Linha D'Água (Online), São Paulo, v. 32, n. 1, p. 39-57, jan.-abril 2019

ENTRIES ON THE HISTORY OF CORPUS LINGUISTICS

SUBSÍDIOS PARA A HISTÓRIA DA LINGUÍSTICA DE CORPUS

*Carlos Assunção** [https://orcid.org/0000-0002-5739-0754]
University of Trás-os-Montes e Alto Douro, Porto, Portugal
*Carla Araújo*** [https://orcid.org/0000-0002-5318-0960]
Polytechnic Institute of Bragança, Bragança, Portugal

Resumo: A Linguística de Corpus ancora-se num paradigma teórico que se caracteriza por uma abordagem empirista e por uma conceção da linguagem como um sistema probabilístico. Em Linguística, o empirismo é uma abordagem que concede estatuto primordial aos dados que provêm da observação da linguagem, geralmente agrupados sob forma de *corpus*, opondo-se ao racionalismo. O racionalismo baseia-se no estudo da linguagem a partir da introspeção, entendida como maneira de averiguar modelos de funcionamento estrutural e a formação do processo cognitivo da linguagem. Por conseguinte, verifica-se um antagonismo entre as posturas filosóficas características da concepção empirista e racionalista da linguagem, representadas pelos seus maiores vultos. Por um lado, Halliday, representante da concepção empirista, e, por outro lado, Chomsky, o maior vulto do racionalismo na Linguística. Há, no entanto, novas abordagens que devem ser consideradas. De todas estas concepções constituiu-se o maior número de trabalhos da linguística de corpus, nas áreas da lexicografia e terminologia – produção de dicionários, glossários, bases de dados terminológicas, etc.

Palavras-chave: História; Corpus; Linguística; Empirismo; Racionalismo.

Abstract: *Corpus linguistics is anchored in a theoretical paradigm characterised by an empiricist approach and as well as by a conception of language as a probabilistic system. In linguistics, empiricism Empiricism is an approach that grants primordial status to data coming from the observation of language, generally grouped together in a corpus, as opposed to rationalism Rationalism. Rationalism is based on the study of language through introspection, which is regarded as a way of assessing models of structural functioning and the formation of the cognitive process of language. As a result, there is a chasm between the philosophical perspectives characteristic of the empiricist and rationalist conceptions of language, represented by its main contributors. On the one hand, there is Halliday, a representative of the empiricist conception, and, on the other hand, Chomsky, the greatest figure of Rationalism rationalism in linguistics. However, new approaches need to be taken into consideration. From all these conceptions the greatest number of works of corpus linguistics has been derived, in the areas of lexicography and terminology – production of dictionaries, glossaries, terminological databases, etc.*

Keywords: *History; Corpus; Linguistics; Empiricism; Rationalism.*

* University of Trás-os-Montes e Alto Douro – UTAD, Porto, Portugal; cassunca@utad.pt
** Polytechnic Institute of Bragança – IPB, Bragança, Portugal, carla.araujo@ipb.pt

http://dx.doi.org/10.11606/issn.2236-4242.v32i1p39-57

Introduction

What is currently called 'corpus linguistics' covers a heterogeneity of theoretical conceptions, fields of study and works: that is, the activities that fall under this label of 'corpus' are singled out due to differing conceptions in the understanding of the notion of corpus itself, and due to the objectives and the fields of language sciences to which they refer, as well as data processing methodologies. This issue has already been the subject of several studies. Researchers such as Chomsky (1959), Jones (1989), Leech (1992), Stubbs (1997), McEnery and Wilson (1996), Kennedy (1991, 1998), McEnery & Wilson (2001), Sardinha (2004), Sinclair (1987), Teubert (2005, 2010), Halliday (1970, 2004) and Rajagopalan (2007), among others, have developed studies in the field of corpus linguistics, as we will outline below.

According to SARDINHA (2004: 3), Corpus Linguistics focuses on the "coleta e exploração de corpora, ou conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística [collection and manipulation of corpora, or a set of textual linguistic data carefully collected, in order to serve as a source for the study of a language or linguistic variety]."

Believing the focus of corpus linguistics to be on meaning, TEUBERT (2005: 2–3) claims that corpus linguistics examines language from a social perspective:

> The focus of corpus linguistics is on meaning. Meaning is what is being verbally communicated between the members of a discourse community. Corpus linguistics looks at language from a social perspective. It is not concerned with the psychological aspects of language. It claims no privileged knowledge of the workings of the mind or of an innate language faculty.

