

Werk

Titel: Report

Ort: Wiesbaden

Jahr: 2015

PURL: https://resolver.sub.uni-goettingen.de/purl?666048797_0019 | LOG_0030

Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

✉ info@digizeitschriften.de

Report

Parallel corpora and Universal Dependencies for Turkic

Éva Á. Csató & Hüner Kaşıkara & Beáta Megyesi & Joakim Nivre

Csató, Éva Á. & Kaşıkara, Hüner & Megyesi, Beáta & Nivre, Joakim 2015. Parallel corpora and Universal Dependency for Turkic. *Turkic Languages* 19, 259–273.

The first part of this paper presents ongoing work on Turkic parallel corpora at the Department of Linguistics and Philology, Uppsala University. Moreover, examples are given of how the Swedish-Turkish-English corpus is used in teaching Turkish and in comparative linguistic studies. The second part deals with the annotation scheme Universal Dependencies (UD) used in treebanks, and its application to Turkic languages.

*Éva Á. Csató, Beáta Megyesi, and Joakim Nivre, Department of Linguistics and Philology, Uppsala University, Box 635, SE-751 26 Uppsala, Sweden. E-mail: eva.csato@lingfil.uu.se, beata.megyesi@lingfil.uu.se, joakim.nivre@lingfil.uu.se
Hüner Kaşıkara, Boğaziçi University, Istanbul, Turkey. E-mail: huner.kasikara@boun.edu.tr*

Turkic languages at Uppsala University

The Faculty of Languages at Uppsala University, Sweden, is recognized for maintaining a rich linguistic environment. Regular courses are offered in over forty languages including both dead languages and large, small and vanishing contemporary vernaculars. Turkic languages are well represented in the curriculum. Besides Turkish, which is the main focus in teaching, there are campus and distance courses in the classical literary language Ottoman, modern Turkic languages such as Uzbek, Uyghur, Kazakh, and Azeri, and the highly endangered Karaim language. The Department of Linguistics and Philology, where the Turkic languages program is situated, has a tradition of supporting scholarly environments for lesser taught languages. For several years the department has received financial aid to build up resources for teaching and research. Thanks to this support, Turkic linguistics at Uppsala can offer a relatively large number of language courses and conduct research in comparative Turkic linguistics.

Some projects at the department aim at giving support to lesser taught languages through developing parallel treebanks. Treebanks are parsed corpora, collections of texts with morphological and syntactic annotation. Treebank projects are carried out by experts in less resourced languages such as Hindi, Persian and Turkic together

with internationally outstanding scholars in computational linguistics working at the department. The collaboration has proved to be mutually advantageous. The focus in language technology has been on English and major Western languages. English is one of the languages that has been included in many of the existing parallel treebanks. Access to data on languages representing language types that are widespread in the world, as for instance Turkic, provides excellent opportunities for computational linguistics to develop their language technological theories and tools.

Part 1: Parallel corpora

A parallel corpus is a collection of multilingual text material containing original texts in one language and translations into one or more other languages, with texts placed alongside their corresponding translations. Parallel texts are usually aligned, meaning that corresponding units (sentences, phrases, or even words) from the different language versions are explicitly linked together. The texts can be linguistically annotated with respect to part-of-speech and morphological features, and on the syntactic level. Syntactically annotated parallel corpora are called parallel treebanks.

A number of parallel corpora are available on the Internet, one example being the Farkas Translations (http://www.farkastranslations.com/bilingual_books.php). Farkas presents out-of-copyright literary works in sentence aligned format without any further linguistic analysis. For instance Mark Twain's novel *The Adventures of Tom Sawyer* can be found in this corpus in the English original and in German, Hungarian, Dutch and Catalan translations. The resource is provided free of charge for the purpose of language learning, language teaching and translation research, and as a demonstration of the text alignment services offered by the website. Farkas has accessed many of the texts published in the Gutenberg Project, a digital library of free books which also can serve as an excellent resource for further projects aiming to build new parallel corpora and other language resources (https://www.gutenberg.org/wiki/Main_Page).

