Language Resources and Evaluation (2023) 57:449–488
https://doi.org/10.1007/s10579-022-09605-4

SURVEY

Resources for Turkish natural language processing: A critical survey

Çağrı Çöltekin[1] · A. Seza Doğruöz[2] · Özlem Çetinoğlu[3]

Accepted: 8 July 2022 / Published online: 26 August 2022

Abstract
This paper presents a comprehensive survey of corpora and lexical resources available for Turkish. We review a broad range of resources, focusing on the ones that are publicly available. In addition to providing information about the available linguistic resources, we present a set of recommendations, and identify gaps in the data available for conducting research and building applications in Turkish Linguistics and Natural Language Processing.

Keywords Turkish · Corpora · Lexical resources · NLP · Linguistics

1 Introduction

As in many other fields of science and engineering, the data-driven methods have been the dominant approach to natural language processing (NLP) and computational linguistics (CL) for the last few decades. The recent (re)popularization of deep learning methods increased the importance and need for the data even further. Similarly, the other subfields of theoretical and applied linguistics have also seen a shift towards more data-driven methods. As a result, availability of large and high-quality language data is essential for both linguistic research and practical NLP applications. In this paper, we present a comprehensive and critical survey of linguistic resources for Turkish.

✉ Çağrı Çöltekin
ccoltekin@sfs.uni-tuebingen.de

A. Seza Doğruöz
as.dogruoz@ugent.be

Özlem Çetinoğlu
ozlem@ims.uni-stuttgart.de

[1] University of Tübingen, Tübingen, Germany

[2] Ghent University, Ghent, Belgium

[3] University of Stuttgart, Stuttgart, Germany

Turkish is a language spoken by over 80 million people mainly in Turkey, also having a significant number of speakers in Cyprus, Europe, and Central Asia (Eberhard et al., 2020).[1] It exhibits a number of interesting linguistic characteristics that are often challenging to handle in NLP applications in comparison to the well-studied languages.

As a result, the linguistic resources for Turkish are important for building practical NLP applications for a large speaker community as well as for quantitative and computational approaches to linguistics, including multilingual and cross-linguistic research. Furthermore, since Turkish is one of the largest and most well-studied languages in the Turkic language family, the resources we review below are potentially useful for language transfer in NLP applications, and as examples for resource and tool creation efforts for the other Turkic languages.

Our survey mainly focuses on currently available resources (see Aksan & Aksan, 2018, for a more historical account of Turkish corpora). We also introduce a companion webpage which we update as new linguistic resources become available.[2] Our survey provides an overview of the available resources, giving details for the major ones, and aims to identify the areas where more effort is needed. To our knowledge, this is the first survey of its kind on Turkish resources. The most similar work is an edited volume of papers on various NLP tasks for Turkish (Oflazer & Saraçlar, 2018). Unlike our work, however, the focus is not the linguistic resources but NLP techniques and tools, and most of the contributions are updated descriptions of the research published earlier. A similar initiative to our companion website is the recently announced Turkish Data Depository (TDD) project (Safaya et al., 2022),[3] which aims to build a repository of data and models for Turkish NLP. Our aim is collecting a more comprehensive list of pointers which can be useful for both NLP and linguistic research, while the TDD intends to store the actual data and the models for NLP with a more practical purpose.

Our focus in this survey is linguistic data, in particular, corpora and lexical resources. We do not aim to describe the research questions, methods and/or the results of these studies but focus on describing the resources in detail. We include resources that are potentially useful for NLP applications, as well as for linguistic research. We also do not focus on NLP tools explicitly, such as data-driven part-of-speech (POS) taggers or parsers and higher level tools or services that target non-technical audience such as the web-based NLP pipelines (e.g., Çöltekin, 2015b; Eryiğit, 2014).

The main contribution of the current paper is a broad, comprehensive overview of the linguistic data available for Turkish to enable linguists and NLP researchers/practitioners to locate these resources easily. We also identify missing or incomplete

[1] Throughout this paper, we use Turkish only for referring to the language variety spoken in modern Turkey and use of this variety in other countries/regions. Hence, this count does not include other Turkic languages, including ones mutually intelligible with Turkish.

[2] The web page is publicly available at https://turkishnlp.github.io. The current list was compiled mostly by our own efforts. However, we also welcome suggestions through a simple web-based form, and also through the GitHub repository associated with this URL.

[3] https://tdd.ai/.

resources, suggesting potential areas for future resource creation efforts. We do not only offer a static survey, but we intend to maintain a 'living list' of resources and a repository of publicly available linguistic data.

2 Corpora

This section surveys corpora available for Turkish. We start with general-purpose, linguistically motivated corpora, followed by corpora used for more specific purposes.

2.1 Balanced corpora

Since corpora collected from a single source (genre, domain) contain many idiosyncratic aspects of its source, the creation of balanced or representative corpora has been a major activity in computational/corpus linguistics since the earliest examples of linguistic corpora (e.g., Francis & Kučera, 1979). There are two well-known balanced corpora for Turkish, the Middle East Technical University (METU) corpus (Say et al., 2002) and Turkish National Corpus (TNC, Aksan et al., 2012).

The METU corpus is the first balanced corpus released for Turkish. The corpus consists only of written modality sampled from 14 different text types including novels, essays, research articles, travel articles, interviews, news, newspaper columns, biographies and memoirs. The corpus contains approximately 1000 documents and 1.7M tokens.[4] The original release does not contain any linguistic annotations. However, a number of annotation projects were carried out on parts of this corpus (e.g., Oflazer et al., 2003; Zeyrek et al., 2013, both discussed in Section "Treebanks and corpora with morphosyntactic annotation"). It is available free-of-charge for research purposes after signing a license agreement.

The second balanced corpus is the Turkish National Corpus (TNC, Aksan et al., 2012). The TNC follows the design principles of the British National Corpus (BNC, Burnard, 2000). The corpus consists of 50M words from texts collected from books, periodicals, and various published and unpublished material. It also includes a small 'spoken text' portion that consists of political speeches and news broadcasts. The TNC contains texts from nine different domains (e.g. fiction, scientific articles, art, opinions and editorials) and includes morphological annotations. The corpus is not available for download but it is accessible through a web interface.[5] A small part of the TNC is also used in constructing the BOUN Treebank (Türk et al., 2022, described below).

[4] The numbers are based on a version obtained in 2015, which includes minor fixes to the first original release.

[5] Further information and the query interface is available from https://www.tnc.org.tr/.

Table 1 A summary of currently available Turkish treebanks

Treebank	Type	Sentences	Tokens
METU-Sabancı (Oflazer et al., 2003)	dep	5 635	56 396
ITU Web (Pamay et al., 2015)	dep	5 009	43 191
UD-GB (Çöltekin, 2015a)	dep	2 880	16 803
UD-PUD (Zeman et al., 2017)	dep	1 000	16 536
UD-BOUN (Türk Utku et al., 2022)	dep	9 761	121 214
TWT (Kayadelen et al., 2020)	dep	4 851	66 466
Turkish-Penn-CS (Yıldız et al., 2014)	con	9 560	81 419
UD-Turkish-Penn	dep	9 560	87 367
UD-Tourism	dep	19 750	92 200
UD-Kenet	dep	18 700	178 700
UD-FrameNet	dep	2 700	19 221

The numbers in the table are based on our own counts on the most recent versions of the datasets. Not all information is reported in the respective papers, and there may be mismatches between the numbers reported in the papers and the released datasets

2.2 Treebanks and corpora with morphosyntactic annotation

This section reviews primarily manually-annotated Turkish corpora with general-purpose *linguistic* annotations, as opposed to corpora annotated for a particular NLP task. The majority of the corpora discussed below are treebanks, however we also include a few other corpora with morphosyntactic annotations.

Treebanks are important resources for linguistic research and applications. Although they have been primarily used for training parsers in CL, multiple levels of linguistic annotations available in treebanks have also been beneficial for other NLP applications and linguistic research. There has been a surge of interest in creating new treebanks for Turkish in recent years. Table 1 presents the currently-available treebanks, along with basic statistics.[6] Below, we provide a brief historical account of treebanks for Turkish.

The first Turkish treebank is the METU-Sabancı treebank (Atalay et al., 2003; Oflazer et al., 2003). The METU-Sabancı treebank is a dependency treebank including a selection of sentences from the METU corpus discussed in Section "Balanced corpora", and includes different text types of the original resource. As an early effort with relatively low funding, the treebank had various issues with formatting and data quality (Say, 2011). Despite these issues, the METU-Sabancı treebank was the only Turkish treebank over a decade. There has been a large number of reports of fixes over the years, but most fixes remained unpublished, or even introduced other errors

[6] We only include manually annotated treebanks. All treebanks listed in the table are directly available for download, with the exception of ITU Web treebank, which requires a signed license agreement to be sent to the maintainers. All UD treebanks can be obtained through the project's webpage at https://universaldependencies.org/. Automatic conversion efforts or parsed corpora are not listed in the table.

or unclear modifications to the annotation scheme. The most up-to-date version of this treebank is made available through Universal Dependencies (UD, De Marneffe et al., 2021; Nivre et al., 2016) repositories based on a semi-automatic conversion (Sulubacak et al., 2016) of a version from Istanbul Technical University (ITU) and hence, named UD-IMST (ITU-METU-Sabancı Treebank). Even the latest version is reported to have a large number of errors, carried over from earlier versions or introduced along the way by many automated conversion processes (see, e.g., Türk et al., 2019). Burga et al. (2017) present a conversion of the same treebank into another related framework, namely Surface-Syntactic Universal Dependencies (SUD, Gerdes et al., 2018). The paper states the intention to publish the resulting treebank, but it is not available at the time of this writing.

After a long time gap, a growing number of new dependency treebanks have recently been released. One of the new treebanks, ITU-Web treebank (Pamay et al., 2015), contains user-generated text from the web. It was annotated following the METU-Sabancı treebank annotation scheme, and later converted to the UD annotation scheme automatically. The first treebank annotated directly using the UD framework is by Çöltekin (2015a). This treebank contains linguistic examples from a grammar book to increase the coverage of different morphosyntactic constructions while minimizing the annotation effort. Two relatively larger and more recent dependency treebanks are the Boğaziçi University (BOUN) treebank (Türk et al., 2022) and the Turkish web treebank (TWT, Kayadelen et al., 2020). The BOUN treebank annotates a selection of sentences from the TNC (Aksan et al., 2012, see Section "Balanced corpora") covering a number of different text types. The BOUN treebank is directly annotated according to the UD annotation scheme. The TWT includes sentences from the web and Wikipedia. The annotations in TwT deviate from the UD and the majority of the existing Turkish dependency treebanks.

Besides the monolingual treebanks above, there have also been a few parallel treebanking efforts. Megyesi et al. (2008, 2010) report automatically annotated parallel dependency treebanks of Turkish, Swedish and English, containing texts published in the forms of popular literature books. However, they have not been released publicly. Another early attempt of parallel treebanking is the constituency treebank described by Yıldız et al. (2014) and Kara et al. (2020b). This treebank includes translations of short sentences (less than 15 words) from Penn Treebank (Marcus et al., 1993). The UD-PUD (Zeman et al., 2017) is part of a parallel dependency treebank effort including 20 languages so far, built on sentences translated predominantly from English. The dependency annotations were performed by Google with their own annotation scheme and automatically translated to UD for the CoNLL multilingual parsing shared task (Zeman et al., 2017). A different type of multilingual treebanking effort is the UD-SAGT treebank, which annotates 2184 spoken language utterances containing Turkish–German code-switching treebank (Çetinoğlu & Çöltekin, 2019, 2022). The treebank follows the UD framework. Section "Code-switching corpora" provides further details about the underlying dataset.

Version 2.8 of the UD treebanks, released in May 2021, introduced four new Turkish treebanks from the same group. One of these treebanks is the dependency version of the Penn treebank translations (Yıldız et al., 2014). Others include a domain-specific tourism treebank, and two treebanks annotating example sentences

from two lexical resources discussed in Section "Lexical Resources" below. The descriptions of the treebanks in the UD repositories indicate that all four treebanks are manually annotated. However, no formal descriptions of these treebanks have been published at the time of writing.

As described above, Turkish is relatively rich with respect to the quantity of available treebanks. However, the need for improvement in terms of the quality of annotations, establishing standards and resolving inconsistencies within and across treebanks has been emphasized by multiple researchers (see, for example Çöltekin, 2016; Say, 2011; Türk et al., 2022, for earlier discussions).

An unusual, yet potentially useful freely-available dataset with morphosyntactic annotation is ODIN (Lewis, 2006), a multilingual collection of examples from linguistics literature with interlinear glosses. Although ODIN does not include full or uniform morphosyntactic annotations, the glossed example sentences can be useful for linguistic research; they may serve as test instances with interesting or difficult linguistic constructions; and they can be converted to a treebank with less effort than that is required for annotating unanalyzed text.

There are also a few corpora that include only morphological annotations. The most popular corpus with morphological annotations is a 1M token corpus disambiguated semi-automatically. The exact procedure used for the disambiguation is unclear. The corpus was introduced by Hakkani-Tür et al. (2002), and made publicly available by later studies on morphological disambiguation (Dayanık et al., 2018; Sak et al., 2011; Yüret & Türe, 2006). Another fully manually disambiguated dataset consisting of 25098 words is reported in Kutlu and Çiçekli (2013), which can be obtained from the authors via email.

2.3 Large-scale (unannotated) linguistic data collections

Although well-balanced, representative corpora have been at the focus of building corpora in corpus linguistics, opportunistic large collections of linguistic data have also been useful in CL/NLP tasks that require large datasets. Furthermore, the size and distribution restrictions on balanced corpora often limits their use both for NLP applications, and research on some linguistic questions (e.g., if the questions are concerned with rare linguistic phenomena). In this section, we review some of the unannotated or automatically annotated corpora that are either used in earlier literature, or publicly accessible without major limitations.

