DIANA OȚĂT

DIANA OŢĂT

CORPUS LINGUISTICS. INTEGRATIVE APPROACHES

EDITURA UNIVERSITARIA

Craiova, 2019

Referenți științifici:
Prof.univ.dr. Titela VÎLCEANU
Conf.univ.dr. Ana Maria TRANTESCU

Descrierea CIP a Bibliotecii Naţionale a României
OŢĂT, DIANA
Corpus linguistics : integrative approaches / Diana Oţăt. - Craiova : Universitaria, 2019
Conţine bibliografie
ISBN 978-606-14-1544-1

81

FOREWORD

Let us put Corpus Linguistics on the map!

Charted as one of the youngest branches of Linguistics, albeit one with ancient roots, Corpus Linguistics has undergone a systematic reconfiguration with the advent of personal computers, back in the 1990's. Powered by the technological advances for over 30 years, computer corpora have been constantly upgraded as to store many millions of running words and to facilitate in-depth language-oriented analyses.

Put simply, Corpus Linguistics is "the study of language based on examples of real life language use" (McEnery and Wilson 2001: 1) that has been reshaping the landscape of empirical research on languages over the past decades.

The late 20th century Digital Age and the progressive upgrading of cutting-edge technologies led modern linguists link this relatively new approach in Linguistics to the experimental study of *real life* language use aided by dedicated software, computer-assisted tools and, more recently, virtual environments and cloud-based resources.

Adopting an integrative approach, the book aims at framing some main functional dimensions of Corpus Linguistics, raising awareness of their interactive features, geared to validate cross-linking developmental patterns of the communication event.

Divided into six main chapters, the book outlines the evolution of Corpus Linguistics, diachronically and synchronically.

In CHAPTER 1 we set out to define and feature the interdisciplinary dimension of Corpus Linguistics by tracing back the "archetypical corpus work". Kenedy (2001: 952) Browsing through the evolution stages of Corpus Linguistics, we aim to shed some light on the manifold research perspectives that place Corpus Linguistics at the crossroads of various disciplines. CHAPTER 2 raises the question of whether Corpus Linguistics is an

independent branch of Linguistics or a research method, in an attempt to frame some main principles which underpin the position adopted by prominent scholars with a view to the scope and status of Corpus Linguistics.

Channelling the focus on the main types of corpora, design parameters and selection criteria, CHAPTER 3 tackles the topic of corpus design, a mandatory condition to secure the functional construction of the corpus under investigation and the validity of corpus study outcomes.

Aiming at highlighting the broad application range of corpora investigations within various cross-disciplinary research approaches, CHAPTER 4 and 5 address the interdisciplinary nature of Corpus Linguistics as approached from the fields of Linguistics and Translations Studies.

With a view to future developments of Corpus Linguistics, CHAPTER 6 is devoted to some state-of-the-art research trends and interactive design and analysis models carried out in user-friendly virtual environments.

The Author

CHAPTER 1

CORPUS LINGUISTICS OUTSET

Acknowledging the importance of Corpus Linguistics and the prospective input this field of research propagates among different academic subjects, numerous scholars set out to feature its interdisciplinary dimension that for "the casual observer or new arrival might also appear to be a bewildering variety of definitions and descriptions". (Taylor 2009: 179)

Witnessing the steady multiplication and diversification of the current research perspectives, we adhere to the recent views on corpus studies as *something more than a collection of almost anything*. Within this framework, we highlight the contemporary academic cooperative effort within the field of Corpus Linguistics to put on the map language-oriented interdisciplinary issues based on samples of naturally occurring texts (written or spoken) in an attempt to investigate "what people do with language and how they view the world". (Mahlberg 2015: 2)

1.1. Defining Corpus Linguistics

Described by various authors as "the study of language based on examples of real life language use" (see McEnery and Wilson 2001: 2), *Corpus Linguistics* is a relatively new term within the field of Linguistics, concerned with the empirical investigation of "language use patterns, based on analysis of large collections of natural texts." (Biber and Reppen 2011: 3)

Branching out from what McEnery and Wilson (2001: 2) would outline as a "marginalised approach used largely in English linguistics, and more specifically in studies of English grammar", Corpus Linguistics has gradually widen its scope towards exhaustive linguistic-oriented analyses of machine-readable language corpora, taking up "the mainstream of research

on the English language" (Mair and Hundt 2000: 2). The development and multiplication of such research studies has further led to the accumulation of an impressive body of results which, "over and above the intrinsic descriptive interest it holds for students of the English language, forces a major and systematic re-thinking of foundational issues in linguistic theory". (ibidem)

Keen on the study of large bodies of texts (corpora), corpus linguists set out to research and apply customised analysis methods to investigate language use patterns empirically, while Corpus Linguistics unfolded as a multilingual field of research, focusing on the development and manifestation of various languages and, gradually, on many varieties of different language systems.

