INTRODUCTION TO THE ISSUE ON CORPUS LINGUISTICS

Marco Rocha
Universidade Federal de Santa Catarina

Since the book named *Corpus linguistics* (Aarts and Meijs 1984) appeared in the mid-eighties, the idea of using a machine-readable corpus to do research related to language spread gradually to become in time a fairly well established approach in a number of areas within the fields of linguistics, human language technology, translation studies and lexicography, to mention just a few. To a certain extent, corpora are now seen as essential by many researchers in these areas because insights provided by the analysis of a substantial amount of authentic data are hard – some would say impossible – to come by through other methods. The community of scholars who regularly use corpus analysis as a standard fact-of-life tool of investigation is large and growing, signalling that corpus-based approaches seem to be here to stay and are likely to interact with a very broad range of language-related research in coming years.

This is not to say that the observed trend is dominant or even stable enough to ensure that a thoroughly empirical approach to language research – which might be seen as corpus linguistics in its essential form – is now fully established, especially in Brazil. The share

of corpus-based studies appearing in events such as the recently held three-day V ABRALIN (Associação Brasileira de Lingüística) International Congress is thus larger now, with at least one session per congress day dedicated to Corpus Linguistics and Computational Linguistics put together. The number of works added to twelve presentations in all three sessions. Several of these combine corpus-based approaches with systemic-functional linguistics, a dialogue that has been under way for a while and is likely to prove fruitful. Sessions dedicated to sociolinguistics also rely on evidence from authentic language data and often resort to systemic-functional linguistics for theoretical background as well.

Other areas of language research represented in dedicated sessions of the ABRALIN Congress use authentic language data to develop their approaches, but these are often structured around previously defined theoretical tenets. Therefore, evidence from language data are searched in order to find examples or data to corroborate a given choice of theory, rather than analysed so as to yield results which are then organised into a theory. Notions such as frequency and probability of occurrence, central to corpus linguistics, may play a diminished role in studies that are selective in regard to the data considered. Moreover, previous affiliation to a theory group may prevent the analyst from perceiving evidence that does not confirm assumptions subsumed in a chosen theoretical approach. In short, it is certainly possible to use a corpus in research without actually embracing an empirical approach to investigations on language or doing corpus linguistics as such.

Nonetheless, there seems to be good reasons for optimism regarding the future of corpus-based approaches, if one thinks of the time when phrases like *empirical research*, *corpus data* and *language in use* were promptly associated to minor forms of investigation which did not quite deserve the name of linguistics, nor were seen as scientific language studies. Thus, it seems safe to say that corpus linguistics has become part of the choices in postgraduate programmes in a number of institutions in Brazil, no matter that the actual spirit of corpus linguistics may occasionally appear somewhat distorted by a still ingrained habit

of resisting evidence from corpus data whenever such evidence does not confirm a priori theoretical definitions.

Abroad, as a second dialogue trend in course, a number of studies seek to develop a usage-based approach within the field of cognitive linguistics. Such studies (see Tummers et al. 2005; Mukherjee 2004) focus on methodological aspects of data collection and analysis, as well as in the "status of empirical data in linguistic research" (Tummers et al. 2005). The use of quantitative analysis is an essential aspect of these developments. Not surprisingly, thus, a lively discussion on the use of statistical techniques in general in language studies, with particular concern for significance and hypothesis testing (Yeh 2000; Evert 2004; Kilgarriff 2005; Gries 2005), has become one of the major issues regarding the use of corpora in linguistics, since language, as an object of statistical analysis, seems to challenge standard methods of statistics for the social sciences with non-trivial difficulties that demand creative solutions. It is to be expected, therefore, that discussions on statistics for language studies with the use of often very large corpora are likely to remain a central issue for the years to come.

Other investigations broadening the scope of interactions between corpus linguistics as such and other approaches in linguistic research include Biber and Jones (2005), who explore the use of corpus data in discourse analysis; the use of corpora as part of experimental designs within empirical methods of language research (Wulff 2005), including language acquisition (Diessel and Tomasello 2005; Theakston et al. 2002) and phenomena such as linguistic variation and change; and Stefanowitsch and Fries (2005), where the core notion of collocation is explored within a construction-grammar approach, so as to develop the idea of collostructional analysis, which integrates both concepts in a framework for the study of 'co-occurrence patterns'. These various interactions with other approaches in linguistics raise again the question of the actual status of corpus linguistics, as a methodology or an approach, which has been around for a while. This issue is succintly developed in the following section, but for a thorough discussion see Berber Sardinha 2004.

This introduction moves on then to examine relevant matters concerning the corpus itself, since decisions involving the nuts and bolts of corpus compilation remain a challenging aspect of corpus-based approaches for researchers who choose to use such collections of authentic data. The third section presents the articles in this issue and tries to contextualise the material presented in these works as part of the effort of a research community to define more precisely the fundamentals of a way forward in corpus linguistics.