After presenting 25 theses that define corpus linguistics, TEUBERT (2005: 2–8) points out that the non-verbal context is beyond the scope of corpus linguistics:

> Language is, first of all, a collective activity. As Wittgenstein has argued, there can be no private language. Speech is the primordial form of language. Why should we restrict corpus linguistics to the investigation of written (or transcribed or otherwise recorded) language? In an informal conversation, the verbal interaction

> normally involves other elements, such as deixis, gestures, and facial expressions. At stake is not only the communication of content, but also the intention to contribute to a group feeling, to create an atmosphere of trust, to attempt to step up on the ladder of social hierarchy. The conversation can be embedded in some other interaction like walking in the park, watching TV on the sofa, or standing alongside an open grave. What is being said cannot be easily dissociated from the situation in which it takes place. It only makes sense within the context in which it has been said. However, to make claims, specific or general, about this non-verbal context is outside of the remit of a corpus linguist.

Rethinking corpus linguistics, TEUBERT (2010: 19) also highlights the co-existence of different conceptions of corpus linguistics:

> Over the summer of 2008, there was, on the Corpora-List under the name of "Bootcamp", a lengthy and quite controversial discussion about the role of corpus linguistics within the discipline of linguistics. It does not matter that the differences were not resolved; what is important about this dispute is that it demonstrates the co-existence of different conceptions of what is called corpus linguistics.

Taking this observation into account, corpus linguistics is, according to Teubert, more than the application of tools to corpora. In light of the above, he offers the following definition:

> [the] corpus linguistics is more than the application of tools to corpora. Corpus linguistics, as I understand it, is a way to make sense of what is said. This focus on the discourse and the texts of which it consists rather than on the language system is what sets it, as I see it, fundamentally apart from the main paradigms we find in the 20th century (TEUBERT 2010: 21).

Teubert notes that Sinclair, a pioneer in corpus linguistics, has always insisted on the idea that corpus linguistics is more than just a set of methods or tools for extracting linguistic facts from a corpus: "he insisted that corpus linguistics is more than just a bunch of methods or a toolkit to extract linguistic facts from a corpus. For him, it was a new and different way to look at language" (Teubert 2010: 24).

SARDINHA (2004: 15) argues that "Linguística de Corpus trabalha dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem como sistema probabilístico [Corpus Linguistics acts within a conceptual framework formed by an empiricist approach and a view of language as a probabilistic system]."

Following these perspectives, we would like to highlight two points that we consider to be of utmost importance to an understanding of the research developed over time in this field, and that are the scope of this study: on the one hand, the relevance of the corpus as a source of information, because it corresponds to a natural language storage used by its native speakers in real situations; and, on the other hand, the relevance of the research on the frequencies of linguistic features, taking into account that confirmation of the proven frequency will lead the researcher to the theoretical probability.

From Tradition to Modernity

The term 'corpus linguistics', coined in Great Britain, was described as a new paradigm in language sciences, due to its base on theoretical and historical knowledge. This knowledge base focuses on its oppositional positioning against generative grammar. Actually, within the British tradition there is, on the one hand, a position that aims at granting corpus-based studies the status of a new paradigm, based on the advances of corpus-based research and on Chomsky's criticism dating back to the 1950/1960s. On the other hand, there is another position that does not intend to summon a historical reconstruction, nor a total theoretical rupture, because it lies mainly in the continuity of the tradition of British empirical linguistics.

Kennedy (1998) claims that the advances have not been as far-reaching as some say. Therefore, corpus linguistics certainly cannot be considered a new paradigm, as he puts it:

> Although there have been spectacular advances in the development and use of electronic corpora, the essential nature of text-based linguistic studies has not necessarily changed as much as is sometimes suggested. Corpus linguistics did not begin with the development of computers but there is no doubt that

> computers have given corpus linguistics a huge boost by reducing much of the drudgery of text-based linguistic description and vastly increasing the size of the databases used for analysis (KENNEDY: 1998: 2).

In fact, Kennedy (1998: 13) believes Alexander Cruden to have been the author of the most well-known edition of a biblical concordance,[1] first published in 1737.

However, Leech (1992) argues that corpus linguistics is a recent methodology and that it introduces a new approach: "I wish to argue that computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject" (LEECH 1992: 106–107). In the same vein, Stubbs (1997) points out that corpus linguistics is not merely a tool, but an important concept in linguistic theory: "First, corpus linguistics is a view about data: many different methods can be used to analyse corpus data. Second, a corpus is not just a tool, but a major concept in linguistic theory" (STUBBS 1997: 300).

