Chapter 14
Linguistic Corpora: A View from Turkish

Mustafa Aksan and Yeşim Aksan

Abstract Usage-based linguistic studies have gained new insights as corpus-based and corpus-driven analyses have advanced in recent years. Linguists working in different domains have turned to corpora as a major source in their study of language at all levels of representation. Currently, corpus linguistics is evolving into a sophisticated methodology in extracting and analyzing data. Building and using corpora in Turkish linguistics is a recent undertaking, initially motivated by work on natural language processing (NLP) research. The number of available corpora is increasing and linguistic research has come to test hypotheses on attested data, or uncover more lexical and grammatical patterns of use that have gone unnoticed in the absence of corpus data. Advances in NLP research and tools provided for corpus building and annotation further contribute to corpus studies in Turkish linguistics.

14.1 Introduction

In his comment on the state of American linguistics in the mid-1950s, Newmeyer (1986, p. 2) defines the period as "a period of optimism." The common understanding among the linguists of the time was that the field had achieved a level of sophistication in which major problems were solved and all that was left to do was to provide details: "…punch the data into the computer and out would come the grammar!" The success of linguistics was beyond doubt as other social sciences were imitating linguistics in adopting its methods in their research. However, the introduction of generative grammar in the late 1950s marked the end of classical structuralism and changed the course of the field. This new revolution in linguistics also "curtailed" early corpus-based theoretical frameworks and introduced "idealizations and abstractions" which had little to do with the empirical methodologies of corpus studies (Barlow 2011).

M. Aksan (✉) · Y. Aksan
Mersin University, Mersin, Turkey

K. Oflazer, M. Saraçlar (eds.), *Turkish Natural Language Processing*,
Theory and Applications of Natural Language Processing,
https://doi.org/10.1007/978-3-319-90165-7_14

Slowly but steadily, empirical linguistics made an impressive comeback, especially after the early 1990s, "... when computational linguistics embraced corpora as the automated analysis of large quantities of text data started to make serious impact in the development of speech recognition, machine-aided translation, and other natural language processing tasks ..." (Leech 2011, p. 157). The approach of linguists toward usage-based studies and the recognition of the role of frequencies and patterns determined via corpus-analytic tools resulted in a significant increase in the number of linguistic corpora. The intricate relationship between data, theory, and methodology is now being discussed in a new perspective motivated by the extensive use of corpus data in all fields of linguistics.

The use of corpora in Turkish linguistics studies is a comparatively recent enterprise. One major reason for this late involvement is the fact that no linguistic corpora was available until the early 2000s. As is well known to many involved in the process, corpus building is a labor-intensive and time-consuming activity that requires committed institutional backing. The small number of linguistics departments in Turkey and the lack of appreciation and funding of such work were the main reasons for lack of such corpora.

In this chapter, we present a brief review of the available Turkish linguistic corpora and corpus-based and corpus-driven linguistics: We will review the evolution of linguistic corpora in general, and corpora as a source in linguistic analysis and corpus linguistics as a method in linguistics along with arguments concerning the nature of the object, addressing the relationship between research questions and the typology of corpora. Then we will discuss the use of corpora in different fields of linguistics and the defining standards of representative and balanced corpora. The final section will give a brief review of the available linguistic corpora in Turkish, some linguistic work using corpus data, and their evaluation.

14.2 Brief History of Corpus Linguistics

It is by now customary to distinguish between the preelectronic and post-electronic eras in the development of corpus linguistics. Svartvik (2007), for example, notes that the initials BC, for a corpus linguist, stands for Before Computers. The preelectronic period refers to corpus studies that were predecessors of contemporary work and which were mostly done before the 1960s. For some, the early studies go back to the thirteenth century indexing work on the Bible and for others, to recent times as recently as the beginnings of the twentieth century work of American Structuralism in collecting textual samples of language use (Leech 1992).

Advances in computer technology, such as increase in storage capacities and the sophistication of available software, had a major impact on the progress of corpus linguistics. In fact, it is such advances that have empowered corpus linguistics to achieve its status today. Equally, we may say that linguistics also provided a strong impetus in developing many practical applications in computing in general because

it demanded new types of software in processing natural language for its complex manifestations at different levels.

Apart from the concordances derived from data stored on punch cards that appeared in the late 1950s, Francis and Kučera (1964) constructed the first ever electronic corpus of written English at Brown University in 1961. The Brown Corpus set the standards for corpus design with a size of one million words. The developments following the Brown Corpus are described as five phases or stages in Renouf (2007, p. 28). The stages are determined on the basis of the periods in which a specific corpus was constructed as well as the "types, styles and design" of the corpora of the time.

1. 1960s onwards: the one-million-word (or less) Small Corpus (standard, general and specialized, sampled, multimodal, multidimensional)
2. 1980s onwards: the multimillion-word Large Corpus (standard, general and specialized, sampled, multimodal, multidimensional)
3. 1990s onwards: the 'Modern Diachronic' Corpus (dynamic, open-ended, chronological data flow)
4. 1998 onwards: the Web as corpus (Web texts as sources of linguistic information)
5. 2005 onwards: the Grid (pathway to distributed corpora, consolidation of existing corpus types)

The impact of computer science and computer technology became more significant in the second and third stages. The development of desktop computing freed many corpus developers and corpus users from mainframes and the rapid growth and expansion of the Internet and data storage capacities helped to store and share data efficiently, and a new generation of scanners increased the capabilities of data entry processes. In linguistics, it became clear that some questions of lexis and grammar cannot be pursued properly in small-size corpora. Thus, the demands of corpus-based analyses and the appealing developments in computer technologies gave way to corpora of a different generation. Some of the resulting multimillion-word super-corpora of these two stages include the *Birmingham Corpus* (1980–1986, 20 million words), the *Bank of English* (1980 onwards, 650 million words as of 2012), and the *British National Corpus* (1991–1995, 100 million words).

In the next stage of development, after the advance of super corpora, almost the same motivations and emerging technological potentials contributed not only to size, but also to types of corpus construction. The development of the *monitor corpus*, or *modern diachronic corpus*, as recalled by Renouf (2007), goes back to 1982. It was observed that language is changing and language change can be captured and observed in corpora. The idea that innovations, variations, and changes in lexis and grammar can be followed in "dynamic corpus unbroken chronological text" resulted in a distinct type of corpus, continuously adding texts from the *Times* (starting in 1988), followed by the monitor corpora of the *Independent* and the *Guardian* journalistic texts.

The expansion of the Web in the 1990s led to a new era in corpus linguistics research and introduced new and improved corpus tools. The World Wide Web itself has become an online corpus in that some of the texts stored on the Web appear only

in electronic form and never in any other format, more varied language is manifested on the Web, and there are citations of new and rare lexical items and patterns that are not found in ordinary available corpora. Furthermore, the Web provided a cheap and easy means of building corpora with huge amounts of accessible texts, representing present-day language, updated continuously. The advantages that the Web as corpus presented were soon overshadowed by the problems observed by researchers. It was argued that the data on the Web is too heterogeneous and unstructured ("cheap and dirty") to derive any reliable conclusions when corpus linguistics methods are put to use.

14.3 Linguistic Corpora and Corpus Linguistics

A Glossary of Corpus Linguistics (Baker et al. 2006, p. 48) defines corpus as

> **corpus** The word *corpus* is Latin for body (plural *corpora*). In linguistics, a corpus is a collection of texts (a 'body' of language) stored in an electronic database. Corpora are usually large bodies of machine-readable text containing thousands or millions of words. A corpus is different from an archive in that often (but not always) the texts have been selected so that they can be said to be representative of a particular language variety or genre, therefore acting as a standard reference. Corpora are often annotated with additional information such as part-of-speech tags or to denote prosodic features associated with speech. Individual texts within a corpus usually receive some form of meta-encoding in a header, giving information about their genre, the author, date and place of publication, etc.

A review of other existing definitions suggests that there are rarely disagreements among researchers in the field. This consensus is captured by McEnery et al. (2006, p. 5):

> . . . a corpus is a collection of (1) machine readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety.

In the same *Glossary*, (Baker et al. 2006, p. 50) defines corpus linguistics as

> **corpus linguistics** A scholarly enterprise concerned with the compilation and analysis of corpora (Kennedy 1998, p. 1). According to McEnery and Wilson (1996, p. 1) it is the 'study of language based on examples of "real life" language use' and 'a methodology rather than an aspect of language requiring explanation or description.'