A computational linguist at Uppsala, Jörg Tiedemann, has built a much more advanced collection of multilingual parallel corpora, OPUS, providing various tools and interfaces and a growing collection of language samples collected from the web for building parallel corpora and related tools (<http://opus.lingfil.uu.se/>). OPUS provides freely available data sets in various formats together with basic annotation to make it useful for applications in computational linguistics, translation studies and cross-linguistic corpus studies. OPUS is probably the largest collection of freely available parallel corpora in the world. It covers over 90 languages and includes data from several domains. Altogether, there are over 3,800 language pairs with sentence-aligned data comprising a total of over 40 billion tokens in 2.7 billion parallel units (aligned sentences and sentence fragments). Including a parallel corpus for Romanian-Turkish. Another improvement of recent versions of OPUS is the availability of various download formats for all sub-corpora. Different tools are available on the OPUS webpage for language-specific taggers and parsers, including for Turk-

ish, as well as alignment tools to automatically link sentences, phrases and words. The texts are processed by Uplug (<https://bitbucket.org/tiedemann/uplug>), a collection of tools and scripts for processing text-corpora to create (annotated) parallel corpora (Tiedemann 2003, 2012).

Building a Swedish-Turkish-English parallel corpus

The Uppsala parallel corpus containing parallel texts in Turkish, Swedish and English is the first of its kind (<http://www2.lingfil.uu.se/corpora/>) and was created between 2006 and 2010, as previously described (Megyesi et al. 2006; Megyesi & Dahlqvist 2007; Megyesi et al. 2008; Megyesi et al. 2010). The corpus contains syntactically annotated parallel texts with various annotation layers ranging from part-of-speech tags and morphological features to dependency annotation. Each layer is automatically annotated and the sentences and words are aligned, and the results are partly manually corrected.

In order to build the treebank automatically, a basic language resource kit (BLARK) was used for the included languages. This consists of (i) tokenizers to segment words and punctuation marks and to mark sentence endings, (ii) taggers for the annotation of part-of-speech and morphological features, (iii) parsers for the annotation of syntactic structures in terms of dependency relations, and (iv) aligners to mark the related sentences and words in the translations. Also, we developed tools for the manual correction of the automatic linguistic annotation and alignment. The tools were included in a pipeline for easier processing. The annotation procedure is shown in Figure 1 below.

First, the original materials received from publishers in various formats were normalized to create a consistent, machine-readable format across the corpus data. For example, rtf, doc, and pdf documents were converted into plain text files. After cleaning up the original data, the texts were processed automatically by using tools for formatting, linguistic annotation and sentence and word alignment. During formatting, the texts were encoded using UTF-8 (Unicode) and marked up structurally using XML Treebank Encoding Standard (XCES). The text files were processed by various tools in the BLARKs developed for each language separately.

A tokenizer was used to split the text into tokens such as words and punctuation marks. Sentence segmentation was also performed to break the texts into sentences. Once the sentences and tokens were identified, the data was linguistically analyzed. The annotation was represented in several annotation layers, first on a morphological level, then on a syntactic level. For the linguistic annotation, external morphological analyzers, part-of-speech taggers and syntactic dependency parsers were used, which were trained on annotated treebanks developed for the specific languages. The annotations and labels for linguistic analysis were the de facto standards for the various languages at that time. For example, for Swedish we used the Stockholm Umeå Corpus tagset (SUC 1997) for the morpho-syntactic annotation and the functional annotation of Talbanken05 (Nivre et al. 2006b), while for the syntactic

analysis of Turkish we derived the linguistic annotation from the METU-SABANCI Turkish Treebank (Oflazer et al. 2003). For English, we used the Penn Treebank tagset.

