Chapter 3

Corpora and corpus linguistics

3.1 Introduction

The aim of the present chapter is to sketch an organic framework within which to understand the materials and methods used in the current research and which are described in Chapter 5. Consequently, the chapter provides an introduction to corpora and corpus linguistics, with an overview of some major issues. This is not intended as a complete list of all possible topics connected to corpora and corpus analysis; in fact, a selection of topics has been made, on the basis of their relevance to the current work.

3.2 What is a corpus and what is corpus linguistics

As any introductory book to corpus linguistics explains, the word *corpus* has always been used by linguists to indicate 'a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study' (Hunston, 2002, p. 2). With the advent of computers and the development of modern corpus linguistics, the word *corpus* has come to acquire the more specialised meaning of a collection of *electronic* texts, selected and collected for a *specific purpose* according to *specific criteria*. Furthermore, as McEnery and Wilson (2001, p. 32) declare, in linguistics "there is also often a tacit understanding that a corpus constitutes a standard reference for the language variety which it represents". And this thought is most certainly what has long driven linguists in the creation of very large corpora, and in the development of coding standards so that their corpora could be shared within the academic community. However, as the review of studies in Chapters 2 and 4 may be taken to show, when the focus is on features that fall outside the realm of purely linguistic events, such as culture, attitude or behaviour, analyses are performed on collections of data (i.e. corpora) which are intended for a one-off use. For this reason, and being my work centred around culture conveyed by language, rather than language *per se*, I would support here a fairly broad definition of corpus, encompassing only two major basic features: an electronic format, connected to the use of computerised analysis tools; and the idea that a corpus should be designed, i.e. planned for some (general or specific) purpose.

Corpus design entails the application of selection and sampling criteria according to the purpose of the analysis, as well as issues of size, balance, and representativeness. These topics will be dealt with in Section 3.3.1. Furthermore,

careful design is an essential prerequisite for the applicability of quantitative methods of analysis and for the generalisability of the results. Key features and methods in the qualitative and quantitative analysis of corpora – such as word lists and frequency; keyword lists and keyness; collocation; semantic preference; and semantic prosody – will be described in Section 3.6.

Finally, the electronic format and the use of computerised tools allow non-linear access to the information in a corpus. This represents a completely different approach to and provides a new perspective on both the language and the content of a corpus. Indeed, while some linguists consider corpus linguistics simply a method of research, others regard it as a new discipline.

In 2004, at the ICAME[1] conference in Verona, during a general meeting on the topic "Corpus linguistics 25 years on", one particular linguist sounded offended by the fact that corpus linguistics was considered a method, and advocated that it was a real discipline. Her reaction might have been due to the fact that to some extent methodological interests are, by some, still considered Cinderellas with respect to theoretical interests (Leech, 1992, p. 105; Aarts, 2000, p. 7). However, it may be explained in more concrete terms by remembering that the advent of corpora and concordancing tools has changed the way we look at language and has deprived the native speaker of his exclusive status of judge and descriptor of the language. Historically speaking, corpus linguistics and the phraseological view of language it carried along was a radical turn from traditional prescriptive grammars, but also from Chomsky's generative grammar and distinction between competence and performance.[2]

A look at some definitions of corpus linguistics provided in books and articles shows that the corpus linguistic community is indeed divided between considering corpus linguistics a method (see, for example, Kennedy, 1998; and Svartvik, 2007) or a discipline (e.g. Mahlberg, 2007).[3]

A fact is that, thanks to corpus studies, new and powerful theoretical views about language have emerged, such as the notions of local grammar (Barnbrook & Sinclair, 1995; Hunston & Sinclair, 2000) and pattern grammar (Hunston & Francis, 2000), or Hoey's (2005) theory of lexical priming. Another fact is that corpus data and computerised analytical methods have been more and more used not only in linguistics, but also in other disciplines, such as the social sciences, psychology,[4] and marketing.

[1] The International Computer Archive of Modern English.

[2] Stubbs (2007a, p. 133) explains this opposition by describing Saussure's and Chomsky's approaches as rationalist deductive views situated within the French tradition of dualism, while the perspective adopted by most text and corpus linguists including Firth, Halliday and Sinclair is an empiricist inductive view which rejects dualism and is situated within the British empiricist tradition.

[3] Interestingly, Mahlberg (2007) equals this debate with Tognini Bonelli's (2001) distinction between 'corpus-based' and 'corpus-driven', whereby a corpus-based approach avails itself of corpus data to exemplify, clarify and illustrate existing linguistic theories, while a corpus-driven approach analyses corpus data and lets the data drive the description of language or of a specific linguistic event.

[4] See for example the high number of publications by psychologists using data from the CHILDES database (http://talkbank.org/usage/childesbib.pdf), or studies such as Hogenraad (2004), or Hogenraad (2005).

3.3 Creating a corpus

The creation of a corpus entails two phases: planning the design of the corpus; and collecting the necessary texts. These two phases, which are equally important and to several extents problematic, are interrelated, and eventually depend on the ways (both technical and theoretical) in which the corpus will be analysed. The following paragraphs will discuss some major theoretical issues, with an eye to the needs of the current research.

3.3.1 Corpus design

The design of a corpus depends on the purpose for which the corpus is created. If a corpus is created with the only purpose of showing students how to use corpus analysis tools, or "to encourage learners to investigate language data for themselves, the precise contents of that corpus may be relatively unimportant" (Hunston, 2002, p. 27). In most other cases, however, and in particular, when a corpus is created for the purpose of investigating a particular 'type of language' or linguistic event (e.g. British English vs. American English; the language of 19th century popular papers; the phraseology used in English civil law; or East London teenage jargon), the contents of the corpus are important and issues such as size, representativeness, balance, and sampling are usually called into play.

3.3.1.1 Size

In the last decades, several 'general purpose' corpora, such as the Brown and the LOB corpora (first generation corpora), the British National Corpus and the Bank of English (second generation corpora), have been assembled. Aiming to be representative of language in general, these corpora were created so as to include a wide variety of texts and text types, both written and spoken, and tended to be as large as possible. First generation corpora reached the important target of 1 million words – a great achievement for the time, given the then limited technological resources. But second generation general purpose corpora aim to be several hundred million words.

Indeed, the larger the corpus, the easier it is to retrieve a reasonable number of hits for infrequent or rare linguistic events (McEnery & Wilson, 2001). Furthermore, a very large corpus may also be required to understand the rationale behind grammatical or lexical forms even when they are highly frequent (see for example Mair, 2006 on the size needed for a full study of *get* in English passive constructions; or Granath, 2007 on the size needed to explain the four possible sentence structures after initial *thus*). On the other hand, extremely frequent events, such as function words and auxiliaries, can be easily retrieved in a statistically significant number of hits even in smaller corpora (see for example Biber & Finegan, 1991; Carter & McCarthy, 1995).

Biber (1993, pp. 253-254), estimated the minimal number of texts necessary for representing specific linguistic features in a corpus (Table 3_1).

