Introduction

Corpus Linguistics (CL) or Computer Corpus Linguistics (Leech 1992) has been one of the most interesting approaches to the study of linguistic phenomena to emerge in the latter part of the 20th century. After an initial period in which its aims were often misunderstood, especially from those with a Chomskyan orientation, CL enjoyed considerable development in the 1990s and continues to grow, both at the surface and in depth (Bunge 1968), in the second decade of the 21st century.

Drawing from the many and varied definitions of what a textual corpus is (e.g., Crystal 1991; Francis 1982; Guilquin and Gries 2009; McEnery and Wilson 1996; Sinclair 1991, 1996; Tognini-Bonelli 2001), we will take as our starting point the following: *A corpus is a set of natural texts (or pieces of texts), stored in electronic form, assumed to be jointly representative of a linguistic variety in some of its components, or in all of them, and grouped together so that they can be scientifically studied*. Let us look more closely at some features of this definition:

- The (pieces of) texts (Sinclair 1996) must have a natural character, that is to say, they must have been produced by human beings in real and natural conditions.
- These texts must be stored in electronic format. Only digitized data allows any kind of practical access to the millions of linguistic forms in a corpus.
- Texts in a corpus must together be representative of the linguistic variety from which they were drawn. Furthermore, the corpus should be balanced, reflecting to as great an extent as possible the different types of texts (newspapers, academic, fiction, talks, radio magazines,

To be published in Lacorte, Manel (ed.): *The Routledge Handbook of Hispanic Applied Linguistics*. New York: Routledge, 2014, 371-387.

etc.) produced in the specific linguistic community.

- The corpus must be compiled in a way that makes possible its scientific analysis.
- The ability to enrich texts by encoding, morphosyntactic tagging and syntactical, semantic and pragmatic annotation should be available.

In the following paragraphs the history and consequences of these features of corpora will be explored. Section 2 summarizes both the antecedents of CL as well as the different phases in its development. Section 3 analyzes the current situation with regard to some fundamental topics, and includes reference to Spanish corpora. In Section 4 a number of notions that are likely to be of central concern in the coming years are examined.

Assuming that readers are more familiar with English CL, I will try in what follows to identify at the relevant points the main differences between Spanish and English CL. In general it can be said that Spanish CL began significantly later than in the case of English. Yet over the last twenty years Spanish CL has witnessed notable advances in the volume and characteristics of corpora compiled. Second, there seems to have been a specific interest with Spanish CL in middle and large sized corpora, and that, as in the case of *Corpus de referencia del español actual* (CREA), *Corpus diacrónico del español* (CORDE) and *Corpus del español* (CE), these may show, through selective recuperation of data, differences in the diachronic, diatopic and diastratic axes. Third, Spanish CL works primarily with very large corpora in which occurrences of expressions can be accessed via the internet, but for which the complete texts of the corpus cannot be obtained.

Historical perspectives

Antecedents

Although the evident dependency of CL on computers may lead us to think of a timeframe of fifty or sixty years for the development of corpora, disciplines and methodologies rarely emerge suddenly and in a vacuum. Although not the principal focus of the present study, it

might be useful here to clarify some general issues, not least to enhance and even counter the usual views on the history of corpora, which are often somewhat superficial and are nearly always framed from the sole perspective of English language and linguistics (Francis 1992; Meyer 2008; Svartvik 2007).

The basic meaning of the Latin word *corpus* (pl. *corpora*) is 'body' (cf. sp. *cuerpo*, fr. *corps*, it. and port. *corpo*, etc.). Yet it also had other, secondary meanings, as *Oxford Latin Dictionary* notes, these denoting 'any structure comparable to a body, a fabric framework' (ac. 6) and 'a comprehensive collection of facts on a given subject; a compendium of scientific, literary or other writings, an encyclopaedia, etc.' (ac. 16). This latter sense continued to be used in Western Europe long after the fall of Rome to refer to a set, a collection of texts assembled in order to make searches easier and to ensure the unity and reliability of its contents. Hence the *Corpus Iuris Civilis*, dating from the time of Justinian, was a compilation of legislative texts; somewhat closer to our time, the *Corpus Inscriptionum Latinarum* is a huge compilation of surviving Latin inscriptions, arranged by country of origin. A corpus, then, consists of a set of objects (mainly texts) collected with the intention of facilitating their examination and study.

However, the current concept of textual corpus derives from a far richer and wider tradition, and begins with concordances (McCarthy and O'Keefe 2010). Thought to date from as early as the 12th century, concordances take the form of references to the same words or concepts (punishment, salvation, etc.) in different texts (chapters and verses in biblical texts, for instance). This initial, topic-based configuration (*concordantiae rerum*) is indeed the origin of the term used today for those text fragments returned during computerized corpus searches as instances of the expression or word under analysis. Soon, early concordances moved on to include the precise location of a certain word or expression in one text or a set of texts, and usually with the inclusion of enough material context to make recourse to the

original text unnecessary. Following this, concordances began to be developed which focused on the work of authors considered to be of special cultural significance.

A second source of modern corpora is that of traditional lexicography based on real texts, in which hundreds of texts were used to compile representative examples of words in their different meanings and uses. Thousands of citations, usually (but not exclusively) from notable authors, were typically transcribed onto slips and filed, to be used as the point of departure for the organization and construction of a dictionary's lexical entries. The two main classical works of this kind in Spanish are the so-called *Diccionario de Autoridades*, published by the Real Academia Española between 1726 and 1739, and the *Diccionario de construcción y régimen*, conceived by Rufino José Cuervo, who published its first two volumes in 1886 and 1893. The same basic approach of collecting representative fragments of texts can be used in order to compile real instances of grammatical phenomena, Jespersen's monumental *Modern English Grammar on Historical Principles* of course being the standard reference on these lines. No comparable work exists for Spanish grammar, although Salvador Fernández Ramírez spent many years in the organization of a huge file of instances of many different grammatical phenomena, now available as *Archivo general de la lengua española* (AGLE).

A different approach is that in which a set of texts is exhaustively analyzed in order to obtain statistical information considered relevant for some specific purpose. In the lexical field, this can result in frequency lists (of lemmas and/or forms), used for second language teaching or some more general purpose. García Hoz (1953), Juilland and Chang-Rodríguez (1964), Alameda and Cuetos (1995), Almela *et al.* (2005) and Davies (2006) are examples of works for the Spanish language. *Mutatis mutandis*, the same perspective can be applied to grammatical phenomena, but here the identification and selection of instances is more complex. Keniston's (1937a, 1937b) lists of grammatical constructions in classical and

modern Spanish are among the very few such works in any language.

