4 Corpus Linguistics Outcomes and Applications in the Digital Era

4.1 Introduction

Prominent scholars highlight the key role of word use in the development of corpus linguistics as the study of linguistic phenomena by means of extensive collections of machine-readable texts, i.e., by means of corpora. The steady evolution of corpus linguistics has been primarily motivated by the linguists’ need to understand how words are actually used in natural languages, which most common words tend to be used in certain contexts, what is common and what is uncommon for certain language variations (including specialisms), thus leading to the first outcomes provided by corpus-based approaches, i.e., words lists and synonymous terms. Hence, mainstream literature pinpoints the emergence of corpus-based investigations as early as 1755, when, as endorsed by Biber et al. (2006: 22), Johnson used a corpus of texts to gather authentic uses of words that he then included as examples in his dictionary – a first step made towards the understanding of the patterns of use associated with a word.

A less popular branch of linguistics, though widely explored in the 1940s and 1950s, statistical linguistics, has also contributed to the development of what nowadays corpus linguistics has become through innovative mathematical theories of information. Yet, due to the lack of computer-assisted processing tools back at that time, it proved to lack productivity and effectiveness.

Another branch of linguistics closely related to the current state-of-the-art of corpus linguistics was what ‘older linguists, of the heyday in the 1950s’, such as Harris, Fries or Hill and other American structuralists regarded as ‘descriptive linguistics’ (Leech, 1992: 106), i.e., the scholars’ aim at describing the corpus under investigation. Accordingly, endorsing the flexible typology of descriptive linguistics towards theory construction, Leech (ibidem) highlights the less abstract nature of its outcomes, particular to one language, where linguistic phenomena are more easily to localise, observe and analyse. In this context, we grow aware of one main characteristic of corpus linguistics, namely that most corpus-based analyses are applied to data inquiries specific to individual languages.

In the 1980s, linguists registered a rebirth of corpus linguistics, which by then had already indicated a close connection with quantitative linguistics, a specialised research branch that promotes the need for quantitative methods in language study, which, according to various linguists, are also frequently used in most other disciplines as they can provide reliable outcomes when it comes to the description of language in terms of frequency and infrequency rates.
However, this diachronic examination of the development of corpus linguistics differs considerably from the status it has acquired in the digital era. For contemporary researchers, corpus-based analysis does not only serve the mere purpose of dictionary making. Theorists and practitioners alike go far beyond the normative function of corpus-based analyses, concerned with simple inventories of linguistic structures, charting new territories in the development of corpus linguistics oriented towards qualitative and functional interpretations of quantitative research. It represents, as postulated by Svartvik (1992: 8), a way ‘to take a look at real manifestation of language when discussing linguistic problems’, for as ‘corpus linguistics is not the heaping of data for its own sake, but rather the investigation of data for scientific purposes’ (ibidem).

4.2 Corpus Linguistics in the Digitalised Era

It is common knowledge that the dominant role in corpus linguistics is played by modern technology. As previously mentioned, corpus linguistics has been regarded as an operational framework in language study, rather than an isolated domain of study. It is not a monolithic system providing fixed sets of homogenous methods and procedures applied to language investigation, but, as defined by McEnery and Hardie (2012: 5), ‘a heterogeneous field’. Applied to the study of the language, corpus-based analysis aims at investigating particular linguistic structures, the way they occur in different contexts and the functions they acquire. The current perspectives in corpus linguistics indicate the extensive use of machine-readable texts as the appropriate resources, the raw material on which to study specific linguistic issues and phenomena.

As put forward by scholars, such as Kennedy (1998) or Biber et al. (2006), computer technology advances not only have multiplied the investigation perspectives where to apply corpus-based analysis, but have also provided novel benefits and backups in comparison to, previous research studies. Undoubtedly, the first significant advance that upgraded this hybridised field is the storage capacity of very large databases of natural language that can be compiled from a wide range of sources. Today, we can save, store and organise on our own computers ample writings or large chunks of texts, being, thus, able to carry out thorough linguistic analyses that are no more limited to sentence-length excerpts.