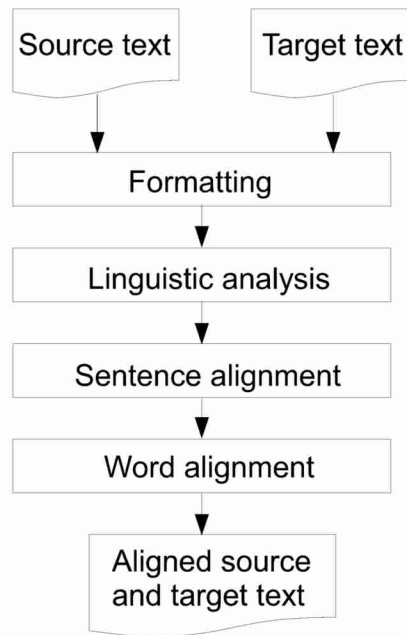


Figure 1. Annotation procedure

The Swedish and English texts were morphologically annotated with the open source HunPoS tagger (Halácsy et al. 2007). The tokens were annotated with parts of speech and morphological features and then disambiguated. The results for the morphological annotation of Swedish show an accuracy of 96.6% (Megyesi 2008). The Turkish material was morphologically analyzed and disambiguated using a Turkish analyzer (Oflazer 1994) and a disambiguator (Yuret & Türe 2006) with an accuracy of 96%. The English data contained less error, approximately 2–3%.

The other linguistic layer contains information about the syntactic analysis. We use dependency rather than constituent structures, as the former has been shown to be well suited for both morphologically rich and free-word-order languages such as Turkish, and for morphologically simpler languages like Swedish. The English, Swedish and Turkish data were annotated syntactically using MaltParser (Nivre et

al. 2006a), and trained on Penn Treebank, Talbanken05 (Nivre et al. 2006b), and the Metu-Sabancı Turkish Treebank (Oflazer et al. 2003) respectively. The annotation includes approximately 15–20% errors, depending of the language, which need to be manually corrected.

The processing tools were implemented in a framework with a graphical user interface, UplugConnector (Megyesi and Dahlqvist 2007), which is based on the modules in the Uplug toolkit (Tiedemann 2003). Our goal was to produce user-friendly tools to make annotation, alignment and correction easy for people with little computer skills.

Resources in the corpus

The corpus includes texts representing both fiction—novels and short stories—and non-fiction, such as information material for immigrants, a book on political history, and texts issued by international organizations. Most of the original texts are in either Swedish or Turkish accompanied by a translation into the other languages. English translations are in several cases missing. The corpus consists of 288 701 tokens in Swedish, 162 302 in Turkish and 140 848 in English, comprising both fiction and technical documents. In order to make the corpus useful for users, search functions are included. Users can search for complete words, beginnings of words, ends of words, and parts of words, being case-sensitive or not. The target language can be Swedish or Turkish and the corpus can be limited to one specific text or text type. Turkish letters are provided.

Another tool produces frequency lists of a freely selectable Turkish text, which has to be copied into a window. The frequency list may be displayed in different formats and contains the number of words, number of different words, average length of words, and type-token ratio in the given text.

Using the corpus in teaching and research

The aim of the Swedish-Turkish treebank is to provide Swedish speaking students and researchers with easily accessible annotated linguistic data on Turkish. The corpus is now being complemented with English texts. The web-based corpus can be used both by regular and distance students in their data-driven acquisition of new vocabulary items and usage. It also functions as a learning platform and for testing hypotheses concerning the morphological and syntactic aspects of Turkish grammar. It further helps students to practice translation between Swedish and Turkish. All this is possible because the Swedish-Turkish parallel texts are available in annotated form. On request, the annotations are visualized in pop-up windows. The morphological analyses are currently given in clumsy, parser-generated formulas but in the near future they will be replaced by more intelligible labels based on the grammatical terms employed in textbooks and in the Turkish Suffix Dictionary (Csató and Nathan 2003). The interface for displaying syntactic information is not ready yet. A search tool assists the students in creating concordance lists. The concordance lists

display whole sentences where the target item appears (and is highlighted). The selected sentences are aligned with their translational equivalents, as illustrated in Figure 2.