While linguists do not diverge in defining corpus, they disagree in defining the field itself. With respect to the corpora themselves, the arguments commonly concern the typology of corpora, the methods by which they are designed and constructed, and the extent to which they should meet the now-standard criteria to count as linguistic corpora. The major disagreement concerns the very nature of the field. Put simply, a group of corpus linguists conceptualize their enterprise as a "methodology" in doing any type of linguistic analysis in which corpus tools provide special qualitative and quantitative methods for the questions at hand.

For another group of linguists, the so-called "neo-Firthians," corpus linguistics is a "theory". A neo-Firthian corpus linguist asserts that corpus linguistics is "a theoretical approach to the study of language" (Teubert 2005).

The ever-increasing use of corpora in linguistic research introduced new concepts and methods that helped uncover many aspects of language structure and language use, which ultimately lead to new theories of language. In a recent introduction to the field, (McEnery and Hardie 2012) argue that corpus linguistics "... is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon a set of procedures, or methods, for studying language." Accordingly, they also argue that

> The procedures themselves are still developing, and remain an unclearly delineated set—though some of them, such as concordancing, are well established and are viewed as central to the approach. Given these procedures, we can take a corpus-based approach to many areas of linguistics. ... it may refine and redefine a range of theories of language. It may also enable us to use theories of language which were at best difficult to explore prior to the development of corpora of suitable size and machines of sufficient power to exploit them. Importantly, the development of corpus linguistics has also spawned, or at least facilitated the exploration of, new theories of language—theories which draw their inspiration from attested language use and the findings drawn from it (McEnery and Hardie 2012, p. 1).

A linguistic corpus is designed by a set of *external* and *internal* criteria. External criteria (*situational*) relate to the selection of texts on the bases of registers, genres, and time span, among others. Internal criteria (*linguistic*) are concerned with the distribution of linguistic features across texts that make up the corpus. It is evident that external criteria do not take into account the linguistic characteristics. Internal criteria, on the other hand, present a problematic situation in which a corpus builder decides in advance which linguistic features are to be represented in the corpus. However, it is helpful in selecting text types with different linguistic features to be added next to the corpus.

The defining features that stand out as the most significant in measuring a corpus as a reliable source for linguistic analysis are *representativeness* and *balance*. In an earlier study on corpus representativeness, Biber (1993, p. 243) explains the standards:

> Some of the first considerations in constructing a corpus concern the overall design: for example, the *kinds* of texts included, the *number* of texts, the *selection* of particular texts, the selection of text samples from within texts, and the *length* of text samples. Each of these involves a sampling decision, either conscious or not. [emphasis added]

Representativeness is a much-debated feature that sets a linguistic corpus apart from an archive or collection of texts. In other words, representativeness makes a corpus a reliable source for any linguistic analysis to derive valid conclusions on language structure or use. Despite its importance in corpus design, there exists little agreement about representativeness. Leech (2007) indicates that for some researchers, if a corpus lacks representativeness, any conclusion derived from such a corpus will be confined to that particular corpus only, and cannot be extended or generalized to language.

Balance is another hard-to-define requirement for linguistic corpora. Leech (2007, pp. 136–138) indicates

> An obvious way forward is to say that a corpus is 'balanced' when the size of its subcorpora (representing particular genres or registers) is proportional to the relative frequency of occurrence of those genres in the language's textual universe as a whole. In other words, balancedness equates with proportionality. . . . There is one rule of thumb that few are likely to dissent from. It is that in general, the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be.

It is expected that a balanced corpus covers as much variety of text categories as possible to represent the language. At present, there is no concrete measure to judge the balance of a corpus other than informed and intuitive judgments. The research interests and their extent determine the type of the corpus to be built. A common typology of corpora include the following:

- *General Corpora*: The driving force in the construction of a general corpus is to produce a *reference corpus* of language use that would be balanced and representative. A general corpus may contain written or spoken texts or may contain texts from both media. The major aim is to represent texts from different genres, domains, and types in a balanced manner so that the conclusions drawn from quantitative and qualitative analyses of corpus data will hold true for language use in general. The British National Corpus (BNC) is one such general reference corpus of modern English having 100 million words and comprising 4048 written texts and ten million words of transcribed spoken data. It is a balanced and representative corpus of modern English as it includes texts sampled from national and regional newspapers and journals, popular and academic books, university essays, e-mail samples, unpublished letters, and reports from different ages, institutions, and readerships. The success of the BNC as a representative and balanced general corpus led others to adopt its basic design principles, including the American National Corpus, the Korean National Corpus, the Polish National Corpus, and recently, the Turkish National Corpus.
- *Specialized Corpora*: Relatively small-sized and specialized in terms of genre or domain, these types of corpora are more varied and available in greater numbers. The current tendency in specialized corpus creation is mostly observed in professional and academic domains. Some representatives of such specialized corpora include the Corpus of Professional Spoken American English (CPSA)[1] and the Michigan Corpus of Academic Spoken English (MICASE).[2] A specialized corpus can also be created by extracting relevant text data from a larger general corpus.
- *Written Corpora*: The Brown Corpus is not only the first corpus, but it is at the same time the first written corpus of English in modern times. The texts that make up the corpus data are collected from written media, sampled from

[1] Athelstan Corpus of Spoken, Professional American-English: www.athel.com/cpsa.html (Accessed Sept. 14, 2017).

[2] quod.lib.umich.edu/m/micase (Accessed Sept. 14, 2017).

15 categories. A counterpart of the Brown Corpus, the Lancaster-Oslo-Bergen Corpus of British English (LOB), is constructed following the same principles of the Brown Corpus, and thus they have collectively come to represent varieties of the same language, providing a reliable means of comparison between two varieties of English. Later, in the early 1990s, the Freiburg-LOB Corpus of British English (FLOB) and the Freiburg-Brown Corpus of American English (Frown) were developed to represent written American and British varieties of English. Furthermore, comparisons of these two Freiburg corpora with previous Brown/LOB corpora revealed data on language change in the time span between the 60s and the 90s.

- *Spoken Corpora*: Compared to general or written corpora, it is harder to construct and annotate the spoken corpora of a language. Only recently, we witnessed an increase in the number of spoken corpora due to improvements in recording technologies and automated transcription software. Pioneering corpora for spoken English were built in the late 1960s, such as the London-Lund Corpus (LLC) (Greenbaum and Svartvik 1990), followed by others, including the Lancaster/IBM Spoken English Corpus (SEC),[3] the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) (Carter and McCarthy 2004), the Santa Barbara Corpus of Spoken American English (SBCSAE) (Du Bois et al. 2005), and the Wellington Corpus of Spoken New Zealand English (WSC) (Holmes et al. 1998). The only existing and linguistically reliable new-generation spoken corpus of Turkish is the Spoken Turkish Corpus (STC) (Ruhi et al. 2010a). The Turkish National Corpus (TNC) (Aksan et al. 2012) also has a spoken component of one million words as a reflection of its adherence to the design principles of the BNC.
- *Synchronic Corpora*: Linguists build synchronic corpora in order to observe language change and language variation in corpus data, primarily for the purpose of providing a "snapshot" of language use at a certain point or period of time. In such corpora, all the texts should be selected from the same time period to account for varieties of the language synchronically present. The International Corpus of English (ICE) is built for the synchronic analysis of the English spoken in Britain, the USA, Australia, Canada, and other countries where English is the first language (Greenbaum 1991). It consists of twenty corpora of one million words each, with samples of both written and spoken English.
- *Diachronic Corpora*: Corpora that are constructed for a linguistic account of language in time commonly contain texts representing language use during different periods of the language under investigation. Given the recent history of sound recording technologies, diachronic corpora represent written language over time, for example, the Helsinki Diachronic Corpus of English Texts (Rissanen et al. 1991)

[3]ICAME Corpus Collection: Information: clu.uni.no/icame/lanspeks.html (Accessed Sept. 14, 2017).