1. Corpus linguistics as a research programme

One can in brief terms define corpus linguistics as the study of human languages on the basis of authentic examples of language in use. Thus defined, corpus linguistics clearly incorporates a methodological element to its very definition. In consequence, it is not surprising that interactions with every form of empirical research in linguistics occur. On the other hand, corpus linguistics proper claims a status of an approach to language studies of its own accord, within which issues such as language acquisition, linguistic variation and discourse are analysed having the core notion of collocation – or, more broadly, patterning – as the essential paradigm of a linguistic theory.

A number of attractive theoretical issues seem indeed to be waiting for wider and deeper investigations in what regards corpus-based approaches, now that the use of a technological tool – that is, the computer – appears to have been accepted as a fact of life, and not a threat to the development of linguistics, although this sort of antitechnology response to the use of computers occasionally still pops up. It is revealing that a specific journal – *Corpus Linguistics and Linguistic Theory* – has been created with this particular title and the manifest purpose of "publishing high-quality original corpus-based research focusing on theoretically relevant issues in all core areas of linguistic research", as defined in the journal's homepage (see http://www.degruyter.com/rs/384_7546_ENU_h.htm).

Some further consideration seems in order regarding the definition of corpus linguistics. Options appearing in a number of works and web pages vary somewhat around terms such as discipline, methodology and approach. Thus, McEnery and Wilson (1996) bring up the question "Is corpus linguistics a branch of linguistics?" at the very beginning of their textbook on corpus linguistics, but the authors' answer is "both yes and no". The negative is justified on the grounds that corpus linguistics does not focus on one specific "aspect of language use". It would be then a "methodology", since it can be applied to "almost any area of linguistics". On the other hand, yes, it is possible to separate corpus-based approaches to a given area of investigation from non-corpus-based approaches. Corpus linguistics could then be understood as a feature on the basis of which one could "discriminate between methodological approaches".

However, as pointed out by Berber Sardinha (2004), it is useful to be clear on what is meant by methodology. According to the Concise Oxford Dictionary, methodology is defined both as "the science of method" and "a body of methods used in a particular branch of activity". As the former definition does not apply, it is the latter sense that should be chosen, but the body of methods may or may not be theoretically motivated. If understood as a set of tools, it is certainly true that the techniques typically used by corpus linguists may be part of methods within a variety of other theoretical frameworks, a process which, as references above show, is already under way, whether corpus linguists are happy about it or not. However, corpus linguistics does have a theoretical motivation behind the choice of methods, which might be summed up by the collocational principle, as formulated by Sinclair (1987).

It is argued, thus (Berber Sardinha 2004), that the word *approach* expresses more accurately the fact that there are theoretical tenets for corpus linguistics which explain why the body of methods associated with corpus-based approaches is as it is, regardless of the fact that research efforts exist that, although a computer and a corpus are included as part of their methodology, are not, strictly speaking, corpus linguistics. Other phrases include "a new philosophical approach" (Leech 1992)

and "the route for linguistics" (Hoey 1997). A search in the Web found several definitions for corpus linguistics, which in many cases seem not to convey essential concern with the issues discussed so far. The Wikipedia definition, for one, focuses basically on the fact that corpus linguistics challenges Chomsky's views, although the study of language as observed in corpora is mentioned, thus again stressing methods. The methodological emphasis is present in many other definitions, sometimes, as in the Chemnitz University page, highlighting "the principles and practice of using corpora".

It is perhaps possible to integrate a good share of these various attempts to define corpus linguistics by resorting to an idea, developed by Imre Lakatos (1978), of scientific theories as research programmes which necessarily imply methodological features crucial to their development. According to Lakatos, scientific achievement is not attained by the development of a single hypothesis, but as a research programme. A research programme is made up by a "hard core" of essential beliefs which are axiomatic for those who work within the programme. These tenets are sustained at all costs in all situations. The hard core is further developed into auxiliary hypotheses, named the "protective belt" by Lakatos, which block refutation of the hard core. Moreover, scientific research programmes also have a "heuristic", consisting of methodological orientations to solve problems and explain observations that apparently do not fit the assumptions in the hard core, including the very choice of prime research focus or object of study.

Applied to corpus linguistics, this view may result – of course other interpretations are possible – in a research programme with a hard core based on Sinclair's collocational principle. Thus, language is organised on the basis of two active principles invariably at work. The open-choice principle operates exclusively on grammatical restrictions to the choice of words to fill slots in a text, whereas the idiom or collocational principle, in Sinclair's words, relates to "the large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments". These choices involve lexical, semantic and grammatical elements, and therefore must

take into account the linkage of patterns of use to meaning as a crucial aspect of the scientific effort to understand language.

Auxiliary hypotheses stemming from the core of the research programme and protecting it from refutation challenge the dualistic approach which separates competence and performance, or, alternatively, structure and use, proposing the notion of lexicogrammar. The association of meaning and pattern relies on evidence that distinctions of meaning are often made as a result of differences in patterns; conversely, similar patterns tend to express shared meanings. The fact that the correspondences of pattern and meaning are not perfect, that is, cannot be expressed in simple one-to-one mappings, is further evidence of the adaptive nature of such correspondences in authentic uses of language for communication. This would define an approach from both lexical and grammatical viewpoints, as described in Hunston and Francis (2000). Notions such as colligations, semantic prosody, schemata and other attempts to capture the combinatory effects that relate patterns to meanings (see Partington 1998 for a full discussion) can be adequately incorporated to this set of auxiliary hypotheses involving the notion of collocation.

