scott 2004-2007: 124).

17 Due to their above-average use in some of the texts, these words may appear key when compared with the words in the reference corpus. They can, however, be excluded from the terminological analysis.
This list does not, as one might assume, represent a final list of terms that require terminological definitions or that are suited for inclusion in a terminological dictionary. It can, however, be extremely useful as it offers an overview of the main subjects covered in the texts and also provides the starting point for further analysis, in particular in connection with the calculation of word clusters.

Word clusters may be defined as "words which are found repeatedly together in each others' [sic!] company, in sequence" (Scott 2004-2007: 225). While forming a tighter relationship than collocates, clusters merely represent repeated strings which may or may not turn out to be true multi-word units (Scott 2007: 19). Biber et al., who refer to clusters as *lexical bundles*, describe them as sequences of words that show a statistical tendency to co-occur in a register (2000: 989). WordSmith offers two approaches to the identification of word clusters, using either Concord or WordList. They vary in that Concord only processes concordance lines, whereas WordList processes whole texts (Scott 2004-2007: 225), and their results therefore also differ to some extent. Both approaches require the user to specify the cluster size (between two and eight words) and a minimum frequency, i.e. a minimum number for the cluster to appear in the results. In this analysis 18, WordList is used to generate the word clusters, with a cluster size of two to six words and a minimum frequency of five as the key parameters.

As the calculation of clusters only yields sequences of words that tend to co-occur, the results have to be revised. This step includes the elimination of those clusters that are clearly nothing more than repeated strings, and the identification of related clusters which Scott describes as clusters "which overlap to some extent with others" (2004-2007: 89). Related clusters that form part of more comprehensive clusters are removed unless they are considered to have a meaning that is independent from the meaning of the latter and occur in the corpus at least five times.19 The aim of this procedure is to generate a list of multi-word units which represent term candidates in the sense that they are relevant from a terminological point of view and considered to have a separate meaning.

---

18 See Section 5

19 In cases where word sequences that form part of more than one cluster are included in the results, the frequency counts are distorted as the same word sequence is counted more often than once.
Along the lines of Mahlberg (2007: 198-199), who establishes groups in order to categorise concordances, the word clusters can be divided into several categories, each of which characterises a particular theme prevailing in the corpus texts. Despite being a rough approach to analysing clusters, this step facilitates the identification of the main characteristics and themes of the underlying texts and makes it easier to grasp the plurality of terms which include the main key words. Moreover, the establishment of groups enables a focused view of the various word clusters and assists in raising issues and questions that would not have come to mind otherwise. Mahlberg refers to these groups as *functional groups*, admitting that these categories are neither watertight nor absolutely clear-cut (2007: 199). She also points out that the labels introduced for the functional groups represent so-called *ad hoc labels*, which aim at nothing more than showing the typical characteristics of the group (Mahlberg 2007: 199-200). Unlike Mahlberg, who is interested in the features of discourse rather than the terminology used in texts, in this study only those multi-word units are categorised that can be considered to have a separate meaning and appear – to varying extents – useful from a terminological angle. Thus, the term *functional group* is replaced with the expression *terminological domain*.

Another key step in the analysis of corpora is the generation of concordance lines, for which the WordSmith concordancer (Concord) will be used. To this end, it is necessary to specify a search word and select a corpus, i.e. the text files in which the program will perform its search. The result is presented in the form of a concordance which displays all the occurrences of the search word in the corpus (Scott 2004-2007: 79). The idea of a concordancer is to see numerous examples of a word or phrase in context. In a concordance the search term is usually displayed in the centre with context to the left and to the right of the key word, therefore revealing collocational information on the words in a corpus (Kennedy 1998: 251). Collocations may be described as the "meaning relations that a word contracts with other words occurring in the same sentence or text" and are concerned with "meaning arising from predictable co-occurrence" (Jackson and Zé Amvela 2007: 131), or, as John Rupert Firth so ingeniously puts it: "You shall know a word by the company it keeps" (Firth 1957: 11, quoted in Stubbs 1996: 35). Collocations may be observed informally but it is sensible to rely on statistical calculations which WordSmith can also provide (Hunston 2002: 68). This supports the theory that a corpus can provide conceptual information, for the context surrounding a term may hold a definition or a description of the key characteristics of the
concept behind the term (Bowker and Pearson 2002: 38). As discussed above, a
corpus-based approach to terminology offers several advantages over
conventional methods of terminology compilation. In summary, it gives a
greater guarantee of thematic completeness and coherence, reveals
information about the linguistic behaviour of terms and presents textual
variants by showing terms in several contexts (Sager 1990: 132 and 142). Three
types of term-specific information can be derived from the corpus: information
about the concept, i.e. the definition; information about the term, e.g.
grammatical features, collocations; and information about usage, e.g. contexts
(Sager 1990: 133). The definition represents the description of a concept
specifying the characteristics which convey the meaning of the concept (Sager
1990: 39). Following Temmerman (2000: 73 and 123), what has traditionally
been referred to as the concept can be seen as a unit of understanding which, in
order to be understood, requires a definition made up of information that is
essential for understanding. As "what is more essential information for the
understanding in one situation or domain [...] is less essential information in
other circumstances" (Temmerman 2000: 123), this study adopts a flexible
approach to definition, bearing in mind the objective of achieving optimal
understanding.