	Mean score in pilot corpus	Standard deviation in pilot corpus	Tolerable error	Required *N*
Nouns	180.5	35.6	9.03	59.8
Prepositions	110.5	25.4	5.53	81.2
Present tense	77.7	34.3	3.89	299.4
Past tense	40.1	30.4	2.01	883.1
Passives	9.6	6.6	0.48	726.3
WH relative clauses	3.5	1.9	0.18	452.8
Conditional clauses	2.5	2.2	0.13	1,190.0

Table 3_1. Biber's (1993, p. 254) estimates
of required sample sizes (number of texts) for a general corpus

His estimates are based on a sample corpus of "481 texts taken from twenty-three spoken and written registers" (Biber, 1993, p. 253), and calculations are made considering mean score, standard deviation and tolerable error of each individual feature

Specialised corpora can usually afford to be smaller than general corpora. On the basis of their experience, Bowker & Pearson (2002, p. 48) declare that "*well-designed* corpora that are anywhere from about ten thousand to several hundred of thousands of words in size have proved to be exceptionally useful in LSP studies".[5] Indeed several recent corpus studies in LSP are based on small-medium sized corpora (see for e.g. Warren, 2007; Gledhill, 2000; Luzon Marco, 2000; Heyland & Tse, 2005; Banks, 2005, just to mention a few). Scientific support to these empirical habits could come from studies on closure measurements. A corpus can be said to reach closure as regards a particular type of linguistic feature when an increase in the size of the corpus does not bring in new instances of the given feature. In a comparative study of closure in a sublanguage corpus – namely the IBM corpus – and two unconstrained language corpora – the APHB (American Printing House for the Blind) corpus and the Canadian Hansard corpus –, McEnery and Wilson (2001, p. 176) found that the IBM sublanguage corpus showed a very high degree of lexical closure; in fact "the lexicon used by the language of the IBM manuals nearly enumerates itself within the first 110,000 words of the corpus".[6] A corpus that reached closure in a specific feature could be considered representative of the given feature.

Other aspects that are frequently taken into consideration when talking about size are: speed and efficiency of the access software, and the human ability to deal with great amounts of data. Not only the computer might be unable to process great amount of data, but also the human brain. This consideration may lead to the creation of smaller corpora, to the use of sub-corpora, or to the selection (either manual or automatic; randomised or reasoned) of the concordance lines on which analysis is carried out. The last two solutions are suggested by Sinclair (1991, 1992), among others.

[5] Emphasis added.
[6] The IBM corpus showed a greater tendency towards closure also at morphosyntactic and sentence-type levels.

3.3.1.2 Representativeness, sampling and balancedness

Although size matters, to quote the title of Granath's (2007) paper, this is not the only important issue in corpus design. Another feature to consider is 'representativeness' (or 'representatitivity', as some seem to prefer),[7] generally seen as necessary feature for any attempt at drawing generalisations from corpus data. Unfortunately, after decades of corpus building, representativeness is still a highly controversial and debated issue, at least when talking of general corpora aiming to be representative of the language in question.

Biber (1993) suggested language-internal criteria – such as situational (register) and linguistic (lexical and morphosyntactic features) variability – as essential elements for a corpus to be representative.[8] Váradi (2001), in strong critical opposition to Biber, advocated the use of language-external (i.e. sociolinguistic) criteria and proportional sampling based on objective demographic data.[9] Leech (2007), suggested that "the representation of texts [in a corpus] should be proportional not only to their initiators [i.e. speakers and writers], but also to their receivers" (*ibid.*: 138), as the importance of a text depends on the number of receivers it has. Furthermore, Leech (*ibid.*) sees representativeness and balancedness are scalar values; consequently some degree of representativeness and balancedness should be pursued and aimed at by corpus compilers, though attaining these desiderata to the full might be impossible. Finally, other linguists such as Kilgarriff and Grefenstette (2003) argue that every corpus is representative of nothing else but itself. Interestingly, even a strong supporter of representativeness in corpus design such as Leech (2007, p. 145) accepts that "even without such qualities as representativeness, a corpus retains the merit [...] in showing up 'language as it is actually attested in real life'".

This debate leads me to believe that different possible views of representativity can be considered and applied depending on the purpose for which a corpus is created. Most of the considerations by Biber, Váradi, and Leech reported above, and certainly nearly all of Biber's calculations, are based on the assumption that a corpus is built for the purpose of linguistic analysis. But corpora can be created also for other purposes and in these cases corpus creation may require the application of other internal or external criteria. In particular, as the cultural theories reviewed in Chapter 2 suggest, the parameters that have a major impact on the cultural core that is common to all members of a single culture[10] are neither register, nor text type, nor specific linguistic phenomena, but rather time (the year or decade when the texts were written), and authorship (intended as knowing that the author belongs to the single culture under investigation). Furthermore, a preliminary experiment by Bianchi (2007) – briefly summarised in Chapter 4 – seems to suggest that a relatively large size and

[7] While Renouf (2007, pp. 33-34) argues that the term 'representativity' has replaced in popularity the older form 'representativeness' when talking about this issue in corpus linguistics, Leech seems to make a subtle distinction between the two terms, defining 'representativity' as "the degree to which a corpus is representative" (Leech, 2007, p. 133).

[8] Hence his calculations of sample and corpus size based on calculations of distribution of morpho-syntactic and lexical phenomena such as prepositions, sentence types, and hapax legomena. See Section 3.3.1.1 and Table 3_1.

[9] Such criticism mines the ground upon which corpus linguistics studies have been carried out so far, as no corpus exists that matches the level of statistical representativeness advocated by Váradi.

[10] Here and elsewhere in this work I will use Fleischer's terminology.

heterogeneity may be sufficient when the corpus is created around a specific word/concept.

3.3.2 Text collection

Text collection is usually the most time-consuming part in corpus creation. In fact, depending on the planned design, texts are to be searched for, selected or sampled and last but not least acquired in a suitable electronic format. Despite the recent advances in OCR (Optical Character Recognition) technologies, scanning printed texts still requires careful revision and correction of the acquired text, which can only be carried out manually. This, along with the fact that an increasing number of publications are now available on-line (Meyer, 2002), has gradually led researchers to look at the Web as a potential source of corpus data. The Web lends itself to the creation of different types of corpora, using different types of 'collection methods' that range from manual download of individual web pages to full automatic download of automatically selected sets of pages. An overview of major issues in the use of the Web in corpus studies is provided in Section 3.4.[11]

3.4 Corpora and the Web

The last decade has seen a rise in interest towards the Web and its potential in corpus studies, as testified, for example, by the growth of several study and research groups on the topic,[12] and dedicated conferences.[13]

The most commonly used expression to refer to this area of interest is 'Web as corpus'. However, as Bernardini, Baroni, and Evert (2006, p. 10) interestingly point out, this expression subsumes at least four separate senses: 1. querying the Web via commercial search engines and using the retrieved data as concordance lines (i.e. using the web as corpus surrogate); 2. creating corpora from the Web (i.e. using the Web as a corpus shop); 3. considering the Web as a corpus proper; 4. creating a new object, a sort of mini-Web (or mega-corpus) adapted to language research.

The first two scenarios have been fostered by the enormous growth of a multilingual Web, the development of search engines offering rather easy-to-use and flexible text search features, the linguists' need for ever larger text corpora, and an expanding use of corpora in teaching as well as research environments. These two scenarios, though seemingly rather well accepted in the scientific community at large, still require much explicatory and descriptive efforts. The third scenario is still a much debated issue. The last scenario – a development of the second one – is very recent and still under investigation.