Modern software and computer-assisted tools secure comprehensive linguistic analyses that are more accurate and reliable. In this climate of opinion, Biber et al. (2006) also state that ‘unlike human readers that are likely to miss certain occurrences of a word, computers can find all the instances of a word in a corpus and generate an exhaustive list of them’. However, it is worth mentioning that corpus linguistics is not only concerned with words frequency, for as other branches of linguistics, such as phonetics, syntax and any another aspect of linguistics may be investigated via corpora analysis while applying and combining a series of corpus-based investigation methods. In line with the prominent figures in the field of linguistics, we can endorse
that computer-assisted analysis methods can be applied to various types of corpora, aiming at investigating particular patterns of word associations on a far more complex model than it is possible manually.

Among the core features of computer-assisted corpora analysis applied to language study, we can mention the empirical character of corpus-based analysis, i.e., the investigation of actual language structures in authentic texts, namely in corpora, which represent the basis for the analysis designed and implemented. As previously indicated, dedicated software and computer-based analysis tools have been extensively applied within this field of research, which led to the implementation and operation of hybridised techniques, which rely heavily on quantitative and qualitative methods alike. As endorsed by Biber et al., computer-assisted investigation has provided not only reliable analysis methods, but also consistent outcomes, enabling individuals actively interact in validating complex linguistic findings, while ‘the computer takes care of record-keeping’ (ibid: 6).

Two main corpus-based research directions have been indicated by the authors mentioned above, where linguistic features are not investigated as isolated occurrences, but as systematic associations with other features. Accordingly, we distinguish between ‘linguistic associations of the features’ that encompass lexical and grammatical associations and ‘non-linguistic associations of the features’, i.e., distribution across registers, dialects and time periods (ibidem).

Concerned with the study of linguistic features by means of corpus-based analyses, it is worth mentioning that corpus-based research can be applied to grammar on the word level as well as on the sentence and discourse levels. Also, register and text typology can be investigated by means of corpora study.

As far as grammar study is concerned, corpus-based analysis has recently registered fruitful advances towards novel interdisciplinary areas of investigation. Thus, even though former linguists highlighted the descriptive approach applied to the grammar study, prescriptive methods tend to be used more frequently, for example, in order to establish the variables framing the syntax of language. By applying corpus-based analysis, new research perspectives envisaged the patterned use of grammatical features in texts or the variation of language in use. The investigations of natural language texts have enabled linguists to register and further investigate the routinised ways that individuals tend to use the grammatical resources of a language. Linguists postulated that, by investigating the distribution of various structures, the association patterns between grammatical structures and other linguistic and non-linguistic factors that influence the individuals’ linguistic choices, all of them recorded in authentic texts, considerable outcomes and research advances have been achieved. Such fruitful headways were not only reached at the grammatical level, but also in terms of language evolution. Its dynamics and the implementation of the most appropriate strategies to bridge socio-cultural and linguistic differences is an enduring benefit for closely related fields of research, such as translation studies, sociolinguistics or cultural studies.
4.3 Corpora Design and Compilation in the Digital Era

The central aim of the present paper is not only to highlight the considerably value-added and reliable outcomes provided by computer-assisted corpus-based analysis, but also to pinpoint the highly effective dedicated software and computer-assisted tools applied in corpora design and compilation.

We should first mention that, paradoxically, the concept of corpus was initially used to designate various non-linguistic collections, such as the Corpus Juris Civils introduced by the emperor Justinian in the sixth century and regarded by numerous scholars as a compilation of early Roman law and legal principles, which also illustrated particular cases and provided clarifications of new laws and future legislation to be put in effect (see Jan Svartvik, 1992). However, linguistics corpora are regarded as collections of texts ‘assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis’, as defined by Francis (1992: 17). Thus, in language study, corpora have been primarily used for linguistic analysis, a feature that according to Francis (ibid:19), differentiates them from other types of corpora or large text collections, such as anthologies, for example, the Oxford Book of English Verse, whose purpose is literary.