 Sökresultat Text: Vår Dag Sökräng: 0:00 Antal funna meningar: 88 Antal förekomster: 82		
num	swedish	turkish
28	Jag följer dem en tid, men blev uttröad, och kom svar från italienska universitet: som gjorde slut på mitt hopp. Även efterforskningarna, jag gjorde på kyrkogårdarna i Gebze, Cennethisar och Uskudar där, jag letade efter författarens namn på övfruktanerna. Jag slutade jaga spår och tog med författaren i önskelöbet med hjälp av uppgifterna i boken.	Bir süre on anın peşinden gittim, ama bıkmıştım çünkü, mekûp yağmuruna tutuşum, İtalyan üniversitelerinden umut kırıcı cevaplar geliyordu. Gebze, Cennethisar ve Uskudar mezarlıklarında yazarın kitabın kendisinden çıkarttığı ama üzerinde yazmayan adına dayanarak yaptığım araştırmalar da başarısız çıkmıştı. İz sürmeyi bıraktım, ansiklopedi maddesini hikâyesinin kendisine dayanarak yazdım.
55	Vår kapten började hoppas när han såg hur de två andra skeppen svingade sig fram mellan de turkiska fartygen och försvarshinderna, och till slut fick han, efter våra påtryckningar, med att låta piska slavarne, men nu var det för sent; ossuom kunde inte ens piskorna ta de av frihetslängtan uppnådda slavarne att yda.	Ötek iki geminin Türk gemilerinin arasından sıyrılıp sisin içinde kayboluşunu görünce kapitanımız umutlandı, bizim de zorumuzla esirleri sıkıştırmaya cesaret ocebildi, ama geç kalmıştık çünkü, üstelik özgürlük tutkusuyla hayecanların kalelere kırılganlar da söz geçiremiyordu.
105	Folk hade hört att jag var läkare, jag behandlade inte bara savares som ruttrade, vårt förgelse utan även andra.	Yalnız zindanda çürüyen köslere değil, hekim olduğumu işiten başka anıta da bak yordum çünkü.
141	Jag fick fortsätta arbeta men nu behandlades jag förmånligt av s avdrivarna.	Gere işe çıkarılıyordum, ama esirbaş anıta kayıyerardı beni.

Figure 2. Concordance list

Such lists are used to find frequent patterns of usage, transformational equivalents, different meanings of polysemic words, translational equivalents of Turkish grammatical categories, etc. Different types of exercises are designed and published on the Internet. Students of Turkic languages also use the corpus while writing their theses. Bergdahl (2006) studied the meanings of the Turkish word *gölge* ‘shadow’ and the corresponding Swedish word *skugga*. Dadasheva (2005) investigated how the Turkish indirective category marked by *-miş / imiş* is translated into Swedish and Russian. Hedman and Uyghur (2009) compared the meanings of the Swedish and Turkish verbs ‘to give’, ‘to do’ and ‘to make’. Haktanır (2006) reviewed the ambiguous Turkish morphological forms in one of the parallel texts and described different types of morphological ambiguities.

Apart from being used in learning environments, the corpus is also used in research. One example is the article “Rendering evidential meanings in Turkish and Swedish” (Csató 2009), which examined the Turkish evidential category of indirectivity and the less grammaticalized or lexical strategies used in Swedish to render evidential nuances. The description of the strategies used in the two languages was complemented with an analysis of data from one of the parallel Turkish-Swedish texts. It was found that although Swedish has means to express evidential

nuances, they were used much less in the Swedish translations than expected. The article describes several possible reasons for this. One is that the Turkish category allows three different types of reading. This ambiguity is significant in certain texts. The Swedish devices can render a particular evidential nuance, but not the whole range of ambiguity of the Turkish forms.

Turkic-Turkic corpora

The development of parallel Turkic-Turkic corpora is a project we would like to carry out in an international collaboration. Our corpus includes a Turkmen novel, Ak Welsapar's *Kepjebaş*, in the original and in Turkish and Swedish translations. The texts have been entered into the database and the sentence alignment is currently underway. The Turkic-Turkic parallel corpora would mostly serve Turcological research interest. Turkic languages are relatively similar, and their morphologies and syntax can easily be annotated in a coordinated way. This raises the need to implement some measures of standardization with regard to transliteration/transcription, and annotation.

Part 2. Universal Dependencies

The annotation of the Uppsala Parallel Corpus as described above was based on de facto standards in 2010. Since then, new annotation standards have been developed for the included languages, especially concerning the syntactic annotation in terms of dependency structures. A recent development is our effort to contribute to the Universal Dependencies (UD) for Turkic languages project. UD aims to develop cross-linguistically consistent treebank annotation for many languages, that is, syntactically annotated corpora that can be used in natural language processing and linguistics, and that allow for meaningful comparisons across typologically diverse languages.