One obvious development which is central to the protective belt is the hypothesis that establishes the analysis of authentic language, as observed in real-life situations, as the main concern of linguistic research. The study of language as a biologically determined mental artifact cannot reveal the actual semantic and functional features of a grammar that interacts with these aspects to form a whole. The heuristic of corpus linguistics, thus, is the analysis of corpus data. Explanatory theories should be elaborated on the basis of language-in-use samples collected and organised under the form of corpora. Such an approach leads to a concept of language knowledge which fits into the notion of situated cognition within cognitive sciences (Suchman 1987), the corpus being the research tool *par excellence* of an envisaged situated linguistics.

Lakatos' formulation is an alternative to the views of both Popper and Kuhn regarding the advance of scientific knowledge. Scientific revolutions, according to Lakatos, are a result of the contrast between

"two rival scientific research programmes", in which one is degenerating and the other is progressive, since scientists tend to join the progressive programme. Progressive research programmes are able to propose theories that discover or explain facts unknown or unexplained before these theories were proposed. Degenerating research programmes produce theories "only in order to accommodate known facts". Thus, the hallmark of science, as compared to pseudoscience, is neither the Popperian falsifiability criterion – which naively ignores the "remarkable tenacity of scientific theories" – nor the change in commitment advocated by Kuhn as the crucial feature in scientific revolutions, which, according to Lakatos, would amount to accepting that such revolutions are irrational and a "religious conversion". The true nature of scientific advancement is the replacement of degenerating research programmes by progressive research programmes.

Deciding whether a given scientific research programme is truly progressive may not be so easy, especially in a branch of science such as linguistics. Still according to Lakatos, new programmes may not "become empirically progressive" before a relatively long timespan after their appearance. Thus, it is not instantly obvious that a new programme is progressive. In addition, linguistics is a particularly difficult case among the various scientific fields, since it is unclear what counts as empirical evidence of progress in linguistics. For many scientists working in the field, language as used in real life for the purposes of communication is not the essential focus of investigation. As a result, advances in knowledge about facts in everyday language are not considered by those researchers to be any actual evidence of progress. It would be necessary therefore to define what is seen as progress in scientific knowledge regarding the study of human languages, a goal that linguistics has not fully achieved so far.

Nonetheless, the assumption in the proposed hard core of corpus linguistics, namely, that there are two principles invariably at work in language production and understanding, results in a far more satisfactory explanation for the relationships holding between the system of language and its actual use for communication. The

collocational principle puts the phenomenon of patterning and meaning at the centre of the investigation in linguistics, challenging the dualistic views of language as flawed in the sense that a clear mapping of connections between an assumed mental faculty of language and real-life usage is simply not provided. The theoretical proposal which places the syntactic and collocational principles at the core of a research programme narrows the gap between language system and use of language by predicting that meaning is to be explained by considering the operation of both. As a result, it responds to a number of questionings lingering in the mind of linguists uncomfortable with this gap, which more often than not meant working with highly abstract constructs freely concocted on the basis of invented examples; producing arbitrary grammaticality judgements; and blatantly disregarding language-usage evidence as unnecessary or even detrimental to language research.

It seems reasonably safe to say thus that corpus linguistics is a progressive scientific research programme, in the sense that it predicts accurately that patterns detected in language corpora can be used effectively to explain language facts such as meaning, syntax and discourse relations. The fact that many researchers from various stances in linguistics and also from other areas are attracted to the methods of corpus linguistics may signal that the research programme is still in the early stages of development. Consequently, it is not a fully specified theory and naturally allows input from other approaches to play a role in possible further elaboration. One may envisage a merger with one or more of these interacting approaches or, alternatively, a new usage-based development that would theoretically connect functional, cognitive and construction grammars to patterning and the collocational principle, a process that seems to be under way anyhow.

It also seems true that the element which most attracts researchers to corpus linguistics is the fact that the research programme gives linguists a good reason to analyse real language, since such research initiatives are now part of an approach which claims that this is theoretically fruitful and does not demean those linguists that choose

to do so. Moreover, applications of linguistic knowledge have also been brought back into the picture as evidence that a given approach to language studies may be useful. Thus, elaborating better dictionaries and grammar books, creating language-teaching methods that improve learning, and building computer systems that obtain better results in tasks involving human language technology also count now as evidence that a given analysis of a linguistic phenomenon or level of analysis may be scientifically sound (see Stubbs 1996).