The constant development of the research spectrum has successfully integrated Corpus Linguistics in almost every branch of knowledge, while, according to Hornero et al. (2006), the contemporary growth of corpus-based research and analysis has become increasingly observable in almost every branch of Linguistics. Thus, "while linguistics divides up into many research areas depending on complexes of research questions, corpus linguistics in essence behaves diametrically: it offers a set of methods that can be used in the investigation of a large number of different research questions." (see Lüdeling and Kytö 2008: iii)

Under these circumstances, Römer and Wulff (2010: 100), advocate that those who adhere to corpus practice:

> (...) share the common assumptions that linguistic theorizing should be driven first and foremost by (representative samples of) authentic language data, and that a solid linguistic hypothesis and theoretical claims should be based on a thorough description of these data with regard to the phenomenon under investigation.

Taking into account the driving elements that feature Corpus Linguistics, i.e. *research methodology*, *data description*, and *language*, Laviosa (2014: 14) takes a step further and claims that Corpus Linguistics can

be labelled as an independent sub-discipline, a branch of Linguistics that undergoes "a continual process involving corpus creation, discovery, hypothesis formation, testing and evaluation."

1.1.1. Zooming in Corpus Linguistics

Outlining the evolution journey of Corpus Linguistics, as a "method of exegesis based on detailed searches for words and phrases in multiple contexts across large amounts of texts", McEnery and Wilson 1996: 5) take us back to the thirteenth century, "when biblical scholars and their teams of minions pored over page after page of the Christian Bible and manually indexed its words, line by line, page by page." Similarly, Kenedy (2001: 952) claims "that archetypical corpus work existed well before the modern digital era" and pins down the early attempts of word indexing and concordance of the Christian Bible in the thirteenth century. From this moment on, the concept of *concordancing* emerged out of an empirical necessity to pass on to future generations alphabetic lists of the words contained in the Bible, along with citations of where and in what passages they occurred. We grow aware that via early times concordances, scholars succeeded in compiling, by hand, laborious works; the same way nowadays software packages perform similar tasks, able to "replicate the work of 500 monks in micro-seconds." (Baker 2010: 7)

Rooted in the Latin word *concordantia*, the contemporary term *concordance* derives from the Latin *cum,* meaning 'with', and *cor* meaning 'heart', which "ties in with the original ideological underpinning of this painstaking endeavour, namely to underscore the claim that the Bible was a harmonious divine message rather than a series of texts from a multitude of sources." (ibid: 3)

Diachronically, the terminus a quo of *concordance* was first associated with *Concordantiae Morales*, namely with the writings of Anthony

of Padua (1195–1231)[1], based on the *Vulgate*, the fifth-century Latin version of the Bible. Prominent scholars mention the first uses of the Latin terms *concordare* and *coincidence* that Anthony of Padua used to express the conviction that the several parts of the Bible are consistent with each other, as parts of a divine revelation, and may be combined as harmonious elements in one system of spiritual truth.

Another well-documented complementary work dating back to 1230 was compiled by Cardinal Hugo of St Caro[2] (also referred to as St Cher), who, according to Bromiley (1997: 757), together with a 500-strong team of Dominican monks at St James' convent in Paris, put together *a word index* of the Vulgate."

Tribble (1990) advocates that the works of Shakespeare were also the subject of *concordancing*, as a means of assisting scholars; in this respect the author mentions Becket's 1787 *A Concordance to Shakespeare*. Tribble (1990) argues that Becket's concordance illustrates the Shakespeare canon by the way the words and the linguistic setting is given.

We share the perspective put forward by Johansson (1998: 22) that the starting point of Corpus Linguistics can be traced back by considering the recordings of various observable data and how they have been handled throughout different periods of time and across different theoretical schools.

1.1.2. The interdisciplinary nature of Corpus Linguistics

Of all branches and sub-branches of Linguistics, Johansson (1998: 29) pinpoints the durable connection between Historical Linguistics and corpora investigations, as it always resorted to corpus-based analysis in order to register

[1] Saint Anthony of Padua (Portuguese: Santo António de Pádua), born Fernando Martins de Bulhões (15 August 1195 – 13 June 1231) – also known as Saint Anthony of Lisbon (Portuguese: Santo António de Lisboa) – was a Portuguese Catholic priest and friar of the Franciscan Order, commissioned to teach theology to the friars.