Finally, in the years immediately preceding the emergence of CL, the idea arose of compiling a set of texts that could be considered representative of the real situation of language use in a given context. The originality here lies not in the idea of compiling such a corpus *avant la lettre* (Julio Cejador [1905-1906], for instance, analyzed the vocabulary and grammar of the complete works by Cervantes) but in the clear orientation as to the type of texts collected. The *Survey of English Usage* (SEU), designed and developed by Randolph Quirk, comprised mainly oral texts transcribed from taped recordings made in the fifties. Ten years later, Lope Blanch initiated a large project centered around the collection of oral texts produced by people of different age, gender and sociocultural level in the great cities of the Hispanic world (Lope Blanch 1986).

The arrival of computers

The arrival of computers led to radical changes in the way of working in many disciplines. Although it is a common view that there exists a huge separation between the technologically advanced world of computation and the 'humanities', we know that Roberto Busa contacted IBM with the idea of developing electronic concordances of the works of Thomas Aquinas as early as 1949 (cf. Hockey 2000: 5-6).

From a theoretical and methodological perspective, the impact of computers on linguistics can be seen very clearly with Freeman Dyson's notion of tool-driven revolutions. For Dyson modern science arises from the fusion of two great traditions: 'the tradition of philosophical thinking that began in ancient Greece and the tradition of the skilled crafts, that began even earlier and flourished in medieval Europe' (1999: 7-8). Changes in the former produced the scientific revolutions with which Kuhn (1962) radically changed the very conception of scientific progress. The idea of the replacement of a paradigm in crisis, no longer able to explain the anomalies accumulated over a period of scientific endeavor, by

another is taken up by Dyson and called ‘concept-driven revolution’. Its effect ‘is to explain old things in new ways’ (Dyson 1997: 50), as well as to address phenomena that previously could not be adequately understood. The best known example of this, of course, is the change from a geocentric to heliocentric model. ‘The concept-driven revolutions are the ones that attract the most attention and have the greatest impact on the public awareness of science, but in fact they are comparatively rare’ (Dyson 1997: 50). Much more frequent, and in most cases with greatest impact on everyday life and scientific work, are those secondary innovations, the result of changes in tools (not only physical) and the emergence of new instruments. Dyson called these ‘tool-driven revolutions’, whose effect ‘is to discover new things that have to be explained’ (Dyson 1997: 50-51). This happened, for instance, at the time when Galileo looked at the sky with the rudimentary telescope he had constructed, thus observing a far richer and far more complex panorama than had until then been possible with the naked eye.

The integration of the entire texts which constitute a corpus – and not only of selected instances of statistical information – allows the recovery of what is going to be analyzed in a fast and convenient way. Naturally, in the early years, when computers were relatively slow and lacked today’s processing power, things were slower, more expensive, and less immediate. Yet computer technology has evolved rapidly, ‘becoming even faster, smaller yet more capacious, and cheaper in relation to what it can do’ (Svensén 1993: 250).

Phases in CL

The arrival of computers led to a significant change in the way in which linguists access their data. It is therefore useful to look at the different phases in the short history of CL, the main reference here being the evolution of speed, capacity and cost of computing (Renouf 2007; Tognini Bonelli 2010). The first electronic corpora, in the 1960s, typically had a size of around one million words (Brown Corpus, Lancaster-Oslo/Bergen corpus [LOB]); the normal

size for a reference corpus today is four, five or even six hundred million words. Moreover, the recent approach known as ‘Web as Corpus’ maintains as its potential corpus the huge collection of publically available texts to be found on the World Wide Web.

The corpus size is not only a function of the speed and capacity of computers, but has other motivations. First, the evolution of technologies used for introducing text in the computer has itself been significant; texts might previously have been typed-up manually or entered via scanners and optical character recognition (OCR) programs, but the current means of acquiring texts in electronic format are far quicker and more direct, with source material available from newspapers, blogs, e-books, etc. Second, formerly a corpus would be installed on one specific computer, in a certain place, making it necessary to travel there to use the corpus; clearly, today’s corpora can offer virtually instantaneous access from anywhere in the world. Third, and associated with the previous point, data from a corpus can now be extracted using only a standard web browser (and, of course, the search application running on the server). Fourth, the texts in corpora now also carry information on the parameters used in the corpus implementation (country, year of production, text type, sex and age of speaker when applicable, etc.), and hence it is possible to be highly selective in the recovery of information (data only from texts produced in a certain country, between year *x* and year *y*, etc.). Indeed, in a great majority of studies, the frequency of occurrence of a form or expression is assessed not with respect to what happens in the whole corpus, but in terms of possible differences between two or more different subcorpora. Finally, the development of computational linguistics made possible the automatic tagging of all forms within the corpus with their lemma and morphosyntactic characteristics, thus allowing for the use of a corpus for research into aspects of grammar, and also permitting the construction of syntactic analyzers, machine translation systems, etc.

Core issues and topics

Guillermo Rojo

A new theory, a new methodology, or a new discipline?

The computer's increasing processing speed and capacity for data storage lead to linguists being able to analyze ever greater volumes of data in a more reliable, rapid and convenient way than was previously possible. Yet in that it is a tool-driven revolution, this does not itself imply accompanying shifts in linguistic theory or practice. Indeed, whereas CL spreads to a great variety of fields, there is at the same time a lack of agreement about the exact character of this new methodology, new theory or new discipline (e.g., Gries 2009; Kennedy 1998; McEnery and Wilson 1996; McEnery *et al.* 2006; Parodi 2010a).

It seems clear that CL is not a theory, in that corpus data can be analyzed from different theoretical orientations, although it is evident that theories that typically do not take the analysis of external data as a cored procedure are less inclined to adopt CL; nor is it a free-standing discipline, in that corpora are used in the study of grammar, the history of language, phonology, sociolinguistics, lexicography, language teaching and many other specialized fields. Yet CL cannot really be considered as a methodology in the usual sense of the word. Indeed, Leech considers CL 'a new research enterprise, and in fact a new philosophical approach to the subject' (1992: 106), and Gries sees it as 'a method(ology), no more, but also not less', although he does not think that 'this difference would result in many practical differences' (2009: 1).

CL is a different way of analyzing linguistic phenomena, and can lead to a variety of different assumptions as to which aspects of the analysis are relevant. Tognini Bonelli characterizes CL along three lines: 'it is an *empirical approach* to the description of *language use*; it operates within the framework of a *contextual and functional theory of meaning*; it makes use of the *new technologies*' (2001: 2; cf. also Gries 2006, Guilquin and Gries 2009).