The current research project will take advantage of large corpora composed of text retrieved from the Web using spidering tools,[14] and subsequently POS-tagged and

[11] For a wider discussion of corpora and the Web see Gatto (2009).
[12] See for example: the Web as Corpus Special Interest Group of the Association for Computational Linguistics (ACL SIGWAC) (at http://www.sigwac.org.uk/); and the WACKY project (http://wacky.sslmit.unibo.it/doku.php).
[13] The Web As Corpus workshop is now at its seventh edition.
[14] Spidering tools are Web crawling scripts, i.e. programs which browse the Web in a methodical way and retrieve text.

lemmatised and made available to the general public. As such, it falls within the Web as a Corpus Shop and the mini-Web/mega-corpus scenarios.

However, the four scenarios, though different one from the other, are not completely apart, since any use we make of the Web strongly depends on the type and quantity of text available on the Internet, not to mention the methods to access it. For this reason, the following paragraphs will introduce issues related to the Web as corpus proper, before discussing the automated processes for querying the Web and creating corpora from it.

3.4.1 The Web as Corpus proper

As Kilgarriff (2001a, sec. 1) vividly puts it, the Web is an anarchic object showing several features that seem to row against its scientific use as a corpus:

> "First, not all documents [on the Web] contain text, and many of those that do are not only text. Second, [the Web] changes all the time. Third, like Borges's Library of Babel, it contains duplicates, near duplicates, documents pointing to duplicates that may not be there, and documents that claim to be duplicates but are not. Next, the language has to be identified (and documents may contain mixes of language). Then comes the question of text type: to gain any perspective on the language we have at our disposal in the web, we must classify some of the millions of web pages, and we shall never do so manually."

Nevertheless, Kilgarriff and Grefenstette (2003) argue that the Web can still be considered a corpus in the broad sense of the word, i.e. a collection of (electronic) texts for language or literary – and, I add, also cultural – study. And the Web is just as representative as any other corpus. In fact, as these authors argue, currently available general corpora such as the British National Corpus[15] – created according to shared and accepted criteria –, though 'balanced', are always arbitrary selections of text, and their concept of 'balance' is an internal, rather than external one. For example, in the general world, speech events exceed writing events, while the reverse is true in the currently available general-purpose corpora, due to the fact that transcribing speech events is still a problematic and time-consuming task. Similarly, due to both technical as well as theoretical issues, such as size and time limits, fuzziness of text type classifications, and continuous emergence of new genres, the current general corpora only include a selection of text types. On the other hand, the Web contains all traditional text types as well as some emerging ones (Yates, 1996; Leech, 2007). Furthermore, online texts are an excellent resource for the study of emerging usage and current issues, as the Web is "a self-renewing linguistic resource [that] offers a freshness and topicality unmatched by fixed corpora" (Fletcher, 2004, p. 1).[16]

This last point is particularly relevant when analysing culture, since, as we have seen in Chapter 2, culture and cultural associations may change very quickly over time (Nobis, 1998), and considerations about when the corpus and the texts it includes were created are of paramount importance (Bianchi, 2007). Commercially

[15] See Chapter 2, Note 31.

[16] Here and in the following quotes of Fletcher (2004), page numbering refers to the paper retrieved from the Web.

available corpora, be they of a closed or monitor type,[17] become quickly obsolete and tend to include texts from a wide time-span. Therefore, for a synchronic study of current cultural features, the Internet can be seen as an essential resource for the creation of an up-to-date general or specialised corpus.

Besides the issue of representativeness, another concern which usually arises when suggesting the use of the Web as corpus is size. Size is an important issue in corpus linguistics for three main reasons, namely comparing corpora, performing quantitative analyses and statistical calculations, and establishing representativeness. Calculating the size of the Web or of the published web pages in any given language is an insidious task, as the Web is constantly updated and grows almost by the second. Any calculation, therefore, would be almost immediately out of date.[18]

However, if we use the Web simply as a source of data to download for the creation of a corpus, the size issue is at least partially downgraded. Our corpus will have a finite number of words that depends on the purposes for which the corpus is created.[19] Such a corpus could be easily compared to other corpora, and quantitative analyses will be possible on the data of the corpus. Furthermore, as we will see in the following section, Web corpora tend to be more varied in content than traditional corpora, which may have an important impact on the size needed for a corpus.

3.4.2 The Web as Corpus Shop: Creating corpora from the Web

A major issue in the Web as Corpus Shop scenario is representativeness. This explains the effort that, at least in these initial phases of studies in the field, those developing and using automated procedures for creating corpora from the Web put in assessing the representativeness of the retrieved corpora.

Fletcher (2004) compared a small corpus of online documents in English – including only about 11,000 running words – to the written texts in the BNC. The comparison showed differences between the two at the level of spelling (US vs. British), register (interactive vs. narrative style), and type of language (prominence of the language of news and politics vs. prominence of academic language). Furthermore, the Web corpus was more varied as far as frequent lexis is concerned. In fact, the most common 5000 words in the BNC were all present in the Web corpus, while the reverse was not true, and this despite the much smaller size of the web corpus (1/16 of the BNC). Thus, the Web corpus could be considered more representative in terms of the most frequent words.

Studies on several-million-word Web corpora for general purposes created using spidering tools showed that Web corpora assembled following a few reasoned basic criteria concerning preliminary choice of query words and size could be considered comparable to standard balanced hand-collected corpora, in terms of coverage of various text types and topics (see Sharoff, 2006; Ueyama, 2006), though not of register (Baroni & Ueyama, 2006).

---

[17] Monitor corpora, also called open corpora, are corpora that are constantly being expanded through the addition of new texts. On the other hand, closed corpora, once compiled, are no longer expanded.

[18] It must be said, however, that estimates of the size of the Web at given points in time are possible and have been computed. See for example Lawrence and Giles (1999), Lyman, Varian *et al.* (2003), Kilgarriff and Grefenstette (2003), Grefenstette and Nioche (2000), Gulli and Signorini (2005).

[19] See Section 3.3.1.1 for issues relating to corpus size.

Finally, in a preliminary experiment to the current study which focused on the semantic associations of key word *cioccolato* in Italian, Bianchi (2007) compared a specialised corpus manually created around the key word using the Web as source for text retrieval to a general corpus (CORIS) of about 100 million words created according to more 'traditional' methods and criteria, such as sampling and representativeness (Rossini Favretti, Tamburini, & De Santis, 2002). The two corpora were compared in terms of semantic and conceptual categories. The comparison showed a limited number of differences, and – by applying the Mann-Whitney test – it was verified that those differences were not statistically significant, as if the two corpora, though constructed with different criteria and purposes in mind, included samples from the same population. Furthermore, the differences could be explained by the time gap between the two corpora.

This preliminary experiment suggested that suitable data for cultural analysis can equally be retrieved from a very large general corpus, or a small-to-medium-sized specialised corpus, provided the latter has been created including a wide variety of texts by different authors. Furthermore, it confirmed that, for this type of cultural analysis, the major concern in corpus creation, along with text variety, seems to be time-coverage, and this is precisely where the Web comes to an aid.