As far as Turkish is concerned, there are two existing treebanks: METU-SABANCI and ITU-METU-SABANCI. Both use dependency structure annotation and the latter builds on the former. Dependency parsing of Turkish has been studied by Eryigit et al. (2008); part-of-speech tagging for Turkish has been investigated by Dincer et al. (2008); and morphological processing of Turkish has been pursued in the work of Sak et al. (2011). Apart from Turkish, work on morphological tagging has been carried out for: Uyghur (Altenbek 2006) and Kazakh (Makazhanov et al. 2014). The listed works target the treebanks and the annotation methods but none have used Universal Dependencies.

The UD subproject has two goals. The first is to investigate the methods for implementing UD for Turkish and provide guidelines. The second is to help with the ongoing development of conversion methods from the ITU-METU-SABANCI treebank to the UD scheme. In order to discuss the issues of conversion and methods for implementing UD to Turkish, a workshop has been arranged involving the team(s) at ITU (Istanbul Technical University) and Francis Tyers from The Arctic Universi-

ty of Norway and interested parties at Uppsala University. We hope that in the future, the discussions will be broader and the UD scheme will be extended to Turkic languages in general.

What is Universal Dependencies?

Universal Dependencies (UD) is an annotation scheme the main goal of which is to provide guidelines for consistent annotation of similar constructions across languages while also allowing language specific extensions, where necessary. Currently 33 languages have been annotated according to the UD scheme and work is ongoing on Turkish and Kazakh. The goals and characteristics of UD have been discussed in Nivre (2015). The following brief introduction to Universal Dependencies is based on Nivre (2015) and the UD documentation.¹

In order to achieve the goal of cross-linguistically consistent treebanks, UD works on different layers. These are tokenization, morphology and syntax. Words are the basic units of annotation in UD and they are considered as syntactic words. The goal of tokenization is to segment the words between which the dependency relations hold. Thus, syntax represents the relations between words. Each word is either a dependent of another word in the sentence or the root of the sentence. The morphological specification of a syntactic word or a unit in the UD scheme consists of lemma, part-of-speech (POS) tag, and a set of features. The lemma represents the semantic content of a unit. The POS tag represents the abstract lexical category associated with the unit. Finally, the features represent the lexical and grammatical properties that are associated with a particular word form or lemma. Every feature has the form Name=Value where Name is the feature name and Value is the value of the feature. If a feature is not mentioned in the data, this implies an empty value. When there is more than one feature for a word, they are ordered alphabetically. A fully annotated text in UD contains the right features and categories of dependency relationships between segmented words.

Application of Universal Dependency (UD) to Turkish

The first part of this section will list the problems and possible solutions related to the verbal lexical category of Turkish when we implement the UD scheme. We have chosen first to implement the scheme for one lexical category and make an exhaustive study of it rather than giving brief descriptions of each lexical category. It seems appropriate to start with verbs. In the second part we will report on the decisions made during the workshop at Uppsala University regarding a conversion method for the ITU-METU-Sabancı treebank.

1 For the documentation on Universal Dependencies together with the languages that are annotated with this scheme; see <http://universaldependencies.org/>

Universal Dependency annotation of Turkish verbs

We will here deal with the subcategorization of verb forms. We first deal with finite verb forms and then with non-finite verb forms without their nominal morphology.

Finite verbs

The following morphological issues, which are problematic for the annotation schemes will be presented: the segmentation of an orthographic word, multiple voice, multiple tense, multiple aspect and multiple modality.

Segmentation of an orthographic word: In UD, clitics are separated from their host and treated as separate words in most but not all cases. We will discuss whether the clitics in Turkish should be separated or not. We will look at three types of clitics, namely; (i) the clitics *idi* (denoting past tense), *imiş* (denoting indirectivity/evidentiality), *ise* (denoting hypothetical mood); (ii) the generalized modality marker *-Dir* which conveys different modal notions such as presumption; and (iii) the question particle *mi*.

The clitics in (i) and (ii) add grammaticalized meanings to the proposition. UD represents the features of words (or units) in the form of a morphological feature list. We choose not to split these types of clitics because they can be represented in the morphological features of the host as a feature value.

The third type of clitic, the question particle, is orthographically separated and the existing treebanks consider it as a separate token. We propose that it can be kept separate from the verb and have an *aux* dependency relation with the verb. Consequently, it carries the feature “question” and in cases where it bears any agreement feature, these can be represented in its feature list.