In the 1940s and 1950s, American structuralists greatly contributed to the flourishing of corpus analysis. However, between 1950 and 1980, these analyses lost their significance due to Chomsky's criticism, but this was regained with the emergence of computers in 1980–1990s. Chomsky's criticism was so influential that McEnery and Wilson (2001), in an attempt to exemplify research prior to Chomsky and whose methodological approach is based upon corpus linguistics, use the term 'early corpus linguistics':

> Early corpus linguistics is a term we will use here to describe linguistics before the advent of Chomsky. In this examination of early corpus linguistics, we will imply that linguistics before Chomsky was entirely corpus-like. This is both true and untrue. The dominant methodological approach to linguistics immediately prior to Chomsky was based upon observed language use. The debate that Chomsky reopened in linguistics was, however, a very old one, as will be seen. Part of the value of looking in some detail at Chomsky`s criticisms of corpus data

[1] For a more thorough overview of Alexander Cruden's work, please see http://www.unz.org/Pub/CrudenAlexander-1858

> is that they in part represent a compendium of centuries of debate on the nature of data in linguistics (MCENERY; WILSON: 2001: 2).

In early corpus linguistics, McEnery and Wilson (2001: 2–4) point to several studies, including Boas' work on American Indians (1940); Harris' structuralist linguistics (1951); language acquisition research based on children's utterances recorded[2] by their parents on a daily basis (1876–1926); Kading's work (1897), which calculated the frequency of distribution and the sequence of occurrence of letters in German using a corpus of nearly 11 million words; Fries and Traver's (1940) studies and those of Bongers (1947), within the frame of foreign language pedagogy;[3] and Eaton's (1940) works, comparing the frequency of word meanings in Dutch, French, German and Italian.

According to Kennedy (1998: 13), there was a tradition of linguistic analysis based on corpora prior to the nineteenth century, long before the arrival of computers, in the context of biblical and literary studies, in lexicography and dialectology, in language education studies and in grammar studies. In fact, if we understand corpus linguistics to be the study of language through its samples, we understand it as a practice that has existed for a long time.

Actually, reflections on language through the analysis of authentic linguistic data inherent to corpora are not an innovation of corpus linguistics, because that attitude, as Rajagopalan (2007: 33) puts it, "é tão antiga quanto o surgimento do empirismo como método alternativo de fazer ciência [is as old as the emergence of empiricism as an alternative method for doing science]". Specifically, the change brought about by corpus linguistics is due to the use of computers, which have

2 Cf. MCENERY; WILSON (2001: 3): "These primitive corpora, on which later speculations were based by the researchers of the period such as Preyer (1889) and Stern (1924), are still used as sources of normative data in language acquisition research today, for example, Ingram (1978)".

3 Cf. MCENERY; WILSON (2001: 4): "Indeed, as noted by Kennedy (1992), the corpus and second language pedagogy had a strong link in the early half of the twentieth century, with vocabulary lists for foreign learners often being derived from corpora. The word counts derived from such studies as Thorndike (1921) and Palmer (1933) were important in defining the goals of the vocabulary control movement in second language pedagogy".

completely revolutionised the manner in which corpora are organised and analysed, enabling a completely different view on language.

Sardinha (2000) points out that a large part of the twentieth century shed light on researchers who developed language description practices through corpora, among them educators such as Thorndike and field linguists such as Boas. However, Sardinha (2000) highlights two essential differences between that period and the present:

> A primeira, obviamente, é que os corpora não eram eletrônicos, ou seja, eram coletados, mantidos e analisados manualmente. A segunda é que a ênfase destes trabalhos era em geral o ensino de línguas. Atualmente o que prepondera na literatura é a descrição de linguagem e não a pedagogia, embora recentemente tenha ressurgido um interesse no emprego de corpora na sala-de-aula e na investigação da linguagem de alunos de língua (SARDINHA 2000: 325).
>
> [Obviously, the first is that corpora were not electronic, that is, they were collected, kept and analysed manually. The second is that the focus of these works was, generally speaking, on the teaching of languages. Nowadays, what prevails in literature is the description of language and not pedagogy, although there has recently been a renewed interest in the use of corpora in the classroom and in the investigation of the language used by students of language sciences.]