- *Learner Corpora*: Corpus use in the language classroom has found its place in teaching and learning contexts. For example, the International Corpus of Learner English (ICLE) (Granger 2003) and as its sub-corpus Turkish International Corpus of Learner English (TICLE) (Kilimci and Can 2009) have been a source of research in teaching contexts in recent years.
- *Monitor Corpora*: A monitor corpus is different from the (static) others presented here in the sense that it is constantly growing (dynamic) with the addition of new material. The Bank of English (BoE)[4] and Corpus of Contemporary American English (COCA) (Davis 2008) are well-known corpora of this type for English.

14.4 Use of Corpora in Linguistics

A corpus is constructed primarily to represent language use in a balanced manner in order to study language empirically on the basis of real data. The role and function of corpora in linguistic analyses can be viewed from different perspectives, depending on the research questions at hand. Lüdeling and Kytö (2008, p. ix) summarize the use of corpora in linguistic analyses for three major purposes: (1) empirical support, (2) frequency information, and (3) meta-information.

The corpus query tools help researchers in finding examples of real language use that are relevant to their questions, that is what they now have as an example is a citation of actual language use rather than the alternative—a made-up example or a sample derived by chance and most often de-contextualized. Providing evidence for language structure and use from corpora is not limited to a specific level of linguistic analysis but works at all levels, from sound to form and to function. The data in corpora are tagged and annotated and thus provide the exact type of sampling that empirically supports the hypotheses. As a repository of real language samples, a corpus query returns citations of language use that had not been envisaged before. Additionally, the empirical nature of corpora makes it possible to replicate the analysis conducted, which is not possible with data based on introspection.

Citations retrieved from a corpus do not simply represent a particular linguistic manifestation, but also provide quantificational information. The occurrences in the data and the patterns in which they occur also provide evidence for their distribution. Depending on the level of analysis and particular research questions at hand, linguists may derive various conclusions regarding different aspects of natural language use. The frequency information concerning distribution of units and patterns may have practical as well as theoretical implications in linguistics.

The language use captured in linguistic corpora further incorporates "meta" information for its users in terms of major participants or components of a communication event. These include the gender of the participants, their age as well as their dialectical background, the medium of the text and its specific genre, among

[4]Titania, The Bank of English: www.titania.bham.ac.uk (Accessed Sept. 14, 2017).

others, all of which provide significant information to a linguist in an analysis of natural language use in context.

When we narrow down the actual corpus linguistic work conducted over the years, we observe that they cover major areas. Meyer (2004) lists these general areas which further include many other subfields of linguistics: Grammatical studies of specific linguistic constructions, lexicography, language variation, historical linguistics, contrastive analysis and translation theory, natural language processing, language acquisition, and language pedagogy.

The ever-growing number of publications and the appearance of special journals in the field clearly underline the increasing importance of corpora in linguistics. It is evident that linguists with different interests will continue to build and use corpora in the future. As before, contributions from neighboring disciplines like computational linguistics and natural language processing research will continue to play a significant role in the future of corpus linguistics. As observed by Sampson (2013), there is currently a rising trend in linguistic analyses to adopt empirical approaches.

14.5 Turkish Linguistic Corpora

We may argue that there are at least three different kinds of corpora in Turkish today: (1) large-sized general linguistic corpora that are constructed and made available for users with proper corpus tools, (2) small-sized specialized corpora that are constructed for the study of specific research questions and are confined to the builders only, and (3) NLP corpora built with no linguistic criteria in mind, but rather as tools for testing algorithms devised for different applications.

We cannot say Turkish is a well-studied language when compared to other languages, for which there are well-documented histories and grammars. In other words, there exist catalogs of constructions or structures that have been collected and documented; however, the number of grammars or general descriptions of Turkish at different levels of representations are quite limited in number. Most linguistic works in current Turkish studies concentrate on a small number of fields like discourse analysis, pragmatics, or syntax. Rarely do we find works on semantics or lexicology or in any other domains, probably because they require enriched datasets. A well-balanced and representative corpus of Turkish is thus a necessity in studying the language where the accumulated and documented potentials of the language and its representative datasets are relatively small in number.

What may be called preelectronic corpora of Turkish are, in fact, not collections of texts, but rather compilations of lexemes. As early as the tenth century, we find the very first dictionary of Turkic languages, namely, the *Compendium of the Languages of the Turks*,[5] compiled by Mahmud al-Kashgari in 1072. Two major undertakings

[5]Divânu Lügati't-Türk.

of the Turkish Language Institute (TDK) in the early 1930s may also be considered early examples of data compilations. The monumental *Derleme Sözlüğü* (Dictionary of Compilations), motivated by the Turkification of lexis during the early years of the newly founded Republic, aimed at compiling vocabulary from the existing dialects of the time. From printed material, a number of themes were listed and then collectors were recruited from village intelligentsia to record samples of lexis. Initially printed in four volumes, this huge dictionary reached its current twelve volumes over the years 1963–1982. The second dictionary, *Tarama Sözlüğü*, also aimed at finding and revitalizing native lexical stock, was published in eight volumes in 1977. The dictionary compiled lexical items of Turkish origin from about 160 different historical texts starting as early as the thirteenth century. In both cases, however, the linguistic material is not extracted from a specially constructed corpus.

The pioneering work and current studies suggest that the role of data seems to be well appreciated in Turkish linguistic work. Apart from very few theoretical studies, almost all linguistic analyses are empirical and data-based. A typical research in Turkish linguistics gathers a "data base" or a "data set" in the analysis of the question at hand. We may say that there are very small-sized special corpora employed in almost all usage-based empirical research. However, these are severely confined in their form and size, they are not available for other researchers, and the data was collected with a specific problem at hand. Such work does not preprocess the data or use corpus-analytic tools.

Work in computational linguistics in Turkish has a longer history than Turkish corpus linguistic studies. The early beginning of corpus research in Turkish was in fact prompted by NLP research and computational linguistics analyses. In computational linguistics and in NLP, large-scale corpora are constructed for "practical" purposes. In a very reductionist manner, it is possible to say that researchers in these domains built corpora first and foremost to evaluate the algorithms that they had developed and to use corpora as a testing ground.

A comprehensive history of computational linguistics in Turkey has yet to be written; however, there are occasional references to earlier work in the field. The first known electronic corpus for linguistic analysis was constructed by Köksal (1976) for "automatic morphological analysis." Köksal tested and evaluated his algorithm over a corpus of 1534-word text sample randomly selected from daily newspapers. Even at this very early stage, some degree of representativeness and balance was sought: "…materials have been selected from the most important six daily newspapers representing different political views and linguistic trends." (Köksal 1976). Köksal's work recognizes the rich morphology of Turkish and possible morpheme combinations, and also points to major challenges further ahead, noting potential fields of application, urging building larger corpora for automated language analyses.

14.5.1 METU-Turkish Corpus

The first electronic linguistic corpus designed and compiled to represent modern Turkish is the Middle East Technical University (METU) Turkish Corpus. The developers of the METU Turkish Corpus (hereafter MTC) note this fact and state that the basic aim was to design a balanced written corpus on Turkish with the hope that it will prove useful to descriptive and theoretical studies alike (Say et al. 2004).

The MTC is also a mother corpus from which two subcorpora are derived. The first one is a morphologically and syntactically annotated treebank of Turkish, namely, the METU-Sabancı Turkish Treebank (Oflazer et al. 2003) (see also Chap. 13), which contains almost 7260 sentences and 65,000 words, and syntactic annotation is realized in a dependency-based XML-compliant format. The genre distribution in the treebank follows the MTC. The METU-Sabancı Turkish Treebank has served as a significant electronic resource for many studies for a long time (see, e.g., Kırkıcı (2009) for realizations of nominal compounds; Çetinoğlu (2014) for developing morphological disambiguators on the basis of the Turkish Treebank). The METU-Turkish Discourse Bank (METU-TDB) (Zeyrek et al. 2013) (see also Chap. 16), which is the first attempt to develop discourse annotation procedures in Turkish, is the second sub-corpus. In order to build an annotated discourse resource for Turkish, an approximately 400,000-word sub-corpus was extracted from MTC datasets, and discourse connectors (i.e., coordinating conjunctions, subordinating conjunctions, discourse adverbials, and phrasal expressions) were annotated manually, sharing the same principles as the Penn Discourse Treebank (Zeyrek et al. 2009). The METU-TDB project has so far developed the sub-corpus, the annotation tool, and the TDB query browser as its products that are freely distributed to academic users.[6]

In introducing the design decisions and principles of the MTC and the processes that led to its construction, the builders are not only confronted with issues facing "trailblazers" in general, but also are faced with many standard problems that corpus builders had to tolerate during construction. The constant reference to "limited resources" by the builders in presenting their construction process and its effects on the final product can be observed in a number of places as we will note below.