CL, rationalism and traditional descriptive linguistics

It is a commonplace in discussions such as these to mention the difficulties encountered by

CL in its early years. The Brown corpus appeared at the moment in which generative linguistics was taking off, and the differences in approaches were so great that it seemed impossible to find any common ground. Chomsky in particular voiced strongly critical considerations to corpus implementation, corpus use, and the role of statistics in grammar. However partially valid his arguments here, we should recall that what Chomsky had in mind was mainly the conception and use of corpora by distributionalists (Caravedo 1999), and also that CL has changed in many significant respects since these criticisms were made. In fact, generative linguistics and CL have both changed immeasurably since then, rendering the great majority of these points either irrelevant or at least of only a secondary nature (Rojo 2010a).

The difference between these two great paradigms in current linguistics is clear and can probably be founded on the conception and utilization of data. According to Aarts (2000, 2002), data may be intuitive or non-intuitive. The former are the result of introspection or of judgments made by other speakers. Non-intuitive data ‘are provided by what people actually say and write’ (2002: 4) and can be either anecdotal or drawn from corpora. The use of intuition-based data typifies Chomskyan linguistics. On the other hand, the essential characteristic of relying on non-intuitive data is not the type of data itself, but rather the way of collecting and analyzing that data, and is the main difference between traditional descriptive linguistics and CL. It has often been argued that linguistics has always employed the systematic collection and use of real data in diachronic studies, for example. CL in this sense follows the same basic procedure, and would not thus constitute a new methodology. Yet this is only partially true, because the use of data in these traditional areas of linguistic studies falls within the category of anecdotal data indicated by Aarts, in the sense that their collection is neither systematic nor exhaustive, and that these data are initially selected in function of their assumed relevance for the specific analysis in question. The use of

computers to store and retrieve information means that CL can aim for what Leech (1992) and Quirk (1992), among others, have called ‘total accountability’, that is, exhaustive analysis, without prior selection, of all the data in a corpus.

Size of corpora and representativeness

Relatively few years ago – as a consequence of the limitations of computers at the time – textual corpora were either small (one million words, following the Brown corpus model) or relatively small (one hundred million forms, following the BNC model). In both cases, compiled texts were carefully selected for representativeness, balanced and highly encoded. These days, reference corpora often contain hundreds or even thousands of million words, mainly downloaded from the internet or directly incorporated from existing electronic formats, and almost always paying only passing attention to the old questions of representativeness and balance. Indeed, a more radical formulation, ‘web as corpus’, defends the direct use of all available material on the web, with a volume clearly far greater than anything that might be integrated into a reference corpus, and with no cost of implementation or software issues, in that commercial search engines can be used as the user-interface.

Size and representativeness, at least in CL, are closely related. Hence, if you need to get a representative sample of a certain linguistic variety in a corpus of only one million words, it is necessary that the corpus be composed of many short texts, carefully selected for origin, type, topic, and so on. Such characteristics assume greater importance if your searches are global, involving the whole of the corpus. However, constructing a representative sample of a linguistic variety is itself problematical, in that we don’t know the real quantitative characteristics of the linguistic universe we are trying to capture. Decisions as to the percentage of oral texts, of different types of written texts, and texts from different countries, for example, will always be approximate, in that no strict criteria can exist (Baker 2010a).

Fortunately, the difficulties arising from issues of representativeness and balance in

textual corpus are now very much reduced. The great size of today's reference corpora itself solves many of these problems. However, and most importantly, it is the current corpus size together with the processing capacity of modern computers that has radically changed issues here. In general, it is not the total frequency of a word or other form that interests the corpus linguist, but rather its frequency in different types of texts, texts produced at different times, in different countries, etc. Given that normalized frequencies are generally used, differences in the sizes of sets involved in comparisons are no longer of great concern, as long as corpus and subcorpora sizes satisfy requirements as to balance.

Texts and corpus

A corpus consists of a series of (pieces of) texts, generally a very large number of them. However, seen as a whole this is much more than simply a collection of texts. Indeed, from the very beginning of CL, the difference between the simple accumulation of texts in electronic form (an *archive*) and the integration of a series of texts according to a certain design has been noted (Atkins *et al.* 1992).

One distinguishing feature of corpora is the very existence of a design. Texts are selected based on their compatibility with criteria relating to text type, temporal distribution, the relative weight of different elements included, etc. and should reflect the representative and balanced character of the corpus. Encoding (cf. below) is added to the text, and can subsequently be used as a means of identifying subcorpora from the whole. All of which implies that 'Web as Corpus' is, in a strict sense, not an adequate expression. The plethora of texts found on the web have no design and no general unified purpose (Sinclair 2005). Using the traditional terms, we might say *Web as Archive,* in that the web can indeed be searched to find occurrences of different expressions, but this can go no further in terms of the analysis of each occurrence. Only a corpus, compiled according to a design, allows us to move from individual level of a text to an understanding of it from a broader, systematic perspective.

Guillermo Rojo

As Tognini Bonelli (2010: 18-20) observes, there are many differences between reading texts and reading a corpus. A text is read line by line, whereas a corpus is typically analyzed by looking at concordances of specific forms across a variety of sources. The text 'is an instance of *parole* while the patterns shown up by corpus evidence yield insights into *langue*'

Corpus encoding

The encoding of corpora has undergone important modifications in recent years. Given their electronic format, character representation in texts needs to be encoded according to a specific system. Even here there are difficulties, with many current programming languages not able to cope with so-called 'special characters', that is, those not belonging to the set used in the alphabet normally used in standard American English.

But 'encoding' has two further senses in CL. The first we might call extra-textual encoding. This consists of the indication, in a way that the search application can handle, of the bibliographical data of every text, and including at least the features of year of publication or production, the name, nationality (and, if possible, gender and age) of the author, in addition to all features used in the basic corpus configuration. Naturally, it is extra-textual encoding that makes possible the selective recuperation of data.

Second is intra-textual encoding, by which we mean encoding that refers to the text structure and other possible factors. Encoding text structure has limited importance. The rise in size of corpora, and the increasing use of texts in electronic format for which no prior printed versions exist, has rendered the indication of features such as page number irrelevant. Of greater importance are aspects related to the internal structure or the text or the characteristics of the edition, mainly in corpora with a diachronic orientation or with data on spoken language: citations, errata, turns, overlaps and additional interventions in oral texts, etc.

Much of this was simply impossible until the introduction of SGML (*Standard Generalized Mark-up Language*) and its derivations (XML mainly). The *Text Encoding Initiative*, in its successive editions, established a *de facto* standard for corpus projects. From today's perspective, it seems that in some aspects at least, text encoding was a target in itself and was not always seen as a means of facilitating the extraction of information from corpora.