As to the corpus-based research conducted manually, Sardinha praises Thorndike's data gathering, claiming that

> um trabalho fenomenal, dadas as condições da época, foi a identificação das palavras mais frequentes da língua inglesa, feita por Thorndike há mais de 75 anos atrás (Thorndike, 1921). O levantamento foi feito manualmente em um corpus de nada menos de 4,5 milhões de palavras, e, quando publicado, impulsionou mudanças no ensino de língua materna e estrangeira, tanto nos Estados Unidos quanto na Europa. As abordagens baseadas no controle do vocabulário, nas quais os alunos têm contato em primeiro lugar com as palavras mais frequentes, devem sua inspiração a estudos como o de Thorndike. Quase 25 anos mais tarde, Thorndike revisou seu levantamento inicial e, tomando como base um corpus maior, com impressionantes 18 milhões de palavras, publicou uma obra listando as 30 mil palavras mais comuns da língua inglesa. Logo depois, em 1953, veio o

> 'General Service List of English Words' de Michael West (West, 1953), talvez a mais famosa descrição do léxico inglês pré-computador. A pesquisa de West dá detalhes do que seriam as 2 mil palavras mais frequentes do inglês e baseou-se no trabalho de pioneiros como Thorndike e Lorge (SARDINHA 2000: 325-326).

> [a phenomenal work, given the prevailing conditions at the time, was Thorndike's identification of the most frequent words in the English language, more than 75 years ago (Thorndike, 1921). The data gathering was conducted manually in a corpus of no less than 4.5 million words, and, when published, it led to changes in the teaching of first languages and foreign languages, both in the United States and in Europe. The approaches based on vocabulary control, in which students first come into contact with the most frequent words, owe their inspiration to studies such as Thorndike's. Almost 25 years later, Thorndike reviewed his initial word-list and, relying on a larger corpus with an impressive 18 million words, published a book listing the 30 thousand most common words in English. Immediately afterwards, in 1953, the 'General Service List of English Words', by Michael West (West, 1953), was released, a work that is perhaps the most well-known description of the English pre-computer lexicon. West's research provides further details on what the 2 thousand most frequent words in English would be and it was based on the work of pioneers such as Thorndike and Lorge.]

Randolph Quirk's text, "Towards a description of English usage" (1960), introduces the project work to the Survey of English Usage (SEU),[4] which was compiled in London from 1959 onwards; it consists of a corpus still in non-electronic format, comprising 1 million words and made up of 200 texts, each consisting of 5000 words. From this corpus, typed slips, each containing 17 lines of context and detailed grammatical annotations, were organised for every single word. From the data collected, a monumental grammar entitled *A Comprehensive Grammar of the English Language* was produced, comprising 1779 pages (Quirk, Greenbaum, Leech and Svartvik, 1985). The SEU corpus was fully computerised in 1989, but the spoken part had been previously recorded and given the name London-Lund Corpus (Sardinha 2000: 326).

[4] For an overview of SEU today, please check http://www.ucl.ac.uk/english-usage

The first electronic language corpus, The Brown corpus[5], was released in 1964. At this time, the computerisation of texts was a very laborious task. As SARDINHA (2000: 324) clarifies, "os textos tiveram de ser transferidos para o computador por meio de cartões, perfurados um a um, tal era a tecnologia da época. Este feito, por si só, já traria respeito e admiração à empreitada [the texts had to be transferred to the computer by means of cards which were punched one at a time, such was the technology of that period. This achievement, on its own, would already bring respect and admiration upon this great deed]."

However, the release of the Brown corpus came at a time when its value was contested, due to the influence of Noam Chomsky's work entitled *Syntactic Structures*, published seven years before the Brown corpus, in 1957.

According to RAJAGOPALAN (2007: 24), after the publication of *Syntactic Structures* and Chomsky's relentless criticism[6] of Skinner's[7] book *Verbal Behavior*, in 1959, "a palavra empirismo tornou-se amaldiçoada (junto com a palavra behaviorismo), pois [...] o espírito do racionalismo cartesiano começou a varrer a Linguística [the word empiricism has become cursed (along with the word behaviourism), because [...] the spirit of Cartesian rationalism started to take hold of Linguistics]."