The MTC is a two million-word general corpus, composed of post-1990 written texts representing different genres. It includes texts from ten different genres and consists of 520 sample texts from 291 different sources (Table 14.1). The corpus does not have a spoken component, the lack of which is explained by the limitation of resources and experience required to process spoken language data at the time of the design process (Say et al. 2004).

As for representativeness of the corpus, the developers suggest that they preferred an "opportunistic" approach. It appears that within the severely limited prospects of accessing and digitizing the data sources (restricted permissions granted by the

[6] www.medid.ii.metu.edu.tr/ (Accessed Sept. 14, 2017).

Table 14.1 Genre distribution of the MTC (Say 2006) (reprinted with permission)

Genre	%
News	42
Novels	13
Stories	11
Articles	8
Op-ed columns	8
Essays	7
Research reports-surveys	5
Others (e.g., memoirs, course books)	3
Travel essays	2
Interviews	1
Total	100

publishers at the time and limited resources in terms of budget and workforce), the developers collected samples of electronic texts mainly from daily newspapers in the form of news and opinion columns. They were, however, very careful to maintain balance by selecting texts with no bias toward a particular genre or a writer. The corpus consists of texts dated between 1990 and 2002.

MTC is tagged by XCES style annotation using special software developed by the members of the project group as well as its corpus query workbench. A graphic-based browser, aimed at ordinary users of the corpus with its user-friendly features, was developed to be multi-platform compatible (see Özge and Say (2004) for a detailed description of the corpus workbench). The MTC remains today the only linguistically sound, freely distributed written corpus of modern Turkish.

From today's perspective and taking into account recent advances in corpus linguistics, the MTC is a less adequate source in meeting the demands of linguistic research. As of today, any general reference corpus is expected to be no less than 50 million words in size (Teubert and Cermakova 2004, p. 67). The defining aspects of balance and representativeness, as they have been discussed in recent years, became more and more important in evaluating a reference corpus as a reliable data source in the analyses of patterns emerging in language use in different genres, in varied contexts, and by users of different ages and genders, among many others. Even though the internal balance of the MTC is maintained to a certain degree, almost half of the corpus consists of texts from newspapers (single medium) and represents mainly news and columns (limited genres); therefore, its overall balance and representativeness fall short in meeting the standards set for current linguistic corpora. As emphasized by Lew (2009), the text types most commonly overrepresented in reference corpora are newspaper archives and fictions. In the MTC, as indicated above, newspapers as a text type are overrepresented.

It is evident that despite technological advances in capturing data via sophisticated scanning tools and software, an increase in digitization capacities, the ease of finding texts in corpus construction, and common data management in building processes, corpus development is still a very laborious undertaking. The developers

of the MTC should be considered as forerunners who have successfully achieved their goals in the face of huge limitations in the resources allocated.

The number of linguistic analyses taking the MTC (also other corpora derived from it) as the major resource grew rapidly in the years following its construction. It has proven its usefulness and still continues to do so for researchers, as a wealth of studies (numerous graduate dissertations and academic articles) in NLP and linguistics make use of the MTC in their analyses (to name a few, see for example, Kawaguchi (2005) for the analysis of participle and infinitive nominalizations; Karaoğlan et al. (2013) for testing metrics in corpus normalization). Given that the MTC is a written corpus with no spoken component and its limitations in extracting quantitative outputs, linguistic studies conducted over the data clustered mostly in the fields of semantics, pragmatics (Ruhi 2009), and language acquisition (Sofu and Altan 2009). Most of these studies simply use the MTC as a naturally-occurring database of Turkish to obtain either sample extracts or frequency counts of linguistic items to validate their hypotheses. There is hardly a linguistic study (e.g., Işık-Güler and Ruhi 2010; Zeyrek 2012) that follows quantitative methods of corpus linguistics and exploits the MTC to describe any issue in Turkish linguistics thoroughly on the basis of a corpus-driven or corpus-based approach.

14.5.2 Turkish National Corpus (TNC)

In the years following the construction of the MTC, the need for a large-scale general reference corpus of Turkish has become more obvious. To meet the challenge, a group of linguists at Mersin University decided to build a reference corpus of Turkish.[7] The project team followed right from the start the so-called best practices at all stages of corpus development. The end product is the Turkish National Corpus (TNC),[8] a well-balanced, representative, and large-scale (50 million words) free resource of a general-purpose corpus of contemporary Turkish.

The design decisions in the construction of the TNC benefited entirely from previous practices. Major design principles were adopted from the experiences of the British National Corpus (BNC) with minor modifications. The simple idea was to follow the BNC model in constructing a linguistic corpus that would represent the language in a well-balanced manner. Considering the labor-intensive nature of the corpus construction task and limitations on reaching and finding relevant data sources, the size of the corpus was decided to be reduced to half of the BNC size where the distribution of the corpus content is proportionally preserved. The number of words in the corpus is distributed proportionally for each medium, time span, and text domain. Since the BNC is commonly accepted as a balanced corpus, many

[7]This was supported by the The Scientific and Technological Research Council of Turkey (TÜBİTAK) (Grant no: 108K242).

[8]www.tnc.org.tr (Accessed Sept. 14, 2017).

Table 14.2 Composition of the written component of the TNC (Aksan et al. 2012) (reprinted with permission)

Domain	%	Medium	%
Fiction	19	Books	58
Social sciences	16	Periodicals	32
Art	7	Misc. published	5
Commerce-Finance	8	Misc. unpublished	3
Op-ed pieces	4	Spoken texts[a]	2
World affairs	20		
Applied sciences	8		
Natural sciences	4		
Leisure writing	14		

[a] Material that is written to be spoken, such as political speech, news broadcasts, etc.

other currently available large-sized reference corpora (e.g., the *American National Corpus*, the *Korean National Corpus*, and the *Polish National Corpus*) also adopt the BNC model to achieve balance and representativeness (McEnery et al. 2006, p. 17).

The selection of texts is based on three criteria: text domain, time, and medium. Put simply, the imaginative domain includes mainly works of fiction (novels, short stories, poems, drama) and the informative domain includes texts representing the sciences, the arts, commerce-finance, belief-thought, world affairs, and leisure. Imaginative texts constitute 19% and informative texts 81% of the TNC, following the distribution adopted in the BNC.

The time span of the texts in the TNC covers a 20-year period between 1990 and 2010. The distribution of sample texts from each medium and domain with respect to years in the period is also carefully calculated (Table 14.2). As for matters of size, the time period covered was also decided on the basis of the volume of publications produced in Turkish and consumed by language users in different genres and text types (see Aksan et al. (2012) for more details of text type choices according to the domains and mediums).

The spoken component of the TNC is composed of orthographic transcriptions of spoken language compiled from formal and informal communicative settings. These include spontaneous, everyday conversations on a variety of topics by users of different ages and genders, and samples of spoken communicative events collected from meetings, lectures, and speeches. A total of one million words in the spoken component represent 2% of the TNC.

Morphological analysis and part-of-speech annotation of the TNC has been done by developing an NLP dictionary based on the NooJ_TR module (Aksan and Mersinli 2011). The unique semiautomatic process of developing the NLP dictionary includes the following steps: (1) automatically annotating the type list with the NooJ_TR module, which follows a root-driven, non-stochastic rule-based approach to annotating the morphemes of the given types by using a graph-based finite-state transducer; (2) manually checking and revising the output and eliminating artificial ambiguities and non-occurring, theoretically possible multi-tags.

The TNC lexicon files containing linguistically motivated tag sets were constructed from scratch. Optimization of the NLP dictionary was conducted manually. Unlike previous studies, the remaining ambiguities do not contain artificial ambiguities and thus serve as a good basis for their documentation (Aksan et al. 2012). Unlike the available taggers, the resulting TNC tagger does not include artificial or theoretically possible but non-occurring ambiguities. Additionally, the number of affixes and the assigned tags for them are all valid in terms of current linguistics literature.