The rapid increase in the number of researchers interested in corpus analysis, now that it is not only feasible – since computers have become commomplace in offices, universities and homes – but also scientifically valued, does seem to be in part a consequence of methodological aspects of corpus linguistics. Researchers are clearly happy that the analysis of real language and the results of applications involving research findings are again part of life as a linguist. Some of them may adopt the heuristic of corpus linguistics without formally accepting the hard core. In the overwhelming majority of cases, however, theoretical elements of the various approaches interacting with corpus linguistics or simply using its tools are not incompatible with the assumptions in the hard core, and may prove to be, as in the case of lexicogrammar, major aspects of a corpus-based linguistic theory still in preparation.

The occasional uneasiness of researchers in "hard-core" corpus linguistics, resulting from the perception of corpus linguistics as "just" a methodology by people that adopt the heuristic of the research programme without explicit commitment to the hard core, should perhaps be mollified. After many years of a dominant trend to create one's own data, it seems only normal that methodology is focused upon in the context of a contrasting approach that values the analysis of real language as observed in machine-readable corpora. Empirical evidence is back in the game, and methods of collecting and analysing it have been neglected for a long time. It is in fact quite positive that corpus linguistics brings to the fore methodological issues of daily research practice. It seems reasonable to expect that this analytical machinery

will eventually produce a fully defined approach in linguistics, whether or not as a result of interactions with distinct research programmes.

2. Corpus, linguistic data and the Web

Several definitions of corpus are found in the literature on the subject, with varying degrees of detail and complexity, although two main contrasting tendencies are noticeable in these definitions. There are definitions that seek to simplify the matter, reflecting the fact that, in actual research practices, a considerable diversity of distinct textual material bears the name of a corpus. It seems fair therefore to elaborate definitions that avoid disqualifying useful bodies of linguistic empirical evidence as acceptable corpora. A second tendency appears to be concerned with a more complete, theoretically motivated treatment of what a corpus is in the field of linguistics. A short selection of definitions is discussed below, covering both tendencies. The issue of the Web as corpus is then introduced and examined as a way to enrich the analysis and broaden the notion of corpus, since nowadays the use of the Web as a corpus in linguistic studies can hardly be ignored. A few general elements for a definition of corpus are systematised by way of conclusion.

Starting with a non-specialised source, the Oxford Concise Dictionary (1990) defines corpus as a "a body of writings, texts, spoken material, etc.". The etcetera is of course particularly inadequate in a scientific context, so, although not wrong and possibly appropriate for a dictionary such as the OCD, this sort of definition is too unspecific to be useful for linguists. Two other definitions shall be considered also as starting points. These come from works which are meant for the person who is interested in the study of languages, but is a beginner or not a specialist. Thus, in David Crystal's Encyclopedic Dictionary of Language and Languages (1992), the definition is a "…collection of linguistic data, either compiled as written texts or as a transcription of recorded speech", whereas, in The Oxford Companion to the English Language (McArthur 1992), a corpus is, in "...linguistics and lexicography, a body of texts, utterances or other specimens considered

more or less representative of a language, and usually stored as an electronic database".

The comparison between the generic definition in the OCD and those shown in works at an intermediate level of specialisation points to an increase in the amount of information included. Crystal adds that the data are linguistic and is more precise about the spoken language materials, mentioning that these are transcriptions of recorded speech. McArthur situates the object in two fields of scientific investigation and, apart from terminological differences that are not relevant, meaningfully includes in the definition that the body in question is "more or less representative" and "usually" stored in electronic format. The elements added to the most basic definition in the OCD are revealing in the sense that they signal which aspects may still demand explicit clarification in the form of elements relevant to a satisfactory definition of corpus.

One of these aspects had already been highlighted in what is perhaps the first formal definition of a corpus since the renaissance of corpus linguistics in the eighties, the one elaborated by Francis (1982), one of the pioneers who, with Henry Kucera, compiled the Brown Corpus: "a collection of texts assumed to be representative of a given language, dialect or other subset of this language, to be used in linguistic analysis". Therefore, the term corpus refers, in linguistics, to a collection which has been assembled for investigative purposes and should be representative. As pointed out in McEnery and Wilson (1996), the name corpus applies, in principle, to "...any collection of more than one text...", but, whenever considered as the basic source of data for linguistic investigation, a corpus should satisfy four specific expectations: sampling and representativeness; finite size; machine-readable form; and a standard reference.

The first characteristic is essential but particularly controversial, having been abandoned by many as an unattainable Holy Grail, if one has in mind a given language "in general". The idea of representativeness is central to statistical analysis. Typically, a representative sample of a given population is collected so as to make

predictions about the population on the basis of effects observed in the sample. Thus, the representativeness of the sample allows analysts to specify which research questions may be answered as a result of investigations carried out on the sample data. More specifically, representativeness determines which generalisations regarding features of a given population are trustworthy, often expressed in terms of populations to which generalisations apply. In a textbook-like example, one should not make generalisations about the height of Brazilian women on the basis of a sample collected exclusively in Jaraguá do Sul. It is consequently the researcher's job to make sure that the sample is representative of the population about which inferences are to be made.