The largest Turkish corpora available are two large multilingual web-crawled datasets: supplementary data released as part of CoNLL-2017 UD parsing shared task (Ginter et al., 2017; Zeman et al., 2017), and the OSCAR corpus (Ortiz Suárez et al., 2019, 2020). Both corpora are sentence shuffled to comply with the copyright laws. The Turkish part of the CoNLL-2017 dataset contains approximately 3.5 billion words. The data is deduplicated, and automatically annotated for morphology and dependency relations. The data can be downloaded directly from the LINDAT/CLARIN repository. The OSCAR corpus is available as raw, and deduplicated versions. The Turkish section contains over 3 billion words after deduplication. The OSCAR corpus can be obtained after creating an account automatically. The

publicly available data does not include any meta information, and the order of the sentences is destroyed by shuffling. However, the webpage of the OSCAR corpus includes a form to request original data without sentence shuffling.

Another popular, relatively large Turkish corpus is the BOUN corpus (Sak et al., 2008). The corpus contains approximately 500M tokens collected from two major online newspapers and other webpages. Although it is used in many studies, it is not clear how to access this corpus.

A relatively large, and easily accessible data source is the multilingual Leipzig Corpora Collection (Quasthoff et al., 2014). The Turkish section contains over 7M sentences (approximately 100M words) of news, Wikipedia and web crawl. The Leipzig corpora are also sentence shuffled. Web-crawled data also contains smaller parts crawled from Turkish-language web sites published in Cyprus and Bulgaria.

The Turkish parliamentary corpus released as part of the ParlaMint project (Erjavec et al., 2021, 2022) contains the transcripts of the Turkish parliament (2011–2021), including approximately 43M words from 303505 speeches delivered at the main proceedings of the parliament. The data also contains speaker information (name, gender, party affiliation) and automatic annotations including morphology, dependency parsing and named entities.

Another relatively large (approximately 10M words), freely accessible corpus is the Kaggle old news dataset.[7] This is a multilingual collection from well-known news sites. The data also includes publication date of the article and the source URL of the document.

The TS Corpus (Sezer, 2017; Sezer & Sever Sezer, 2013) is also a large collection of corpora with a web interface. The collection contains some corpora released earlier (e.g., the BOUN corpus discussed in Section "Balanced corpora") as well as sub-corpora collected by the authors. The authors report over 1.3 billion tokens in 10 sub-corpora from various text sources and various levels of (automatic) annotation. The corpus is served via a web-based query interface, and, to our knowledge, the full corpus is not publicly available for download.

Another relatively small, but potentially interesting unannotated dataset is a compilation of 6844 essays on creative writing classes by Turkish university students between 2014–2018. The essays (approximately 400K words) are published on the course webpage as PDF files.

2.4 Corpora with discourse annotation

There are two corpora that are annotated for discourse markers in Turkish. The first one, Turkish Discourse Bank (TDB, Zeyrek et al., 2013), includes roughly 400K words across various written genres in the METU corpus (Section "Balanced corpora"). The corpus is annotated based on explicit connectives and their two arguments. The TDB is available for academic use through email. Zeyrek et al. (2018,

[7] The corpus is not described in any earlier publication. Throughout this survey, we cite the papers describing each resource, if one is available, otherwise provide a hyperlink to the resource. Links to all available resources are provided in the companion webpage at https://turkishnlp.github.io.

2010), on the other hand, focus on annotating discourse markers in the transcripts of TED talks in six languages (i.e., English, German, Polish, European Portuguese, Russian and Turkish). The Turkish corpus measures 5164 words. The annotation tasks in each language were carried out according to the Penn Discourse Treebank (PTDB) guidelines. The corpus was annotated for five discourse relation types (i.e., explicit connectives, alternative lexicalizations, implicit connectives, no relation) and five top-level senses (i.e., temporal, comparison, expansion, contingency, hypophora). The annotated corpus is freely available.

2.5 Word sense disambiguation corpora

The word sense disambiguation (WSD) task has been defined in two ways: lexical sample and all-words. The lexical sample task aims to disambiguate a restricted set of ambiguous words in their context. The all-words variant, on the other hand, disambiguates all words of a given input. Turkish has resources for both variants.

The first WSD dataset for Turkish is created as part of a SemEval 2007 task and opts for the lexical sample variant (Orhan et al., 2007). 26 unique lexical samples are tagged for their senses, and each sample is tagged in about 100 sentences. The corpus used for the annotation is the METU-Sabancı Treebank, hence the WSD dataset is already accompanied with morphosyntactic annotations. The WSD annotation adds fine-grained senses from the dictionary of Turkish Language Association (TDK), coarse-grain senses, which are a set of semantically closest fine-grained senses, and three levels of ontology. The website link provided in the paper for obtaining the resource is not accessible.

İlgen et al. (2012) also employ the lexical samples approach but choose their words among the most ambiguous words based on a frequency list (Göz, 2003). There are 35 lexical samples in total and each sample is annotated in at least 100 sentences. The corpus was collected from well-known websites on news, health, sports, and education in Turkish. The word senses come from the TDK dictionary (though the authors eliminated some senses that are infrequent in online resources). The availability of the resource is unclear.

The first all-words WSD resource for Turkish annotates a set of sentences that contains translations of Penn Treebank sentences up to 15 tokens (the treebank is described in Section "Treebanks and corpora with morphosyntactic annotation"). Akçakaya and Yıldız (2018) annotates the dependency version of the treebank as an all-words WSD resource. Therefore, the sentences also include morphosyntactic annotations. As in other resources, the sense information comes from the TDK dictionary.[8] In total, there are 7595 unique lexical samples to disambiguate in a corpus of 83473 tokens. 77% of these unique samples are nouns, followed by verbs and adjectives. The website link provided in the paper for obtaining the resource is not accessible. The statistics for WSD resources are given in Table 2.

[8] Note that it is the same dictionary, yet different versions.

Table 2 A summary of WSD resources

Resource	Type	Additional	Samples	Sent.
METU (Orhan et al., 2007)	Lexical sample	morph, dep	26	5 385
ITU (İlgen et al., 2012)	Lexical sample	–	35	3 616
Işık (Akçakaya & Yıldız, 2018)	All-words	morph, con	7595	83 474

The 'Additional' column mentions additional annotations, namely, morph: POS tags and morphology, dep: dependency, con: constituency

2.6 Corpora of parent-child interactions

Language acquisition has been a major interest in modern linguistics, where Turkish also received a fair amount of attention because of a rather interesting learning course observed by young learners, for example, an early and error-free acquisition of case morphology (Xanthos et al., 2011). The CHILDES database (MacWhinney & Snow, 1985) contains two freely-available Turkish datasets with transcriptions of parent–caregiver interactions. The first dataset (Aksu-Koç & Slobin, 1985) contains transcripts of 54 sessions consisting of interactions with 33 children between 28 to 56 months of age. The second dataset (Altıntaş, 2005, 2012) contains transcriptions of 15 recordings with the same child between ages 16 months to 28 months. Both corpora mark speakers, and include some extra-linguistic information. The latter corpus also includes morphological annotation of a subset of the child utterances. A larger and more recent child-language dataset is reported in Moran et al. (2015). However, the Turkish section of this corpus was not released as of this writing. Rothweiler (2011) has also released a 'Turkish-German successive bilinguals corpus' which contains 94 longitudinal spontaneous speech samples by Turkish-German bilingual children (7–28 months-old) recorded between 2003–2008. Part of the data could be viewed for research purposes after obtaining a password.

2.7 Social media text normalization corpora

Normalization of social media text is an important first step in many NLP applications, where ill-formed words or phrases are replaced (or associated) with their normal forms. The definition of 'ill-formed' text is debatable and text normalization in social media hinders analyzing social aspects of language use from a computational sociolinguistic point of view (Eisenstein, 2013; Nguyen et al., 2016). However, normalization datasets enable the use of tools created for formal/standard language, and non-destructive text normalization is also helpful in analyzing interesting aspects of non-standard language use by individuals or groups. We review corpora for normalization purposes here, for lexical resources for the same purpose, see Section "Sentiment, emotion and other application-specific lexicons".

Eryiğit et al. (2017) report a 'big Twitter dataset' (BTS) for normalization which consists of 26149 tweets, as well as using IWT (see Section "Treebanks and corpora with morphosyntactic annotation") as a source of normalization data. The BTS

contains 57088 manually normalized tokens out of a total of 385568. In IWT, 5101 tokens (out of 39152 are normalized. The datasets are available from the group's webpage after signing a license agreement. Çolakoğlu et al. (2019) introduced another normalization test set of 713 tweets (7948 tokens, 2856 normalized). The dataset is available via W-NUT 2021 Shared Task on Multilingual Text Normalization. A more recent Twitter normalization data consisting of 2000 sentences was introduced in Köksal et al. (2020). 6488 out-of-vocabulary (OoV) tokens (out of 16878) identified using lexical resources were manually annotated (below 10% of the OoV tokens are well-formed, e.g., foreign names or neologisms). The dataset is available through a GitHub repository. Besides these monolingual resources, a normalization dataset for Turkish–German is also available (Van der Goot & Çetinoğlu, 2021). This dataset is a revised version of the data from Çetinoğlu and Çöltekin (2016) for normalization by employing token-level alignment layers and adapting existing language IDs and POS tags for these new layers.

2.8 Corpora for named entity recognition

Named entity recognition (NER) for Turkish has been studied by diverse groups of researchers with a few publicly available datasets. Tür et al. (2003) is one of the first to study NER in Turkish with a dataset compiled from newspaper articles over approximately one year (1997–1998). The dataset is annotated for ENAMEX (person, location, organization) named entity types. The dataset has been the standard benchmark for many subsequent studies, with some changes along the way. Original article reports a dataset of approximately 1M words. The version of the dataset as used by Yeniterzi (2011) consists of approximately 500K words with 37189 named entities (16291 person, 11715 location 9183 organization). This version of the data can be obtained through email. Çelikkaya et al. (2013) report three additional datasets covering different text sources, namely, a computer hardware forum, orders to a speech assistant, and Twitter. The data is also annotated for NUMEX entities (numerical expressions). Şeker and Eryiğit (2017) report an annotation effort partially based on the datasets reported in Çelikkaya et al. (2013) and Tür et al. (2003), but also annotating the IWT (described in Section "Treebanks and corpora with morphosyntactic annotation"). The datasets are available from the group's webpage after signing a license agreement. Eken and Tantuğ (2015) also report additional 9358 tweets annotated similar to Çelikkaya et al. (2013). However, availability of this dataset is unclear.

Küçük et al. (2014) and Küçük and Can (2019) report two Twitter datasets of 2320 and 1065 tweets, respectively. These datasets are annotated for person, location, organization, date, time, money and misc (e.g., names of TV programs, music bands), and publicly available through the authors' GitHub repositories. Another, more recent, NER data set annotating 5000 tweets was released by Çarık and Yeniterzi (2022).

2.9 Code-switching corpora

Code-switching refers to mixing more than one language in written and spoken communication and it is quite common in multilingual settings (e.g., immigration contexts, India, Africa etc.).

Nguyen and Doğruöz (2013) and Papalexakis et al. (2014) report analyzing code-switching (e.g., Turkish-Dutch) in online fora for automatic language identification and a prediction task but this data set is not publicly available.

Çetinoğlu (2016) released a Turkish–German Twitter corpus which is annotated with language IDs. The dataset consists of 1029 tweets that are automatically collected, semi-automatically filtered, and manually annotated. Each tweet contains at least one code-switching point, the tweets are normalized and tokenized before adding language IDs. Çetinoğlu and Çöltekin (2016) added POS tag annotations to the same dataset following UD guidelines. A spoken corpus of interviews with Turkish–German bilinguals was presented by Çetinoğlu and Çöltekin (2019, 2022). The audio files are annotated with sentence and code-switching boundaries. Sentences that contain at least one code-switching point are transcribed and normalized to their orthographic representation. The resulting 2184 sentences are annotated with language IDs following (Çetinoğlu, 2017), and with lemmas, POS tags, morphological features, and dependency relations following the UD framework. The treebank version of the dataset is available in the Universal Dependencies repositories, the audio files and aligned transcriptions are available to researchers after signing a license agreement. Yirmibeşoğlu and Eryiğit (2018) worked on detecting code-switching in Turkish–English social media posts. The data is claimed to be available but it was not found on the website link suggested in the paper.

The MULTILIT project (Schroeder et al., 2015) focuses on multilingual children and adolescents of Turkish and Kurdish background living in Germany and France. The corpora they collected include Turkish oral monologues (and their transcription), and written text produced by bilingual students. A subset of the corpus is annotated with POS tags, morphological features and partial syntactic structures, as well as markers showing deviations from standard language use. The data is not publicly available. The RUEG project aims at similar goals at a larger range of age groups, and investigates bilingual speakers of Russian, Turkish and Greek background in Germany and the U.S., bilingual speakers of German in the U.S., as well as monolingual speakers of these languages in respective countries. As part of their collection there are Turkish corpora collected in Germany (1197 sentences) and in Turkey (1418 sentences), publicly available as audio files and annotated transcriptions (Wiese et al., 2020). The lemmas, POS tags, and morphological features are manually annotated, dependencies are automatically predicted. All layers follow the UD framework except the fine-grained POS tags which follow the MULTILIT project.