The TNC has a platform-independent, user-friendly Web-based user interface for making queries. It provides for multitude of features for the analysis of corpus texts including concordance display, sorting concordance data, creating descriptive statistics for query results over the language-external restriction categories of texts via distribution, and compiling lists of collocates for node words on the basis of several statistical methods. With 48 million words, the TNC-Demo Version represents 4438 different data sources over 9 domains and 34 different genres, and was published as a free resource for noncommercial use in October 2012. The morphologically annotated, complete version of the TNC v3.0 is planned for release in 2018, offering new query options for linguistic analyses.

The number of users and the number of studies using the TNC as the major electronic resource is increasing. While some of the studies use the TNC for compiling naturally-occurring language evidence and for hypothesis-testing (e.g., Sebzecioğlu 2013; Akşehirli 2014), there are still others following a corpus-driven approach that attempts to build hypotheses and describe Turkish on the basis of the TNC (see, e.g., (Erköse and Uçar 2014) for the cognitive semantic analysis of posture verbs in TNC). Since the TNC is a linguistic corpus, and because it is well-balanced and representative, the conclusions based on TNC data provide valid linguistic descriptions of Turkish, both qualitatively and quantitatively. For example, for the first time in Turkish linguistics, we are able to account for patterns of language use that would give hints for formulaicity in Turkish (see Uçar and Kurtoğlu (2012) for semantic patterning of polysemous verbs; Aksan and Aksan (2013) for genre specification through multiword units). It is now possible to derive frequency information of Turkish lexical items and affixes (Aksan and Yaldır 2012; Aksan and Aksan 2014) as well as multiword units (Aksan and Aksan 2012).

14.5.3 Spoken Turkish Corpus (STC)

The Spoken Turkish Corpus (STC) is the only corpus of its kind that is available for linguistic analyses. Given that the challenges faced by builders of spoken corpora are demanding and that they require a different set of measures for the creation of the resource, maintenance, and dissemination (see e.g., Ruhi et al. (2014) for recent debates on best practices for spoken corpora in linguistic research), the STC is a pioneering work undertaken to create and sustain a multimodal spoken corpus that overcomes most of these challenges in order to be published in its demo version. It

is also the product of a team of linguists at METU, constructed with contributions from international collaborators.

The STC is the first general-purpose, large-scale corpus of present-day spoken Turkish. The ultimate aim is to reach the size of ten million words, so the corpus is designed accordingly. Ruhi (2011) states that the raw database of the STC currently contains three million words of audio and video recordings from a variety of geographical and social settings and domains. About 440,000 words of these recordings are under transcription control, with partial morphological and speech act annotation processing in the corpus management system. The STC Demo Version consists of 23 communications and represents 2.4 h of interaction, with 18,357 tokens having been published. It is freely available for nonprofit research purposes.[9] Since the STC is a multimodal corpus, the transcriptions are presented in a time-aligned manner with audio and video files. It uses EXMARaLDA (Extensible Markup Language for Discourse Analysis), an open-source system of data models, formats, and tools for the production and analysis of spoken language corpora (Schmidt 2004).[10] Transcriptions are created with EXMARaLDA's Partitur Editor. The project team adapted a revised form of HIAT for the transcriptions (Ruhi et al. 2010b). The partial morphological analysis of the STC data is done with TRmorph (Çöltekin 2010), and the annotation of requestive/directive speech acts is implemented with Sextant (Wörner 2009) (see Ruhi et al. (2011) and Ruhi (2014) for retrieving requestive/directive speech acts). The final aim is to create a spoken resource annotated for morphology, the socio-pragmatic features of Turkish (e.g., address terms, [im]politeness markers, and a selection of speech act realizations), anaphora, and gestures (Ruhi et al. 2010b).

Among its notable features, the STC's pragmatically informed metadata fields make the sociocultural situatedness of communication visible to researchers. While determining the metadata features, the STC has scrutinized and considered the text classification and other metadata parameters proposed in standardization schemes and features implemented in other spoken corpora (e.g., the BNC). At the same time, in order to achieve pragmatically more fine-grained text descriptors, the STC implements a two-layered scheme regarding text type and discourse content.

On the first level, texts are classified according to speaker relations and the major social activity type. The domains for speaker relations are family/relatives, friend, family-friend, educational, service encounter, workplace, media discourse, legal, political, public, research, brief encounter, and unclassified conversations (Ruhi et al. 2010b). These domains are then subclassified according to activities. The class of workplace discourse includes, for instance, meetings, workplace cultural events (e.g., parties), business appointments, business interviews, business dinners, shoptalk, telephone conversations, and chats.

The second layer of metadata annotation is implemented at the corpus assignment stage and involves the annotation of speech acts based on Searle (1975) (e.g.,

[9] std.metu.edu.tr (Accessed Sept. 14, 2017).

[10] exmaralda.org (Accessed Sept. 14, 2017).

Table 14.3 Distribution of domains planned for the STC

Domain	%
Conversations among family and/or relatives	25
Workplace conversations	20
Education	15
Broadcasts	15
Conversations among friends and/or acquaintances	12
Service encounter	5
Natural sciences	4
Other	4

offers and requests), on the one hand, and, on the other hand, the annotation of conversational topics (e.g., child care), speech events (e.g., troubles talk,[11]), and ongoing activities (e.g., cooking)—all encoded under the super metadata category, Topic, in the current state of STC. Speech act and Topic annotation are thus two further metadata parameters in STC (Ruhi et al. 2012).

It is possible to overview the content of the corpus in terms of text categories and the distribution of gender and age at the website of the STC and in its demo version. Table 14.3 displays the STC domains and the planned proportion of the samples from them.[12]

With the publication of the STC Demo version, spoken Turkish discourse has been investigated from different perspectives. The *Journal of Linguistics and Literature* published a special issue on corpus-based analysis of interactional markers (e.g., *tamam* 'okay,' *şey* 'thing,' *hayır* 'no') in the demo version and a selection of the publishable version of the STC (Ruhi 2013). The studies in the collection highlight the significance of "corpus-based perspective to analyzing spoken Turkish and to explore the affective dimension of a number of markers especially in regard to relational management in the tradition of (im)politeness theories" (Ruhi 2013, p. 2). Since the STC consists of data collated from a relatively wide range of domains and genres, the articles explore the pragmatic functions of a number of interaction markers in these domains and genres, and thus they display a depth of discourse domains in the analysis of spoken Turkish. Another comprehensive study, Çelebi (2014) aims to develop a methodological framework to analyze impoliteness in a corpus-driven approach. To attain this goal, the study investigates the STC demo and its publishable data thoroughly by emphasizing the empirical and explanatory power of a corpus approach in pragmatics studies. Lastly, the STC demo version is

[11]Tannen defines troubles-talk as a conversational event where interlocutors "share their moments of frustration and irritation, but without expecting a solution"—see The Art of Talking and Listening (Philosophy on the Mesa, November 22, 2010): philosophyonthemesa.com/tag/deborah-tannen/ (Accessed Sept. 14, 2017).

[12]See Spoken Turkish Corpus. Main Features of STC Demo Version: std.metu.edu.tr/en/main-features-of-stc-demo-version (Accessed Sept. 14, 2017).

also utilized to annotate explicit discourse connectives of spoken Turkish in line with the Turkish Discourse Bank's style of annotation (Demirşahin and Zeyrek 2014).

It is worth mentioning another attempt to construct a spoken corpus of Turkish. As a product of two research projects conducted at the Institute of Global Studies and Tokyo University of Foreign Studies,[13] a Corpus of Spoken Turkish containing 514,400 tokens compiled from free conversations on a variety of topics is published and distributed freely for academic research purposes.[14]

This second kind of corpora that we have noted above are the small-size specialized corpora or datasets, each designed for the study of a specific problem identified by the researcher/builder. The existence of such corpora can only be discovered when a particular study appears in publication, announcing the results of the analysis based on a special corpora built for that particular problem only. This is a more common practice in discourse analysis (see e.g., Özyıldırım (2010) for genre analysis on a 160,000-word corpus; Oktar and Cem-Değer (1999) for a critical discourse analysis on 15 newspaper articles) or pragmatics studies where the researcher gathers data either for citing natural language use that would provide evidence for a particular type of a text or speech act (see e.g., Ruhi (2006) for politeness in compliment responses on a spoken Turkish dataset) or to document context-specific preferences in confined contexts of use (see e.g., Çubukçu (2005) for constructive back-channels in 30 Turkish conversations recorded during everyday conversations, business, and formal discussions). There are also small-size sub-corpora that are extracted from the datasets of already existing larger corpora. For instance, the spoken sub-corpus of the TNC containing private and public speeches and conversations is used to investigate discourse analytic and corpus-driven features of requests (Aksan and Mersinli 2015) and thanking (Aksan and Demirhan 2015) speech acts in Turkish.