Multiple voice: Turkish allows verb to carry more than one voice morpheme. The reflexive suffix is unproductive and the reciprocal suffix combines with very few stems. Thus they can be viewed as separate lexical entries. The double passive usage is also very limited and, when present, is perceived as a single passive. The extreme cases of multiple voice are observed when there are multiple causative voices or when a causative and passive suffix are observed together, example (1). When some features are marked more than once on the same word, UD suggests the usage of *layered features*.² We propose to apply the layered features idea to multiple voice marking in Turkish verbs. In this case, the voice features of example (1) will be: Voice[1]=Caus|Voice[2]=Pass. Note that the numbering in the brackets indicates multiple layers of the feature. However, the numbering does not necessarily indicate an ordering between the morphemes.

2 *Layered features:* According to the UD documentation; “when some features are marked more than once on the same word, it is said that there are several layers of the feature”. This type of representation has been implemented for Basque verbs in the agreement paradigm.

- (1) *Giy-dir-il-di.*
 wear-CAUS-PASS-PAST-3SG
 ‘It was made put on.’

Multiple tense marker: The only tense that is marked by the morphemes is the past tense morpheme *-DI* (TAM1) or clitic *idi* (TAM2). In the absence of the past tense clitic *idi*, the TAM1 morphemes have other meanings than tense. As a result, the multiple tense issue is resolved trivially.

Multiple aspect/viewpoint: No two morphemes come together in such a way that both represent aspect/viewpoint syntactically thus cause a multiple aspect/viewpoint problem in the Turkish verbal system. Postverbs that mark actionality modification, such as *-(y)Adur-* marking durativity or repeatedness, *-(y)Iver-* marking ‘to do something uncontrolled or quickly’, should be treated as derivations.

Multiple modality: In the current version³ of UD, evidentiality is analyzed under modality. However, for Turkic languages evidentiality has to be analyzed as a separate feature just like tense, aspect and modality. We propose that an evidentiality feature can/should be added to the feature list in UD to cover languages like Turkish that mark evidentiality separately.

Another issues is the marker *-(y)Abil*, originally a combination of a converb and a postverb, but which has been further grammaticalized as marker of possibility/probability/capability. Possibility can also be combined with the necessitative, presumptive, and hypothetical markers. The negated form of the possibility marker can be combined with its affirmative form; see (2).

- (2) *Birinci ol-ama-yabil-ir.*
 first be-NEGATED.POSSIBILITY-POSSIBILITY-AORIST3SG
 ‘S/he might not be able to be first.’

Each case can be resolved in a similar fashion as multiple voice marking, by using layered features. The feature representation of example (2) is Mood[1]=Pot|Mood[2]=Pot.

To sum up, we have gathered all of the proposals that we have made in this section in Table 1. On the left side, the issues are listed, and on the right side, under Comments, our suggested proposal within the UD scheme can be found.

Issues on Finite Verbs		Comments
Segmentation of clitics	Clitic Type	
	TAM2	Not segmented
	Presumptive	Not segmented
	Question Particle	Orthographically segmented. It has an aux-dependency relation with the host
Tense	Morpheme(s)	
	<i>-mİş</i> or <i>imiş</i>	These forms refer to evidentiality (Turkic indirectivity). A new feature that represents this is needed.
	<i>-(y)AcAkl</i>	The combination of the two morphemes corresponds to future past tense (English)
Actionality (lexical aspect)		Lexical aspect should be viewed as derivation. In this case there is no multiple aspect.
Modality		Layered features
Voice		Layered features

Table 1. Summary of the proposal for finite verbs

Non-finite verbs

Between verb root and subordinating suffixes the following morphology could be found: voice, negation and mood/aspect. This means that the voice problem we have observed in finite verbs will be repeated for non-finite verbs. However, the layered feature solution is still applicable.

Another issue with the non-finite verbs relates to terminology. Studies on Turkic languages group the subordinating suffixes into three categories. These are: verbal nouns where the subordinated verb functions as predicate in a nominal clause, participles where it functions as predicate in a relative clause, and converbs where it functions as an adverbial clause.

The definition of transgressive in the UD documentation corresponds to that of the converbs. Since the goal of UD is not to define similar things differently, we will be using transgressive instead of introducing a new value under verbform. The basic syntactic dependency relationship for transgressives is the *advcl*. When there is a postposition, there should be a *case* relation between the converb and the postpositional element. Some of the complex verbal forms are treated as multiword expressions in the existing treebanks.