Nowadays things are simpler and more efficient. The degree of extra-textual and intra-textual encoding depends on the characteristics of a text and the objectives established for the corpus. Thus, a small corpus composed of texts of a very specific type (say, medieval bibles in Spanish) must have a high degree of encoding so that all textual and hypertextual information relevant to this text type is available. On the other hand, huge corpora compiled from material downloaded directly from the web can include only data that can be obtained and integrated automatically. Between these two extremes are reference corpora, normally containing hundreds of million words, and including information relevant to the construction of the corpus which will allow for diverse forms of selective searches.

Corpus parsing and tagging

What is first seen with a corpus is usually an electronic version of a written text, that is, a set of orthographic words with some typographical marks expressing additional information. This extremely useful resource is, nevertheless, limited in its uses by the orthography itself: what can be recovered relates only to the graphic presentation of any elements we might be interested in. Of course, depending on the morphological characteristics of the languages, it is possible to simulate morphological characteristics via orthographic forms. A search using the expression *cant**, for example, will recover all forms belonging to the paradigm of the regular verb *cantar*. But the problems are immediately evident: such a search will also return every other form whose first four characters are *cant* (*cantera*, *cantuesa*, *cantina*, *cantimplora*, etc.), forms linked to the verb *cantar* but not belonging to its paradigm (*cantor*, *cantante*,

etc.), and, of course, cases of homographs of forms of the verb *cantar* (as *canto* 'stone'). Furthermore, if a corpus were to contain only the electronic equivalent of written forms, searches for abstract grammatical features, such as noun + adjective + adjective or verb + preposition, would not be possible.

Solving this problem involves the addition of this lexical and grammatical information. So, the form *llegaremos* will be associated with the set of indications 'first person, plural, indicative, future of the verb *llegar*'. This can only be done automatically when working with a corpus of hundreds of million of words, and hence computational linguistics is drawn into the process (see Martí and Taulé, this volume). The objective of morphosyntactic tagging is to associate the corresponding tag (in a system with an acceptable degree of generalized usability) to every form in a text. This task presents major difficulties, the specific nature of which depend on the morphosyntactic characteristics of a particular language. The size and complexity of such a task is well exemplified by the form *la*, which is the second most frequent orthographic word in current Spanish (4,11% of all words in the CREA without taking into account its appearance as an enclitic form -*míra*la, *mirar*la, *mirándo*la, etc.-): *la* can be an article (*la lámpara*), a personal pronoun (*la trajeron*) or a noun (*la* the note of the musical scale). Indeed, one of the main problems with Spanish in this sense is the prevalence of homographs. Such cases require disambiguation (not an especially appropriate term, in that there is normally no real ambiguity in the text) and the use of the correct tag in each context.

At this level of analysis, taggers handle contextual information introduced via statistical considerations, via contextual rules, or via a combination or both systems. It is necessary to assume that no automatic system can be 100% successful in this task. Indeed, even two linguists working on the same text might well have certain not trivial differences about the attribution of tags at various points in a text. Furthermore, automatic tagging often

relies on contextual information within a range of a few words on either side of the word in question, whereas in many cases the appropriate information might be found several sentences away. Finally, the degree of success depends on the granularity of the tag system and the objectives established. For example, you might try to assign only the indication of part of speech and lemma (with no grammatical features), decide which cases of *cantaba*, *decía*, etc. correspond to the first or to the third person, or try to clarify whether the orthographic form *decírselo* must be analyzed in the same way as expressions like *decirse algo a sí mismo*, *decirle algo a alguien* or *decirles algo a algunas personas*.

Prior to tagging a text, it must be parsed, involving as a first step the segmentation and identification of tokens (usually known as tokenization), generally orthographic words. Following this, it is necessary to segment and identify the fragments of the text that will be used as the context for tagging. Morphosyntactic tagging is then necessary as a means of allowing for syntactic, semantic and pragmatic analysis and annotation. Many of these tools (or even all of them) are necessary for specific practical applications, such as machine translation, opinion mining, automatic summarization, natural language understanding, natural language generation, etc. (Lavid 2005).

Corpus typology

The evolution seen in CL over the last half century has included its integration into mainstream linguistic studies. The compilation and use of corpora is now a common practice in many different fields of study, not restricted to linguistics, and includes corpora of phone calls, patient-doctor conversations, the language production of foreign language learners, etc. In all these cases, as well as many others, the technical component and point of departure are the same, with differences lying only on the specific characteristics of the texts involved. They usually share also a secondary, but important, feature of linguistic corpora: their public character, that is, the fact that corpora are built, encoded and tagged with the clear intention

of providing access to the totality of the corpus to those professionals interested in it or, when owners rights make this impossible, they allow for searches of its contents through concordances, the analysis of collocates, phraseology, etc.

From a general point of view, a corpus can have different orientations or characteristics depending on the types of texts it contains and their mutual relations. So, a corpus may be built with a synchronic or a diachronic orientation, may look for the existence of diatopic, diastratic or diaphasic differences or it might focus on what can be considered the standard variety of a language. The former is the case of GRIAL, in which the differences related to genres of texts are the objective of study (Parodi 2010b), the *Corpus oral de lenguaje adolescente* (COLA, cf. Hofland *et al.* 2005), or the corpus *Iberia*, integrated by scientific texts (Porta Zamorano *et al.* 2011). And, of course, attending to the processes developed with the texts, a corpus might be encoded or not, morphosyntactically tagged or not, and syntactically analyzed or not.

Leaving aside these general issues, the first factor with regard to corpus typology is the nature of the texts included: novels, oral texts, foreign language learners' production, child language, newspapers, technical writing, parliamentary speeches, or indeed combinations of several of these. A second, associated factor is that of the difference between a reference corpus (that is, a general corpus built with the objective of representing the general characteristics of a certain linguistic variety at a certain moment or over a given period) and corpora with specific purposes (learner language, technical corpus, training corpus, corpus for the study of specific types of texts such as bibles, goliardic poetry, works from one author, cultural trends and movements, etc.).

In the early days of CL the difference was established between open and closed corpora. A closed corpus is designed with a specific size and distribution of the different types of texts it can include; when these objectives are reached, the corpus is complete, as is

the case with the CE or AnCora-ES. An open corpus, on the contrary, has a general design, but new texts can always be added to it. Thus, a closed corpus, once finished, remains identical, but becomes obsolete for many of its initial purposes after only a few years. On the other hand, an open corpus is potentially in a state of constant renewal, and these changes, although enriching the content, present difficulties in the reproduction and comparison of search findings at different points in time. Sinclair (1991) modified this 'classical' view by introducing the concept of 'monitor corpus', a corpus which would process continuously a great amount (at the time of the proposal) of text, processing the information contained in them and storing the results (Teubert and Čermáková 2007). A different but related type of corpus is the intermediate solution devised for *Corpus of Contemporary American English* (COCA, cf. Davies 2009) and the *Corpus del español del siglo XXI* (CORPES). In its first phase, the CORPES is intended to contain 25 million words for each of the years 2001 to 2012 (300 million words in total) and will continue to grow in annual increments of 25 million words thereafter.