Table 3 A selection of parallel corpora available for Turkish

Corpus	Text type	Languages	Sentences
Bianet (Ataman, 2018)	News	English, Kurdish	61 472
Bible	Religious	Multiple (102)	48 500
EU book shop	EU texts	Multiple (48)	33 398
GlobalVoices	News	Multiple (92)	8 796
JW300 (Agić & Vulić, 2019)	Religious	Multiple (380)	535 353
OpenSubtitles	Subtitles	Multiple (62)	173 215 360
QED (Abdelali et al., 2014)	Educational	Multiple (225)	753 343
SETimes (Tyers & Alperen, 2010)	News	Balkan (10)	1 776 431
TED talks	Subtitles	English	746 857
Tanzil	Religious	Multiple (42)	105 597
Tatoeba	Misc	Multiple (359)	746 857
Wikipedai (Wołk & Marasek, 2014)	Wikipedia	English, Polish	175 972
infopakki	Informational	Multiple (12)	50 909

The third column lists the languages in each corpus (numbers include Turkish), for massively parallel corpora Turkish may not be aligned to all languages. The number of sentences indicates the number of Turkish sentences in the particular corpus. The number of actual aligned sentences vary depending on the target language. All numbers are based on the corpora as available from OPUS parallel corpora collection http://opus.nlpl.eu/

2.10 Parallel corpora

Parallel, aligned corpora in multiple languages are essential for machine translation (MT) as well as multilingual or cross-lingual research. A number of parallel corpora including Turkish have been reported in some of the earlier works on MT between Turkish and mainly English (e.g., Durgar et al., 2010, 2019; Oflazer et al., 2018). Similarly, shared tasks which included Turkish as one of the languages, such as two IWLST shared tasks (Cettolo et al., 2013; Paul et al., 2010), and WMT shared tasks between 2016 and 2018 (Bojar et al., 2016), also provided data for use during the shared tasks. However, none of these resources are available, nor are there clear procedures to obtain these datasets. In this review we only list the resources available (for at least for non-commercial, research purposes) in detail.

Almost all publicly available parallel corpora that include Turkish are available from the OPUS corpora collection (Tiedemann, 2012). A selection of publicly available corpora are listed in Table 3 (except the parallel treebanks discussed in Section "Treebanks and corpora with morphosyntactic annotation"). The table does not list corpora of public software localization texts and some of the other small corpora available through OPUS. The sizes, text types and the target languages vary considerably. This list of resources, to our knowledge, are not used widely by researchers interested in machine translation to/from Turkish.

Another active area of machine translation is translation between Turkic languages (e.g., Altıntaş, 2001; Hamzaoğlu, 1993; Gilmullin, 2008; Gökırmak et al., 2019; Tantuğ et al., 2007; see Tantuğ and Adalı (2018) for a recent summary). Similar to the Turkish–English translation studies, the resources specifically built for the

purpose are scarce, and even if they are reported in the literature, to our knowledge, no specific corpora build for translation between Turkic languages were released.[9] Except for small samples in Apertium repositories (Forcada et al., 2011), the corpora build with large-scale parallel text collections (e.g., ones listed in Table 3) seem to be the only easily obtainable resource for studies requiring parallel corpora between Turkic languages.

2.11 Corpora for sentiment and emotion

Demirtaş and Pechenizkiy (2013) introduced two Turkish datasets consisting of movie and product reviews. The movie reviews, scraped from a popular Turkish movie review site, contain 5331 positive and 5331 negative sentences. The product reviews data, scraped from an online retailer web site, consists of 700 positive and 700 negative reviews. The labels are assigned based on the scores assigned to the movie or the product by the reviewer. The datasets are available at the author's web site.

Kaya (2013) used a balanced corpus of 400 newspapers columns from 51 journalists labeled for positive and negative sentiment. The study also reports a Twitter corpus of 123074 tweets (not labeled). Türkmenoğlu and Tantuğ (2014) also report multiple datasets, consisting of 20244 movie reviews, 4324 tweets and 101346 news headlines. The tweet dataset was annotated with three-way classes (positive, negative, neutral). Similar to other studies, the movie reviews are labeled them based on the scores assigned by the reviewers. However, it is not clear how the authors labeled the headlines corpus and used it for the presented research. Yıldırım et al. (2014) report another manually annotated Twitter dataset of 12790 tweets, labeled as positive (3541) negative (4249) and neutral (5000). None of these publications indicate the availability of the corpora introduced. Hayran and Sert (2017) present another dataset of 3200 tweets. The data is labeled (negative or positive) based on the emoticons in the messages. The dataset is available through email.

Boynukalın (2012) has investigated emotions in Turkish through two datasets. The first dataset is a translation of a multilingual emotion corpus (ISEAR, Scherer & Wallbott, 1994) into Turkish where the participants are asked to describe experiences associated with a given set of emotions (e.g., joy, sadness, anger). Although the original study describes seven emotions, the authors focused on four of them in Turkish and they have identified 4265 short texts in total. The second dataset consists of 25 fairy tales in Turkish collected across various websites on the web. The emotions in this dataset were labeled based on intensity (low, medium, high) at the sentence and paragraph levels. Demirci (2014) analyzed the emotions in a dataset of 6000 tweets, and labeled based on the hashtags they contain as anger, fear, disgust, joy, sadness, surprise. The availability of these two datasets is unclear. A more recent emotion dataset, TREMO, based on the ISEAR corpus is presented by Toçoğlu and Alpkoçak (2018). Instead of translating the original texts, Toçoğlu and Alpkoçak (2018) follow the methodology used to collect the ISEAR corpus, and collect 27350 entries from 4709 individuals describing memories and experiences

[9] Except Gökırmak et al. (2019), who state the intention to release their data pending copyright clearance, most papers do not include intentions of sharing their data.

related to six emotion categories. Toçoğlu et al. (2019) built a dataset consisting of 195445 tweets automatically labeled with these emotion categories based on a lexicon (see Section "Sentiment, emotion and other application-specific lexicons") extracted from the TREMO dataset. Both of these datasets are available online for non-commercial use.

2.12 Speech and multi-modal corpora

As in other languages, speech corpora or other forms of multi-modal datasets (e.g., video) are scarce in comparison to text corpora. The only linguistically motivated speech corpus creation effort seems to be the Spoken Turkish Corpus (STC, Ruhi et al., 2010, 2012). Although an initial sample consisting of 20 recordings, 4514 utterances and 16107 words was released in 2010, the full corpus is still not available.

Easily-accessible Turkish speech corpora are generally parts of multilingual corpus creation efforts. Notable examples include Common Voice (Ardila et al., 2020), and MediaSpeech (Kolobov et al., 2021). The Common Voice dataset is an ongoing data collection effort by Mozilla Foundation. The project collects audio recordings of a set of sentences and phrases in multiple languages. The January 2022 release includes over 68 hours of recordings from 1228 Turkish speakers. The MediaSpeech dataset includes 10 hours of speech recordings (2513 short segments less than 15 seconds each) with transcriptions from two news channels. MuST-C (Cattoni et al., 2021; Di Gangi et al., 2019) is a multilingual corpus of TED talks including Turkish transcripts, but the audio data is only in English.

The majority of the other speech datasets are collected/created within practical speech recognition/processing projects (see Arslan et al., 2020, for a recent review of Turkish speech recognition). The speech corpus introduced in Mengüşoğlu and Deroo (2001) consists of broadcast news and a set of sentences from news read by multiple speakers. Another early speech corpora collection is OrienTel-TR (Çiloğlu & Tokatlı, 2004), Turkish part of the multilingual OrienTel project (Draxler, 2003), collecting phone recordings of pronunciations of a selected set of words and phrases. Arısoy et al. (2009) report a larger dataset of broadcast news, and a dataset of 38000 hours of call center recordings is reported by Haznedaroğlu and Arslan (2014). A recent speech corpus, consisting of movies with aligned subtitles, and read speech samples are reported by Polat and Oyucu (2020). The availability of corpora listed in this paragraph is unclear.

Salor et al. (2007) report a spoken corpus of 2462 sentences, read by 193 speakers with varied ages and backgrounds. Another, similar but smaller set of recordings are available through GlobalPhone corpus (Schultz et al., 2013), which is a collection of parallel sentences from 20 languages including Turkish. Another interesting dataset where native speakers were recorded while reading parts of dialogues in the ATIS corpus (Hemphill et al., 1990) is reported in Upadhyay et al. (2018). These corpora are available for purchase through the LDC or the ELRA.

Topkaya and Erdoğan (2012) report a dataset of audio/video recordings in which 141 Turkish speakers pronounce selected numbers, names, phrases and sentences in a controlled environment. Finally, it is also worth mentioning the

Turkish–German spoken code-switching treebank described in Section "Code-switching corpora" contains aligned audio recordings of Turkish–German bilinguals. Both datasets can be obtained by contacting the authors.

2.13 Corpora for question answering

Although a highly applicable and popular area, there have been relatively few Turkish resources available for question answering (QA) until recently. Early QA work on Turkish include short lists of question–answer pairs without the context including the answer. For example Amasyalı and Diri (2005) report the use of a 524 question–answer pairs. However, to our knowledge none of these datasets are made available. Similarly, Pala Er (2009) includes 105 factoid questions and their answers as part of her thesis manuscript. Longpre et al. (2020) present a freely-available dataset containing human translations of 10000 question–answer pairs sampled from the Natural Questions dataset (Kwiatkowski et al., 2019) to 25 languages including Turkish. Another multilingual QA set released by Artetxe et al. (2020) includes 1190 human-translated question–answer pairs from Stanford Question Answering Data Set (SQuAD, Rajpurkar et al., 2016). In a more recent study, Gemirter and Goularas (2020) report both a new domain-specific dataset as well as an automatic translation of SQuAD. The availability of this dataset is unclear.

2.14 Other corpora for specific applications

The subsections above survey the areas where a relatively large number of resources are available. In this subsection, we review other areas where there are fewer resources, either because it is a new area, or because there has not been enough interest in the Turkish CL community.

Offensive or aggressive language online has been a concern since the early days of the Internet (Lea et al., 1992). With the increasing popularity of social media, and because of the regulations introduced against certain forms of offensive language such as hate speech online, there has been a surge of interest in automatic detection of various types of offensive language. Currently, there are four Turkish corpora related to offensive language. The cyberbullying corpus by Özel et al. (2017) is a manually annotated corpus of 15658 comments collected from multiple social media sites. This dataset is not available. The corpus reported in Çöltekin (2020) is a general offensive language corpus hierarchically annotated according to OffensEval guidelines (Zampieri et al., 2019). This corpus is publicly available and consists of 36232 manually annotated tweets. In addition, two recent hate speech date sets were released by research groups at Aselsan (Toraman et al., 2022), at the Sabancı University (Beyhan et al., 2022).

Natural language inference (NLI) attracted considerable interest recently. The cross-lingual NLI dataset (XNLI, Conneau et al., 2018), includes 7500 premise–hypothesis pairs created for English, and translated to Turkish as well as 13 other languages. More recently, Budur et al. (2020) released a dataset consisting of

automatic translations of Stanford NLI (SNLI, Bowman et al., 2015) and MuliNLI (Williams et al., 2018) datasets, consisting of approximately 570000 and 433000 sentence pairs, respectively. A small part of the SNLI data (250 sentence pairs) was also translated to Turkish earlier for a SemEval-2017 task (Cer et al., 2017). The data is available from the SemEval-2017 multilingual textual similarity shared task website. All NLI datasets listed above are publicly available.

Summarization datasets for Turkish are also mostly from multilingual corpora collection efforts (e.g., Ladhak et al., 2020; Scialom et al., 2020). Almost all work on summarization of Turkish texts we are aware of (e.g., Kutlu et al., 2010; Özsoy et al., 2011) rely on automatic ways to obtain texts and their summaries. However, the availability of these corpora is not clear.

Paraphrasing corpora have interesting applications such as machine translation and determining semantic similarity. Two paraphrasing corpora in Turkish are introduced in Demir et al. (2012) and Eyecioğlu and Keller (2016). The former study reports an unpublished (work-in-progress) corpus of 1270 paraphrase pairs and it can be obtained by contacting the author. The latter study reports a publicly-available corpus of 1002 paraphrase pairs which also includes human-rated semantic similarities of the sentence pairs. Another textual similarity dataset created by automatic translation of the English STS benchmark (Cer et al., 2017) is published by Beken et al. (2021).

Text categorization or topic modeling studies in Turkish often use opportunistic labeling of the topics published in newspaper sections (e.g., politics, economics, sports). Although there are many studies reporting such datasets, they are rarely made publicly available. We only note one publicly available corpus by Kılınç et al. (2017) which has become a common benchmark data for later studies. This corpus consists of 3600 news feeds (RSS) obtained from online newspapers in 6 categories.

Similar to text categorization, stylometry related studies also typically use newspaper columns scraped from online newspapers, and the corpora are not made available publicly (possibly also due to copyright restrictions). Exception we are aware of are a few datasets available from Yıldız Technical University NLP group (Amasyalı & Diri 2006; Türkoğlu et al., 2007) and the publicly available dataset of Twitter gender identification corpus by Sezerer et al. (2019), which contains 5292 users with more than 100 tweets each manually labeled for gender.

Coreference resolution is another task for which the quantity of resources available is rather small. Earlier work on coreference resolution (Küçük & Yöndem, 2007; Küçük & Yazıcı, 2008) report the use of annotated corpora without indication of availability. In the only publicly available corpus with coreference annotation, Schüller et al. (2018) annotate all sentences of METU-Sabancı treebank (described in Section Treebanks and corpora with morphosyntactic annotation) for coreference.

We also note two large multilingual COVID19-related tweet collections by Qazi et al. (2020) and Abdul-Mageed et al. (2021). The first corpus focuses on tweets geolocation in many languages. Although the number of tweets in Turkish is not specified, the total number of tweets is about half a billion. The second corpus includes 28.5M Turkish tweets with COVID-19 related keywords. Both COVID-19 datasets are available as tweet IDs. Kartal and Kutlu (2020) presents a dataset of 2287 Turkish tweets labeled whether they are worth fact checking or not. The dataset is available through a GitHub repository.