Similarly, SARDINHA (2000: 326) points out that "No final dos anos 50 apareceria 'Syntactic Structures', de Chomsky, e com ele uma mudança de paradigma na linguística: saía de cena o empirismo e com ele a sustentação dos trabalhos baseados em corpora, tomando lugar central as teorias racionalistas da linguagem [In the late 1950s, Chomsky's 'Syntactic Structures' would come up, accompanied by a paradigm shift in linguistics: empiricism and, with it, the theoretical underpinning of corpora-based works would fade out, giving way to rationalist theories of language]."

Chomsky's work quickly fostered a paradigm shift in linguistics. Therefore, empiricism, dominated by the observation of data through the medium of a corpus,

5 As Sardinha (2000: 324) points out, the pioneering status of the Brown corpus is related to the fact that it is a corpus of written language, because, as to the spoken language, the first electronic corpus, comprising 220 thousand words, is attributed to John McH. Sinclair (cf. Sinclair 1995: 99).

6 For further reading, please see Chomsky (1959).

7 For further reading, please see Skinner (1957).

gave way to rationalism, which aimed at developing a linguistic theory focused on competence rather than performance.

Within this approach, the human being has a capacity for language (linguistic competence) which gives him or her the possibility of producing and/or understanding an infinite number of utterances from a finite number of rules (performance). In this light, the mental processes that lead to performance should be the focus of research, while the study of the external result of the corresponding mental processes is of no interest and therefore the corpora, as sources of research and information, lose their relevance.

In addition to Chomsky's criticisms, corpus linguistics was the target of further criticism related to the manual handling of data, as the thorough manual procedures and large-scale corpora-based work (developed by Thorndike or Kading, for instance) were criticised as leading to inaccuracies. This possible lack of rigour would discredit corpora-based approaches. As SARDINHA (2000: 327) states,

> Além do apelo natural da linguística Chomskyana, outro fator que contribuiu para a perda de fôlego de abordagens baseadas em corpus foi uma crescente leva de críticas sobre o processamento manual de corpora. Uma das críticas mais contundentes era exatamente que o processamento de corpora gigantescos, como o de Thorndike, com 18 milhões de palavras, por meios manuais, não era confiável. O ser humano não é talhado para tarefas deste tipo. Não seria o caso de simplesmente aumentar a equipe de analistas para resolver o problema, pois este trabalho já era realizado com grandes contingentes de assistentes. A pesquisa de Kading, por exemplo, sobre a ortografia do alemão, consumiu a mão-de-obra de 5000 analistas! Os problemas da possibilidade de erro e de falta de consistência persistem, ou até pioram, com grandes equipes. A outra alternativa era diminuir o tamanho dos corpora para facilitar a inspeção manual, mas isto atentava contra a própria natureza da pesquisa.
>
> [Besides the natural appeal of Chomskyan linguistics, another factor that contributed to the shortness of breath of corpus-based approaches was a growing body of criticism concerning manual corpora processing. One of the strongest criticisms was exactly that the gigantic corpora processing, such as that of Thorndike, comprising 18 million words, conducted manually, was unreliable. The human being is not meant for such tasks. Strengthening the team of analysts would

> just not be enough to solve the problem, because this work was already being carried out by a large contingent of assistants. For example, Kading's research on German spelling consumed the workforce of 5000 analysts! Problems arising from the possibility of error and lack of consistency persist, or even get worse, with large teams. The other alternative was to reduce the size of corpora to enable manual examination, but this was against the very nature of the research.]

In order to change this situation, a resource that would enable the analysis of large amounts of data in a reliable way would be needed, but the technology of the time did not allow that possibility. The invention of the computer changed this scenario, enabling the production of works involving information storage, indexation and word counting in an increasingly fast, reliable and accessible way.

In 1957, John W. Ellison started integrating the use of computer resources in the study of texts, producing the first computer-generated concordance. As JONES notes (1989: 131),

> In 1957 the Reverend John W. Ellison gave the word its first computer-generated concordance, to the Revised Standard Version of the Bible. The event was sufficiently significant to deserve a full-page article in the February 18, 1957 issue of LIFE Magazine. LIFE reports that Ellison`s RSV concordance was completed in "only" 400 hours, i.e. the time required of the Remington Rand Univac for the processing of the 80 miles of tape. No mention is made of how much time was spent in entering and proofreading the text or writing the computer program. But in spite of the many hours required to produce the RSV concordance it still represented a significant saving over previous hand produced biblical concordances.