However, the notion of representativeness does not apply so easily to language samples, such as corpora are intended to be. Differently from the height of human beings, the full extent of variation in language is unknown, and there is no agreement as to how to measure it, although text genres are often seen as an appropriate attempt. Variables to be considered as measures of existing variation in texts are also undetermined, a fact that poses rather serious difficulties to a specification of how many and which genres are to be included in a classification model. Many researchers in the field now cautioulsy prefer the term balanced corpus, which avoids terminology that implies that a given corpus **represents** the language or a subset of it in any relevant sense. Biber et al. (1998) discuss the notion of a balanced corpus, attempting to specify what should be included in it. The authors explain that typical sampling techniques used in statistical studies are only useful to linguistics to a limited extent. Thus, a proportional sample of a language, as registered through a group of language users in their daily activities, would result in a rather homogeneous corpus, in which conversation would dominate, along with a limited variety of additional text genres, such as TV programmes, some journalistic reporting and billboard texts. Moreover, textual features of these dominant conversations would be mostly quite similar, if contrasted to other text genres.

The actual challenge in corpus studies is to find ways to cover all the range of linguistic variation. A proportional sample of Brazilian Portuguese, for instance, based on the linguistic experience of native speakers of the language in their daily business, would not include legal texts or scientific articles, simply because most speakers of Brazilian Portuguese never read such texts. On the other hand, excluding these texts would seriously jeopardise representativeness, if the importance of these text genres in society is taken into account. Moreover, contrastive descriptions demand representative samples of each variant, so as to enable the investigation of similarities and dissimilarities between them. It is therefore crucial to corpus building that the range of linguistic variation exisitng in a language is covered, but proportions for each variant in the experience of the average speaker are not as relevant as a goal in the design of a balanced corpus.

It is certainly true, however, that the problem of proportions in a sample is well known in statistics. In social sciences, populations are very often too large to be fully surveyed. In what regards language, the very notion of a population of texts in a language is hard to envisage, since the population of texts grows without interruption as a result of daily human activities. Sampling is thus a particularly intricate problem in linguistic studies, and it is never too much for researchers in the area to keep in mind that corpus-based approaches have been challenged in the past – as well as in the present - with the allegation that every corpus is skewed. The main difficulty in sampling is variability, which is particularly high and complex to measure in language studies.

The typical solution for statistical studies in social sciences is the stratified random sample, in which a population is divided in strata, deemed as homogeneous subsets of this population. The subsets may be proportional, but, in the case of language, this is not an adequate choice, as shown in Biber et al. shortly summarised above. The first step would be therefore to define these subsets, often called text genres, although also referred to as registers or text categories, but in practice this has proved to be a daunting task. Criteria for the categorisation should be clear, but often they are not. As pointed out by Santini and

Sharoff in the webpage for the colloquium named *Towards a reference corpus of Web genres*, held in Birmingham on July 27 in conjunction with *Corpus Linguistics 2007*, the typology of genres included in large corpora "varies widely". Santini and Sharoff mention the Brown Corpus and the British National Corpus as examples of a noticeably subjective treatment of the genre problem.

As reflected in the subject focused upon by the colloquium, the text genre issue grows to new dimensions with the use of the Web as a corpus. The Web is a huge source of linguistic data ready to be mined as a corpus, and the potential of the Web for use as a source of linguistic data has already been noticed by many, especially by those who do not have the means to buy corpora in the market. Since the Web is free and uncontrolled, in the sense that no single institution oversees what is or is not included, the share of text in a given language it contains has as good a claim – some would say better – for being a random sample of this language as any other collection of texts. However, the ignorance concerning linguistic text genres is replicated in the process of deciding on methods for the mining of Web data. Questions regarding text formats, genres specific to the Web and genre granularity require precise specifications, particularly if, as it is obviously desirable, a degree of automation is to be part of the collecting. The representativeness issue again looms in the background as a possible challenge to the validity of conclusions reached on the basis of Web linguistic data.

One solution is of course simply to treat the whole Web as a corpus, regardless of any critical considerations as to its representativeness, an approach that is in fact used quite often nowadays, for the obvious reason that data are hard to come by, particularly if a specific kind of text is required for research. In his introduction to Computational Linguistics' *Special Issue on the Web as Corpus*, Kilgarriff and Grefenstette (2003) argue that representativeness – or its weaker version, balance – is a poorly defined concept in corpus linguistics that in fact complicates the matter unnecessarily, and so do the other "connotations" listed by McEnery and Wilson and quoted above. It is therefore more useful to be concerned with determining if a given corpus

is good for a given research purpose, rather than speculating in abstract whether or not a certain body of data is a corpus.

Routine classwork in corpus linguistics courses shows that students fruitfully develop projects which treat various bodies of text as corpora, including: a collection of business letters; compositions by students when tested for their writing skills in their own language; mechanical engineering trainee reports; a single book; a single book and a translation of this same book into another language; various translations of a book into one or more languages; and many different compilations analysed according to the methodology of corpus linguistics. Not only the treatment of the Web as a corpus, but a wide variety of grassroots projects involving corpus-based approaches seem to point to the fact that a broad definition of corpus, as the one proposed by Kilgarriff and Grefenstette – *a corpus is a collection of texts when considered as an object of language or literary study* –, avoids restricting research initiatives and is likely to be the best way forward. It also accomodates well the remark, in Manning and Schütze (1999), that "one should use all the text that is available" when building general models of language, a common goal in human language technology.