3.4.3 Further issues and comments

An issue that is certainly relevant when dealing with Web corpora retrieved using spidering tools, is that of authorship. In fact, a large quantity of Web text does not bear the author's name, and once a page has been automatically retrieved and included in a Web corpus any possibility of recovering information about the author is lost. The most common solution to work around this problem is limiting Web searches to a specific language and Internet domain. Almost all the Web corpora created so far, and certainly the ones which will be used in the current research (and which are described in detail in Chapter 5), were created following this procedure. However, for some languages, such as English, which includes several different international varieties and which has been gradually establishing itself as a lingua franca and as 'the' language of the Internet (Crystal, 2003), the sole fact that a page is written in a specific language or appears in a geographically located web site (e.g.: .uk) does not guarantee that the author is native to that language. For other languages, including Italian, whose use is still limited to Italy and a very small area in Switzerland, the chances that a piece of text in that language has been written by a non-native are few. Some attempts have been made to sieve out text by automatically detecting spelling and grammar mistakes (see for example Fletcher, 2004; and Ringlstetter, Schulz, & Mihov, 2006). These methods, however, seem to be still in their infancy, and have not been applied to the Web corpora used in the current study. Nevertheless, no spelling or grammar mistakes which might suggest that the texts were not written by native speakers were noticed while performing manual coding of the Web data. We will come back on this issue later on in the work, after the analyses on the Web corpora have been accomplished and the results have been compared to those of the elicited data.

A second issue is that of readership. Web pages are a form of public communication and, when they are written in an 'international' language such as English, the (perspective) audience is international. However, every culture has specific values and beliefs, and the average speaker is absolutely unaware of that. In fact, as already noticed in Chapter 2, values, value orientations, beliefs, and judgments belong to the informal level of culture. This informal level of culture is where people normally react in everyday life and communication (Hall, 1982). Indeed, adaptation of discourse to target readers is only performed by experts in cross-cultural communication, such as professional translators and marketing experts. Consequently, only a specific part of Web communication can be expected to have been adapted to the values of a perspective audience belonging to a different culture from the author's one. As regards the current research, the authorship and readership issues are of no relevance, given that use of Italian is generally limited to Italy and its native residents. The two issues, however, might bear relevance in the discussion of the English Web data, and in comparing them to the results of the elicited data which were clearly written by native speakers with a native audience in mind.

Finally, although the semantic associations that are common to a whole single culture emerge in language regardless of register (i.e. formal vs. informal language) and text type (poem vs. letter vs. blog vs. news article, etc.), the communicative purpose for which a specific text has been written and the audience to which the text is targeted may influence the semantic content of the text. The Web as a whole is an immense box containing varied but unspecified material which cover all aspects of society and range from scientific papers to gossip news, from marketing advertisements to personal narratives (e.g. blogs), from official legal documents to transcripts of songs and films, from religious text to every day news. But every single document in the Web mirrors only one of those aspects. Unfortunately,

> "automated methods of corpus construction allow for limited control over the contents that end up in the final corpus [and] the actual corpus composition needs therefore to be investigated through post-hoc evaluation methods" (Baroni, Bernardini, Ferraresi, & Zanchetta, 2008, Sec. 3).

The Web corpora chosen for the current experiments – described in Chapter 5, Section 5.2.2.1 – were compared to general reference corpora by their authors. The Web data extracted from those corpora in the current research will be compared to elicited data in Chapter 10.

3.5 Annotating a corpus

Annotation (or markup) is the act of adding explicit (meta-)information to a corpus. Different types of information can be added: textual, such as part of speech information (POS tagging), syntactic annotation (parsing), semantic annotation; and meta-textual, such as sociolinguistic information. Depending on the type of information, annotation takes place at word, sentence, paragraph, or file level. Furthermore, annotation can be done manually, automatically by means of specific software tools, or semi-automatically.

An annotated corpus can be queried and analysed starting from the annotated information, as well as from words in the corpus; this is what Hunston (2002, p. 79) calls 'category-based' methodology. Though annotation is not a compulsory step for carrying out corpus investigation involving categories,[20] it certainly makes category-based investigation easier, and is generally considered added value to a corpus (*ibid.*, pp. 79-80). However for annotation to be usable, it has to be systematic, precise, and intelligible to the end-user.

The following paragraphs provide an introductory overview of the annotation processes which will be used in the current work, namely POS tagging, lemmatisation, and semantic annotation. Details of the annotation systems of the corpora used in the current work will be provided in Chapter 5.

3.5.1 Part-of-speech tagging

Part-of-speech tagging – usually called POS tagging, or simply tagging, but also known as grammatical tagging or morphosyntactic annotation (McEnery & Wilson, 2001, p. 46) – takes place at word level and adds morphosyntactic information next to each word in the corpus. The information added makes the grammatical category to which each word belongs explicit, by adding codes such as: adjective, comparative; noun, countable, singular; verb, simple present, 3rd person, etc. Punctuation is also tagged. Different tagsets may distinguish a different number of categories, and consequently include a different number of tags, and they may use very different codes for the same categories.

Deciding the number and types of tags to use is not the only issue in POS tagging. Other issues include how to deal with multi-word units which function as a single grammatical unit (e.g.: *so that*, or *such as*) and contracted forms (e.g.: *don't*, or *it's*).[21]

As Hunston (2002, p. 82) points out, "tagging needs to be done automatically […] otherwise the labour of adding tags by hand would outweigh the advantages of having them". POS tagging was the first type of tagging to be accomplished automatically, and with relatively good results; in fact, in 1971 the TAGGIT program (developed at Brown University) already achieved an accuracy of 77% (Green & Rubin, 1971). Currently, POS tagging techniques have reached excellent levels of accuracy. CLAWS, the tagger developed at UCREL – Lancaster University and which will be used in the current research, can boast an error rate as small as 4%-2% (Rayson, Archer, Piao, & McEnery, 2004). Furthermore, taggers have been created for languages other than English. The most famous and popular language independent tagger is certainly Tree Tagger (Schmid 1997), developed at the University of Stuttgart. The accuracy of this tagger in its English version is over 96% (*ibid.*), in its Italian version it seems to be around 91% for known words and 86% for unknown words according to Sogaard (2009) and about 96% according to Schmid, Baroni, Zanchetta, and Stein (2007). The general Web corpora which will be used in the

[20] A short list of category-based studies carried out on unannotated corpora is offered by Hunston (2002, p. 80).

[21] For a detailed description of how these issues were solved in CLAWS, the POS tagger used in the current research, see Garside and Smith (1997).

current research to create specialised corpora about given key words have been POS tagged by their authors using Tree Tagger.[22]

Finally, the POS tagging process could be finalised with post-editing, i.e. detection and correction of tagging mistakes in the tagged corpus. Post-editing has traditionally been a manual, time consuming, and costly task. Recently, computational linguists have been experimenting with methods for automatic post-editing of POS tagged corpora (see for example Loftsson, 2009). However, neither manual nor automatic methods seem to guarantee an error-free tagged corpus, especially when the corpus is rather large.

POS-tagged corpora allow corpus linguists to perform advanced searches in the corpus, based on POS tags, and are used by computational linguists to train and develop POS taggers. Furthermore, part-of-speech tagging is the first necessary step for other types of annotation, such as lemmatisation, semantic annotation and parsing.

The following sections introduce some basic issues in lemmatisation, and semantic annotation. Parsing, i.e. syntactic annotation, will not be used in the current research; consequently it will not be discussed here.