The existing treebanks treat *-mAk*, *-mA*, *-(y)İş* as infinitives. We would like to preserve the insight of these treebanks and keep them separate from gerunds.

The issues related to non-finite verbs are summarized in Table 2, including other types of verbforms in this category.

Issues on non-finite verbs	Comments	
VerbForm	Terminology	
	Infinitive	<i>-mA, -mAk, -(y)Is</i>
	Gerund (verbal noun)	<i>-(y)AcAk, -DIK</i>
	Participle	<i>-(y)AcAk, -DIK, -(y)An</i>
	Transgressive	converbs such as <i>-(y)Ip, -(y)ArAk</i> etc.
Voice	Voice problem is similar to that with finite verbs and is resolved by the same method.	

Table 2. Summary of the proposals for non-finite verbs

Conversion

A workshop has been arranged to discuss the details of the conversion of the existing ITU-METU-Sabancı treebank to the Universal Dependency scheme. The workshop was organized by the Department of Linguistics and Philology, Uppsala University in 26–27 November 2015. The final version of the conversion will be contributed by the team at Istanbul Technical University, who are working on the software that converts each particular treebank scheme to the UD format. The following are the results of the discussions.

During the workshop, mainly the syntax-related issues were discussed. These involve the types of dependency relationships. Some differences can be solved by changing labels. PREDICATE, PUNCTUATION, DETERMINER, COORDINATION and, Intensifier in ITU-METU-SABANCI treebank are re-labeled as *root*, *punct*, *det*, *conj* and *advmod*, respectively.

The basic difference between the annotation scheme used in the ITU-METU-SABANCI treebank and the UD scheme is that UD makes a distinction between clausal and non-clausal dependency relations. For example, a MODIFIER in ITU-METU-SABANCI annotation may correspond to a clausal adjective or just a lexical adjective. In light of this the following conversion table, Table 3, has been agreed upon. Under the ITU-METU-SABANCI label, the names of dependency relations are listed as they are found in the treebank. Under the UD label are the corresponding relations in the UD listed. Each label in the ITU-METU-SABANCI treebank has more than one corresponding UD label. The definition of the relation determines what type of UD label will be used.

The ITU-METU-SABANCI annotation method uses umbrella terms which correspond to more detailed descriptions in UD. APPPOSITION, POSSESSOR, ARGUMENT, MODIFIER and VOCATIVE in ITU-METU-SABANCI treebank correspond to a variety of relations in UD.

ITU-METU-SABANCI label	UD label	Definition of the relation
RELATIVIZER	Ccomp	Complement clause
	advcl	Adverbial clause
	acl	Adjective clause
MODIFIER	acl	Adjective clause
	amod	Adjective
	advcl	Adverbial clause
	advmod	Adverb
OBJECT	ccomp	Complement clause
	dobj	Direct object
SUBJECT	csubj	Subject clause (Any clause that functions as the subject)
	nsbj	Subject

Table 3. Conversion table

UD and Turkic

Applying the UD scheme to Turkic languages is a new movement. There are ongoing projects with Turkish and Kazakh. Makazhanov et al. (2015) is an example of applying UD to a Kazakh treebank. Çöltekin (2015) lists the issues related to Turkish and annotation schemes. Again Çöltekin provides an annotation tool for UD and Apertium⁴ has a Turkic lexicon which may be useful.

Most Turkic languages do not have a treebank, but this will change in the future. In this report, we have aimed to highlight issues that may be problematic in Turkish, and suggest methods to solve them in a cross-linguistic annotation scheme.

References

- Altenbek, G. 2006. Automatic morphological tagging of contemporary Uyghur corpus. *Proceedings of the 2006 IEEE International Conference on Information Reuse and Integration*. 557–560.
- Apertium—Turkic lexicon. http://wiki.apertium.org/wiki/Turkic_lexicon# (accessed 19 March 2016).
- Bergdahl, E. A. 2006. Shadow. From a relaxing spot to darkness and death. A semantic study of how the word shadow is used in Swedish and Turkish. [Term thesis]. Batı Dilleri ve Edebiyatı Bölümü, Bogaziçi University & Department of Linguistics and Philology, Uppsala University.
- Çöltekin, Ç. 2015. A grammar-book treebank of Turkish. In: Dickinson M. et al. (eds.) *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*. 35–49.
- Çöltekin, Ç. 2016. *Brat Annotation Tool*. <http://coltekin.net/cagri/ud/#/> (accessed 19 March 2016).