Moving in the other direction, definitions can be found in the literature that explicitly include reference to representativeness, to size and to the fact that the corpus is stored in electronic format, with varying degress of detail. The latter, for instance, seems to be true in most cases nowadays, since the rebirth of corpus linguistics in the eighties was directly influenced by the availability of computers. In other words, corpus linguistics as known at present would not exist without the computer. However, including electronic format in a definition may exclude significant research, rare as it may be. References to size are also potentially challenging to research in restricted domains and do not seem to be of major importance. Stating that a corpus should be "large" or "vast" is not saying much. If a corpus is used as the source of data for the elaboration of a corpus-based general dictionary of a given language, it is assumed that it needs to be large enough for the purpose,

but it is the nature of the task at hand that allows the corpus to be evaluated as to adequacy of size.

3. Articles in this issue

Although the issue contains a relatively small number of papers, aspects of corpus-based research included cover a considerably broad diversity of investigations, ranging from language acquisition and human language technology to translation studies and metaphor. In the process, relevant matters within corpus-based studies, such as annotation, non-linguistic factors in corpus compilation and grammaticalisation were discussed as part of the material presented. This is particularly fortunate for the present-day Brazilian reader, who may have a relatively narrow view of corpus linguistics as basically a matter of counting and tabulating corpus information. It is undeniable that counting and tabulating are part of the business of corpus linguistics, but so are the fundamental questions in the study of language, such as linguistic variation, linguistic change, and the nature/nurture debate with regard to linguistic knowledge. The latter is the theme of the opening article.

The author focuses on Steve Pinker's thesis of a language instinct encoded in the genes of human beings, which became widely known in the nineties. Sampson has published extensively on the subject, providing an alternative view of languages as a product of culture evolution rather than of a genetically-encoded instinct. Arguments in favour of language inatism by Chomsky and other authors, developed in the sixties and seventies, are analysed to show that they do not hold when submitted to examination. The analysis is then extended to Pinker's own contributions and that of other prominent authors, such as Derek Bickerton and Ray Jackendoff, who added new arguments to the "language instinct" debate during the nineties.

Aspects of language acquisition examined include intensively discussed issues such as poverty of data, age dependence, and evidence stemming from sign language, to mention just a few. Since the advent of Chomskyan linguistics in the late fifties, language acquisition

became an integral part of most attempts to formulate linguistic theories. Many researchers would consider a lack of explanatory material regarding language acquisition as a major flaw in a given approach to the investigation of human language. Developing an adequate explanation within the empirical view of linguistic phenomena which is the hallmark of corpus linguistics is thus of utmost relevance for the advancement of the scientific research programme.

Orasan, Ha, Evans, Hasler and Mitkov present a broad overview on the use of corpora in computational linguistics. The importance of corpora for three specific fields within research aimed at having computers process human languages – namely, anaphora resolution, text summarisation and term extraction – is discussed to a greater extent, although other aspects of work with corpora for the focused purpose are also examined. It should be highlighted that the article contains a useful review of methods to annotate corpora, including elements such as annotation schemes, standardisation of annotation schemes through XML, and ways to measure and improve inter-annotator agreement in particularly difficult annotation tasks, which are, roughly, those areas of linguistic analysis in which consensus is low.

The authors point out that corpora have been used in computational linguistics for a long time, but far more intensively nowadays, for reasons that are not dissimilar to those which are true for linguistics in general: the availability of computers capable of handling efficiently the amount of textual data typically included in corpora; and the "paradigm shift", as the authors define it, which occurred in artificial intelligence, causing research initiatives to rely on empirical evidence to build computer systems able to carry out tasks commonly pertaining to the field, rather than on a knowledge basis assumed to be adequate descriptions of processes occurring in the human mind. Previous appearances notwithstanding, corpora may be said to have officially become a regular part of research in computational linguistics after the substantive impact of the special issue on large corpora of Computational Linguistics (Church and Mercer 1993), probably the most influential journal in the field.

The paper by Signe Oksefjell is a corpus-based contrastive study which relies on the COMPARA corpus as a source of data for tokens of the English verb form *I think* and its translations into Portuguese. The Portuguese sources of *I think* in translations of Portuguese originals into English are also included in the investigation, taking full advantage of the bidirectional nature of data contained in COMPARA. Some of the recurring themes in corpus-based translation studies, such as complex polysemy in the source language, the variability in renderings into the target language which results from polysemous lexical items, interference of the source language, and translationese, are put into perspective through the analysis of corpus data. Most interestingly, grammaticalisation processes involving the lexical item studied are analysed so as to draw conclusions regarding the diachronic elements of the polysemy, which point to a development from main verb to epistemic modal and finally to epistemic adverb-like form, although all meanings and uses have been preserved in present-day usage.