The fact that the completion of this work took "only" 400 hours – the time in which the Remington Rand Univac[8] processed the 80 miles of tape on which it had been recorded – represents a significant time saving over that required for the hand production of this type of practice. Nowadays, considering the greater speed that characterises the information technologies of the twenty-first century, these 400 hours can represent a lot of time. However, in 1957, the outcome of the work

[8] For a more detailed overview, please see http://www.computerhistory.org/revolution/early-computer-companies/5/100.

carried out in that period of time represented the beginning of a new era for the study of language based on samples of language use.

Concurrently with Chomsky's rationalist theory, corpus linguistics was gradually building up its pace to a crescendo, namely in Europe and, particularly, in Great Britain. Some scholars, such as John Sinclair and Geoffrey Leech, now acknowledged as the greatest representatives in the field, continued his work, publishing outstanding works between 1960 and 1970.

The history of corpus linguistics has a very close relationship with technology, because the latter spawns new forms of action. However, for many decades, access to computers was not easy, mainly due to the fact that the computers of the time, which we can call mainframe, were very large machines that worked in a very complex way, requiring the aid of very specialised technicians to put them into operation.

Difficulties in data collection on mainframe computers were overcome in the 1980s with the emergence of personal computers, which contributed to the increasing popularity of corpora and of new processing tools. Hence, linguistic research based on corpora regained importance. A pioneering partnership between the University of Birmingham and the Collins publishing house, aiming at working on the first dictionary compiled according to the principles of corpus linguistics, the *Cobuild English Dictionary*,[9] also contributed to the changing framework that was being shaped. This partnership became known as the COBUILD project[10] (Sinclair 1987). The project saw the production of several dictionaries, grammars and didactic books geared towards the teaching of English. COBUILD is a reference point in the development and application of corpora-based researches for business purposes.

Highlighting the business potential that lies within corpora works, SARDINHA (2000: 329) states that

> há um desenvolvimento crescente de centros de pesquisa mantidos por empresas. Estes centros utilizam-se de pesquisas baseadas em corpus para várias finalidades comerciais, como o processamento automático de textos, informatização de

9 http://dictionary.reverso.net/english-cobuild/

10 Collins Birmingham University International Language Database.

> grandes bases de dados e a montagem de sistemas inteligentes de reconhecimento de voz e gerenciamento de informação. As grandes empresas de telecomunicações investem nestas áreas, reconhecendo o potencial econômico deste campo. Outras empresas de produtos de informática como a Xerox, Microsoft e Canon também possuem centros desenvolvidos de pesquisa de corpus e Processamento de Linguagem Natural.
>
> [there is a growing development of research centres held by companies. These centres use corpus-driven research for a variety of business purposes, such as the automatic processing of texts, computerisation of large databases, and the assembly of intelligent voice recognition and information management systems. Large telecommunications companies invest in these areas, recognising the economic potential of this field. Other companies specialised in computer products, such as Xerox, Microsoft and Canon, have also developed centres for corpus research and Natural Language Processing.]

There was a resurgence of empiricism against rationalism in the 1990s. Corpora studies were then referred to in terms of the resurgence of the empirical and statistical methodologies of the 1950s. Chomsky is singled out as the sole culprit for the negligence of corpora for 25 years: "The impact of Chomskyan linguistics was to place the methods associated with CCL [computer corpus linguistics] in a backwater, where they were neglected for a quarter of a century" (Leech 1992: 110).

Chomsky's argument against corpora and statistical methods was called forth in the 1990s to acclaim corpus linguistics as a renewing force in the empirical corpus research of the 1940–1950s and as a new paradigm in language sciences. According to Sinclair (1991: 4), "First and foremost, the ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before".

In the 1990s, Stubbs (1996: 231) viewed corpus linguistics as a theoretical framework still in its preliminary stage: "Corpus linguistics has as yet only very preliminary outlines of a theory".

The popularity and development of corpus linguistics stem from the establishment of partnerships between universities and private entities. On the one hand, in Europe, dictionary publishing houses such as Longman, Cambridge,

Collins and Oxford joined projects in the field of lexicography carried out by corpus linguists. On the other hand, in the United States, corpus linguistics is associated with natural language processing and computational linguistics; so, we notice a strong investment in the sector on the part of telecommunication and IT companies such as Xerox, Microsoft and Canon.

Corpus linguistics has exerted substantial influence on linguistic research in several international centres. In the United Kingdom, one of the most developed centres in the world, several universities, such as Birmingham, Brighton, Lancaster, Liverpool and London, conduct corpus-based research to describe various linguistic aspects. The research that has been carried out by British entities has enabled the development of theories and the production of corpora and of supporting tools in several areas.