Other examples of corpora types are lexicographical corpora, compiled in the process of making dictionaries, for example, the Oxford English Dictionary, edited in the 19th century by Murray or the Merriam-Webster Dictionary edited in the 20th; dialectological corpora compiled for the purpose of designing dialect atlases, for example, in the Middle English Dialect Atlas, issued in 1981 by Benskin, or grammatical corpora, among which Quirk’s Survey of English Usage published in the 20th century is the most well known.

Starting with the early 1960s, modern concepts of corpora generally indicate the use of large collections of texts available in machine-readable forms, of which the most representative one is the Brown University Standard Corpus of Present-Day American English compiled by Henry Kucera and W. Nelson Francis and which remains a sample of present-day English for use with digital computers. In the next two decades, it was followed by the creation of the British National Corpus (BNC), started in 1991.

Given the broad and multi-faceted research directions that emerged in the field of linguistics, as well as in previously mentioned close related domains, where corpora investigations are applied in order to achieve novel quantitative and qualitative outcomes, McEnery and Hardie (2012: 12) put forward a series of criteria applied to distinguish between the types of corpus-based investigations:

-  mode of communication
-  corpus-based versus corpus-driven communication
-  data collection regime
-  the use of annotated versus unannotated corpora
-  total accountability versus data selection
-  multilingual versus monolingual corpora
By establishing these criteria, the authors aimed at featuring a typology of corpus linguistic research framed by the principles of corpora use.

Albeit corpus-based analyses have by no means been restricted to the English language, it is common knowledge that corpora investigations applied to the study of the English language have provided the most relevant and significant language study perspectives in corpus linguistics during the past period, leading to a boost and proliferation in English studies generally. Among the most frequently approached research perspectives carried out by means of corpus analysis, we could mention the study of language variation, dialect, register and style, where authentic samples of different areas of language use have been compiled and investigated in order to validate the diverse range of language users or the close analysis of frequency rates of particular linguistic structures in different language varieties or in certain specialised languages.

Thus, corpora are far from being considered mere language samples aimed to provide useful illustrative examples, but genuine theoretical resources, which, implemented in a series of applied fields of research, such as language teaching, translation studies or even machine translation and dedicated processing software (spelling, grammar and style), have provided essential sources of information leading towards generalisations about the language and language use. Furthermore, online freely available corpora are meant to ensure linguists from all over the world a user-friendly access to language materials that would otherwise be difficult or impossible to obtain. Conversely, linguists who are non-native speakers can use such corpora for further research and practice.

Having established that corpora design and compilation in the context of new cutting-edge technologies is a key concern of the present paper; it is also noteworthy mentioning that corpora fall into two major categories, i.e., general corpora – designed to investigate a given language as a whole – and specialised corpora – designed to answer more specific research questions, mainly used in the study of special languages or closely connected applied fields of research. Accordingly, while general corpora are thoroughly organised in order to have a long ‘shelf-life’, as claimed by Lüdeling and Kytö (2008: 154), specialised corpora aimed for the study of certain linguistic items in certain contexts or circumstances are much more rapidly constructed. The authors consider that most of the corpora can be understood as a collection of sub-corpora which display relative homogenous features. Inevitably the current corpora design is based on ‘a template of variables that creates a number of cells, each of which constitutes a sub-corpus’ (ibidem). If a linguist aims at investigating the specialised domain of the written legal language, s/he may start designing her/his corpus by collecting, storing and organising various types of legal documents, such as contracts, laws, treaties, etc. that may constitute the cells or the sub-corpora.