Last but not the least, we note two sign-language corpora. The first corpora of Turkish sign language was introduced by Camgöz et al. (2016), and contains sentences and phrases from finance and health domains. Eryiğit et al. (2020) present a Turkish sign language corpus with morphological and dependency annotations, as well as parallel sentences in Turkish. The availability of these two corpora is unclear. Sincan and Keleş (2020) describe a publicly available sign language corpus. However the link provided in the article is not active at the time of this writing.

3 Lexical Resources

In this section we describe large lexicons and lexical networks that are built either as standalone projects or as part of multilingual collections. The majority of these lexicons also provide various levels of annotations and in multilingual cases, they usually have a mapping to the other languages of the collection.

3.1 Lexicons, word lists

Inkelas et al. (2000) aim at creating a Turkish Electronic Living Lexicon (TELL) that reflects actual speaker knowledge. The lexicon they built consists of 30000 lexemes from dictionaries and place names. Nouns are inflected for five forms and verbs are for three, more than half also have morphological roots. All entries have phonemic transcriptions, 17500 of them also have pronunciations. Moreover, 11500 entries are annotated with their etymological source language. It is possible to search the whole lexicon via a webpage which also offers an email address to access the database. LC-STAR (Fersøe et al., 2004) is a collection of lexicons for speech translation between 13 languages including Turkish. The Turkish lexicon consists of 59213 common words (in sport, news, finance, culture, consumer information, and personal communication domains) and 43500 proper names of persons, places, and organizations. The data has been originally released via ELRA but currently it is not available in their catalog.

BabelNet (Navigli & Paolo Ponzetto, 2012) is a semantic network covering 284 languages, It is created using WordNets, Wikipedia, and machine translation. The project's webpage offers a search interface for end users and APIs for programmers. PanLex (Kamholz et al., 2014) builds translation lexicons for over 5700 languages by utilizing their dictionaries and other multilingual resources such as WordNets. The project's webpage lists collected lexicons and available resources for each language. However, most links for Turkish seem to be broken. While PanLex is the largest among mentioned lexicons, it should be noted that some non-Turkish entries are marked as Turkish. The lexicons, their number of lexemes, and additional annotations are summarized in Table 4.

Inflectional and derivational lexicons focus on the morphosyntactic representations of words. The UniMorph project (Sylak-Glassman et al., 2015; Kirov et al., 2016) aims at building a universal schema for morphological representation of

Table 4 The statistics for Turkish large-scale lexicons

Lexicon	Lexemes	Additional
TELL (Inkelas et al., 2000)	30 000	phonemic transcriptions, roots, inflected forms, etymo.
LC-STAR (Fersøe et al., 2004)	104 513	phonetic transcriptions
BabelNet (Navigli & Paolo Ponzetto, 2012)	?	translations, semantic relations
Panlex (Kamholz et al., 2014)	242 635	translations

The 'Additional' column mentions additional annotations. 'etymo.' stands for etymological source

inflected forms. So far, over 120 languages are annotated (based on their webpage) with their features in a combination of automatic extractions from Wiktionary and collaborative efforts. For Turkish, there are 275460 inflected forms of 3579 unique entries (some are multiword expressions). The data is publicly available.

TrLex (Aslan et al., 2018) converts the word entries of the Turkish Language Association (TDK) dictionary into an XML format with separate fields (e.g., lemma, POS tag, origin, meaning, example) and annotates them with morphological segmentation for derivational suffixes. In addition, there is a phonological representation that encodes how entries undergo Turkish morphophonemic rules. There are 110960 entries in total. It is possible to obtain the version with morphological segmentation and POS tags through email communication with the authors.

Universal Derivations (UDer, Kyjánek et al., 2019) proposes a unified scheme for derivational morphology. The Turkish part of the project uses EtymWordNet (De Melo & Weikum, 2010) as a resource. The unified resources of 20 languages are currently available online. In the Turkish part, there are 1937 unique entries and it adds up to 7774 derived word forms. However, there are also errors (e.g., most of the derivational entries are inflectional forms).

Oflazer et al. (2004) built a multiword expression extraction tool that exploits the morphological analyzer lexicon of Oflazer (1994) for non-lexicalized and semi-lexicalized multiwords. The lexicalized multiwords collected in this study are publicly available.

Zeyrek and Başıbüyük (2019) built a lexicon of discourse connectives extracted from Turkish discourse corpora (Zeyrek et al., 2013; Zeyrek & Kurfalı, 2017; Zeyrek et al., 2018). The lexical entries are annotated with a canonical form, orthographic variants, corpus frequency and POS tags. The data is part of a publicly available multilingual connective lexicon database.

3.2 Morphological analyzer lexicons

Since Turkish is a morphologically rich language, morphological analysis and lexical resources related to morphological analyzers have been a central component of Turkish NLP. Early attempts of building morphological analyzers date back to Köksal (1975) and Hankamer (1986). The first practical and most influential morphological analyzer is by Oflazer (1994). This analyzer has been used in a large number of studies. It is also extended by Oflazer and Inkelas (2006) to produce pronunciations

as well as the written forms. However, these resources are developed using non-free Xerox tools, and their availability and license is unclear. More recently, increased availability of free finite-state tools [e.g., SFST (Schmid, 2005), HFST (Lindén et al., 2009) and Foma (Hulden, 2009)] resulted in a relatively large number of freely available morphological analyzers during the last decade. The free/open-source morphological analyzers written in conventional finite-state tools include Çöltekin (2010), Kayabaş et al. (2019), and Öztürel et al. (2019). Another popular tool is Zemberek (Akın & Akın, 2007) which is an open-source application written in Java for various NLP tasks including morphological analysis.

3.3 WordNets and other lexical networks

A WordNet is a lexical database where lexical items (words and phrases) are grouped into synonym sets ("synsets"). All synsets are organized in a tree structure with the hypernymy relation. Some synsets also bear additional semantic relations such as antonymy. The original WordNet for English was built at Princeton University starting in 1990 (Fellbaum, 1998) and over the years WordNets have been developed for more than 200 languages (Global Wordnet Association, 2020).

The first Turkish WordNet (Bilgin et al., 2004; Çetinoğlu et al., 2018) is developed as part of the BalkaNet project (Stamou et al., 2002), which has a direct influence on the selection of synsets. As the main goal of the project was to ensure parallelism among six Balkan WordNets as well as direct mapping to Princeton WordNet and to the eight WordNets of EuroWordNet (Vossen, 1998) the majority of the synset concepts are translated from Princeton WordNet. The remaining synsets are comprised of Balkan-specific concepts and frequent Turkish words. Synonyms of translated synsets and their semantic relations are populated by exploiting the TDK dictionary. The Turkish WordNet is publicly available.

KeNet (Ehsani et al., 2018), on the contrary, follow a bottom-up approach for creating their version of the Turkish WordNet and take the concepts in the TDK dictionary as their starting point. These concepts are semi-automatically grouped into synsets and verified manually. They also exploit Turkish Wikipedia for hypernymy relations. The resulting WordNet is standalone. This is partially improved by Bakay et al. (2019) who match 4417 of most frequent English senses from Princeton WordNet to KeNet synsets. KeNet is also publicly available.

Another popular lexical network is a PropBank that annotates semantic relations between predicates and their arguments. The first example is the English PropBank (Palmer et al., 2005) and several PropBanks followed over the years, including Turkish ones. The first Turkish PropBank is annotated by Şahin and Adalı (2018) on top of the IMST dependency treebank. Later, it was adapted to the UD version of the same treebank. The annotation scheme includes numbered arguments (up to six), which correspond to the core arguments of a verb (e.g., `Buyer` is Arg0 for the predicate *buy*), and 14 temporary roles that represent adjunct-like arguments (e.g., DIR for direction) of a verb. The resource is available by requesting it via a license form.

Table 5 Turkish PropBanks and their basic statistics.'Avg. arg/prd' stands for average arguments per predicate

PropBank	Sentences	Avg. arg/prd
Turkish PropBank (Şahin & Adalı, 2018)	5635	1.80
Turkish PropBank (Ak et al., 2018b)	9560	–
TRopBank (Kara et al., 2020b)	?	1.68

Another PropBank for Turkish is constructed by Ak et al. (2018b) on top of the constituency treebank of Turkish (Yıldız et al., 2014). In this case, numbered arguments are up to four and nine temporary roles are employed. Ak et al. (2018a) compare their PropBank to that of Şahin and Adalı (2018). The same group has continued working on PropBanks and released TRopBank (Kara et al., 2020a) which employ numbered arguments up to four and a different set of semantic role labels. While the former paper has a broken link, the latter version is publicly available online. The number of sentences that are annotated and the average of arguments per predicate are provided in Table 5 for all PropBanks.

ConceptNet (Speer et al., 2018) is a semantic network that creates knowledge graphs from several multilingual resources such as infoboxes of Wikipedia articles, Wiktionary, and WordNets. The concepts are connected with intralingual and interlingual links. 304 languages take part in the project with varying vocabulary sizes. Turkish is in the mid-range with a vocabulary size of 65892. As a follow-up project, Speer and Lowry-Duda (2017) have developed multilingual embeddings based on ConceptNet. Both resources are available for download.

FrameNet (Baker et al., 1998) is a lexical database that structures predicates and their arguments as frames. The first FrameNet is developed for English and over the years other languages have built their FrameNets. A Turkish FrameNet was recently introduced (Marşan et al., 2021). It is designed to be compatible with KeNet (Ehsani et al., 2018; Bakay et al., 2019) and TRopBank (Kara et al., 2020b) by using the same lemma IDs. In total there are 139 frames that include 2769 synsets, which corresponds to 4080 predicates. The FrameNet is available online.

3.4 Word embeddings and pre-trained language models

Word embeddings have gained substantial ground with the rise of neural models. As a consequence, several pretrained models for Turkish were released, as well as multilingual models. For Turkish, there are Word2vec (Şen & Erdoğan, 2014; Güngör & Yıldız, 2017),[10] GloVe (Ferreira et al., 2016), fastText (Grave et al., 2018), ELMo (Che et al., 2018), and several BERT (Schweter, 2020) models available for download. Kuriyozov et al. (2020) created cross-lingual fastText embeddings aligned to English embeddings for five Turkic languages. The embeddings as well as the dictionaries they used for alignments are publicly available. Turkish is also part of the

[10] Also at https://github.com/akoksal/Turkish-Word2Vec without an associated publication.

Table 6 The statistics for Turkish sentiment lexicons. For SentiTurkNet, each synset member is counted as one token

Sentiment Lexicon		Tokens	Polarity
Tr SentiStrength	Vural (2013)	1366	Pos (1-5), Neg (1-5)
Multilingualsentiment	Chen and Skiena (2014)	2500	Pos, Neg
SentiTurkNet	Dehkharghani et al. (2016)	21623	Pos (0-7),Neg (0-7),Neut

multilingual embeddings such as MUSE (Conneau et al., 2017), mBERT (Devlin et al., 2019), and XLM-R (Conneau et al., 2020).

3.5 Sentiment, emotion and other application-specific lexicons

Emotion and sentiment lexicons play an important part for emotion and sentiment analysis approaches. Çakmak et al. (2012) has created an emotion words lexicon for Turkish by translating EMO20Q's list of English emotions (Kazemzadeh et al., 2011) and adding synonyms for some translations. The total list of 197 words is not publicly available. A more recent emotion lexicon is introduced by Toçoğlu and Alpkoçak (2019), which contains scores for six emotion categories across 4966 lexical entries. The lexicon is available online for non-commercial use.

Vural (2013) has translated SentiStrength (Thelwall et al., 2012) to obtain a sentiment lexicon. SentiStrength assigns positive and negative scores to a set of words as well as creating lists of booster words, negation words, idioms, and emoticons. All lists are created also for Turkish. The paper does not provide information about the availability of the dataset.

Chen and Skiena (2014) have automatically generated sentiment lexicons for 136 languages including Turkish, using English as the source language. They used Wiktionary, Google Machine Translation API, and WordNets as mapping resources. About 60% of the words are negative in the Turkish lexicon. The dataset is accessible via the authors' webpage.

Dehkharghani et al. (2016) utilize Turkish WordNet (Çetinoğlu et al., 2018) to create a sentiment lexicon named SentiTurkNet. They first manually label each synset with positive, negative, and neutral polarity. Then they make use of the synset mapping between Turkish and English WordNets (Fellbaum, 1998) so that by transitivity SentiTurkNet can inherit the polarity strength scores of SentiWordNet (Baccianella et al., 2010), a sentiment lexicon which is built on top of the English WordNet. The dataset is publicly available online (Table 6).

A normalization lexicon for social media text normalization is presented in Demir et al. (2016). The lexicon is demonstrated to provide accurate normalization, but statistics of the lexicon are not specified. The paper notes that the resource is publicly available without indicating a method for obtaining it.

4 General discussion

The focus of our survey is exploring data sources for Turkish NLP applications, computational/quantitative linguistics research, as well as (digital) humanities research that may benefit from linguistic data. In this section, we list some of our observations, followed by a short list of recommendations for future efforts on creating language resources. Although we found them to be more prevalent in comparison to efforts for resource rich, well-studied languages, most of the observations and recommendations are not specific to Turkish language resource creation efforts. We believe these recommendations could particularly be useful for linguistic resource creation efforts for languages for which there are relatively few data-driven studies, and the conventions and traditions in the field are not yet well established.

4.1 Availability and maintenance of resources

Although it is not unique to Turkish resources, we have encountered difficulties about finding and/or confirming the availability of the data sources. The locations of published resources are not always stable and/or permanent. The URLs indicating the location of the resources in papers or on the webpages of the authors or institutions are not always maintained and the resources often disappear after publication. Although our efforts to reach out to the authors/creators of the resources often yielded positive results, it is desirable to diminish these barriers to keep up with the fast-paced research community.