A corpus can be complete (containing the complete works by an author, members of a literary school or movement, the whole print history of a newspaper, etc.) or can comprise a sample of the productions of the members of the linguistic community (which leads to problems on representativeness and balance noted above). Depending on factors related to its objectives, scheduled distribution, rights and size, a corpus can be composed only of fragments of texts, complete texts, or indeed a combination of both. Finally, a corpus can be monolingual or multilingual, and the latter may be comparable (texts of the same type in different languages) or parallel (the 'same' text in different languages, i.e. translations).

The influence of CL

As described in previous sections, the third phase in the history of CL includes its integration in almost every area of linguistic study. In very many technical studies in linguistics, one or

more textual corpora are used, at least as a data source, irrespective of whether the work might be considered to form part of CL in a strict sense.

Despite this general extension, the influence of CL is currently felt more strongly in some fields of linguistics than others. The most important field here is practical lexicography (or 'dictionaristics'). It constitutes the linguistic area in which CL has been most thoroughly integrated, and also serves to refute the well extended (yet false) argument that computers and the empirical cultural sciences are at best distant relations.

It is difficult, perhaps even impossible, to conceive nowadays of the existence of a lexicographical project without the use of computers in each of its components and phases, and more specifically the use of corpora (either of general use or built for that project) in the first two phases recognized by Zgusta (1971) (cf. Rojo 2009): the collection of materials, and the determination of the relation of lemmas that the dictionary will contain. We might recall that the first corpora were used mainly for the analysis of the frequency of words and the determination of their contexts of occurrence and meanings. Following the construction of the Brown corpus and its British counterpart (the Lancaster-Oslo-Bergen corpus), John Sinclair began to develop the COBUILD project, whose main characteristic was specifically the determination of words and their senses as they were found in 'real English' through their presence and behavior in a textual corpus (Sinclair 1987). Some years later, the Longmann corpus adopted the same orientation, and the BNC can be considered as a natural consequence of both these. In the Spanish context, the Real Academia Española took the decision in 1995 to compile the *Corpus de referencia del español actual* (CREA) and to adopt it as the main source of documentary data for the DRAE, its general dictionary. Some months later, a similar decision was adopted with respect to the documentation for the *Diccionario histórico*, leading to the *Corpus diacrónico del español* (CORDE) and, more recently, to the more specific *Corpus del nuevo diccionario histórico del español* (CDH).

The existence of diachronically oriented corpora has exerted a great influence on the study of the history of languages, especially in terms of their grammatical component. There is a large body of work on grammatical changes in very different languages. It is clear that the construction of a corpus containing texts from different periods, in which even the characters used may differ, and with important problems regarding authorship, date of composition, authentication of the text, edition, etc. implies many additional problems. Indeed, CORDE, containing 280 million words from the origins of Spanish up to 1974, is the exception and not the rule. Recently, the approaches known as 'modern diachronic corpus' (cf. Mair 2009) and 'comparative corpus linguistics' or 'short-term diachronic comparable corpus linguistics' (cf. Leech *et al.* 2010) has arisen with the study of changes produced in the last thirty or fifty years of a language as its objective .

A third area in which CL has led to significant changes in the way of working is sociolinguistics (cf. Baker 2010b; Kendall and van Herk 2011; Romaine 2008). Returning to Lope Blanch's proposal for developing the project on the *norma culta* (cf. section 2.1), we note that the project was not conceived of as a textual corpus *per se*, yet the great number of texts collected could have constituted a fine corpus, comparable to the *International Corpus of English* (ICE). In fact, the *Asociación de Lingüística y Filología de América Latina* (ALFAL) made a selection from the interviews from *Norma culta*, normalizing their distribution among cities, genders, ages and sociocultural levels, and used them to compile the ALFAL corpus (Samper 1995; Samper *et al.* 1998), subsequently distributed on CD. The ALFAL corpus and many of the other original interviews of the *Norma culta* have also been included in the CREA. Moreover, many of the interviews from the *Proyecto para el estudio sociolingüístico del español de España y de América* (PRESEEA), currently in its last phase will form part of the oral component of the CORPES.

CL is widely used in applied linguistics (Hunston 2002) and in every field of

linguistic research (Meyer 2002), but it is no coincidence that practical lexicography, historical linguistics and sociolinguistics can these days be considered the three main areas in which CL shows the greatest development. Given the tendency in traditional lexicography to write new dictionaries based on older ones, with scant attention to real language use, the availability of data on meanings, frequency, contexts for use, distribution, etc. constitutes a decisive change, with important repercussions in every field of practical lexicography, including the historical area. Historical linguistics, sociolinguistics, and historical sociolinguistics (Conde Silvestre 2007) constitute what, from an integrated perspective, is now known as variation and change. Textual corpora allow access to data on the distribution of linguistic phenomena in different time periods, countries, genres of texts, etc., and the ability to relate findings to speaker-characteristics such as age, gender, and sociocultural level or linguistic registers (Parodi 2007). Clearly, the advantages of selective recovery, as mentioned above, are crucial here.

Quantitative analysis

As has been commonly observed, quantitative analysis is an important characteristic of CL. Naturally, this implies far more than simple counts of words and their meanings or of grammatical forms. What is really relevant is not the general frequency, but the differences in frequencies among different corpora or different subcorpora of the same corpus. Hence some statistical techniques must be employed, which itself entails that a grounding in linguistic studies should include a theoretical and practical knowledge of statistics. Indeed, Gries (2010 and many other places) argues that statistics ought to be a part of university linguistics programs, just as is the case in psychology, sociology, etc.

In support this argument we might bear in mind the enormous change that the consideration of frequency in linguistic studies has brought about (Bybee and Hopper 2001). Undervalued in traditional linguistics, and hardly even addressed in the initial stages of

Chomskyan linguistics, frequency has become a central aspect of all areas of linguistic and related analytical fields, from sociolinguistics to historical linguistics, and thoroughly embracing phonology, morphology and syntax. The change in attitude with respect to frequency has a complex set of causes (Bybee 2007), but what cannot be ignored is the importance of being able to obtain and analyze data from huge textual corpora and making comparisons with the frequencies from different parts of the same corpus.