[2] Cardinal Hugo of St Caro - a Dominican monk (1263), who, in preparing for a commentary on the Scriptures, found the need of a concordance, and is reported to have used for the purpose the services of five hundred of his brother monks.

and display language dynamics within different periods and settings. Along the same lines, Lüdeling and Kytö (2008: 6) outline the methodological nature of modern Corpus Linguistics and its initial link to the history of Linguistics as an empirical science, arguing that the contemporary dimension of Corpus Linguistics relies on much older methods than computers, many of them rooted in the tradition of the late eighteenth and nineteenth century, when Linguistics was for the first time claimed to be a "real", or empirical, science.

Allegedly, the evolution of modern Corpus Linguistics is endorsed by some main contributions brought by Comparative and Historical Linguistics, since that time linguists used to rely on text samples and text collections as their raw investigation sources.

Once the nineteenth century language variation studies and the reconstruction efforts carried out by linguists such as Jacob Grimm[3] and the later Neogrammarians[4] expanded, many corpus-design and corpus-analysis techniques developed as well; back in the day, such research studies depended on early texts or corpora (see *Sprachdenkmäler*/*Language Monuments*). Many of the nineteenth-century techniques were adopted and further developed in modern Corpus Linguistics. Maybe this is the reason large historical corpora were first addressed, also among the first electronically available corpora. Lüdeling and Kytö (ibid: 8) highlight Roberto Busa's (1973) pioneering work on the writings of St Thomas Aquinas and Louis Milic's *Augustan Prose Sample*.

Depending on the complexity level of the research questions and objectives set for different types of corpora, corpus investigations have branched out to various research areas. New hybrid and/or niche research fields such as recent developments in the area of Forensic Linguistics were

[3] Jacob Ludwig Karl Grimm, also known as Ludwig Karl, was a German philologist, jurist, and mythologist. He is known as the discoverer of *Grimm's law of linguistics*, the co-author of the monumental *Deutsches Wörterbuch*, the author of *Deutsche Mythologie*, and the editor of *Grim's Fairy Tales*.

[4] The Neogrammarians (German: Junggrammatiker, "young grammarians") were a German school of linguists, originally at the University of Leipzig, in the late 19th century who proposed the Neogrammarian hypothesis of the regularity of sound change.

generated from the attempt to identify distinctive writing characteristics of a particular author, such as the investigation of monogeneric corpora that focus on style and authorship-oriented studies. (Coulthard 1993, Baker 1996)

Similarly, Political Sciences have exploited meticulously various corpora to study different particular situations. Inter alia, Partington (2003) mentions the development of *corpus-driven studies* aimed to establish specific language patterns, metaphors and motifs used by participants in a number of political/institutional spheres, e.g. journalists and spokespersons in US press conferences, and how they reflect their world-views. Likewise, Garzone and Santulli (2004) would focus on the rhetoric of Berlusconi's electoral speeches, while Vaghi and Venuti (2004) have addressed the issue of data quality among the UK newspapers in their stance on European Monetary Union. Based on Corpus Linguistics approach, Walsh (2004) was keen to find out how prediction is established in economic texts. Moreover, Bayley (2004) carried out intensive corpus-based studies on the language of representative assemblies in order to feature the particularities of the special discourse communities that activate within specific political institutions.

1.2. Mapping the Routes of Corpus Linguistics

Unquestionably, corpus studies have provided remarkable support for various branches of Linguistics enabling researchers and scholars to widen their research interests and concerns. Taking advantage of a more systematised compilation of larger corpora displaying more accurate and multifaceted annotation, contemporary theorists and practitioners could access much easier a large number of written and/or spoken samples with chronological dialectal or psycholinguistic significance, thus contributing to extensive analyses that "radically improved the replicability of research results" (Rissanen 2008: 54) and legitimated claims about language use on the basis of corpora material.

Beyond the technological breakthrough that enabled linguists to access and investigate large corpora at a touch of a button, and, moreover to store and reuse the data for further investigations, language researchers began to manifest an increased hands-on approach to language use as opposed to language systems from a theoretical perspective.

Browsing through the evolution stages of Corpus Linguistics, we grow aware of the manifold research perspectives that place Corpus Linguistics at crossroads of various disciplines, as it exhibits an interconnected developmental pattern, departing from its dominant sources, i.e. Historical and Comparative Linguistics (see Figure 1.1 below).