Another difficulty about the availability and maintenance of the resources is related to the publication traditions in other fields outside computational linguistics. In particular, most papers published in general computer science venues (e.g., in ACM conferences or journals) do not include information about the availability of their data sources. In some fields (e.g., speech processing), it is more common to make the resource available for a fee which reduces their accessibility especially for early stage researchers or researchers with limited research budgets. In addition, the majority of published resources for Turkish do not include an explicit license or ethical statement concerning collection, distribution and use of the data.

4.2 Awareness of earlier work

Although it is not unique for the research papers in Turkish Computational Linguistics, earlier research/resources (either for Turkish or other languages) are not cited or there is only a short list of references ignoring other relevant research. This results in many repetitions and inconsistencies in the newly created resources.[11] For example, the inconsistencies and the lack of communication during the creation of

[11] This criticism does not refer to the creations of similar resources from multiple independent groups. As the CL and NLP become more and more data driven, we definitely benefit from more data, and well-informed and yet different approaches to the same problem.

different treebanks for Turkish have been brought up by multiple researchers (see Section Treebanks and corpora with morphosyntactic annotation).

Another, related, observation is the tendency to create new resources rather than improving the existing ones. This leads to substantial effort put into the same work, without clear improvements over the earlier systems. For example, despite the fact that some of the earlier morphological analyzers reviewed in Section Morphological analyzer lexicons have been available with free licenses, a large number of new ones were created without a clear statement of difference or comparison. Similar observations can be made for other resources (e.g., WordNets) and annotation tools as well, e.g., improving existing annotation tools could be more useful than creating new tools which are often used in a single project.

Although most research in computational linguistics is publicly available, there is also a need for better communication among scholars to inform each other and collaborate on the ongoing projects, efforts and plans for building and maintaining linguistic resources. In addition, there is a need for more communication and collaboration between linguists and computational linguists for creating, annotating and analyzing language related data and resources.

4.3 Issues about multilingual resources

There is a rapid increase in the efforts of building massively multilingual resources for various tasks and applications. We covered some of these efforts in our survey as well. By necessity, these efforts involve either opportunistic annotations (e.g., use of already existing information for other purposes, like word lists in Wiktionary), or rely heavily on crowd sourcing and/or automatic annotations. However, a potential pitfall is the lack of quality checks for these resources which do not necessarily involve linguistic expertise in each language included in the resource. For example, there are serious issues about the inflectional and derivational lexicons discussed in Section Lexicons, word lists. Although these multilingual resources are useful in many tasks, one should be aware of potential quality issues as well.

4.4 Issues about translated resources

Like for other languages, automatic or manual translations of large datasets created originally for English are also translated to Turkish. Although this approach is interesting as it yields parallel resources, the resource created in this manner includes effects of ‘translationese’, as well as additional errors that may be introduced during the translation process. Translated datasets may even include correct translations that are not appropriate for a particular task. For example, as noted by Budur et al. (2020), the inferential relation for two English sentences may be reversed when translated to Turkish, because Turkish pronouns are gender-neutral. In general, the same type of inference in the original language may not be applicable in the translation. Similar problems are difficult to prevent with automatic translations or non-expert human translations performed without paying attention to the purpose of the dataset.

4.5 Issues about quantity and quality

With respect to the quantity of resources, Turkish may be considered close to a 'resource-rich' language. For example, Turkish has the largest number of treebanks (together with English) in the Universal Dependencies repositories (as of UD version 2.10). However, most Turkish treebanks are smaller in size in comparison to treebanks in other languages, and quality and inconsistency issues have been raised in multiple earlier studies (see Section Treebanks and corpora with morphosyntactic annotation for a short discussion and pointers to relevant papers). The same trend can be observed in other types of resources as well. For example, Aksan and Aksan (2018) report partial results of a questionnaire conducted in 2011, where Turkish NLP specialists were asked to rate the quantity and quality of the available corpora on a scale of 0 to 6. The results indicate rather low judgments, 1.9 for quantity and 2.9 for quality.[12] Although the quantity issues seem less of a problem currently, the number of linguistic resources for Turkish are still relatively low compared to well-studied European languages.

Overall, it is difficult to qualify Turkish as a 'low-resource language' based on the breadth and depth of the resources available. However, the resources are rather scattered across different fields, and there are issues of availability and quality. In sum, it is probably apt to classify Turkish as a 'resource poor' language (following the terminology used by Zaghouani (2014) for Arabic).

4.6 Descriptions of datasets

A related problem in the publications introducing resources is the lack of sufficient descriptions. In some cases, even the basic statistics about the data are not presented or it is difficult to interpret the statistics due to unclear units of measurements. There is also a need for better descriptions of proper quality assurance procedures, metrics and inter-annotator agreements (IAA). Lack of proper linguistic glosses and translations in the provided examples also create extra barriers for readers without any Turkish background to understand and evaluate the research article and/or the data resource.

4.7 Gaps in the existing resources

Although there are a number of sources for (social media) text normalization, we are not aware of any publications on datasets of spelling or grammar errors.[13] Similarly, there is no known learner corpus or resources that can help second language research and practice for Turkish.

[12] The complete results of the questionnaire are not published. Hence, the wording of the questions, and the type of corpora queried are not clear.

[13] A new spelling dictionary with an associated tool has been announced https://extensions.libreoffice.org/en/extensions/show/20565 during the final revisions of this paper.

Another general area with no or little resources is semantics. Except for the lexical resources listed in Section Lexical Resources, we are not aware of any semantically annotated corpora (e.g., one that would be used for semantic parsing). There is also a lack of benchmark datasets for assessing pre-trained word or text representations (word embeddings, or pre-trained language models). So far, most linguistic resources available for Turkish aim to be domain independent. If a resource is domain-specific, it is often due to practical reasons rather than a specific interest in this particular domain. On the other hand, domain-specific data is crucial for NLP applications. Although the uses of unpublished datasets were reported in earlier literature (e.g., a corpus of radiology reports by Hadımlı & Turhan Yöndem, 2011), there is a big gap in domain-specific datasets for critical domains or sub-fields like biomedical, legal or financial NLP.

There is also a need for more systematic data collection and analysis of dialectal and sociolinguistic variation with easy-to-access language resources (Doğruöz forthcoming).

4.8 A concise list of recommendations

The issues raised above in this section have some rather obvious solutions. Nevertheless, the concise list below may be beneficial for future resource creation efforts.

- *Publish your corpora, and publish it on permanent (or long-lasting) venues.* Beyond the value of the published data and code for reproducibility, published data allows others to study the data in ways creators of the data cannot possibly foresee. Furthermore, growing evidence suggests that the papers that publish their data get more recognition (Colavizza et al., 2020; Wieling et al., 2018). It is also important to publish the data in locations that would not disappear shortly after the publication. Our experience in this survey shows that the data shared through personal and also institutional webpages often become inaccessible as authors move to other institutions, or their research interests change. As a result, publishing the data in general repositories like Zenodo and OSF, or CLARIN repositories that are more specialized for language resources is a better choice than personal and institutional webpages. Similarly, to our experience, software development infrastructures like GitHub also provide stable locations for publishing linguistic data.
- *Describe all aspects of the corpora adequately.* As we occasionally noted above, a large number of papers we reviewed do not describe the resources introduced sufficiently. It is important for a paper to include information on aspects of the corpora such as, size, label distribution, source material, sampling method, as well as indications of annotation quality (e.g., IAA) in proper units and using proper metrics for the task at hand. Being aware of the earlier recommendations (e.g., Ide et al., 2017; Bender & Friedman, 2018; Gebru et al., 2020) for resource creation efforts and their descriptions would be useful for any annotation or curation project.

– *Be explicit about the licensing and potential ethical issues.* Although major computational linguistics venues started to require statements about legal and ethical aspects of data collection and sharing, not all the venues require such statements. It is important to be aware of the existing guidelines, such as ACM code of ethics (Gotterbarn et al., 2018), or the guidelines adapted by major CL conferences,[14] as well as the recent discussion in the field (e.g., Rogers et al., 2021; Šuster et al., 2017). Even though the common guidelines may not fit every task, or every legal jurisdiction, being aware of potential issues, and being explicit about the legal and ethical considerations during data collection and annotation is important. The lack of clarity around these issues may also reduce the usability of the data (and hence, the recognition the creators may receive).
– *Before creating a new resource, perform a thorough literature review of the relevant research, consider improving existing resources, and collaborating with other scholars in the field.* As evidenced by the lack of citations in published papers, most resources are built from scratch, not paying attention to the lessons learned in the earlier work. The quality of linguistic resources could be improved by awareness of earlier work and more collaboration between different groups. Besides individual efforts from researchers and reviewers, a regular meeting of CL/NLP researchers and practitioners working on Turkish (and possibly Turkic languages) may help alleviate this problem. Although a number of 'first attempts' were made for such meetings, unlike many other CL communities, no regular/stable meeting has been established so far.
– *Contribute to multilingual resource creation efforts.* One of the issues we observed above with large-scale, multilingual resources is the lack of quality in Turkish data in these efforts. Bringing the language expertise of Turkish (computational) linguists in these projects would definitely improve the quality of these efforts, which, in turn, would be beneficial to the CL/NLP studies in Turkish.

5 Conclusion

Our goal in this survey was to present a comprehensive summary of language resources NLP and computational/quantitative linguistic research for Turkish. In addition to the resources listed in our survey, we also provide a companion website (https://turkishnlp.github.io) which includes links to even more Turkish resources, and we will update it regularly. In this way, our survey and the companion website will serve as stable and sustainable resources for researchers across disciplines (e.g., linguistics, NLP) who are currently working on Turkish. In addition, researchers who are not currently working on Turkish but who need linguistic resources outside their current expertise and/or those who are interested in including Turkish in multi- or cross-lingual tasks could benefit from our contribution as well.

[14] For example, NAACL guidelines at https://2021.naacl.org/ethics/faq/ which is also adapted by some of the other major CL conferences.

Besides the comprehensive overview of the resources, we have also summarized some of the common problematic issues and gaps in the field and provided a set of short suggestions for future resource creation efforts. We cautiously note that not all the problematic issues could easily be resolved by individual researchers and research groups immediately. Some of these issues require long-term collaborative efforts within the community as well as substantial support from academic funding agencies for further research. The issues we raise in this paper are based on our impression from published papers and cursory inspection of the available corpora. To understand the factors behind these issues better and propose informed solutions, future studies with in-depth analyses (e.g., through questionnaires directed to creators and users of the resources, or more systematic inspection of the available data) can be helpful. Similarly, effectiveness of the guidelines (offered in papers we cite in Section 4) may also be measured in future experimental studies.

In short, we hope that our survey and its companion webpage will serve as a useful reference for locating resources for existing fundamental and applied research and for creating future resources and projects for Turkish and/or other languages.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abdelali, A., Guzman, F., Sajjad, H., & Vogel, S. (2014, May). The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1856–1862).

Abdul-Mageed, M., Elmadany, A., Nagoudi, E. M. B., Pabbi, D., Verma, K., & Lin, R. (2020). Mega-COV: A Billion-Scale Dataset of 100+ Languages for COVID-19. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics* (pp. 3402–3420). https://www.aclweb.org/anthology/2021.eacl-main.298.

Agić, Ž., & Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3204–3232). Association for Computational Linguistics.

Ak, K., Toprak, C., Esgel, V., & Yıldız, O. T. (2018b). Construction of a Turkish proposition bank. *Turkish Journal of Electrical Engineering & Computer Sciences, 26*(1), 570–581.

Akçakaya, S., & Yıldız, O. T. (2018). An all-words sense annotated Turkish corpus. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pp. 1–6. https://doi.org/10.1109/ICNLSP.2018.8374368.

Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. *Structure, 10*, 1–5.

Aksan, M., & Aksan, Y. (2018). Linguistic corpora: A view from Turkish. In *Turkish natural language processing* (pp. 291–315). Springer.

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, U., Demirhan, U. U., Yılmazer, H., Atasoy, G., Öz, S., Yıldız, İ., & Kurtoğlu, Ö. (2012). Construction of the Turkish National Corpus (TNC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3223–3227). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/991_Paper.pdf.

Aksu-Koç, A. & Slobin, Dan I. (1985). The acquisition of Turkish. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (Vol. 1, pp. 839–878). Lawrence Erlbaum Associates.

Altınkamış, F. 2012). Turkish Altınkamış Corpus. https://doi.org/10.21415/T5H89W. http://childes.talkbank.org/access/Other/Turkish/Altinkamis.html.

Altınkamış Türkay, F. (2005). Children's early lexicon in terms of noun/verb dominance. PhD thesis. Çukurova University. https://tez.yok.gov.tr/UlusalTezMerkezi/TezGoster?key=vbVkXe1KChYWNElr1MuLZkSZIFvXBJpcL-G5wtalqSvAlPjIZeecxgYeEKGMm7xZ.

Altıntaş, K. (2001). *Turkish to Crimean Tatar machine translation system*. MA thesis. Bilkent University.

Amasyalı, M. F., & Diri, B. (2005). Bir soru cevaplama sistemi: Baybilmiş. In *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi 1.1*.

Amasyalı, M. F., & Diri, B. (2006). Automatic Turkish text categorization in terms of author, genre and gender. In *International Conference on Application of Natural Language to Information Systems*, pp. 221–226. Springer.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4218–4222. ISBN: 979-10-95546-34-4. https://www.aclweb.org/anthology/2020.lrec-1.520.

Arısoy, E., Can, D., Parlak, S., Sak, H., & Saraçlar, M. (2009). Turkish broadcast news transcription and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(5), 874–883.

Arslan, R. S., & Barışçı, N. (2020). A detailed survey of Turkish automatic speech recognition. *Turkish Journal of Electrical Engineering & Computer Sciences, 28*(6), 3253–3269

Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4623–4637). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.421. https://www.aclweb.org/anthology/2020.acl-main.421.