Aston and Burnard (1998: 29) put forward a series of references applied in corpora design, i.e., field, tenor and mode variables, which, for example, can be easily recognisable amongst the most representative criteria of the British National Corpus.
Corpus Linguistics Outcomes and Applications in the Digital Era

(BNC) that includes field (the subject-matter of written texts), tenor (spoken texts), and mode (books, periodicals).

A central issue in corpora design and particularly in the design of specialised corpora is how to establish the most appropriate variables and how the dynamics of the sub-corpora should function, considering the fact that most of the times such specialised corpora reveal a multi-dimensional character. Nevertheless, as argued by Lüdeling and Kytö (2008), in most situations it is obviously not possible to collect all the texts that constitute the research object into a corpus. In this respect, the authors advocate that the selection of a subset of the texts aimed for analysis stands as a reliable method. Moreover, the scholars recommend that establishing of strict sampling techniques may lead to reliable and operative corpus design. We would tend to emphasise that the contents of a corpus designed for research purposes need to be carefully considered, though some will argue that the design of the corpus will depend more on what is freely available in an easily converted format than on other criteria, thus pinpointing towards the benefits provided by the digital era.

Within this context, a series of other very important aspects that need to be considered in the corpora design can be mentioned, i.e., software limitations, copyright, and text availability.

It is a truism that dedicated software provides an overwhelming capacity to store databases under different shapes (written, spoken, images, graphics, etc.); various linguists have pointed out that readily available software packages, such as WordSmith Tools (Scott, 2004), can deal with a corpus of tens of millions of tokens in size. Lüdeling and Kytö (2008: 157) also mention the existence of larger corpora that often work with software, which demands that the tokens are converted into digits, which would facilitate the software to process more quickly. Be that as it may, specialists from linguistics and IT alike advocate that complex operations on corpora that may involve hundreds of millions of words can take some time to complete.

Furthermore, another aspect that needs to be taken into consideration is that the storage and further use of the electronic version of published texts remains a crime in many of the EU Member States and worldwide as well, if no copyright permissions are legalised. Thus, due to the fact that such permissions are often difficult to gain, the availability of some corpora may be restricted.

Another aspect highlighted also by Lüdeling and Kytö (ibid: 158) refers to the availability of texts. In this respect, the authors mention the design limitations imposed by the corpora designers who wish to include in their corpora ‘spoken texts from an era before the invention of tape recorders’. The accuracy of the data as a representation of the actual speech is always questionable. Even though written texts are easier to obtain, there are some situations when extensive written texts are difficult to be incorporated into a corpus, unless very large storage facilities are available for scanning and keying.
4.4 Application: A Model for Computer-Assisted Corpus Design and Analysis

In what follows, we aim at exemplifying a computer-assisted model for corpus design and corpus analysis that can be applied to specialised corpora used for the investigation of specialised languages as well as to other fields of research, such as language teaching or translation training programmes.

The model proposed is based on the MAXQDA 11 for Windows professional software dedicated to qualitative and mixed methods data analysis. As highlighted by its promotors, MAXQDA provides a variety of research methods and approaches. By means of this computer-assisted tool, we can organise, encode, annotate, and interpret an array of data. Moreover, the analysis outcomes can be generated in easy-to-read reports, visualisations, Excel sheets, while enabling the researchers to work interactively and share the results among each other.

Applying MAXQDA to the study of language, we simulated first the possibilities of the corpus design. Providing the fact that MAXQDA allows the import of various format types, a possible corpus for language analysis may encompass a wide range of written and spoken texts, as well as images, graphics or tables in TXT, RTF, DOC/X, PDF, JPG, GIF, TIF, and PNG format. Media files can be also selected here, enabling the designer(s) to set up even a multimodal corpus.

The software allows the users to design their corpus in a project-based format, where each group member can actively participate, save memos and observations, and apply them both to the design and the analysis of the corpus.

Moreover, the software allows the users design various corpora simultaneously, which can be further structured in a sub-corpora, if organised in separate document groups.