Looking into the future

In two different, but parallel processes, the development of CL has brought with it evident reductions in time, effort and money needed to build a corpus and at the same time the increasing acceptance across all linguistic disciplines of this way of analyzing phenomena. As a consequence, CL has shifted from the periphery of linguistic studies, the place where the Brown corpus was conceived, to its center. In the coming years we will see an extension and intensification of this process. As we can appreciate, corpora have broadened and deepened our knowledge of languages and in doing so have become an essential element in linguistic research. However, besides their main role in theoretical and descriptive linguistics, we might also note that corpora are also seen as more general resources, and as such are available to many different fields of applied linguistics, from the compilation of dictionaries and reference grammars to translation, through a host of other specialties, among them forensic linguistics, stylistics and even cultural studies.

With input methods based on the keyboard and OCR scans now a thing of the past, the work for integrating texts becomes ever simpler and cheaper, since an increasing proportion of texts exist in electronic format or are indeed directly published in this form. Changes with newspapers over the last twenty years perhaps gives us a good taste of what can be expected in the future of CL. (Rojo and Sánchez 2010).

The saving in time, effort and cost is, of course, reflected in the encoding of the texts

in a corpus. We must take into account differences due to the existence of many types of corpus, with a plethora of objectives and characteristics. As already noted, the extreme points here are the use of web as corpus on the one hand, and on the other a corpus of reduced size, composed of texts of a relatively homogeneous nature and with a high degree of encoding marks and complementary utilities (links among different versions, translations, and so on). Reference corpora, occupying the middle ground, typically contain only the encoding marks useful for general uses. Thus, according to one's specific objectives, the size and complexity of the corpus necessarily changes. For example, if one is interested in the relative importance of alternative graphic or morphological variants (such as *zinc* or *cinc*; *asola* or *asuela*), a web search will provide useful and relevant data; but Google will be of little use in the differentiation of linguistic variants by country or text type, and in such a case a reference corpus will be necessary (Kilgarriff 2007).

We can expect that progress in computational linguistics will lead to better morphosyntactic taggers for both general and specific purposes. Such advances will help us to increase our knowledge of grammatical structures, and in the field of computational linguistics, corpora that are syntactically analyzed, and semantically and pragmatically annotated, will serve as the basis for applications such as opinion mining, natural language production, etc.

Finally, search applications will be able to provide users with sets of instances of sequences or cases of grammatical structures extracted from corpora of 500 or 1000 million words almost instantly, will be able to order results with respect to various parameters, and will calculate collocates in subsets of the corpus, giving the corresponding normalized frequencies, etc.

The future for CL lies in the integration of all these different functions. The cost and complexity of each of these tasks currently produces a situation in which a particular corpus

tends to focus on one such function. Thus, the Spanish component of the *Leeds collection of Internet Corpora* has 145 million words downloaded from the Internet and automatically tagged, but only global searches are possible, without the option of selecting by genre of text or country. The CE consists of 100 million words from the earliest texts of Spanish to the end of 20th century, is automatically tagged and partially disambiguated, gives normalized frequencies for centuries, and, in texts from the 20th century, also for text type. But its speed is due to the fact that results are 'frozen', and you cannot select by countries, or by periods other than centuries, or indeed by text genre, other than in the 20th century. CREA and CORDE, on the contrary, are considerably more flexible in terms of the features of texts in searches and also the ordering of instances, although they are not tagged, so searches are restricted to the orthographic forms of the expression (Davies 2008, Rojo 2010b).

The future lies in the possibility of integrating all these capabilities, with corpora of great size supporting lexical and grammatical searches using abstract features, in which subcorpora can be searched, and where results include normalized frequencies, collocates in subcorpus are shown, etc., and, of course, where results can be ordered and reordered in different ways. The *Corpus del español del siglo XXI* (CORPES), the first public version of which is scheduled for the end of 2013, aims for these targets.

The second great challenge for CL in the coming years concerns oral texts (Briz and Albelda 2009). The BNC model established 10% of the total corpus for transcriptions of talk, interviews, radio and TV programs, etc. Such a proportion is evidently small, yet it is very difficult to achieve even this objective due to the high costs of transcription, a general estimate being that it takes about twenty times the length of a recording to transcribe it. Of course, exclusively oral corpora suffer from this problem more than any, and thus are usually small in size. This can only be resolved by the re-use of material such as that from the PRESEEA project and, more importantly, by the possibility of using automatic transcription

programs, capable of transcribing speech from a wide variety of speakers, with different accents, and where sound quality is not optimum.

A substantial part of the work with oral texts comes from the need to complement the strict transcription of the sounds to written text with the encoding of specific phenomena (variants in pronunciation, hesitations, broken words, overlaps, etc.). However, we have now the interesting possibility of aligning the text of the transcription with the corresponding part of the sound chain. The alignment of transcription and sound allows us to reduce the complexity of encoding marks (making easier the localization of textual forms) without renouncing to the study of the sound. A second and important step in this direction comes from the possibility of taking a similar approach to images. Thus, we now have the possibility of aligning textual transcription with sounds and images. Given the relationship between these three components, it is simple to locate the fragment we are interested in through the text of the transcription and at the same time to recover the sound and the images associated with it. The possibility of tagging sound and image will ultimately offer the means of a fully integrated study of linguistic phenomena.

Conclusion

In its fifty years of history, CL has moved from the peripheries of linguistic studies to become a central methodology, used in almost every sub-discipline, and with very different purposes. The evolution of computers has allowed for the growth in corpus size, but also in the typology of texts included (extension) and the richness of information added to the text (depth). At the same time, search tools have offered the linguist a wealth of complexity in what can be searched for, but also a simplification in terms of the special knowledge and resources required for their use.

Increased corpus size and the addition of the relevant features for the classification of the texts in a corpus have radically changed the linguist’s working environment and the way

in which some of the fundamental topics of the field are now understood. Certainly, corpora must be representative and balanced, but we now understand these terms in greater depth, and work is carried out largely through the comparison of what is seen in various subcorpora (dynamically built). Furthermore, advances in computational linguistics have allowed the enrichment of texts with many different informative tags linked to forms and sequences in texts. Morphosyntactic tagging, the most elementary form of tagging, makes possible the formulation of abstract grammatical features in searches. Naturally, successive levels of tagging allows for more complex searches.

It is important to note a change of perspective brought about by the use of corpora, especially with respect to the notion of 'total accountability'. The objective is not, of course, the description of what a corpus contains, but the analysis of these data in order to understand the system. It is true that this objective could also be found in traditional descriptive linguistics, but its attainment was impeded by the fragmentary character of the data used. Corpora are providing new and complete data for a correct understanding of current Spanish and how it changes through time and space.

Finally, the quantitative component is a crucial aspect of CL. It is no coincidence that the development of CL has seen a parallel rise in the general use of frequency in all lexical and grammatical studies. Frequency analysis is impossible without using the type of data corpora can provide.