In addition to the major linguistic corpora we have reviewed above, there are also specialized corpora, as we have noted previously. These are constructed to serve as a comprehensive resource for the particularly specified aims of the researchers. Uçar (2014) built a 713,000-word corpus of the popular comedy show *Komedi Dükkanı* (Comedy Shop) to analyze the semantic and pragmatic properties of conversational humor in Turkish (see also Uçar and Yıldız 2015). To examine lexico-grammatical differences and similarities in predicate uses among disciplinary discourses, Yıldız and Aksan (2014) compiled data from the introduction and conclusion sections of 1178 scientific articles published in the humanities, applied sciences, and basic sciences, and built a one- million-word specialized corpus of Turkish scientific text. Similarly, Uzun et al. (2014) conducted their rhetoric structure analysis on a one-million-word corpus of social science academic articles obtained from the Social Science Database of TÜBA ULAKBİM. Here, we should note that these corpora

[13]The twenty-first COE Program "Usage-Based Linguistic Informatics" 2002–2006 and the Global COE Program "Corpus-based Linguistics and Language Education" (2007–2011).

[14]Global COE Program, Corpus-based Linguistics and Language Education: cblle.tufs.ac.jp/en/ (Accessed Sept. 14, 2017).

are not available for other users and do not provide any interface for access. They solely provide linguistically significant outcomes for their specialized domains.

The NLP corpora in Turkish easily outnumber the available linguistic corpora.[15] As we have noted above, a corpus linguistic analysis of Turkish in fact was initiated by the work of NLP researchers. Such corpora cannot be defined as "linguistic" corpora and can by no means function as a representative and a well-balanced resource for linguistic analyses. The main reason why this web-harvested collection of texts is not considered linguistically significant corpora is that they lack design principles or a rationale (Wynne 2005) in their creation. The following points specify the results of this shortcoming on the basis of the principles of corpus design:

- They are not representative and balanced in terms of the text samples they contain. A representative and balanced sample of written and/or spoken texts is compiled in a linguistic corpus, and, thus, observations on linguistic behavior of queried items on this corpus constitute both quantitative and qualitative linguistic findings. These findings lead linguists to make generalizations on typical and central properties of that language overall (see Hoffmann et al. 2008). Otherwise, "without representativeness whatever is found to be true of a corpus, is simply true of that corpus—and cannot be extended to anything else" (Leech 2007, p. 135).
- They are not designed and constructed to meet the external criteria (e.g., domain, genre, date of sample texts) of the corpus- creating process. As a result of this, most of them do not carry any metadata information and thus the content of the corpora is not transparent pertaining to documentation. As underscored by Sinclair (2005), the proper stance of corpus compiler is "to be detailed and honest about the contents. From their description of the corpus, the research community can judge how far to trust their results, and future users of the same corpus can estimate its reliability for their purposes" (p. 98).
- Most of them are not available for public use. Even if they are publicly available as datasets (see e.g., Ferraresi et al. (2008) for English ukWaC; Sak et al. (2011) for Turkish BOUNCorpus. Yıldız University provides a variety of Turkish datasets containing Turkish tweets, blogs, poems, etc.[16]), linguists are not able to utilize them as a language resource for their studies since these corpora are not published with user interfaces to process the sample texts they contain and to conduct corpus queries on them.

Obtaining Web content and processing it as an offline, static corpus is described as Web for Corpus (de Schryver 2002). In line with this approach, The

[15] In this chapter, we have strictly confined ourselves to corpora constructed following basic design principles that define the products as corpora in the true sense of the term. There are a number of corpora, some of which are even publicly available; however, they neither provide information regarding their design criteria nor follow the general guidelines of legal issues in corpus construction. Such corpora will not be reviewed here.

[16] www.kemik.yildiz.edu.tr/?id=28 (Accessed Sept. 14, 2017).

BOUNCorpus, constructed to exploit Turkish morphology in natural language processing applications, is the largest web-crawled corpus containing 500 million words. It is composed of NewsCor, which contains texts from three major news portals in Turkish, and GenCor, which includes texts from a general sampling of Turkish Web pages. The corpus is encoded by following the XML Corpus Encoding Standard, XCES[17], and is freely available as a language resource (Sak et al. 2011). Compared to the BOUNCorpus, the relatively small size TurCo is a 50-million-word corpus with 90.40% of it compiled from ten different sites with Turkish content. It is widely used to investigate lexical statistical properties of Turkish (Dalkılıç and Çebi 2002) and to test Turkish word *n*-gram analysis algorithms (Çebi and Dalkılıç 2004).

Along with these web-derived datasets, the 42-million-word TurkishWaC (Ambati et al. 2012), containing texts from Wikipedia entries and built by employing the Corpus Factory Method (Kilgarriff et al. 2010), is accessible through the commercial corpus query tool Sketch Engine.[18] The tool is a web-based program and works on corpora of any language with tokenized, lemmatized, and POS-tagged content. It offers a number of language-analysis functions among which the most significant are concordance outputs and word sketches summarizing the grammatical and collocational behavior of the query items.

It should be noted that the Web is also accessed directly via Internet-based search engines as a dynamic corpus itself and freely available tools like WebCorp[19] (Renouf et al. 2007), providing users options to utilize the Web as a corpus through commercial search engines. WebCorp is developed for studying language on the Web, and in this respect, searches can be performed to find words or phrases, including pattern matching, wildcards, and part-of-speech. Results are given as concordance lines in KWIC format. Post-search analyses are possible, including time series, collocation tables, sorting, and summaries of meta-data from the matched web pages.

14.6 Conclusions

In this short review, we presented the basics of linguistic corpora, and efforts in Turkey in developing different types of linguistic corpora. Still in its infancy, Turkish corpus linguistics is "practical, pragmatic, and opportunistic." The coming years will bring more sophisticated products, tools of analyses, and linguistic research. A thorough evaluation of the current state of research on language technologies on Turkish was previously presented in the final report of a workshop organized

[17]Vassar College, Department of Computer Science, NY, USA: www.xces.org (Accessed Sept. 14, 2017).

[18]Lexical Computing CZ s.r.o.: www.sketchengine.co.uk (Accessed Sept. 14, 2017).

[19]Birmingham City University, Research and Development Unit for English Studies: www.webcorp.org.uk/ (Accessed Sept. 14, 2017).

by the Foundation of the National Speech and Language Technologies Platform in October 2011 (Doğan 2011). Among others, in a separate questionnaire, the participants were asked to evaluate "status of tools and resources for Turkish." On a scale of 0–6 points, "reference corpora" received 1.9 for quantity and 2.9 for quality. The other types of corpora also were assigned scores in the same questionnaire, including treebanks, semantic corpora, discourse corpora, parallel corpora, and speech corpora, and they did not fare much better than the reference corpora. The expert participants, some of whom were corpus builders themselves, agreed to score available corpora below average with respect to measuring criteria. It is no surprise that the final report places the insufficiency of data sources and corpora to the very top of the list of negatively evaluated aspects of the field. We believe that, when asked, the evaluation of the present state of corpus studies would score the same by linguists as well.