Cross-linguistic studies are often seen as more than a comparison between two or more languages. It seems reasonable to believe that certain insights regarding linguistic facts may not be achieved without the analysis of phenomena in more than one language (see Johansson and Oksefjell 1998). Parallel corpora of translations, as well as comparable corpora, are likely to play a major role in the process of establishing methodological standards for cross-linguistic investigations in search of functional correspondences across languages. This may eventually yield results that would allow analysts to separate language-specific facts from broader realities of human languages, empirically verifying the existence of general features which hold for a substantial number of languages. The advance of multilingual corpus-based research is likely to profit greatly from improvements in techniques to mine the Web as a multilingual corpus.

Santos provides a cautionary view regarding the reliability of conclusions drawn on corpus evidence, although the article is none the less explicit about the refreshing power of "bathing" in language data as uniquely supplied by corpora. Santos tackles four particularly

difficult aspects involving the English-Portuguese language pair by simultaneously using both the above mentioned COMPARA parallel corpus and a large monolingual European Portuguese corpus, CETEMPúblico. Specific situations approached include translating from English into Portuguese and vice-versa; teaching English to native speakers of Portuguese; and teaching Portuguese to native speakers as well as non-native speakers of the language. Results show that corpora are invaluable to improve the effectiveness of solutions in this broad range of linguistic activities. They also highlight problems which concern the nature of translated text and the importance of evaluating materials extracted from a corpus.

This introduction is not meant to spoil the pleasure of reading the articles in the issue by offering imperfectly summarised versions of the corpus-based studies described in the actual texts. However, it seems right to give readers a taste of the food for thought served in the analysis by Santos. Consider the English word *ruthless*. Bilingual dictionaries list only correspondences of a negative connotation, such as *cruel* e *desapiedado*, turning *ruthless* into a synonym of the English words *cruel* and *merciless*. Data extracted from COMPARA, in contrast, show that translators, in some cases, choose renderings of a different semantic prosody, namely *infatigável* and *rigorosa*, which are at least non-negative, if not positive. Although left out of dictionaries, the choices seem clearly appropriate in context. The analysis of corpus data, therefore, allows a better understanding of possible senses of the lexical item. On the other hand, there is at least one token of a rendering included in the dictionary list that is obviously inadequate. For teaching purposes, including translator training, this should be pointed out, thus demanding critical evaluation of corpus data.

Fernandes' paper focuses attention on the compilation of a specialised parallel corpus also involving the English-Portuguese language pair. In this case, texts selected for inclusion in the corpus are children's fantasy books originally written in British English and translated into Brazilian Portuguese within a relatively short timespan. The approach broadens design considerations in building parallel

corpora to encompass extralinguistic aspects, such as the time in which the translation activity occurs and possible influences of the situation upon the translator and the translation product. Thus, research carried out on the basis of parallel corpora should not be necessarily restricted to a contrastive orientation focusing exclusively on the systems of the languages involved. Social, cultural and cognitive factors are likely to play a relevant role in determining observable features of a given translation product. Corpus-based translation studies may certainly profit from including such factors in considerations concerning the compilation of a parallel corpus.

Some of the books that are classified as children's fantasy literature are published in quite special circumstances, and that holds true for originals and translations. A new Harry Potter is a world event. It is certainly not surprising that the actual linguistic phenomena possibly to be included in a contrastive analysis of originals and translations are strongly influenced by these circumstances. Neglecting to analyse the impact of these factors upon translations and translators is unlikely to be wise. If one may think of corpus linguistics as a starting point for an approach to linguistic studies that incorporates the notion of situated cognition, it is only natural that corpus-based translation studies develop along similar lines. Ideally the development should eventually become a methodological standard and not be restricted to extraordinary events such as the translation of Harry Potter books.

Berber Sardinha uses a corpus-based approach to explore the rich research area of metaphor analysis. In this article, the corpus itself is a matter of special interest, being a collection of President Lula's official speeches since his inauguration until the moment in which the article has been written. The collection amounts to over two million words and reveals in detail the array of metaphoric resources displayed by a politician who is seen as a master of the trade. Berber Sardinha underscores the difficulty of dealing with the large number of metaphors usually found in corpora and presents a useful solution: a metaphor identifier available on line. The software selects words with a high probability of being used as metaphors but requires human

intervention for the final decision of classifying a given word or phrase as a metaphor.

A number of fascinating questions arise from the discussion of metaphors in use on the basis of corpus data. Understanding language as a probabilistic system entails the evaluation of what words are more likely to be used in a given context. This evaluation may be extended to the probability of a given word or phrase being used metaphorically. The core notion of patterning as an essential principle of corpus linguistics could thus play the role of a probabilistic filter for possible metaphors. This would restrict candidates for metaphoric value to manageable proportions. The collocational principle and the notion of lexicogrammar may then interact with a concept of metaphor so as to process and interpret metaphors in a systematic way arguably integral to linguistic competence with a communicative slant.

The variety of theoretical and methodological points covered in this issue is thus heartening in the sense that it may help provide a wider view of the research possibilities within corpus linguistics to those readers new to the field or working in a related area. This is likely to be useful for the development of a better understanding of corpus-based approaches in linguistics, specially in Brazil, where scholars are often misled by incomplete "hearsay" descriptions of corpus linguistics. Once prejudice is cleared, one might predict a substantial increase in corpus-based research in the country, as more and more scholars make up their minds to give data a chance.