Similarly, for several years in Scandinavian countries (Denmark, Norway and Sweden), centres have been engaged in the study of corpus linguistics.

In the United States, due to the influence of generative-transformational linguistics, corpus linguistics has shown a more timid presence than in Europe, even though it has been doing active work, both academically and commercially. Paradoxically, one of the greatest representatives of corpus linguistics in the world is the American scholar Douglas Biber,[11] who works in an English Department (Sardinha 2000: 328). In fact, the research undertaken in the field of corpus linguistics (Biber, Conrad and Reppen, 1998) is of interest in several domains of empirical studies in which patterns of use of authentic texts are analysed, starting with large quantities of thoroughly collected data and using both quantitative and qualitative techniques.

From the analysis of the corpus-based ups and downs, Gabrielatos, McEnery, Diggle and Baker (2012) set forth the fruitful interface that can be created between qualitative and quantitative techniques. The expansion of corpora-based linguistic research has been supported by the various programmes[12] being devised.

According to Sardinha (2004: 38), we can identify within corpus linguistics three main areas of action: the area of corpora collection and organisation, which

[11] For a more detailed overview, please see http://jan.ucc.nau.edu/biber/

[12] These programmes are available online at http://registry.dfki.de/

compiles and organises the data collected, aiming at its subsequent use; the area concerning the development of computer resources, which aims at corpora analysis (in this area, those taking action are the researchers focused on computational linguistics and on the production of models and algorithms for natural language processing); and the third area, in which the researchers use corpora and computer resources to describe the lexicon and linguistic functioning, based on their use.

The greatest number of works derives from the applied field of corpus linguistics, such as lexicography and terminology (production of dictionaries, glossaries, terminological databases, etc.); the teaching of languages (production of didactic resources, observation of authentic examples of language in actual use); computational linguistics and natural language processing (in machine translation, in speech recognition, in the development of spelling and grammar checkers) (Kennedy 1998: 9). Actually, corpus linguistics has completely revolutionised the manner in which language is studied (McEnery and Wilson, 1996).

At a mega-corpora level, Teubert and Čermáková (2004: 115) refer not only to the Bank of English for the English language, but also to the IDS (Institut für Deutsche Sprache), comprising more than 1 billion words, the Språkbanken Swedish, with 75 million words, and the Czech National Corpus, with 100 million words.

Currently, on the Web, we can find many databases that disseminate and make corpora available, some providing open access and others with an associated cost.

Final Remarks

Corpus linguistics focuses on linguistic analysis, which can be used to conduct research on several issues related to language. Within its field of action, interesting and often surprising knowledge about language is discovered; it is one of the most widespread methods in linguistic research in recent years. Corpus linguistics is based on an empiricist approach and conceives language as a probabilistic system. According to this conception, linguistic features do not occur randomly, making it possible to detect and quantify patterns – that is, regularities. In light of this, language is said to be standardised – that is, there is an interdependence between the

linguistic features and the situational contexts of language use. Standardisation occurs through collocations, coalitions or structures that are significantly repeated.

Some linguists, such as Kennedy (1998), favour a mixed approach, combining intuition and corpus. Moreover, they share similarities in some respects with Chomsky. Kennedy accepts that the functioning of language cannot be fully revealed by the corpora because they do not enable the distinction between possible and impossible structures. In line with many corpora linguists, Kennedy admits that the non-appearance of a certain element in a corpus, even if a large one, does not invalidate its existence. Conversely, the appearance of a structure in a corpus does not automatically determine its grammaticality. Thus, from the outset, the exclusive use of a corpus may limit the range of linguistic data to be studied, or reveal previously well-established data, making its study redundant. In light of the above, the use of a corpus is a key auxiliary device, especially to check examples and validate intuitions.

From a traditionalist point of view, the use of computer science in lexical analysis seems to be unavailing. However, many scholars of the humanities in general, besides revealing the salutary realisation of the inevitable affiliation of the humanities with computer science, recommend the use of the computer as a precious asset to ensure the vitality of the humanities with respect to statistical and lexical analysis.

Through computer science, we can thoroughly observe the frequency with which certain words occur in the text or analyse the theme words, the exclusive forms or frequency forms; we can also use concordances, among other things.

In fact, corpus linguistics has made undeniable progress.