Related topics

computational linguistics, frequency, functional grammar, generative grammar, lexicography, grammar, sociolinguistics, syntax

Further reading

Baker, P. (2010). (An overview of the main topics in which CL and Sociolinguistics coincide and influence each other.)

Gries, S. T. (2009). (A general presentation of the main characteristics of CL from a more theoretical point of view.)

Hockey, S. (2000). (A general view of how and where the widespread adoption of electronic texts has influenced different humanistic disciplines.)

Hunston, S. (2002). (A discussion of how CL has been embraced by applied linguistics, and the changes this has implied for both subdisciplines.)

Lavid, J. (2005). (An exploration of new areas and topics for linguistic research in the 21st century.)

Parodi, G. (ed.) (2007). (A set of corpus-based studies on different aspects of Spanish.)

Sinclair, J. (1991). (The classic work on corpora and the way they can be used for studies on lexis and grammar.)

URLs for corpora and other electronic resources mentioned in the text

AGLE (Archivo general de la lengua española): www.cvc.cervantes.es/lengua/agle/.

AnCora-ES: http://clic.ub.edu/corpus/ancora.

.BNC (British National Corpus): www.natcorp.ox.ac.uk/.

Brown Corpus (The Standard Corpus of Present-Day Edited American English):

www.helsinki.fi/varieng/CoRD/corpora/BROWN/.

CDH (Corpus del nuevo diccionario histórico):

www.frl.es/Paginas/Corpusdiccionariohistorico.aspx.

CE (Corpus del español): www.corpusdelespanol.org/.

COCA (Corpus of Contemporary American English): corpus.byu.edu/coca/.

COLA (Corpus oral del lenguaje adolescente): www.colam.org/om_prosj-espannol.html.

CORDE (Corpus diacrónico del español): http://rae.es/recursos/banco-de-datos/corde

CORPES (Corpus del español del siglo XXI): http://rae.es/recursos/banco-de-datos/corpes-xxi

CREA (Corpus de referencia del español actual): http://rae.es/recursos/banco-de-datos/crea

GRIAL: www.elv.cl/prontus_linguistica/site/edic/base/port/grial.html.

IBERIA (Corpus de español científico): www.investigacion.cchs.csic.es/elci/node/8.

ICE (*International Corpus of English*): http://www.ucl.ac.uk/english-usage/projects/ice.htm

Leeds collection of Internet Corpora: corpus.leeds.ac.uk/internet.html.

LOB (Lancaster–Oslo/Bergen Corpus): www.helsinki.fi/varieng/CoRD/corpora/LOB/.

PRESEEA (Proyecto para el estudio sociolingüístico del español de España y de América): preseea.linguas.net/.

SEU (*Survey of English Usage*): http://www.ucl.ac.uk/english-usage/index.htm

References

Aarts, J. (2000). 'Towards a new generation of corpus-based English grammars'. In B. Lewandowska Tomaszczyk and P. J. Melia (eds.), *PALC '99. Practical Applications in Language Corpora. Papers from the International Conference at the University of Lodz* (pp. 17-36). Frankfurt am Main: Lang.

-----. (2002). 'Does corpus linguistics exist? Some old and new issues'. In L. E. Breivik and A. Hasselgren (eds.), *Language and Computers. From the COLT's Mouth . . . and Others* (pp. 1-17). Amsterdam: Rodopi.

Alameda, J. R. and Cuetos, F. (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano.* Oviedo: Universidad de Oviedo.

Almela, R., Cantos, P., Sanchez, A., Sarmiento, R. and Almela, M. (2005). *Frecuencias del español: Diccionarios y estudios léxicos y morfológicos.* Madrid: Universitas.

Atkins, S., Clear, J. and Ostler, N. (1992). 'Corpus design criteria'. *Literary and Linguistic Computing* 7 (1): 1-16.

Baker, P. (2010a). 'Corpus linguistics', in L. Litosseleti (ed.), *Research Methods in Linguistics* (pp. 93-113). London: Continuum.

-----. (2010b). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Briz, A. and Albelda, M. (2009). 'Estado actual de los corpus de lengua española hablada y escrita: I+D'. In *El español en el mundo. Anuario del Instituto Cervantes 2009* (pp. 165-226). Madrid: Instituto Cervantes.

Bunge, M. (1968). 'The Maduration of Science'. In I. Lakatos and A. Musgrave (eds.), *Problems in the Philosophy of Science. Proceedings of the International Colloquium in the Philosophy of Science (London, 1965)* (pp. 120-147). Amsterdam: North-Holland.

Bybee, J. (2007). *Frequency of Use and the Organization of Language.* Oxford: Oxford University Press.

Bybee, J. and Hopper, P. (2001). 'Introduction to frequency and the emergence of linguistic structure'. In J. Bybee and P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 1-24). Amsterdam: John Benjamins.

Caravedo, R. (1999). *Lingüística del corpus. Cuestiones teórico-metodológicas aplicadas al español*. Salamanca: Ediciones Universidad Salamanca.

Cejador, J. (1905-1906). *La lengua de Cervantes.* Madrid: J. Ratés.

Conde Silvestre, J. C. (2007). *Sociolingüística histórica*. Madrid: Gredos.

Crystal, D. (1991). *A Dictionary of Linguistics and Phonetics.* Oxford: Blackwell.

Davies, M. (2006). *A Frequency Dictionary of Spanish. Core Vocabulary for Learners*. Oxon: Routledge.

-----. (2008). 'New directions in Spanish and Portuguese corpus linguistics'. *Studies in Hispanic and Lusophone Linguistics* 1 (1): 149-186.

-----. (2009). 'The 385+ million word *Corpus of Contemporary American English* (1990-2008+): Design, architecture, and linguistic insights'. *International Journal of Corpus Linguistics* 14 (2): 159-190.

Dyson, F. (1997). *Imagined Worlds.* Cambridge, MA: Harvard University Press.

-----. (1999). *The Sun, the Genoma, the Internet*. Oxford: Oxford University Press.

Francis, N. W. (1982). 'Problems of assembling and computerizing large corpora'. In S. Johansson (ed.), *Computer Corpora in English Language Research* (pp. 7-24). Bergen: Norwegian Computing Centre for the Humanities.

-----. (1992). 'Language corpora B. C.'. In J. Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (pp. 17-34). Berlin: Mouton de Gruyter.

García Hoz, V. (1953). *Vocabulario usual, vocabulario común y vocabulario fundamental*. Madrid: CSIC.

Gries, S. T. (2006). 'Introduction'. In S. T. Gries and A. Stefanowitsch (eds.), *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis* (pp. 1-17). Berlin: Walter de Gruyter.