3.5.2 Lemmatisation

Lemmatisation, i.e. "the reduction of the words in a corpus to their respective lexemes" (McEnery & Wilson, 2001, p. 53), is an important process in corpus linguistic tagging. It differs from stemming as the latter is a semantic process, while lemmatisation is a grammatical one. In a lemmatised corpus, next to each word, its lemma is provided. This entails the automatic recognition of all the inflected forms. In English, inflected forms are found in verbs (e.g.: *plays*, *played*, and *playing* belong to lemma PLAY, and *goes*, *went*, *gone*, *going* to lemma GO), nouns (e.g. *children* belong to lemma CHILD; *flowers* to lemma FLOWER), and adjectives (e.g. *greater* and *greatest* belong to lemma GREAT). In Italian, inflected forms characterize verbs, nouns, adjectives and articles and the variety of forms belonging to a lemma is much greater than in English. In fact, Italian includes 3 different verb conjugations, about 10 simple verb tenses, and different endings for each person in almost all verb tenses; nouns can be modified by suffixes indicating dimension, affection, etc. (e.g. *casina*, *casetta*, *casettina*, *casuccia*, *casona* are different forms of lemma CASA); adjectives are inflected to distinguish masculine/feminine, singular/plural, and degree (e.g. *bella*, *belli*, *belle*, *bei*, *bellissimo* are forms of lemma BELLO);[23] while articles are inflected to indicate masculine/feminine and singular/plural (*il*, *lo*, *gli*, *i*, *l'* are all different forms of the definite article). In both languages, however, there are cases when a decision has to be made about whether two words belong to the same lemma or to different ones. A controversial case is that of the Italian definite article: *il*, *lo*, *gli*, *i*, *l'* are all forms of the masculine definite article, while *la*, *le*, *l'* are forms of the feminine definite article. Should they be considered as two separate groups/lemmas (as the Tree-Tagger does) or should they be considered as forms of one lemma (the definite

[22] A detailed description of these corpora is provided in Chapter 5.

[23] In theory also diminisher *bellino/a*, and comparative forms *più bello/a/i/e* belong to lemma BELLO, but they do not seem to be treated as such by some taggers, such as the Tree-Tagger.

article, as dictionaries do)? As usual, the answer depends on the aim for which lemmatisation is carried out, i.e. on the granularity needed in the research.[24]

A typical problem is represented by use of apostrophes, as in the case of Italian definite article *l'*, or English Saxon genitive *'s*. The Tree-Tagger, for example, does not seem to recognize *l'* as an article and treats it as part of the word that follows it. Analogously, *'s* does not seem to be considered as a genitive, while *child's* and *children's* could legitimately and reasonably be classified under lemma CHILD.

Automatic lemmatisation is usually performed by POS taggers, but this process takes place after POS tagging has been completed. During POS tagging, disambiguation of words like *plays* – verb *play* vs. noun *play* – takes place. Next, the lemmatiser adds lemma information to each word/grammatical_category pair. Usually, lemmatisers are based on lemma dictionaries, but they may also include rules for desuffixation after automatic recognition of suffixes; these apply when a word is not included in the dictionary (Baroni, 2004).

3.5.3 Semantic annotation

By semantic annotation, here, we mean "the marking of semantic features of words in a text, essentially the annotation of word senses in one form or another" (McEnery & Wilson, 2001, p. 61). In other words, with semantic annotation every word in the corpus is attached a label which indicates the semantic field to which the word belongs. Semantic fields[25] are conceptual abstractions which include not only synonyms, but also other words that are in some way logically associated to the given concept, including hypernyms and hyponyms. Indeed, these mental abstractions are determined by the way the world is, the way the human mind works, and the operational context within which the semantic classification is needed. Consequently, the phrase 'Virgin Mary', for example, could be rightfully classified as 'religion', but also as 'woman' or 'mother'. Furthermore, like in hypernymic/hyponymic relations, different 'levels' of abstraction are possible: 'cat' could be tagged as 'feline', 'mammal', 'animal', or even 'living being' if necessary. This issue is sometimes called 'granularity' or 'delicacy of detail', and choice of one level of granularity over another one is a pragmatic rather than theoretical issue (Wilson, 2003).

Following Schmidt (1988), Wilson and Thomas (1997, p. 55) declare that although "there is no such thing as an 'ideal' semantic annotation system", some general criteria can be listed for the creation or selection of a suitable semantic annotation system. Hence they offer the following criteria (Wilson & Thomas, 1997, p. 55-57):

> 1. It should make sense in linguistic or psycholinguistic terms;
> 2. It should be able to account exhaustively for the vocabulary in the corpus, not just for part of it;
> 3. It should be sufficiently flexible to allow for those emendations which are necessary for treating a different period, language, register or textbase;
> 4. It should operate at an appropriate level of granularity (or delicacy of detail);

[24] See Wilson (2003) for an example of a case when limited granularity could be desirable.
[25] Semantic fields are also called semantic domains, conceptual fields, lexical domains, or lexical fields (Wilson & Thomas, 1997).

5. It should, when appropriate, possess a hierarchical structure;
6. It should conform to a standard, if one exists.

Semantic analysis is a complex task with several issues to be considered, including homography, polysemy, sense ambiguity, units of meaning, and figurative language, as a consequence of the complex network of relationships that subtends words in a language. Indeed, in our mind concepts do not appear to form discrete categories, but rather "fuzzy sets", as prototype theories have shown, and it is not infrequent to find words that fall into more than one semantic field (Wilson, 2003).

The following paragraphs summarize how these problems are dealt with in the UCREL semantic analysis system (USAS), a tool for semantic annotation developed at the University of Lancaster and which will be used in the analytical part of the current work. Originally developed for automatic content analysis of elicited data, such as in-depth survey interviews (Wilson, 1993; Wilson & Rayson, 1993), the USAS tagset has been used with interesting results in several corpus linguistic studies on a range of different topics, from stylistic analysis of prose literature to the analysis of doctor-patient interaction, and from translation to cross-cultural comparisons (see http://ucrel.lancs.ac.uk/usas). In particular, in a cross-cultural study on attitude to shoe fashion by Wilson and Moudraia (2006), the results of automatic tagging with the USAS tagset were compared to those of manual semantic coding and the two coding methods highlighted similar between-group differences. Further details about this tagging system and its semantic categories are provided in Chapter 5.

As described in Rayson, Archer, Piao, and McEnery (2004), semantic annotation employs two main lexical resources: a single word lexicon of 42,000 entries and an idiom lexicon of 18,400 entries, plus an extra single lexicon of about 50 words preceded by wildcard characters to match things like weights and measures. The idiom lexicon – aimed to resolve the tagging of units of meaning – includes phrasal verbs, noun phrases (e.g. *riding boots*), proper names, and true idioms. Tagging of idioms takes priority over tagging of individual words, in order to prevent tagging overlap. Disambiguation of homographs and polysemy is resolved resorting to a combination of seven techniques including POS tagging, which is a pre-requisite in automatic semantic annotation processes, as well as frequency and other types of statistic information and context-sensitive rules.

Finally, USAS's solution to the problem of a word falling into more than one semantic field is attaching several separate labels to the same word (when applicable), and then choosing the most suitable one on the basis of frequency or domain considerations. However, there might be cases when selection of one semantic category only is not applicable. Indeed, this is not the only possible solution to this problem: Wilson (2003), for example opted for assigning more than one category to the same occurrence of a word. The multiple-assignment solution will be adopted also in the current research when tagging data manually (see Chapter 5).

3.6 Analysing a corpus: major analytical features and methods

Corpus analysis is accomplished taking advantage of specific software tools, or concordancers. Concordancers may differ in terms of number of features offered, user interface, supported file format, output format, and query language; however, some basic analytical features are common to all of them, namely the frequency word list and concordance features. More advanced concordancers (among which Wordsmith Tools, used in the current research) include other features such as automatic extraction of clusters, collocates, keyword lists, as well as the computation of various types of statistics.