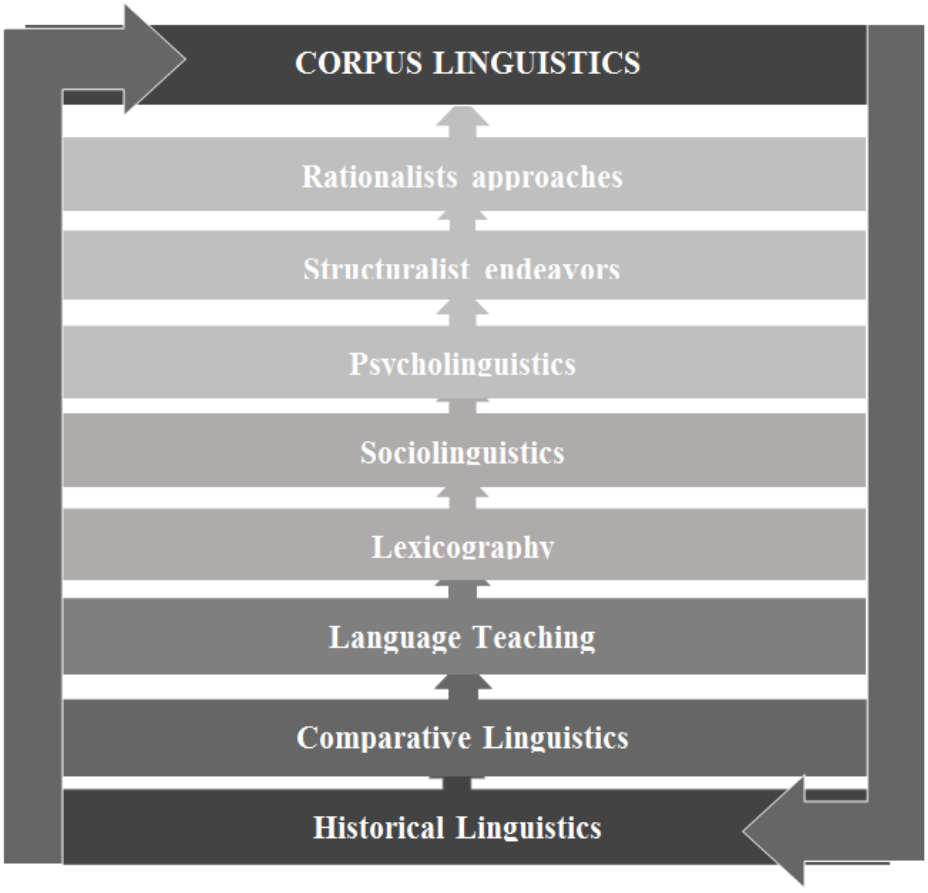

Figure 1.1. – *Corpus Linguistics developmental routes*

1.2.1. Language History, Comparative Linguistics and Corpus Linguistics

As previously mentioned, contemporary scholars link the maturing of modern Corpus Linguistics to Historical Linguistics, namely to the early corpus-based investigations on language change and reconstruction. Hence, Historical and Comparative Linguistics are considered the two major contributors to the development of today's Corpus Linguistics. Most of the contemporary corpus-based analysis methods can be traced back in the nineteenth century, applied in the attempt to reconstruct older languages or to identify relationships among different languages. Large samples of texts were exhaustively investigated to gather information on language change. (see Kytö and Rissanen 1990, Mair 1993, Baker 2010)

Taking over the nineteenth century Indo-European descriptive-oriented tradition introduced by Language History and Comparative Linguistics scholars, the American Structuralists systematically elevated the scope of Corpus Linguistics. On the one hand, the Structuralists set out to collect information on different languages so as to better describe these languages, and, on the other hand, they enhanced corpus-based investigations by transcribing various oral language productions that had not been recorded in written form previously. It is the era when corpus-based investigations generated considerable amounts of information, especially in terms of morpho-syntax and semantics, following various research studies carried out within most branches of the historical study of language. As technology evolved, developing more accurate and multifaceted annotation algorithms, the linguistics-oriented investigation spectrum expanded, focusing on discourse and pragmatics-related issues.

At the same time, some contemporary scholars highlight the solid support corpora investigations generated in terms of research replicability, enabling history linguists to check the correctness and accuracy of the linguistic evidence presented in historical language studies.

Within this context, Dipper (2008) argues that corpora investigations have played a major role, particularly in analysing variation and change for