-----. (2009). 'What is corpus linguistics', in *Language and Linguistic Compass* 3: 1-17.

-----. (2010). 'Methodological skills in corpus linguistics. A polemic and some pointers towards quantitative methods'. In T. Harris and M. Moreno Jaén (eds.), *Corpus Linguistics in Language Teaching* (pp. 121-146). Frankfurt am Maine: Peter Lang.

Guilquin, G. and Gries, S. T. (2009). 'Corpora and experimental methods: A state-of-the-art review'. *Corpus Linguistics and Linguistic Theory* 5 (1): 1-26.

Hockey, S. (2000): *Electronic Texts in the Humanities. Oxford: Oxford University Press..*

Hofland, K., Jørgensen, A., Drange, E-M. and Stenström, A-B. (2005). 'COLA: A Spanish spoken corpus of youth language'. In *Proceedings from the Corpus Linguistics Conference Series*. Birmingham: University of Birmingham Center for Corpus. Retrieved from: www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/SpokenDisclosure/cl-195-pap-COLA.doc.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Juilland, A. and Chang-Rodríguez, E. (1964). *Frequency Dictionary of Spanish Words*. The Hague: Mouton de Gruyter.

Kendall, T. and van Herk, G. (2011). 'Corpus linguistics and sociolinguistic inquiry: Introduction to special issue'. *Corpus Linguistics and Linguistic Theory* 7 (1): 1-6.

Keniston, H. (1937a). *The Syntax of Castilian Prose. The Sixteenth Century.* Chicago, IL: The University of Chicago Press.

-----. (1937b). *Spanish Syntax List: A Statistical Study of Grammatical Usage in Contemporary Spanish Prose on the Basis of Range and Frequency.* New York: H. Holt and Company.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London/New York: Longman.

Kilgarriff, A. (2007). 'Googleology is bad science'. *Computational Linguistics* 33 (1) 147-151.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* Chicago, IL: The University of Chicago Press.

Lavid, J. (2005). *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Madrid: Cátedra.

Leech, G. (1992). 'Corpora and theories of linguistic performance'. In J. Svartvik (ed.),

Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 (pp. 105-122). Berlin: Mouton de Gruyter.

Leech, G., Hundt, M., Mair, C. and Smith, N. (2010). *Change in Contemporary English*. *A Grammatical Study*. Cambridge: Cambridge University Press.

Lope Blanch, J. M. (1986). *El estudio del español hablado culto. Historia de un proyecto.* México: UNAM.

Mair, C. (2009). 'Corpora and the study of recent change in language'. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook. Volume II* (pp. 1109-1125). Berlin: Walter de Gruyter.

McCarthy, M. and O'Keeffe, A. (2010). 'Historical perspective: What are corpora and how have they evolved'. In A. O'Keefe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 3-13). Oxon: Routledge.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006): *Corpus-Based Language Studies*. Oxon: Routledge.

Meyer, C. F. (2002). *English Corpus Linguistics. An Introduction.* Cambridge: Cambridge University Press.

-----. (2008). 'Pre-electronic corpora'. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook. Volume 1* (pp. 1-14). Berlin: Walter de Gruyter.

Parodi, G. (ed.) (2007). *Working with Spanish Corpora*. London/New York: Continuum.

-----. (2010a): *Lingüística de corpus: De la teoría a la empiria*. Madrid: Iberoamericana Vervuert.

-----. (ed.) (2010b). *Academic and Professional Discourse Genres in Spanish*. Amsterdam: John Benjamins.

Porta Zamorano, J., Del Rosal García, E. and Ahumada, I. (2011). 'Design and development of Iberia: A corpus of scientific Spanish'. *Corpora* 6 (2): 145-158.

Quirk, R. (1992). 'On corpus principles and design'. In J. Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (pp. 457-469). Berlin: Mouton de Gruyter.

Renouf, A. (2007). 'Corpus development 25 years on: From super-corpus to cyber-corpus'. In R. Facchinetti (ed.), *Corpus Linguistics 25 Years On* (pp. 27-49). Amsterdam/New York: Rodopi.

Rojo, Guillermo. (2009). 'Sobre la construcción de diccionarios basados en corpus', *Tradumàtica* 7. Retrieved from: webs2002.uab.es/tradumatica/revista/num7/articles/02/02art.htm.

-----. (2010a). 'Aguja de navegar corpus'. In V. M. Castel and L. Cubo de Severino (eds.), *La renovación de la palabra en el bicentenario de la Argentina. Los colores de la mirada lingüística* (pp. 1151-1163). Mendoza: Editorial FFyL-UNCuyo.

-----. (2010b). 'Sobre codificación y explotación de corpus textuales: Otra comparación del *Corpus del español* con el CORDE y el CREA'. *Lingüística* 24: 11-50.

Rojo, G. and Sánchez, M. (2010). *El español en la red.* Madrid/Barcelona: Fundación Telefónica/Ariel.

Romaine, S. (2008). 'Corpus linguistics and sociolinguistics'. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook. Volume 1* (pp. 96-111). Berlin: Walter de Gruyter.

Samper Padilla, J. A. (1995). 'Macrocorpus de la norma lingüística culta de las principales ciudades de España y América'. *Lingüística* 7: 263-293.

Samper Padilla, J. A., Hernández Cabrera, C. E. and Troya Déniz, M. (eds.) (1998). *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo*

hispánico. Las Palmas: Universidad de Las Palmas de Gran Canaria.

Sinclair, J. (1987). 'Introduction'. In *Collins Cobuild English Language Dictionary* (pp. xv-xxi). London: HarperCollins Publishers.

-----. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

-----. (1996). *Preliminary Recomendations on Corpus Typology* (EAGLES document EAG-TCWG-CTYP/P). Pisa: Consorzio Pisa Ricerche. Retrieved from: www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html.

-----. (2005). 'Corpus and text. Basic principles'. In M. Wynne (ed.), *Developing Linguistic Corpora. A Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books.

Svartvik, J. (2007). **'**Corpus linguistics 25+ years on'. In R. Facchinetti (ed.), *Corpus Linguistics 25 Years On* (pp. 11-25). Amsterdam/New York: Rodopi.

Svensén, B. (1993). *Practical Lexicography*. Oxford: Oxford University Press.

Teubert, W. and Čermáková, A. (eds.) (2007). *Corpus Linguistics. A Short Introduction.* London/New York: Continuum.

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

-----. (2010). 'Theoretical overview of the evolution of corpus linguistics'. In A. O'Keefe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 14-27). Oxon: Routledge.

Zgusta, L. (1971). *Manual of Lexicography*. The Hague: Mouton de Gruyter.