Aslan, Ö., Günal, S., & Taner Dinçer, B. (2018). A computational morphological lexicon for Turkish: Trlex. *Lingua, 206*, 21–34.

Atalay, N. B., Oflazer, K., & Say, B. (2003). The Annotation Process in the Turkish Treebank. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*. https://www.aclweb.org/anthology/W03-2405.

Ataman, D. (2018). Bianet: A parallel news corpus in Turkish, Kurdish and English. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by J. Du, M. Arcan, Q. Liu, & H. I. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-15-3.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

Bakay, Ö., Özlem, E., & Yıldız, O. T. (2019). Integrating Turkish WordNet KeNet to Princeton WordNet: The Case of One-to-Many Correspondences. In *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5. https://doi.org/10.1109/ASYU48272.2019.8946386.

Baker, C. F, Charles J. F., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (Vol. 1, pp. 86–90).

Fikri, F. B., Oflazer, K., & Yanikoglu, B. (2021). Semantic Similarity Based Evaluation for Abstractive News Summarization. In *Proceedings of the 1st Workshop on Natural Language Generation,*

Evaluation, and Metrics (GEM 2021). Online: Association for Computational Linguistics, pp. 24–33. https://doi.org/10.18653/v1/2021.gem-1.3. https://aclanthology.org/2021.gem-1.3.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics,6*, 587–604.

Beyhan, F., Çarık, B., Arın, I., Terzioğlu, A., Yanikoglu, B., & Yeniterzi, R. (2022). A Turkish Hate Speech Dataset and Detection System. In *Proceedings of the Language Resources and Evaluation Conference*. European Language Resources Association, pp. 4177–4185. https://aclanthology.org/2022.lrec-1.443.

Bilgin, O., Çetinoğlu, Ö., & Oflazer, K. (2004). Building a WordNet for Turkish. *Romanian Journal of Information Science and Technology, 7*(1–2), 163–172.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, & K., & Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers* (pp. 131–198). https://doi.org/10.18653/v1/W16-2301. https://www.aclweb.org/anthology/W16-2301.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642. https://doi.org/10.18653/v1/D15-1075. https://www.aclweb.org/anthology/D15-1075.

Boynukalın, Z. (2012). *Emotion analysis of Turkish texts by using machine learning methods*. MA thesis. Middle East Technical University.

Budur, E., Özçelik, R., Güngör, T., & Potts, C. (2020). Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8253–8267. https://www.aclweb.org/anthology/2020.emnlp-main.662

Burga, A., Öktem, A., & Wanner, L. (2017). Revising the METU-Sabancı Turkish Treebank: An Exercise in Surface-Syntactic Annotation of Agglutinative Languages. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)* (pp. 32–41). Linköping University Electronic Press. https://www.aclweb.org/anthology/W17-6506.

Burnard, L., (Ed.), (2000). *The British National Corpus users reference guide*. http://www.natcorp.ox.ac.uk/docs/userManual/.

Çakmak, O., Kazemzadeh, A., Yıldırım, S., & Narayanan, S. (2012, December). Using interval type-2 fuzzy logic to analyze Turkish emotion words. In *Proceedings of the 2012 Asia pacific signal and information processing association annual summit and conference* (pp. 1–4). IEEE

Camgöz, N. C., Kındıroğlu, A. A., Karabüklü, S., Kelepir, M., Özsoy, A. S., & Akarun, L. (2016). BosphorusSign: A Turkish sign language recognition corpus in health and finance domains. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1383–1388). https://aclanthology.org/L16-1220.

Çarık, B., & Yeniterzi, R. (2022). A Twitter Corpus for named entity recognition in Turkish. In *Proceedings of the Language Resources and Evaluation Conference*. European Language Resources Association (pp. 4546–4551). https://aclanthology.org/2022.lrec-1.484.

Cattoni, R., Antonino Di Gangi, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). MuST-C: A multilingual corpus for end-to-end speech translation. In *Computer Speech & Language, 66*, 101155.

Çelikkaya, G., Torunoğlu, D., & Eryiğit, G. (2013). Named entity recognition on real data: a preliminary investigation for Turkish. In *2013 7th International Conference on Application of Information and Communication Technologies* (pp. 1–5). IEEE.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 1–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2001. https://www.aclweb.org/anthology/S17-2001.

Çetinoğlu, Ö. (2016). A Turkish-German Code-Switching Corpus. In N. Calzolari (Conference Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language*

Resources and Evaluation (LREC 2016) (pp. 23–28). European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.

Çetinoğlu, Ö. (2017). A code-switching corpus of Turkish-German conversations. In *Proceedings of the 11th Linguistic Annotation Workshop*(pp. 34–40). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-0804. https://aclanthology.org/W17-0804.

Çetinoğlu, Ö., Bilgin, O., & Oflazer, K. (2018). Turkish wordnet. In K. Oflazer, & M. Saraçlar (Eds.), *Theory and Applications of Natural Language Processing* (pp. 317–336). Springer International Publishing. ISBN: 9783319901657.

Çetinoğlu, Ö., & Çöltekin, Ç. (2016). Part of speech annotation of a Turkish-German code-switching corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 120–130). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-1714. https://www.aclweb.org/anthology/W16-1714.

Çetinoğlu, Ö., & Çöltekin, Ç. (2019). Challenges of annotating a code-switching treebank. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)* (pp. 82–90). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7809. https://www.aclweb.org/anthology/W19-7809.

Çetinoğlu, Ö., & Çöltekin, Ç. (2022). Two languages, one treebank: Building a Turkish-German code-switching treebank and its challenges. In *Language Resources and Evaluation*, (pp. 1–35). ISSN: 1574-020X. https://doi.org/10.1007/s10579-021-09573-1.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., & Federico, M. (2013). Report on the 10th IWSLT evaluation campaign. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign*.

Che, W., Liu, Y., Wang, Y., Zheng, B., & Liu, T. (2018). Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 55–64). Association for Computational Linguistics. http://www.aclweb.org/anthology/K18-2005.

Chen, Y., & Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 2: Short Papers, pp. 383–389). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2063. https://www.aclweb.org/anthology/P14-2063.

Çiloğlu, T., Acar, D., & Tokatlı, A. (2004). OrienTel-Turkish: Telephone speech database description and notes on the experience. In *Eighth International Conference on Spoken Language Processing*.

Çolakoğlu, T., Sulubacak, U., & Tantuğ, A. C. (2019). Normalizing noncanonical Turkish texts using machine translation approaches. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 267–272). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-2037. https://www.aclweb.org/anthology/P19-2037.

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLoS ONE 15*(4), 1–18. https://doi.org/10.1371/journal.pone.0230416

Çöltekin, Ç. (2010). A Freely Available Morphological Analyzer for Turkish. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), pp. 820–827. http://www.lrec-conf.org/proceedings/lrec2010/summaries/109.html.

Çöltekin, Ç. (2015a). A grammar-book treebank of Turkish. In M. Dickinson, E. Hinrichs, A. Patejuk, & A. Przepiórkowski (Eds.), *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*, pp. 35–49.

Çöltekin, Ç. (2015b). Turkish NLP web services in the WebLicht environment. In *Proceedings of the CLARIN Annual Conference*.

Çöltekin, Ç. (2016). (When) do we need inflectional groups? In *Proceedings of The First International Conference on Turkic Computational Linguistics*.

Çöltekin, Ç. (2020). A corpus of Turkish offensive language on social media. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 6174–6184). https://www.aclweb.org/anthology/2020.lrec-1.758.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747. https://www.aclweb.org/anthology/2020.acl-main.747.

Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L., & Jégou, H. (2017). Word Translation Without Parallel Data. In: arXiv preprint arXiv:1710.04087.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2475–2485. https://doi.org/10.18653/v1/D18-1269. https://www.aclweb.org/anthology/D18-1269.

Dayanık, E., Akyürek, E., & Yüret, D. (2018). MorphNet: A sequence-to-sequence model that combines morphological analysis and disambiguation. In *CoRR abs/1805.07946*. arXiv:1805.07946.

Dehkharghani, R., Saygın, Y., Yanıkoğlu, B., & Oflazer, K. (2016). Senti- TurkNet: A Turkish polarity lexicon for sentiment analysis. In *Language Resources and Evaluation*, pp. 1–19.

De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics, 47*(2), 255–308. ISSN: 0891-2017. https://doi.org/10.1162/coli_a_00402.

De Melo, G., & Weikum, G. (2010). Towards universal multilingual knowledge bases. In P. Bhattacharyya, C. Fellbaum, & P. Vossen (Eds.), *Principles, Construction, and Applications of Multilingual WordNets. Proceedings of the 5th Global WordNet Conference (GWC 2010)* (pp. 149–156). ISBN: 978-81-8487-083-1. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.194.2529.

Demir, Ş., El-Kahlout, İ. D., Ünal, E., & Kaya, H. (2012). Turkish paraphrase corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 4087–4091. http://www.lrec-conf.org/proceedings/lrec2012/pdf/968_Paper.pdf.

Demir, Ş., Tan, M., & Topcu, B. (2016). Turkish Normalization Lexicon for Social Media. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing*, pp. 418–429.

Demirci, S. (2014). *Emotion analysis on Turkish tweets*. MA thesis. Middle East Technical University.

Demirtaş, E., & Pechenizkiy, M. (2013). Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (pp. 1–8).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Vol. 1, Long and Short Papers, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., & Turchi, M. (2019). Must-c: A multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2012–2017). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1202. https://www.aclweb.org/anthology/N19-1202.

Doğruöz, A. S. (Forthcoming). Documenting sociolinguistic variation in Turkish. In Y. Asahi, A. D'arcy, & P. Kerswill (Eds.), *Routledge handbook of variationist sociolinguistics*. Routledge (Forthcoming)

Draxler, C. (2003). Orientel: Recording telephone speech of Turkish speakers in Germany. In *Proceedings of the Eighth European Conference on Speech Communication and Technology* (pp. 1557–1560).

El-Kahlout, I. D., Bektaş, E., Erdem, N. Ş., & Kaya, H. (2019). Translating between morphologically rich languages: An Arabic-to-Turkish machine translation system. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop* (pp. 158–166). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4617. https://www.aclweb.org/anthology/W19-4617.

El-Kahlout, İ. D., & Oflazer, K. (2010). Exploiting morphology and local word reordering in English-to-Turkish phrase-based statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(6), 1313–1322.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.), (2020). Ethnologue: Languages of the world. Online version: http://www.ethnologue.com. Dallas, Texas.

Ehsani, R., Solak, E., & Yıldız, O. T. (2018). Constructing a wordnet for Turkish using manual and automatic annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17*(3), 1-15

Eisenstein, J. (2013). What to do about bad language on the Internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (pp. 359–369). Association for Computational Linguistics. https://www.aclweb.org/anthology/N13-1037.

Eken, B., & Tantuğ, C. A. (2015). Recognizing named entities in Turkish tweets. In *Proceedings of the Fourth International Conference on Software Engineering and Applications, Dubai, UAE*.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Hansen, D. H., Navarretta, C., Pérez, M. C., de Macedo, L. D., van Heusden, R., et al., (2021). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1431.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., et al. (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0.

Eryiğit, G. (2014). ITU Turkish NLP Web Service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1–4). Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-2001. https://www.aclweb.org/anthology/E14-2001

Eryiğit, G., Eryiğit, C., Karabüklü, S., Kelepir, M., Özkul, A., Pamay, T., Torunoğlu-Selamet, D., & Köse, H. (2020). Building the first comprehensive machine-readable Turkish sign language resource: methods, challenges and solutions. *Language Resources and Evaluation, 54*(1), 97–121.

Eryiğit, G., & Torunoğlu-Selamet, D. (2017). Social media text normalization for Turkish. *Natural Language Engineering 23*(6), 835–875. https://doi.org/10.1017/S1351324917000134.

Eyecioğlu, A., & Keller, B. (2016). Constructing a Turkish corpus for paraphrase identification and semantic similarity. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 588–599). Springer.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Language, Speech and Communication. MIT Press, 9780262061971.

Ferreira, D. C., Martins, A. F., & Almeida, M. S. (2016). Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, Long Papers) (pp. 2019–2028) https://doi.org/10.18653/v1/P16-1190. https://www.aclweb.org/anthology/P16-1190.

Fersøe, H., Hartikainen, E., Heuvel, H., Maltese, G., Moreno, A., Shammass, S., & Ziegenhain, U. (2004). Creation & Validation of Large Lexica for Speech-to-Speech Translation Purposes. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal. European Language Resources Association*. http://www.lrec-conf.org/proceedings/lrec2004/summaries/452.htm.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F. M. (2011). Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation, 25*(2), 127–144.

Francis, W. N., & Kučera, H. (1979). *Brown corpus manual: Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for Use with Digital Computers*. Brown University.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Hal Daumé III, H. W., & Crawford, K. (2020). Datasheets for datasets. arXiv: 1803.09010 [cs.DB].

Gemirter, C. B., & Goularas, D. (2020). A Turkish question answering system based on deep learning neural networks. *Journal of Intelligent Systems: Theory and Applications 4*(2), 65–75.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018). SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (pp. 66–74). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6008. https://www.aclweb.org/anthology/W18-6008.

Gilmullin, R. A. (2008). The Tatar-Turkish machine translation based on the two-level morphological analyzer. In *Interactive systems and technologies: the problems of human- computer interaction*, pp. 179–186.

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., & Zeman, D. (2017). CoNLL 2017 Shared task—Automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-1989.

Global Wordnet Association. (2020). Wordnets in the world. http://globalwordnet.org/wordnets-in-the-world. Accessed: November 30, 2020.

Gökırmak, M., Tyers, F., & Washington, J. (2019). Machine translation for crimean tatar to Turkish. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages* (pp. 24–31). European Association for Machine Translation. https://www.aclweb.org/anthology/W19-6805.