The following paragraphs illustrate the analytical features and methods which have been used or mentioned in the current research. The degree of detail in each paragraph reflects the degree of relevance each feature had in the research. Indeed, my experiment, which focused on semantics but aimed to establish cultural associations of given key words, made ample use of frequency lists and, at least in a preliminary experiment, keyword lists; concordancing was necessary to understand the context of the key words; a look at collocations and semantic preference helped semantic tagging, while colligation was ignored; finally semantic prosody was systematically analysed.

3.6.1 Wordlists and frequency

Wordlists, i.e. lists showing the number of occurrences (raw frequency) of each word in the corpus, provide an overview of the corpus; for this reason they are the first thing that corpus linguists tend to examine, in both quantitative and qualitative studies. As we have seen in Chapter 2, wordlists have also largely been used as a starting point for cross-cultural comparisons.

Wordlists, which can be ordered alphabetically, or by frequency, are always accompanied by information on the total number of running words (tokens), and the total number of word forms (types),[26] in order to allow conversion of raw counts into percentages (normalisation) and comparisons between corpora of different size, as well as calculation of Type-Token Ratio, a measure of lexical variation within the corpus.[27]

If necessary, 'abridged' wordlists can be created by applying a specific 'stop list'[28] which excludes undesired word forms – for example function words – from the wordlist. In the current research, stop lists will be used to filter out function words, as well as other non-desired words such as the various forms of the key word itself, in a series of experiments aimed to explore the possibility of using only the most frequent words in the wordlist to highlight the same cultural traits that would emerge from the analysis of the whole corpus.

---

[26] *Word forms* or *types* are not to be confused with *lemmas*. In fact, the lemma EAT – to quote an example from Hunston (2002, pp. 17-18) – would include word forms such as *eat*, *eats*, *eating*, *ate*.

[27] More sophisticated concordance packages may also provide other types of statistical information, including standardized type frequency, Type/Token Ratio (TTR), average word length, number of sentences, and average sentence length.

[28] A stop list is a list of words that the researcher wants to exclude from the analysis. The list is created by the researcher – usually in the form of a txt file.

If the corpus is not POS-tagged or lemmatised, the information provided by the wordlist is rather rough, since it will not take into account issues such as polysemy, homography and different word-classes: all occurrences of word *bank*, for example, would be listed under the same entry, regardless of their meaning 'bank of the river', or 'financial institution', and of their being noun or verb ('to bank'). Consequently, entries in an untagged wordlist need to be checked against concordances, to see the contexts in which the given tokens appear. The wordlist of a POS-tagged and/or lemmatised corpus, on the other hand, provides the frequency of lemmas and/or words according to their POS category.[29]

Quantitative comparisons between wordlists is only possible when the frequency counts in the two corpora are normalised to the same figure; it also requires that frequency counts have been conducted in the same way as regards stop lists, numbers, hyphenation, apostrophes, and the like.[30] Comparison of normalised figures, however, only tells us where similarities and differences appear, but not whether they are significant, or due to chance (Meyer, 2002, p. 126). To this purpose, statistical procedures should be applied, and several types of statistics have been proposed, including the *chi-square test*, the *chi by degrees of freedom*, the *log-likelihood test* and the *Mann-Whitney test*. None of these tests is exempt from drawbacks and debate over their application seems to be still open. Most concordancers, however, offer only the *chi-square* and *log-likelihood* options.[31]

The current research will take advantage of wordlists, as a starting point for the identification of semantic categories. Consequently, wordlists from different corpora will not be compared as such, but only after applying semantic analysis. The statistics used to perform quantitative comparisons will be described and discussed in the relevant chapters.

A few more interesting comments could be made about frequency in a corpus list. First of all, an almost linear inverse relationship between word frequency and word rank has been noticed, which is described by Zipf's law. In other words:

> "a word list [and this appears to be true of any word list based on at least a few hundreds of words] contains a very small number of very highly used items, and a long declining tail of items which occur infrequently, with roughly half occurring only once as hapax legomena" (Scott & Tribble, 2006, pp. 27-29).

As a result, the most frequent 150 words in a wordlist typically account for about half of the words in the corpus, though this number may vary depending on factors such as corpus size, genre and register (Powers, 1998). A consequence of the Zipfian distribution of the words in a corpus is the fact that, as the size of a corpus increases new vocabulary enters the corpus following a distribution that is marked, after an initial sharp increase, by a gradual reduction in the number of new words; this is known as Heaps' Law (Heaps, 1978). Although this distribution is not really upper-bounded, due to the presence of proper names and typos, if collecting data from the

[29] For details about lemmatisation, see Section 3.5.2.
[30] For a detailed discussion of issues and possibilities in creating a word list, see Scott and Tribble (2006, pp. 13-20).
[31] For a survey of the various statistics used for comparing corpora see Kilgarriff (1996a, 1996b), and Rayson (2003).

same genre and time period, enlarging a corpus over a certain limit will yield diminishing returns in terms of giving new vocabulary.

Furthermore, some words appear consistently in a high number of texts, while others appear frequently only in a limited number of texts or text types (Scott & Tribble, 2006, p. 29). This suggests the importance of an analysis of the distribution of the words across texts, as well as of their frequency.

3.6.2 Keywords and keyness

The term *keyword* (or *key word*) is widely and constantly used in linguistics and other disciplines; however different meanings are given to this term in different contexts and research traditions. Williams, who paved the way to a rich research tradition in the field of cultural analysis, describes keywords as "significant, binding words in certain activities and their interpretation; they are significant, indicative words in certain forms of thought" (Williams, 1976, p. 13). This is a general definition that can easily be understood and shared; but no indication is provided about how to choose keywords in the analysis of specific contexts, such as culture. Indeed, most linguists working in Williams' research tradition have not felt the need to investigate possible scientific methods for the selection of cultural keywords.[32]

In corpus and computational linguistics, on the other hand, the notion of *keyword* includes the idea of statistical significance deriving from frequency comparisons. Corpus linguistics *keywords* are usually obtained by comparing the wordlist of the corpus under investigation with the wordlist of a suitable reference corpus; any word of the given corpus whose frequency is found to be outstanding with respect to the reference corpus is considered a keyword. As Baker (2006, p. 123) states, a keyword list "gives a measure of *saliency*, whereas a simple word list only provides *frequency*".

As was the case with word lists, several statistical methods can be applied for comparing two corpora by (key)word frequency. The chi-square test and the log-likelihood test are frequently used for determining *keyness*, i.e. the degree of outstandingness, or salience, of the specific word in the target corpus. The Wmatrix interface, used in a pilot experiment to the current research, adopts the log-likelihood measure.[33] Keyness can be positive or negative: positive keywords are words that are unusually frequent in the target corpus, while negative keywords are unusually **in**frequent in comparison to the reference corpus.

The reference corpus is usually, but not necessarily, larger and more general than the other one (Hunston, 2002, p. 68).[34] As regards the composition of the reference corpus, Scott and Tribble (2006, p. 65) declare that

> "further research is needed before we can confidently offer a rule of thumb, if one exists. In any case the research purpose is fundamental: in our experience, even the use of a

[32] An exception is perhaps represented by Rigotti and Rocci (2002), who have developed a method for verifying whether selected words are cultural keywords.