References

Aksan M, Aksan Y (2012) Multi-word units in informative and imaginative domains. In: Proceedings of the international conference on Turkish Linguistics, Ankara

Aksan M, Aksan Y (2013) Multi-word units and pragmatic functions in genre specification. In: Proceedings of the international pragmatics conference, New Delhi, pp 239–240

Aksan Y, Aksan M (2014) Frequency effects in Turkish: a study on multi-word units. In: Proceedings of the international conference on Turkish Linguistics, Rouen

Aksan Y, Demirhan UU (2015) Expressions of gratitude in the Turkish National Corpus. In: Ruhi Ş, Aksan Y (eds) Exploring (im)politeness in specialized and general corpora: converging methodologies and analytic procedures. Cambridge Scholars, Newcastle upon Tyne, pp 121–172

Aksan M, Mersinli Ü (2011) A corpus-based Nooj module for Turkish. In: Proceedings of the Nooj 2010 international conference and workshop, Komotini, pp 29–39

Aksan M, Mersinli Ü (2015) Retrieving and analyzing requestive forms: evidence from the Turkish National Corpus. In: Ruhi Ş, Aksan Y (eds) Exploring (im)politeness in specialized and general corpora: converging methodologies and analytic procedures. Cambridge Scholars, Newcastle upon Tyne, pp 173–220

Aksan Y, Yaldır Y (2012) A corpus-based frequency list of Turkish: evidence from the Turkish National Corpus. In: Proceedings of the international conference on Turkish linguistics. Gold Press Nyomda Kft, Szeged, pp 47–58

Aksan Y, Aksan M, Koltuksuz A, Sezer T, Mersinli Ü, Demirhan UU, Yılmazer H, Kurtoğlu Ö, Öz S, Yıldız İ (2012) Construction of the Turkish National Corpus (TNC). In: Proceedings of LREC, Istanbul, pp 3223–3227

Akşehirli S (2014) Dereceli karşıt anlamlılarda belirtisizlik ve ölçek yapısı. J. Lang. Linguist. Stud. 10:49–66

Ambati BR, Reddy S, Kilgarriff A (2012) Word sketches for Turkish. In: Proceedings of LREC, Istanbul, pp 2945–2950

Baker P, Hardie A, McEnery T (2006) A glossary of corpus linguistics. Edinburgh University Press, Edinburgh

Barlow M (2011) Corpus linguistics and theoretical linguistics. Int J Corpus Linguist 16:3–44

Biber D (1993) Representativeness in corpus design. Lit Linguist Comput 8:243–257

Carter R, McCarthy M (2004) Talking, creating: interactional language, creativity, and context. Appl Linguist 25(1):62–88

Çebi Y, Dalkılıç G (2004) Turkish word *n*-gram analyzing algorithms for a large-scale Turkish corpus – TurCo. In: Proceedings of international conference on information technology: coding and computing, Las Vegas, NV, pp 236–240

Çelebi H (2014) Impoliteness in corpora: a comparative analysis of British English and spoken Turkish. Equinox, London

Çetinoğlu Ö (2014) Turkish treebank as a gold standard for morphological disambiguation and its influence on parsing. In: Proceedings of LREC, Reykjavík, pp 3360–3365

Çöltekin Ç (2010) A freely available morphological analyzer for Turkish. In: Proceedings of LREC, Valetta, pp 820–827

Çubukçu H (2005) Karşılıklı konuşmada destekleyici geri bildirim. In: Ergenç İ (ed) Dilbilim İncelemeleri. Doğan Yayıncılık, Ankara, pp 289–305

Dalkılıç G, Çebi Y (2002) A 300MB Turkish corpus and word analysis. In: Proceedings of the conference on advances in information systems. LNCS, vol 2547. Springer, Berlin, pp 205–212

Davis M (2008) The 385+ million word corpus of American English (1990–2008+): design, architecture and linguistic insights. Int J Corpus Linguist 14(2):159–190

Demirşahin I, Zeyrek D (2014) Annotating discourse connectives in spoken Turkish. In: Proceedings of the linguistic annotation workshop, Dublin, pp 105–109

de Schryver G (2002) Web for/as corpus: a perspective for the African languages. Nord J Afr Stud 11:266–282

Doğan M (ed) (2011) Multisaund: Ulusal konuşma ve dil teknolojileri platformu kuruluşu ve Türkçede mevcut durum çalıştayı bildirileri, TÜBİTAK-BİLGEM, Gebze

Du Bois J, Chafe W, Meyer C, Thompson S, Englebretson R, Martey N (2005) Santa Barbara corpus of spoken American English, Parts 1–4, Philadelphia, PA

Erköse Y, Uçar A (2014) Türkçedeki dur- konumlama eyleminin derlem temelli bilişsel anlam çözümlemesi. In: Proceedings of the national linguistics conference, Hacettepe University, Kemer, pp 351–358

Ferraresi A, Zanchetta E, Baroni M, Bernardini S (2008) Introducing and evaluating ukWaC, a very large web-derived corpus of English. In: Proceedings of the workshop on web as corpus workshop – Can we beat Google? Marrakech, Morocco

Francis W, Kučera H (1964) A standard corpus of present-day edited American English, for use with digital computers. Brown University, Providence, RI

Granger S (2003) The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. TESOL Q 37(3):538–546

Greenbaum S (1991) The development of international corpus of English. In: Aijmer K, Altenberg B (eds) English corpus linguistics. Studies in honour of Jan Svartvik. Longman, London, pp 83–91

Greenbaum S, Svartvik J (1990) The London-Lund Corpus of Spoken English. In: Svartvik J (ed) The London-Lund corpus of spoken English: description and research. Lund University Press, Lund, pp 11–45

Hoffmann S, Evert S, Smith N, Lee D, Prytz YB (2008) Corpus linguistics with BNCweb: a practical guide. Peter Lang, Frankfurt

Holmes J, Vine B, Johnson G (1998) Guide to the Wellington corpus of spoken New Zealand English. University of Wellington Press, Wellington

Işık-Güler H, Ruhi Ş Ruhi (2010) Face and impoliteness at the intersection with emotions: a corpus-based study in Turkish. Intercult Pragmat 7:625–660

Karaoğlan B, Dinçer BT, Kışla T, Kumova-Metin S (2013) Derlem normalizasyonu için bir öneri. In: Proceedings of IEEE signal processing and communications applications conference, Magosa

Kawaguchi Y (2005) Two Turkish clause linkages: –DIK- and –mE-: a pilot analysis based on the METU Turkish corpus. In: Takagaki T, Zaima S, Tsuruga Y, Moreno-Fernandez F, Kawaguchi Y (eds) Corpus-based approaches to sentence structures. John Benjamins, Amsterdam, pp 151–177

Kennedy G (1998) An introduction to corpus linguistics. Longman, London

Kilgarriff A, Reddy S, Pomikalek J, Avinesh PVS (2010) A corpus factory for many languages. In: Proceedings of LREC, Valletta, pp 904–910
Kilimci A, Can C (2009) TICLE: Uluslararası Türk Öğrenci İngilizcesi Derlemi. In: Sarıca M, Sarıca N (eds) Proceedings of the national linguistics conference, Yüzüncü Yıl Üniversitesi, Van, pp 1–11
Kırkıcı B (2009) İmparator çizelgesi vs. imparatorlar çizelgesi: on the (non)-use of plural non-head nouns in Turkish nominal compounding. Dilbilim Araştırmaları Dergisi 1:35–53
Köksal A (1976) A first approach to a computerized model for the automatic morphological analysis of Turkish. PhD thesis, Hacettepe University, Ankara
Leech G (1992) Corpora and theories of linguistic performance. In: Svartvik J (ed) Directions in corpus linguistics. Mouton de Gruyter, Berlin, pp 105–122
Leech G (2007) New resources, or just better old ones? The holy grail of representativeness. In: Hundt M, Nesselhauf N, Biewer C (eds) Corpus linguistics and the web. Rodopi, Amsterdam, pp 133–149
Leech G (2011) Principles and applications of corpus linguistics: interview with Geoffrey Leech. In: V V, Zyngier S, Barnbrook G (eds) Perspectives on corpus linguistics. John Benjamins, Amsterdam, pp 155–170
Lew R (2009) The web as corpus versus traditional corpora: their relative utility for linguists and language learners. In: Baker P (ed) Contemporary corpus linguistics. Continuum, London, pp 289–300
Lüdeling A, Kytö M (2008) Introduction. In: Lüdeling A KM (ed) Corpus linguistics: an international handbook. Walter de Gruyter, Berlin, pp v–xii
McEnery T, Hardie A (2012) Corpus linguistics: method, theory and practice. Cambridge University Press, Cambridge
McEnery T, Wilson A (1996) Corpus linguistics. Edinburgh University Press, Edinburgh
McEnery T, Xiao R, Tono Y (2006) Corpus-based language studies. Routledge, London
Meyer C (2004) English corpus linguistics: an introduction. Cambridge University Press, Cambridge
Newmeyer F (1986) Linguistic theory in America. Academic, London
Oflazer K, Say B, Hakkani-Tür DZ, Tür G (2003) Building a Turkish Treebank. In: Treebanks: building and using parsed corpora. Kluwer Academic, Berlin
Oktar L, Cem-Değer A (1999) Gazete söyleminde kiplik ve işlevleri. Dilbilim Araştırmaları Dergisi, pp 45–53
Özge U, Say B (2004) Development of a corpus workbench for the METU Turkish Corpus. In: Proceedings of LREC, Lisbon, pp 223–225
Özyıldırım I (2010) Tür çözümlemesi. Bilgesu Yayınları, Ankara
Renouf A (2007) Corpus development 25 years on: from super-corpus to cyber-corpus. In: Facchinetti R (ed) Corpus linguistics 25 years on. Rodopi, Amsterdam, pp 27–49
Renouf A, Kehoe A, Banerjee J (2007) WebCorp: an integrated system for web text search. In: Hundt M, Biewer C, Nesselhauf N (eds) Corpus linguistics and the web. Rodopi, Amsterdam, pp 47–67
Rissanen M, Kytö M, Kahlas-Tarkka L, Kilpiö M, Nevanlinna S, Taavitsainen I, Nevalainen T, Raumolin-Brunberg H (eds) (1991) The Helsinki corpus of english texts. University of Helsinki, Helsinki
Ruhi Ş (2006) Politeness in compliment responses: a perspective from naturally occurring exchanges in Turkish. Pragmatics 16:43–101
Ruhi Ş (2009) The pragmatics of *yani* as a parenthetical marker in Turkish: evidence from the METU Turkish corpus. In: Working papers in corpus-based linguistics and language education, vol 3, pp 285–298
Ruhi Ş (2011) Creating a sustainable large corpus of spoken Turkish for multiple research purposes. In: Proceedings of Multisaund: Ulusal konuşma ve dil teknolojileri platformu kuruluşu ve Türkçede mevcut durum çalıştayı, TÜBİTAK-BİLGEM, Gebze, pp 70–73
Ruhi Ş (2013) Interactional markers in Turkish: a corpus-based perspective. J Linguist Lit 10:1–7