Gotterbarn, D. W., Brinkman, B., Flick, C., Kirkpatrick, M. S., Miller, K., Vazansky, K., & Wolf, M. J. (2018). ACM code of ethics and professional conduct. https://www.acm.org/code-of-ethics.

Göz, İ., Ed. (2003). Yazılı Türkçenin kelime sıklığı sözlüğü. Türk Dil Kurumu.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Güngör, O. & Yıldız, E. (2017). Linguistic features in Turkish word representations. In *2017 25th Signal Processing and Communications Applications Conference (SIU)* (pp. 1– 4). https://doi.org/10.1109/SIU.2017.7960223.

Hadımlı, K., & Yöndem, M. T. (2011). Two alternate methods for information retrieval from Turkish radiology reports. In *Computer and Information Sciences II* (pp. 527–532). Springer.

Hakkani-Tür, D. Z., Kemal O., & Tür, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities, 36*(4), 381–410.

Hamzaoğlu, İ. (1993). *Machine translation from Turkish to other Turkic languages and an implementation for the Azeri language*. MA thesis. Boğazici University.

Hankamer, J. (1986). Finite state morphology and left to right phonology. In *Proceedings of the West Coast Conference on Formal Linguistics*, Vol. 5. Stanford Linguistic Association.

Hayran, A., & Sert, M. (2017). Sentiment analysis on microblog data based on word embedding and fusion techniques. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4.

Haznedaroğlu, A., & Arslan, L. M. (2014). Language model adaptation for automatic call transcription. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4102–4106.

Hemphill, C. T., Godfrey, J. J., & Doddington, G. R. (1990). The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings of the Workshop on Speech and Natural Language. HLT '90*. Hidden Valley, Pennsylvania: Association for Computational Linguistics, pp. 96–101. https://doi.org/10.3115/116580.116613.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*. Association for Computational Linguistics, pp. 29–32.

Ide, N., Calzolari, N., Eckle-Kohler, J., Gibbon, D., Hellmann, S., Lee, K., Nivre, J., & Romary, L. (2017). Community standards for linguistically-annotated resources. In *Handbook of Linguistic Annotation*. Springer, pp. 113–165.

İlgen, B., Adali, E., & Tantuğ, A. C. (2012, July). Building up lexical sample dataset for Turkish word sense disambiguation. In *2012 International Symposium on Innovations in Intelligent Systems and Applications* (pp. 1–5). IEEE

Inkelas, S., Küntay, A., Orhan Orgun, C., & Sprouse, R. (2000). Turkish Electronic Living Lexicon (TELL): A lexical database. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/86.pdf.

Kamholz, D., Pool, J., & Colowick, S. (2014). PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3145–3150. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1029_Paper.pdf.

Kara, N., Aslan, D. B., Marşan, B., Bakay, Ö., Ak, K. (2018a). Comparison of Turkish proposition banks by frame matching. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pp. 352-356. https://doi.org/10.1109/UBMK.2018.8566426.

Kara, N., Aslan, D. B. , Marşan, B., Bakay, O., Ak, K., & Yıldız, O. T. (2020a). TRopBank: Turkish PropBank V2.0. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 2763-2772).European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.336.

Kara, N., Marşan, B., Özçelik, M., Arıcan, B. N., Kuzgun, A., Cesur, N., Aslan, D. B., & Yıldız, O. T. (2020b). Creating a syntactically felicitous constituency treebank for Turkish. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1–6). https://doi.org/10.1109/ASYU50717.2020.9259873.

Kartal, Y. S., & Kutlu, M. (2020). TrClaim-19: The first collection for Turkish check-worthy claim detection with annotator rationales. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 386–395). Online: Association for Computational Linguistics, . https://doi.org/10.18653/v1/2020.conll-1.31. https://aclanthology.org/2020.conll-1.31.

Kaya, M. (2013). *Sentiment analysis of Turkish political columns with transfer learning*. MA thesis. Middle East Technical University.

Kayabaş, A., Schmid, H., Topcu, A. E., & Kılıç, Ö. (2019). TRMOR: A finite-state-based morphological analyzer for Turkish. *Turkish Journal of Electrical Engineering & Computer Sciences 27*(5), 3837–3851.

Kayadelen, T., Öztürel, A., & Bohnet, B. (2020). A gold standard dependency treebank for Turkish. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5156–5163). ISBN: 979-10-95546-34-4. https://www.aclweb.org/anthology/2020.lrec-1.634.

Kazemzadeh, A., Lee, S., Georgiou, P. G., & Narayanan, S. S.(2011). Emotion twenty questions: Toward a crowd-sourced theory of emotions. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 1–10). Springer.

Kirov, C., Sylak-Glassman, J., Que, R., & Yarowsky, D. (2016). Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3121–3126). Portorož, Slovenia: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1498.

Kılınç, D., Özçift, A., Bozyiğit, F., Yıldırım, P., Yücalar, F., & Borandağ, E. (2017). TTC-3600: A new benchmark dataset for Turkish text categorization. *Journal of Information Science, 43*(2), 174–185. https://doi.org/10.1177/0165551515620551.

Köksal, A (1975). *A first approach to a computerized model for the automatic morphological analysis of Turkish*. PhD thesis. Hacettepe University, Ankara.

Köksal, A. T., Bozal, O., Yürekli, E., & Gezici, G. (2020). #Turki$hTweets: A Benchmark Dataset for Turkish Text Correction. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4190–4198.). Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.findings-emnlp.374

Kolobov, R., Okhapkina, O., Omelchishina, O., Platunov, A., Bedyakin, R., Moshkin, V., Menshikov, D., & Mikhaylovskiy, N. (2021). MediaSpeech: Multilanguage ASR benchmark and dataset. In arXiv preprint arXiv:2103.16193.

Küçük, D., & Can, F. (2019). A tweet dataset annotated for named entity recognition and stance detection. arXiv: 1901.04787 [cs.CL].

Küçük, D., Jacquet, G., & Steinberger, R. (2014). Named entity recognition on Turkish tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 450–454). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/380_Paper.pdf.

Küçük, D., & Yazıcı, A. (2008). Identification of coreferential chains in video texts for semantic annotation of news videos. In *2008 23rd International Symposium on Computer and Information Sciences* (pp. 1–6). IEEE.

Küçük, D., & Yöndem, M. T. (2007). Automatic identification of pronominal Anaphora in Turkish texts. In *2007 22nd international symposium on computer and information sciences*. IEEE.

Kuriyozov, E., Doval, Y., & Gómez-Rodríguez, C. (2020). Cross-lingual word embeddings for Turkic languages. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 4054–4062). European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.499.

Kutlu, M., & Çiçekli, İ. (2013). A hybrid morphological disambiguation system for Turkish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 1230–1236). Asian Federation of Natural Language Processing. https://www.aclweb.org/anthology/I13-1175.

Kutlu, M., Çığır, C., & Çiçekli, İ. (2010). Generic text summarization for Turkish. *The Computer Journal, 53*(8), 1315–1323.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural questions: A benchmark for question answering research.

Transactions of the Association for Computational Linguistics, 7, 452–466. https://doi.org/10.1162/tacl_a_00276. www.aclweb.org/anthology/Q19-1026

Kyjánek, L., Žabokrtský, Z., Ševčíková, S., & Vidra, J. (2019). Universal derivations kickoff: A collection of harmonized derivational resources for eleven languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Charles University, Faculty of Mathematics, Physics, Institute of Formal, and Applied Linguistics, pp. 101-110. https://www.aclweb.org/anthology/W19-8512.

Ladhak, F., Durmuş, E., Cardie, C., & McKeown, K. (2020). WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4034–4048). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.360. https://www.aclweb.org/anthology/2020.findings-emnlp.360.

Lea, M., O'Shea, T., Fung, P., & Spears, R. (1992). 'Flaming' in computer-mediated communication: Observations, explanations, implications. In M. Lea (Ed.), *Contexts of computer-mediated communication* (pp. 89–112). Harvester Wheatsheaf.

Lewis, W. D. (2006). ODIN: A model for adapting and enriching legacy infrastructure. In *2006 Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)* (pp. 137–137). IEEE.

Lindén, K., Silfverberg, M., & Pirinen, T.. (2009). HFST tools for morphology—An efficient open-source package for construction of morphological analyzers. In C. Mahlow & M. Piotrowski (Eds.), *State of the art in computational morphology* (pp. 28–47).

Longpre, S., Lu, Y., & Daiber, J. (2020). MKQA: A linguistically diverse benchmark for multilingual open domain question answering. arXiv:2007.15207

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12*(2), 271–269. https://doi.org/10.1017/S0305000900006449.

Marcus, M. P., Santorini, B., & Ann Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.

Marşan, B., Kara, N., Özçelik, M., Arıcan, B. N., Cesur, N., Kuzgun, A., Sanıyar, E., Kuyrukçu, O., & Yıldız, O. T. (2021). Building the Turkish FrameNet. In *Proceedings of the 11th Global Wordnet Conference* (pp. 118–125). University of South Africa (UNISA): Global Wordnet Association. https://aclanthology.org/2021.gwc-1.14.

Megyesi, B., Dahlqvist, B., Csató, E. Á., & Nivre, J. (2010). The English- Swedish-Turkish parallel Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/116_Paper.pdf.

Megyesi, B., Dahlqvist, B., Pettersson, E., & Nivre, J. (2008). Swedish- Turkish parallel Treebank. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/121_paper.pdf.

Mengüşoğlu, E., & Deroo, O. (2001). Turkish LVCSR: Database preparation and language modeling for an aglutinative language. In *IEEE International Conference on Acoustics Speech And Signal Processing* (Vol. 6. 1999, pp. 4018–4018). IEEE.

Moran, S., Schikowski, R., Pajović, D., Hysi, C., & Stoll, S. (2015). The ACQDIV Corpus: A comparative longitudinal language acquisition corpus. Version 1.0.

Navigli, R., Simone, P. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217–250. ISSN: 0004-3702. https://doi.org/10.1016/j.artint.2012.07.001. http://www.sciencedirect.com/science/article/pii/S0004370212000793.

Nguyen, D. & Doğruöz, A. S. (2013). word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 857–862). Association for Computational Linguistics. https://www.aclweb.org/anthology/D13-1084.

Nguyen, D., Seza Doğruöz, A., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics, 42*(3), 537–593. https://doi.org/10.1162/COLI_a_00258.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 23-28.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing, 9*(2).

Oflazer, K., Çetinoğlu, Ö., & Say, B. (2004). Integrating morphology with multiword expression processing in Turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing* (pp. 64–71). Association for Computational Linguistics. https://aclanthology.org/W04-0409.

Oflazer, K., & Inkelas, S. (2006). The architecture and the implementation of a finite state pronunciation lexicon for Turkish. *Computer Speech & Language, 20*(1), 80–106.

Oflazer, K., & Saraçlar, M. (Eds.), (2018). *Turkish Natural Language Processing*. Theory and Applications of Natural Language Processing: Springer International Publishing. 9783319901657

Oflazer, K., Say, B., Hakkani-Tür, B. Z., & Tür, G. (2003). Building a Turkish treebank. In A. Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora* (pp. 261–277). Springer.

Oflazer, K., Yeniterzi, R., & Durgar-El Kahlout, İ. (2018). Statistical machine translation and Turkish. In K. Oflazer & M. Saraçlar (Ed.), *Theory and applications of natural language processing* (pp. 207–236). Springer. ISBN: 9783319901657.

Orhan, Z., Çelik, E., & Demirgüç, N. (2007). SemEval-2007 Task 12: Turkish lexical sample task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 59–63). Association for Computational Linguistics. https://www.aclweb.org/anthology/S07-1011.

Ortiz, S., Javier, P., Romary, L., & Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1703–1714). Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.156.

Ortiz, S., Javier, P., Sagot, B., & Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In P.Bañski, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen, & C. Iliadi (Eds.), *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019* (pp. 9–16). Cardiff, 22nd July 2019. Mannheim: Leibniz-Institut für Deutsche Sprache. https://doi.org/10.14618/ids-pub-9021. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

Özel, S. A., Öztürk, E., & Eşsiz, E. S. (2017). A new dataset for cyberbully detection from Turkish texts. In *5th International Conference on Natural and Engineering Sciences (ICNES)*. IEEE, pp. 366–370.

Özsoy, M. G., Alpaslan, F. N., & Çiçekli, İ. (2011). Text summarization using latent semantic analysis. *Journal of Information Science, 37*(4), 405–417.

Öztürel, A., Kayadelen, T., & Demirşahin, I. (2019). A syntactically expressive morphological analyzer for Turkish. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing* (pp. 65–75). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3110. https://www.aclweb.org/anthology/W19-3110.

Pala Er, N. (2009). *Turkish factoid question answering using answer pattern matching*. MA thesis. Bilkent University.

Palmer, M., Gildea, D, & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71–106. https://www.aclweb.org/anthology/J05-1004.

Pamay, T., Sulubacak, U., Torunoğlu-Selamet, D., & Eryiğit, G. (2015). The annotation process of the ITU web Treebank. In *Proceedings of The 9th Linguistic Annotation Workshop* (pp. 95–101). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-1610. https://www.aclweb.org/anthology/W15-1610.

Papalexakis, E., Nguyen, D., & Doğruöz, A. S. (2014). Predicting codeswitching in multilingual communication for immigrant communities. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 42–50). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3905. https://www.aclweb.org/anthology/W14-3905.

Paul, M., Federico, M., & Stüker, S. (2010). Overview of the IWSLT 2010 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT)*.

Polat, H., & Oyucu, S. (2020). Building a speech and text Corpus of Turkish: Large corpus collection with initial speech recognition results. *Symmetry 12*(2), 290.