[33] A detailed description of Wmatrix is provided in Chapter 5.

[34] Gledhill (1995, 1996) and Bianchi and Pazzaglia (2007), for example, compared different folders of the same corpus, corresponding to the different sections of research articles. Culpeper (2002) extracted the keywords characteristic of six characters of Shakespeare's *Romeo and Juliet* by comparing the lines spoken by each character to the lines of the remaining five characters (taken together).

> clearly inappropriate reference corpus as in the case of the BNC for studying a Shakespeare play may well suggest useful items to chase up using the concordancer."

To avoid possible terminological confusion, in the current work, the term *keyword* is used when the computational methods described above are applied, while the terms *key word* and *node word* are preferred when a word is chosen according to other, non computational criteria and used as starting point for analysis or for the generation of concordances, respectively. Finally, the term *search word* is used when talking about information retrieval with search engines.

3.6.3 Concordancing

Any word or keyword can be used as starting point (node word) for concordancing. Concordance lines are chunks of text that show the node word in context – hence the term KWIC (Key Word In Context) format. The length of concordance lines depends on the parameters set by the user. In a KWIC concordance, all the occurrences of the node word are displayed one under the other, with the key words vertically aligned and highlighted.[35]

If a corpus is lemmatised, a lemma can be made node word, and the concordancer will search for strings of text containing any of the words belonging to the given lemma. If the corpus is POS tagged, and the software offers specific query options, concordancing can take grammatical category into consideration or even start from a POS tag.

In the current research concordancing will be used at several stages and for different purposes: in the preparatory phases, for extracting sentences containing selected words from general Web corpora (see Chapter 5); and when manually coding wordlists, for seeing the context of each word (see Chapters 7 and 8).

Most software programs allow users to decide the way they want the concordance lines to be shown, in terms of number of words to be displayed, sorting criteria (e.g.: sort alphabetically by node word, by 1 left, and/or by 1 right), and even the presence of specific words in the co-text. KWIC display and a correct use of sorting options facilitate the qualitative analysis of concordance lines and the observation of repeated patterns.

Concordance lines are the typical starting point for the analysis of collocation, colligation, semantic preference and semantic prosody which are usually considered in corpus linguistics the four descriptive components of units of meaning (Sinclair, 2004). As Mahlberg (2007, p. 195)[36] puts it

> "From the level of collocation to semantic prosody the descriptive components of a lexical item become increasingly abstract and move from the fixed core of the item towards its boundaries. Collocation is a very concrete category and accounts for the actual repetition of words on the textual surface around the core. The component colligation introduces a level of abstraction with reference to grammatical categories. Semantic preferences interpret the context of the core in terms of shared semantic

[35] An interesting and precise description of KWIC concordance lines can be found in Tognini Bonelli (2001, 2004) and in Stubbs (2007b, p. 177).

[36] Description of the four levels of analysis as different levels of abstraction is not specific to Mahlberg; in fact, she is following Sinclair and Stubbs.

> features, and finally the semantic prosody accounts for attitudinal or pragmatic meanings."

KWIC displays have greatly changed our way of looking at texts: linguists have passed from linear reading of one text after the other, to non linear and focused access to several texts at once. Also, by looking at chunks of sentences, our attention is necessarily concentrated on the node word and its immediate surroundings, without distractions. On the other hand, 20- or even 50-word chunks can at times be too short to understand all of the semantic components of the given word. A typical case is when a word takes part in an anaphoric chain and its referent can only be understood by going back to the first element of the anaphoric chain; or, as we shall see later, when it comes to analysing semantic prosody. For this reason almost all concordancers allow the user to expand concordance lines to display full sentences, paragraphs or even texts.

3.6.4 Collocation, semantic preference and semantic prosody

As Evert (2007) points out, the term collocation is used in linguistics to refer to various different textual features. In an attempt to make distinctions clearer, he distinguishes between 'lexical collocations' and 'empirical collocations'. Lexical collocations are a series of more or less transparently fixed expressions, ranging from well-known idiomatic expressions and set phrases (e.g. *a school of fish*), to multiword expressions (e.g. *credit card*), to multiword units with mobile elements (e.g. *as far as X is concerned*). The term 'empirical collocations', on the other hand, refers more generally to the fact that some words (collocates) tend to appear more frequently than others in the same linguistic environment, and the study of empirical collocations requires the use of statistical association measures (such as T-score or MI score) to quantify the attraction between co-occurring words.

Although Evert (2007) suggests that this mathematical meaning of 'association' should not be confused with psychological association, a psychological component seems to be present in collocations, alongside a textual and a statistical components (Partington, 1996, pp. 15-16). From a textual point of view, "collocation is the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991, p. 170). From a statistical point of view, it is "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991, pp. 6-7). Finally, from a psychological or associative perspective, collocation is the expectations (or 'expectancies', in Firthian terms) that native speakers have of encountering a given word in the same environment as another one (Leech, 1974). In a study on priming, Durrant and Doherty (2010) provide an interesting review of major issues in assessing the psychological reality of collocations, discuss a few studies which suggest that high frequency collocations are psychologically real, and describe two experiments whose results seem to confirm that high-frequency collocations are likely to have psychological reality, though the models currently used to represent priming may need further elaboration.

Semantic preference and semantic prosody are two separate phenomena, but the boundary between the two is not always clear-cut: they frequently appear together (Bednarek, 2008) and they are frequently discussed together. Indeed, they can both be

considered as an extension of collocation (see for example Baker & McEnery, 2005; Bednarek, 2008).

Stubbs (2001, p. 65) defines semantic preference as "the relation, not between individual words, but between a lemma or word-form and a set of semantically related words", i.e. the tendency of a word to co-occur with words belonging to one or more specific semantic domains. It has been noticed that a word may have different semantic preferences depending on features such as context, genre, domain, but also literal or metaphorical use (Bednarek, 2008). Furthermore, like collocation, semantic preference varies when syntactic patterning (colligation) varies (see for example Partington, 2004 and his discussion of *sheer*). Finally, different word classes tend to have different semantic preferences (O'Halloran, 2007).

As already mentioned, semantic preference entails a greater level of abstraction than collocation, and the semantic categories are decided by the researcher after looking at the concordance data available, on the basis of his/her intuition of what is most suitable in the project at hand. For example, a series of concordance lines where the node word 'sports car' co-occurred with names of famous American actors could lead to identifying as suitable semantic preference ACTORS, MEN, or even AMERICANS. None of these is preferable to the others *a priori*; only the whole context and aim of the research project may lead to a suitable solution.