![Print Screen: Corpus organisation in MAXQDA 11.](image-url)
For example, we could simultaneously design two specialised corpora used for the study of different linguistic issues typical for legal language, i.e., a corpus encompassing legal documents, which can be further organised in sub-corpora, such as contracts, laws, regulations, etc., and a corpus encompassing specialised texts from the automotive industry, i.e., user manuals, technical specifications, manufacturing provisions, etc. Such corpora may be then individually investigated by each member of the project group in accordance with the specific research objectives, or they can even be contrastively analysed, if, for example, certain linguistic features are to be characterised contrastively in particular registers or even along certain time periods, i.e., time series.

By providing a user-friendly interface, MAXQDA reveals similar features to other Windows programs, facilitating quick and effective processing on behalf of the users. After completing the design stage of a corpus, which, as previously indicated, offers innumerable compiling possibilities, we can embark on the linguistic analysis of the corpus by means of the drop-down menus and the various toolbars with buttons that offer quick access to the functions. It is worth mentioning that even at these stages, the investigation possibilities provided by the software are numerous. A general overview of a linguistic analysis model would imply the use of the following menus:

- the Analysis Menu – provides a series of analysis options applied especially to the lexical search and retrieval functions. Thus, the lexical search option enables the user to search within the document, or just in the activated document sets, memos, and retrieved segments. This function facilitates the search for certain words, phrases, or combinations. Also, the keywords in a context can be searched and automatically encoded. As indicated in the user manual, most of this menu functions relate to retrieval. We can chose various criteria for the segments to be found (e.g., OR, AND, logical combinations, or NEAR). Moreover, the retrieved segments can also be filtered based on certain criteria in the ‘Retrieved Segments’ window.

- the Codes Menu – enables the user to create and apply new codes on all the documents or only on the activated ones, or even to create a complete index of all codes assigned to all the document segments.

- the Mixed Methods Menu – is used to process and combine qualitative and quantitative data using documents and variables. Documents or document groups can be investigated based on the assigned variables, limiting the retrievals to certain document segments. The Quote Matrix and Crosstabs functions can be applied in order to indicate connections between the encoded segments and the selected variables.

- the Visual Tools Menu – enables the users to visualise the outcomes by means of seven different visualisation function options. MAXMaps, the tool for qualitative modelling; the Code Matrix Browser; the Code Relations Browser; and the Document Comparison Chart. The Document Portrait and the Codeline functions
also provide further visualisations which can be exported to Excel sheets as well under the shape of tables or graphics.

Figure 4.2: Print Screen: document portrait generation according to the assigned codes.

- The **MAXdictio Menu** – is an optional menu which offers a number of functions for quantitative content analysis, e.g., coding according to created dictionaries and viewing word frequencies.

Figure 4.3: Print Screen: Word frequency list in MAXQDA 11.
Corpus Linguistics Outcomes and Applications in the Digital Era

As we can see, there are countless possibilities of analysis models, depending on the investigated linguistic issues and the specificity of the research fields applied. Of course, the design and implementation of simultaneous and multi-dimensional language analyses is possible, providing reliable and consistent outcomes in just a few minutes.

4.5 Conclusions

We can conclude that the computer-assisted approaches in corpus linguistics may lead, nowadays, to the refining and redefining of a wide range of theories of language. By means of dedicated software and computer-assisted tools, corpus linguistics has broadened its research directions considerably, smoothing the path towards new language explorations and theories.

As previously highlighted, within the context of advanced technologies, corpora are steadily exploited by tools that enable users to search through them rapidly and reliably. Most of such tools allow the production of frequency data, e.g., word frequency lists, document portraits, or comparison charts.

Unquestionably, there is a close link between the current status of corpus linguistics and modern technologies that brought to this field incredible speed, total liability, statistical reliability, sustainable results, and the opportunity to manipulate over considerably substantial and varied databases.

References


**Online resources**