References

BIBER, D., S. CONRAD and R. REPPEN. *Corpus linguistics. Investigating language structure and use. Cambridge: Cambridge University Press, 1998.*

BOAS, F. *Race, language and culture*. New York, NY: Macmillan, 1940.

BONGERS, H..*The history and principles of vocabulary control*. Worden, IL: Wocopi, 1947.

CHOMSKY, N. *Syntactic Structures*. The Hague, Mouton, 1957

CHOMSKY, N. A review of B. F. Skinner's 'Verbal Behavior'. *Language*, Vol. 35, Nº 1: 26-58, 1959.

EATON, H. S. *Semantic frequency list for English, French, German and Spanish*. Chicago, IL: Chicago University Press, 441p., 1940.

FRIES, C. and TRAVER, A. *English word lists: A study of their adaptability and instruction*. Washington, DC: American Council of Education, 1940.

GABRIELATOS, C., T. MCENERY, P. J. DIGGLE and P. BAKER. The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics*, Vol. 17 Nº 2: 151–17, 2012.

HALLIDAY, M. A. K. Language structure and language function. In: J. LYONS (ed.), *New horizons in linguistics*. Harmondsworth: Penguin Books, 140–164, 1970.

HARRIS, Z. *Methods in structural linguistics*. Chicago, IL: University of Chicago Press, 1951.

JONES, R.L. Review of *Micro-OCP. Computers and the humanities*. 23, 2: 131-135, 1989.

KADING, J. *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz bei Berlinn: Selbstverlag, 1897.

KENNEDY, G. 'Between' and 'through': The company they keep and the functions they serve. In: K. AIJMER and B. ALTENBERG (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik*. London/New York: Longman. 239-256, 1991.

KENNEDY, G. *An introduction to corpus linguistics*. London: Longman, 1998.

LEECH, G. Corpora and theories of linguistic performance. In: Jan Svartvik (ed.), *Directions in corpus linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*. Berlin, New York: Mouton de Gruyter, 105–122, 1992.

MCENERY, T.; WILSON, A.. *Corpus linguistics*. Edinburgh: Edinburgh University Press, 1996.

_____*Corpus linguistica: An introduction*. Edinburgh: Edinburgh University Press, 2001.

QUIRK, R. Towards a description of English usage. *Transactions of the Philological Society*. 59 (1) (1960), 40–61, 1960.

QUIRK, R., GREENBAUM, S., LEECH, G. and SVARTVIK, J.. *A comprehensive grammar of the English language*. London: Longman, 1985.

RAJAGOPALAN, K. A Linguística de Corpus no tempo e no espaço: Visão reflexiva. In: R. M. GERBER and V. VASILÉVSKI (eds.), *Um percurso para pesquisas com base em corpus*. Florianópolis: Editora da UFSC. 33-44, 2007.

SARDINHA, T. B. "Lingüística de Corpus: Histórico e Problemática". In: *D.E.L.T.A.* Vol. 16, Nº 2: 323-367, 2000.

SARDINHA, T. B. *Linguística de Corpus: Histórico e Problemática*. São Paulo: Manole, 2004.

SINCLAIR, J. (ed.). *Collins cobuild English dictionary.* London: Williams Collins Sons & Co, 1987.

SINCLAIR, J.. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991.

_____From theory to practice. In: G. LEECH et *al*. (eds.). *Spoken English on computer: Transcription, mark-up and application*. London: Longman, 1995.

SKINNER, B. F. *Verbal behavior*. New York, NY: Appleton-Century-Crofts, 1957.

STUBBS, M. *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell Publishers, 1996.

_____Review of T. McEnery and A. Wilson. 'Corpus Linguistics'. Edinburgh: Edinburgh University Press, 1996. *International Journal of Corpus Linguistics*, Vol. 2, Nº 2, 296–300, 1997.

TEUBERT, W. "My version of corpus linguistics." *International Journal of Corpus Linguistics*, Vol. 10 Nº 1: 1–13, 2005.

_____ Rethinking corpus linguistics. In: Aquilino Sánchez and Moisés Almela (eds.), *A mosaic of corpus linguistics: Selected approaches.* Frankfurt: Internationaler Verlag der Wissenschaften. 19-42, 2010.

TEUBERT, W. and ČERMÁKOVÁ, A.. Directions in corpus linguistics. In: M. HALLIDAY et *al.* (eds.), *Lexicology and corpus linguistics*. London: Continuum.113-166, 2004.

Recebido: 21/02/2019.
Aprovado: 03/04/2019.