Ruhi Ş (2014) Sözlü Türkçe Derlemi'nde temel arama ve edimbilimsel açımlama: Yöntem geliştirme. In: Proceedings of the national linguistics conference, Hacettepe University, Kemer, pp 271–279

Ruhi Ş, Eröz-Tuğa B, Hatipoğlu Ç, Işık-Güler H, Acar G, Eryılmaz K, Can H, Karakaş Ö, Karadaş DÇ (2010a) Sustaining a corpus for spoken Turkish discourse: accessibility and corpus management issues. In: Proceedings of the workshop on language resources: from storyboard to sustainability and LR lifecycle management, Valetta, pp 44–48

Ruhi Ş, Işık-Güler H, Hatipoğlu Ç, Eröz-Tuğa B, Karadaş DÇK (2010b) Achieving representativeness through the parameters of spoken language and discursive features: the case of the spoken Turkish corpus. In: Moskowich-Spiegel F, Isabel CG, Begona I, Lareo M, Sandino PL (eds) Language windowing through corpora. Visualización del Lenguaje a Través de Corpus. Universidade da Coruña, Coruña, pp 789–799

Ruhi Ş, Schmidt T, Wörner K, Eryılmaz K (2011) Annotating for precision and recall in speech act variation: the case of directives in the Spoken Turkish Corpus. In: Proceedings of the conference of the german society for computational linguistics and language technology – working papers in multilingualism, Hamburg, pp 203–206

Ruhi Ş, Eryılmaz K, Acar G (2012) A platform for creating multimodal and multilingual spoken corpora for Turkic languages: insights from the Spoken Turkish Corpus. In: Proceedings of the first workshop on language resources and technologies for Turkic languages, Istanbul, pp 57–63

Ruhi Ş, Haugh M, Schmidt T, Wörner K (eds) (2014) Best practices for spoken corpora in linguistic research. Cambridge Scholar, Newcastle upon Tyne

Sak H, Güngör T, Saraçlar M (2011) Resources for Turkish morphological processing. Lang Resour Eval 45(2):249–261

Sampson G (2013) The empirical trend. Int J Corpus Linguist 18:281–289

Say B (2006) Türkçe için bir derlem geliştirme çalışması. In: Bilgisayar Destekli Dilbilim Çalışmaları Bildirileri, TDK, Ankara, pp 81–88

Say B, Zeyrek D, Oflazer K, Özge U (2004) Development of a corpus and a treebank for present-day written Turkish. In: Proceedings of the international conference on Turkish linguistics, Magosa, pp 183–192

Schmidt T (2004) Transcribing and annotating spoken language with EXMARaLDA. In: Proceedings of the workshop on XML-based richly annotated corpora, Lisbon, pp 69–74

Searle JR (1975) A taxonomy of illocutionary acts. In: Mind and knowledge. Minnesota studies in the philosophy of science. University of Minnesota Press, Minneapolis, pp 344–369

Sebzecioğlu T (2013) Anlık oluşum ve Türkçe anlık sözcüklerin oluşum süreçleri üzerine bir betimleme. J Lang Linguist Stud 10:17–47

Sinclair JM (2005) Appendix to chapter one: how to make a corpus. In: Wynne, M (ed) Developing linguistic corpora: a guide to good practice. ota.ox.ac.uk/documents/creating/dlc. Accessed 3 July 2017

Sofu H, Altan A (2009) Partial reduplication: revisited. In: Proceedings of the international conference on Turkish linguistics, Wiesbaden, pp 63–72

Svartvik J (2007) Corpus linguistics 25+ years on. In: Facchinetti R (ed) Corpus linguistics 25 years on. Rodopi, Amsterdam, pp 11–25

Teubert W (2005) My version of corpus linguistics. Int J Corpus Linguist 10:1–13

Teubert W, Cermakova A (2004) Corpus linguistics: a short introduction. Continuum, London

Uçar A (2014) Özel amaçlı derlemi çeviriyazmak: Bir çeviriyazı modeli. Dilbilim Araştırmaları Dergisi 1:1–30

Uçar A, Kurtoğlu Ö (2012) A corpus-based account of polysemy in Turkish: a case of ver-'give'. In: Kincses-Nagy E, Biacsi M (eds) Proceedings of the international conference on Turkish linguistics. Gold Press Nymoda Kft, Szeged, pp 539–552

Uçar A, Yıldız İ (2015) Humor and impoliteness in Turkish: a corpus-based analysis of the television show Komedi Dükkânı 'comedy shop.'. In: Ruhi Ş, Aksan Y (eds) Exploring (im)politeness in specialized and general corpora: converging methodologies and analytic procedures. Cambridge Scholars, Newcastle upon Tyne, pp 40–81

Uzun L, Erk-Emeksiz Z, Turan ÜD, Keçik İ (2014) Sosyal bilimler alanında Türkçe yazılan özgün araştırma yazılarında uslamlama türlerine göre sav şemaları. In: Proceedings of the national linguistics conference, Kemer, pp 305–321

Wörner K (2009) Werkzeuge zur flachen Annotation von Transkriptionen gesprochener Sprache. PhD thesis, Bielefeld University, Bielefeld

Wynne M (2005) Developing linguistic corpora: a guide to good practice. icar.univ-lyon2.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf. Accessed 14 Sept 2017

Yıldız İ, Aksan M (2014) Türkçe bilimsel metinlerde eylemler: Derlem temelli bir inceleme. In: Proceedings of the national linguistics conference. Hacettepe University, Kemer, pp 247–253

Zeyrek D (2012) Thanking in Turkish: a corpus-based study. In: Ruiz de Zarobe L, Ruiz de Zarobe Y (eds) Speech acts and politeness across languages and cultures. Peter Lang, Bern, pp 53–88

Zeyrek D, Turan ÜD, Bozşahin C, Çakıcı R, Sevdik-Çallı A, Demirşahin I, Aktaş B, Yalçınkaya İ, Ögel H (2009) Annotating subordinators in the Turkish Discourse Bank. In: Proceedings of the Linguistic annotation workshop, Singapore, pp 44–47

Zeyrek D, Demirşahin I, Sevdik-Çallı A, Çakıcı R (2013) Turkish Discourse Bank: porting a discourse annotation style to a morphologically rich language. Dialogue Discourse 4(2):174–184