References

Aarts, J. e Meijs, W. (Eds.). *Corpus linguistics*. Amsterdam: Rodopi, 1984.

Allen, R.E. (Ed.). *The Concise Oxford Dictionary of Current English*, 8th Edition. Oxford: Clarendon Press, 1980.

Berber Sardinha, T. *Lingüística de corpus*. Barueri: Manole, 2004.

Biber, D., Conrad, S. and Reppen, R. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press, 1998.

Biber, D. and James K. Jones. Merging corpus linguistics and discourse analytic research goals: Discourse units in biology research articles. *Corpus Linguistics and Linguistic Theory* 1(2), pp. 151-182, Berlin and New York: De Gruyter, 2005.

Church, Kenneth and Robert L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1), pp.1-24,1993.

Crystal, D. *Encyclopedic Dictionary of Language and Languages*, London: Penguin,1992.

Diessel, H. and Tomasello, M. Particle placement in early child language: a multifactorial analysis. *Corpus Linguistics and Linguistic Theory*, 1(1), pp.89-112, 2005.

Evert, S. Significance tests for the evaluation of ranking methods. *Proceedings of the 20th International Conference on Computational Linguistics (Coling 2004)*, Geneva, Switzerland, 2004.

Francis, W.N. Problems of assembling large corpora. In: Stig Johansson (Ed.). *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for the Humanities,1982.

Gries, S. Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2) , p.277-99, Berlin and New York: De Gruyter, 2005.

Hoey, M. From concordance to text structure; new uses for computer corpora. In: Lewandowska-Tomaszczyk, B. and Mella, P. J. (eds.). *PALC' 97 – Practical applications in language corpora*. Lodz: Lodz University Press, pp.2-22, 1997.

Hunston, S. and Francis, G. *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: John Benjamins, 2000.

Johansson, S. On the role of corpora in cross-linguistic research. In S. Johansson and S. Oksefjell (eds.). *Corpora and cross-linguistic research*. Amsterdam: Rodopi, 1998.

Kilgarriff, Adam. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1:2, Berlin and New York: De Gruyter, 2005, pp.263-276.

Kilgarrif, A. and Grefenstette, G. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*. 29:3, Cambridge, MA: The MIT Press, pp. 333-347, 2003.

Lakatos, I. *The methodology of scientific research programmes. Philosophical papers Volume I*. Cambridge: Cambridge University Press, 1978.

Leech, G. Corpora and theories of linguistic performance. In: J. Svartvik (ed.) *Directions in corpus linguistics. Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991*, , pp.19-39, Berlin and New York: De Gruyter, 1992.

Manning, C. and Hinrich Schütze. *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press, 1999.

McArthur, T. (Ed.). *The Oxford Companion to the English Language*, Oxford: Oxford University Press,1992.

McEnery, T. and Andrew Wilson. *Corpus linguistics*. Edinburgh University Press: Edinburgh, 1996.

Mukherjee, J. Corpus data In a usage-based cognitive grammar.In:Ajmer, Karin and Bengt Altenberg (eds.). *The theory and use of corpora: papers from the 23 rd ICAME conference*. Amsterdam:Rodopi, pp.85-100, 2004.

Partington, A. *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam and Philadelphia; John Benjamins, 1998.

Santini, M. and Serge Sharoff. *Towards a Reference Corpus of Web Genres*. Colloquium in conjunction with Corpus Linguistics 2007, http://corpus.leeds.ac.uk/serge/webgenres, Birmingham, July 27, 2007.

Sinclair, J. Collocation: a progress report. In: R. Steele and T. Threadgold (eds.). *Language topics: essays in honour of Michael Halliday*. Amsterdam and Philadelphia: John Benjamins, 1987.

Stefanowitsch and Fries. Covarying collexemes. *Corpus Linguistics and Linguistic Theory*. 1:1, pp.1-43, Berlin and New York: De Gruyter, 2005.

Suchman, L.A. *Plans and situated action*. New York: Cambridge University Press, 1987.

Stubbs, M. *Text and corpus analysis: computer-assisted studies of language and culture*. Oxford: Blackwell, 1996.

Theakston, Anna L., Elena V.M. Lieven, Julian M. Pine and Caroline F. Rowland. Going, going, gone: The acquisition of the verb 'go'. *Journal of Child Language* 29 (4), pp. 783-811, 2002.

Tummers, J., Kris Heylen and Dirk Geeraerts.Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2), pp. 225-261, Berlin and New York: De Gruyter, 2005.

Wulff, S. Online statistics labs. *Corpus Linguistics and Linguistic Theory* 1(2), pp.303-308, Berlin and New York: De Gruyter, 2005.

Yeh, A. More accurate tests for the statistical significance of result differences. *Proceedings of the 18th conference on computational linguistics (COLING 2000)*, Saarbrucken, Germany, pp. 947-953, 2000.