Qazi, U., Imran, M., & Ofli, F. (2020). GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special 12*(1), 6–15.

Quasthoff, U., Goldhahn, D., & Eckart, T. (2014). Building large resources for text mining: The Leipzig corpora collection. In C. Biemann & A. Mehler (Ed.), Text mining. Theory and applications of natural language processing. Springer (pp. 3–24). ISBN: 978-3-319-12654-8. https://doi.org/10.1007/978-3-319-12655-5_1.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1264. https://www.aclweb.org/anthology/D16-1264.

Rogers, A., Baldwin, T., & Leins, K. (2021). Just what do you think you're doing, dave?' A checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 4821–4833). Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.414. https://aclanthology.org/2021.findings-emnlp.414.

Rothweiler, M. (2011). Turkish-German Successive-Bilinguals Corpus (TÜ_DE_cL2 Hamburg). Version 0.1. Publication date 2011-06-30. http://hdl.handle.net/11022/0000-0000-7D90-1.

Ruhi, Ş., Eröz-Tuğa, B., Hatipoğlu, Ç., Işık-Güler, H., Acar, M. G. C., Eryılmaz, K., Can, H., Karakaş, Ö, & Karadaş, D. Ç. (2010). 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* (Vol. 44).

Ruhi, Ş., Eryılmaz, K., & Acar, M. G. C. (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*, pp. 57–63.

Safaya, A., Kurtuluş, E., Göktoğan, A., & Yüret, D. (2022). Mukayese: Turkish NLP strikes back. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 846– 863). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.69. https://aclanthology.org/2022.findings-acl.69.

Şahin, G. G., & Adalı, E. (2018). Annotation of semantic roles for the Turkish proposition bank. *Language Resources and Evaluation, 52*(3), 673=706.

Sak, H., Güngör, T., & Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *International Conference on Natural Language Processing (GoTAL 2008)* (pp. 417–427). Springer.

Sak, H., Güngör, T., & Saraçlar, M. (2011). Resources for Turkish morphological processing. *Language Resources and Evaluation 45*(2), 249–261.

Salor, Ö., Pellom, B. L., Çiloğlu, T., & Demirekler, M. (2007). Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition. *Computer Speech & Language, 21*(4), 580–593. ISSN: 0885-2308. https://doi.org/10.1016/j.csl.2007.01.001.

Say, B. (2011). To build on the past for a better future in Turkish Natural Language Processing. In: Multisaund: Ulusal Konuşma ve Dil Teknolojileri Platformu Kuruluşu ve Türkçede Mevcut Durum Çalıştayı Bildirileri. Ed. by M Doğan. TÜBİTAK-BİLGEM. Gebze, pp. 54–56.

Say, B., Zeyrek, D., Oflazer, K., & Özge, U. (2002). Development of a Corpus and a TreeBank for present-day written Turkish. In *Proceedings of the Eleventh International Conference of Turkish Linguistics*. Eastern Mediterranean University, Cyprus.

Scherer, K. R, & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 66*(2), 310.

Schmid, H. (2005). A programming language for finite state transducers. In *Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing (FSMNLP 2005)*. Helsinki, pp. 308–309.

Schroeder, C., Schellhardt, C., Akıncı, M.-A., Dollnick, M., Dux, G., Gülbeyaz, E. I., Jähnert, A., Koç-Gültürk, C., Kühmstedt, P., Kuhn, F., Mezger, V., Pfaff, C., & Ürkmez, B. S. (2015). MULTILIT: Manual, criteria of transcription and analysis for German, Turkish and English. Ed. by Christoph Schroeder and Christin Schellhardt.

Schüller, P., Cingilli, K., Tunçer, F., Sürmeli, B. G., Pekel, A., Karatay, A. H., & Karakaş, H. E. (2018). Marmara Turkish Coreference Corpus and Coreference Resolution Baseline. In *CoRR abs/1706.01863*. arXiv: 1706.01863.

Schultz, T., Vu, T., Ngoc, & Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8126–8130). IEEE.

Schweter, S. (2020). BERTurk - BERT models for Turkish. *Version, 1*,. https://doi.org/10.5281/zenodo.3770924

Scialom, T., Dray, P. A. Lamprier, S., Piwowarski, B., & Staiano, J. (2020). MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8051–8067). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2020.emnlp-main.647. https://www.aclweb.org/anthology/2020.emnlp-main.647.

Şeker, G. A., & Eryiğit, G. (2017). Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content. *Semantic Web, 8*(5), 625–642.

Şen, M. U., & Erdoğan, H. (2014). Learning word representations for Turkish. In *2014 22nd Signal Processing and Communications Applications Conference (SIU)* (pp. 1742–1745). IEEE.

Sezer, T. (2017). TS Corpus Project: An online Turkish Dictionary and TS DIY Corpus. *European Journal of Language and Literature, 3*(3), 18–24.

Sezer, T., & Sever Sezer, B. (2013). TS corpus: Herkes için Türkçe derlem. In *Proceedings of the 27th Turkish National Linguistics Conference*, pp. 217–225.

Sezerer, E., Polatbilek, O., & Tekir, S. (2019). A Turkish dataset for gender identification of twitter users. In *Proceedings of the 13th Linguistic Annotation Workshop* (pp. 203–207). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4023. https://www.aclweb.org/anthology/W19-4023.

Sincan, Ö. M., & Keleş, H. Y. (2020). Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access, 8*, 181340–181355.

Speer, R., & Lowry-Duda, J. (2017). ConceptNet at SemEval-2017 Task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. https://doi.org/10.18653/v1/s17-2008.

Speer, R., Chin, J., & Havasi, C. (2018). ConceptNet 5.5: An open multilingual graph of general knowledge. arXiv: 1612.03975 [cs.CL].

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., & Grigoriadou, M. (2002). Balkanet: A multilingual Semantic Network for Balkan Languages. In *Proceedings of the First Global WordNet Conference*.

Sulubacak, U., Gökırmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., & Eryiğit, G. (2016). Universal dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3444–3454). http://aclweb.org/anthology/C16-1325.

Šuster, S., Tulkens, S., & Daelemans, W. (2017). A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 80–87). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1610. https://aclanthology.org/W17-1610.

Sylak-Glassman, J., Kirov, C., Post, M., Que, R., & David, Y. (2015). A universal feature schema for rich morphological annotation and fine-grained cross- lingual part-of-speech tagging. In *International Workshop on Systems and Frameworks for Computational Morphology* (pp. 72–93). Springer.

Tantuğ, A. C., Adalı, E., & Oflazer, K. (2007). A MT system from Turkmen to Turkish employing finite state and statistical methods. In *Machine Translation Summit XI. European Association for Machine Translation (EAMT)*.

Tantuğ, A. C., & Adalı, E. (2018). Machine translation between Turkic languages. In K. Oflazer & M. Saraçlar (Ed.), *Turkish Natural Language Processing* (pp. 317–336). Springer International Publishing.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology, 63*(1), 163–173. https://doi.org/10.1002/asi.21662.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2214–2218). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Toçoğlu, M. A., & Alpkoçak, A. (2018). TREMO: A dataset for emotion analysis in Turkish. *Journal of Information Science, 4*(6), 848–860. https://doi.org/10.1177/0165551518761014.

Toçoğlu, M. A., & Alpkoçak, A. (2019). Lexicon-based emotion analysis in Turkish. *Turkish Journal of Electrical Engineering & Computer Sciences, 27*(2), 1213–1227.

Toçoğlu, M. A., & Öztürkmenoğlu, O., & Alpkoçak, A. (2019). Emotion analysis from Turkish tweets using deep neural Networks. *IEEE Access, 7*, 183061–183069. https://doi.org/10.1109/ACCESS.2019.2960113

Topkaya, İ. S., & Erdoğan, H. (2012). SUTAV: A Turkish audio-visual database. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2334–2337). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/483_Paper.pdf.

Toraman, Ç., Şahinuç, F., & Yılmaz, E. H. (2022). Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Language Resources and Evaluation Conference* (pp. 2215–2225). European Language Resources Association. https://aclanthology.org/2022.lrec-1.238.

Tür, G., Hakkani-Tür, D., & Oflazer, K.. (2003). A statistical information extraction system for Turkish. *Natural Language Engineering, 9*(2), 181–210. https://doi.org/10.1017/S135132490200284X.

Türk, U., Atmaca, F., Özateş, Ş. B., Başaran, B. Ö., Güngör, T., & Özgür, A. (2019). Improving the annotations in the Turkish universal Dependency Treebank. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)* (pp. 108–115). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-8013. https://www.aclweb.org/anthology/W19-8013.

Türk, U., Atmaca, F., Özateş, Ş. B., Berk, G., Bedir, S. T., Köksal, A., Başaran, B. Ö., Güngör, T., & Özgür, A. (2022). Resources for Turkish dependency parsing: Introducing the BOUN Treebank and the BoAT annotation tool. *Language Resources and Evaluation, 56*, 259–307. https://doi.org/10.1007/s10579-021-09558-0

Türkmenoğlu, C., & Tantuğ, A. C. (2014). Sentiment analysis in Turkish media. In *International Conference on Machine Learning (ICML)*.

Türkoğlu, F., Diri, B., & Amasyalı, M. F. (2007). Author attribution of Turkish texts by feature mining. In *International Conference on Intelligent Computing* (pp. 1086–1093). Springer.

Tyers, F. M., & Alperen, M. S. (2010). South-East European times: A parallel corpus of Balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pp. 49–53

Upadhyay, S., Faruqui, M., Tür, G., Dilek, H. T., & Heck, L. (2018). (Almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE In- ternational Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6034–6038.

Van der Goot, R. & Çetinoğlu, Ö. (2021). Lexical normalization for code-switched data and its effect on POS tagging. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 1, Long Papers. Association for Computational Linguistics.

Vossen, P. (Ed.), (1998). *EuroWordNet: A multilingual database with Lexical semantic networks*. Kluwer Academic Publishers. ISBN: 978-94-017-1491-4.

Vural, A. G. (2013). *Sentiment-focused web crawling*. PhD thesis. Middle East Technical University.

Wieling, M., Rawee, J., & van Gertjan, N. (2018). Reproducibility in computational linguistics: Are we willing to Share? *Computational Linguistics 44*(4), 641–649. https://doi.org/10.1162/coli_a_00330. https://www.aclweb.org/anthology/J18-4003.

Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., Jahns, E., Klotz, M., Krause, T., Labrenz, A., Lüdeling, A., Martynova, M., Neuhaus, K., Pashkova, T., Rizou, V., Rosemarie, T., Schroeder, C., Szucsich, L., Tsehaye, W., Zuban, Y. (2020). RUEG Corpus. *Version*(3). https://doi.org/10.5281/zenodo.3765218

Williams, A., Nangia, N., & Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1 (Long Papers), pp. 1112–1122). New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101. https://www.aclweb.org/anthology/N18-1101.

Wołk, K., & Marasek, K. (2014). Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs. In *Procedia Technology 18. International workshop on Innovations in Information and Communication Science and Technology, IICST 2014, 3-5 September 2014* (pp. 126–132). ISSN: 2212- 0173. https://doi.org/10.1016/j.protcy.2014.11.024. http://www.sciencedirect.com/science/article/pii/S2212017314005453.

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., et al. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language, 31*(4), 461-479. https://doi.org/10.1177/0142723711409976.

Yeniterzi, R. (2011). Exploiting morphology in Turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session* (pp. 105–110). Association for Computational Linguistics. https://www.aclweb.org/anthology/P11-3019.

Yirmibeşoğlu, Z., & Eryiğit, G. (2018). Detecting code-switching between Turkish-English language pair. In *Proceedings of the 2018 EMNLP Workshop WNUT: The 4th Workshop on Noisy User-generated Text* (pp. 110–115). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6115. https://www.aclweb.org/anthology/W18-6115.

Yıldız, O. T., Solak, E., Görgün, O., & Ehsani, R. (2014). Constructing a Turkish-English parallel TreeBank. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

(Volume 2: Short Papers, pp. 112–117). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2019. https://www.aclweb.org/anthology/P14-2019.

Yüret, D., & Türe, F. (2006) Learning morphological disambiguation rules for Turkish. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. HLT-NAACL '06. New York, pp. 328–334. https://doi.org/10.3115/1220835.1220877.

Yıldırım, E., Çetin, F. S., Eryiğit, G., & Temel, T. (2015). The impact of NLP on Turkish sentiment analysis. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 7*(1), 43–51.

Zaghouani, W. (2014). Critical survey of the freely available Arabic Corpora. In *Proceedings of the LREC 2014 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, pp. 1–8.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, Long and Short Papers, pp. 1415–1420). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1144. https://www.aclweb.org/anthology/N19-1144.

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinkova, S., Hajic Jr., J., Hlavacova, J., Kettnerová, V., et al. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 1–19). Association for Computational Linguistics. http://www.aclweb.org/anthology/K/K17/K17-3001.pdf.

Zeyrek, D., & Başıbüyük, K. (2019). TCL—A Lexicon of Turkish discourse connectives. In *Proceedings of the First International Workshop on Designing Meaning Representations* (pp. 73–81). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3308. https://www.aclweb.org/anthology/W19-3308.

Zeyrek, D., Demirşahin, I. B. 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.

Zeyrek, D., & Kurfalı, M. (2017). TDB 1.1: Extensions on Turkish Discourse Bank. In *Proceedings of the 11th Linguistic Annotation Workshop* (pp. 76–81). Association for Computational Linguistics https://doi.org/10.18653/v1/W17-0809. https://www.aclweb.org/anthology/W17-0809.

Zeyrek, D., Mendes, A., Grishina, Y. Kurfalı, M., Gibbon, S., & Ogrodniczuk, M. (2020). TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation 54*(2), 587–613.

Zeyrek, D., Mendes, A., & Kurfalı, M. (2018). Multilingual extension of PDTB-Style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language