When the semantic categories adopted fall into evaluative categories (e.g. 'good/positive/healthy/legal' and 'bad/negative/unhealthy/illegal'), then we enter the realm of semantic prosody. Identifying evaluation in text is a problematic issue, since evaluation can be expressed in several ways. Some lexical items, such as words 'wonderful' or 'good' or 'bad', have an evident evaluative component. However, as Hunston (2004, p. 157) notices, "the group of lexical items that indicate evaluative meaning is large and open and does not lend itself to quantification". Despite this, semantic taggers, such as the USAS tagset, which will be used in the current work, show attempts to list evaluative words, and also phrases (e.g. 'a cut above', or 'below standard', 'hand on heart'). The USAS tagset (Archer, Wilson, & Rayson, 2002) includes a specific category for evaluation (A5), subdivided into 4 subcategories: 'A5.1 Evaluation: Good/bad', 'A5.2 Evaluation: True/False', 'A5.3 Evaluation: Accuracy', and 'A5.4 Evaluation: Authenticity'. Within each category, plus (+) or minus (-) signs indicate positive or negative polarity, respectively. Alongside lexis, lexical-grammatical sequences may be indicators of evaluation (e.g. 'there is something x about y'), as suggested by Hunston and Sinclair (2000). Furthermore, frequently words inherit the positive or negative aura of the collocates they co-occur with (see for example Sinclair, 1991 and his analysis of *set in*). Finally, words and phrases may acquire different evaluative meanings depending on context and "the reader assumptions about value" (Hunston, 2004, p. 158) – to make an easy example, word 'low' indicates positive assessment when collocating with inflation, and negative when next to salaries – but also genre and domain (Bednarek, 2008) – corpus analysis has shown, for example, that phrase 'responsibility for' acquires negative connotation in the news, since it always collocates with negative events such as bombings, explosions, or acts of terrorism, but neutral in business texts where it collocates with budgets, outcomes or decisions (Bednarek, 2008, p. 123). Examples of corpus methodologies which may be used to identify and analyse evaluative language

in large collections of texts can be found in Hunston (2004; 2011). A specific area of research concerned with identifying and quantifying expressions of opinion in text is sentiment analysis. Sentiment analysis, a particular type of automatic content analysis focussing on semantic prosody, will be described in Chapter 4, since it frequently used in marketing research.

3.7 Some thorny issues

Not all linguists are in favour of corpus linguistics, and its detractors include very famous names such as Chomsky[37] and Widdowson. Chomsky's criticism to corpus linguistics traditionally revolved around the following two points: the use of texts as the primary source of linguistic information, and the finite nature of a corpus.[38] Indeed the corpus perspective, where the data and their frequency of use are key elements in linguistic description, strongly clashes with Chomsky's distinction between competence (I-language) and performance (E-language) and the former's prioritisation over the latter. Furthermore, Chomsky argued that the finite nature of any corpus, even the largest ones, cannot account for the infinite possibilities of language (Chomsky, 1962). Hence, in his view, introspection and not corpus data is the primary key to linguistic research.

Less radical, but nonetheless critical is Widdowson, who considers corpus linguistics as a 'development in E-language description' (Widdowson, 2000, p. 6). Though agreeing that corpus analysis reveals facts about the way language is used that are not directly accessible by intuition or surveys among speakers, Widdowson (2000) sees serious limitations in corpus linguistics connected to its inability to describe member categories (in ethonomethodological terms), to provide insight into the encoded possible and the contextually appropriate and to its showing decontextualised language.

Criticisms such as the ones above have been taken into serious consideration in corpus linguistics and, rather than defeating it, they have aided the development of this field of enquiry. As McEnery and Wilson (2001, p. 5) observe, "[c]oncepts […] such as balance and representativeness […] are a direct response to some of the criticisms Chomsky made." Similarly, awareness of the need to 'recreate' the socio-pragmatic context of corpus data has led to the development and use of tagging schemes which encode sociolinguistic information.[39]

Modern (as opposed to early) corpus linguists are aware of the limitations of corpora and of caveats in their use. The limitations of corpora are summarised by Hunston (2002, pp. 22-23) and are shortly listed and commented below.

First of all, corpora present language out of its context. The word context is to be interpreted here in many senses that range from social and pragmatic context, to visual and audio context. Despite several possibilities exists to include information about

[37] Chomsky's consideration of corpora, however, seem to have slightly changed in recent times (see Aarts, 2000).

[38] A clear review of Chomsky's criticisms to corpus linguistics can be found in McEnery and Wilson (2001, pp. 5-12); mention of the debate is also present in many papers and books about corpora, such as Leech (1992), and Tognini Bonelli (2001).

[39] See for example the following corpora: ICE-GB; The Wellington Corpus of Spoken New Zealand English; The Limerick corpus of Irish English; The Scottish Corpus of Texts and Speech (Xiao, 2008).

textual and contextual data into the corpus, this kind of annotation is time consuming and consequently relatively little used. Similarly, although some multimodal corpus analysis tools have recently been developed (see for example Baldry & Beltrami, 2005) their use is still extremely limited. Finally, several projects have addressed the issue of 'recontextualisation' by annotating important pieces of contextual information, but, to my knowledge, none of them has ever been able to fully provide all of the contextual elements (from the socio-pragmatic to the audio-video ones).

Second, any corpus is a limited sample of language and can only show its own contents. Therefore the linguist must be very careful at making generalisations from a single corpus, as "conclusions about language drawn from a corpus have to be treated as deductions, not as facts" (Hunston, 2002, p. 23). This is particularly true when

> "evidence from a corpus is used to make statements about 'the way the world is' […]. For example, there are roughly twice as many instances of *left-handed* as *right-handed* in the Bank of English corpus. What is the reason for this? One possible explanation is that there are more left-handed people in the world than right-handed people, but we know that this is not so. Another explanation is that left-handed people are considered to have a higher status than right-handed people, and therefore to be more worth talking about. Most left-handers would argue that this does not accord with their daily experience. A third possibility is that right-handedness is considered to be 'the norm' and left-handedness is 'deviant', and that deviance is more often mentioned than normality. Looking at the lines themselves suggests that this is the most likely interpretation, but it is important to recognise that this is an interpretation of evidence, not 'fact'" (Hunston, 2002, p. 66).

Third, a corpus can only provide information about whether something is used or frequent, but not whether something is correct (from the point of view of 'standard grammar') or impossible. As both Chomsky (indirectly) and Widdowson (directly) noticed, we cannot say that something is not possible simply because it does not appear in a corpus.

Fourth, a corpus can offer linguistic evidence but not linguistic information. The corpus only lists several examples of language in use, or frequency counts, but making sense of them is left to the researcher. Indeed, a corpus does not automatically provide answers to linguistic questions. Analysis and intuition are always necessary to make sense of the data.

Awareness of these limitations is probably one of the reasons (though not the only one) that has led corpus linguists to working more and more on specialised corpora, and use general ones as term of comparison. Highly specialised corpora reduce the problem of decontextualisation. Furthermore, as we have see in Section 3.3.1.1, the more a corpus is specialised the smaller it can be, and a small corpus is easier and quicker to annotate. Finally, it has become frequent practice for corpus linguists to carry out the same type of analysis on several different corpora, or to compare corpus results to other types of empirical data or to a specific theory, before drawing generalised conclusions. In fact, when comparing different corpora or different types of empirical data, conclusions are drawn from the analysis of several different samples of the same population, rather than from one single sample. Furthermore, as we have seen in Chapter 2, interpreting corpus data within a clear theoretical framework may

help formulate sound hypotheses or draw convincing conclusions, and at the same time prove (or disconfirm) a specific theory.

I agree with Baker in believing that any method of research has "associated problems which need to be addressed and [is] also limited in terms of what [it] can and can not achieve" (Baker, 2006, p. 7). Moreover, as suggested by Fillmore (1992) and others, there is no reason why theoretical linguistics could not go hand in hand with corpora, and various 'types' of linguists, including theoreticians, could not make use of corpus data.

The current chapter has outlined some features and key elements of corpus linguistics and has introduced the Web as a source for corpus data. In the next chapter will see how some of the concepts and methods of corpus linguistics recur in or compare to analytical methods in marketing research.