4 http://wiki.apertium.org/wiki/Turkic_lexicon#

- Csató, É. Á. 2009. Rendering evidential meanings in Turkish and Swedish. In: Csató, É. Á. et al. (eds.) *Turcological letters to Bernt Brendemoen*. Oslo: Novus. 77–86.
- Csató, É. Á. & Nathan, D. 2003. Turkish suffix dictionary. <http://www.dnathan.com/language/turkish/tsd>
- Dadasheva, Sabina 2005. Den turkiska indirektiva kategorin. En undersökning av återgivningen av den turkiska indirektiva kategorin i ryska och svenska autentiska översättningar. [Term thesis.] (In Swedish.) Department of Linguistics and Philology, Uppsala University.
- Dincer, T. & Karaoglan, B. & Kisla, T. 2008. A suffix based part-of-speech tagger for Turkish. *Information Technology: New Generations, 2008. ITNG 2008*, 680–685.
- Eryiğit, G., J. Nivre & K. Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics* 34, 3.
- Göksel, A. & Kerslake, C. 2005. *Turkish*. London: Routledge.
- Haktanır, M. 2006. Orhan Pamuk'un Beyaz Kale adlı eserinde çok anlamlılık. [Term paper] (In Turkish.) Department of Linguistics and Philology, Uppsala University.
- Hedman, M. 2009. Verbet 'göra' i svenska och turkiska. [Term thesis.] (In Swedish.) Department of Linguistics and Philology, Uppsala University.
- Halácsy, P. & Kornai, A. & Oravecz, Cs. 2007. Hunpos—an open source trigram tagger. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion volume, proceedings of the demo and poster sessions, Prague, Czech Republic*. 209–212.
- Makazhanov, A. & Yessenbayev, Z. & Sabyrgaliyev, I. & Sharafudinov, A. & Makhambetov, O. 2014. On certain aspects of Kazakh part-of-speech tagging. *Proceedings of IEEE AICT 2014*. 1–4.
- Makazhanov, A. & Sultangazina, A. & Makhambetov, O. & Yessenbayev, Z. 2015. Syntactic annotation of Kazakh: Following the Universal Dependencies guidelines. A report. In: *Proceedings of TurkLang 2015*. Kazan: Academy of Sciences of the Republic of Tatarstan Press. 338–350.
- Megyesi, B. 2008. The open source tagger HunPoS for Swedish. Department of Linguistics and Philology, Uppsala University.
- Megyesi, B. & Sâgvall Hein, A. & Csató, É. Á. 2006. Building a Swedish-Turkish parallel corpus. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy.
- Megyesi, B. & Dahlqvist, B. 2007. A Turkish-Swedish parallel corpus and tools for its creation. In: *Proceeding of Nordiska Datalogistdagarna, NoDaLiDa 2007*.
- 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)*, Morocco. 470–473.
- Megyesi, B. & Dahlqvist, B. & Csató, É. Á. & Nivre, J. 2010. The English-Swedish-Turkish parallel treebank. In: *Proceedings of Language Resources and Evaluation (LREC 2010)*, May 2010.
- Nivre, J. 2008. Treebanks. In: Kytö, M. & Lüdeling, A. (eds.) *Corpus linguistic: An international handbook*. Berlin: Mouton de Gruyter. 225–241.
- Nivre, J. 2015. Towards a Universal Grammar for natural language processing. In: Gelbukh, A. (ed.) *Computational linguistics and intelligent text processing. 16th International Conference, Cairo, Egypt, April 14-20, 2015, Proceedings 1*. Heidelberg: Springer.

- Nivre, J. & Hall, J. & Nilsson, J. 2006a. MaltParser: A data-driven parser-generator for dependency parsing. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. 2216–2219.
- Nivre, J. & Hall, J. & Nilsson, J. 2006b. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. 1392–1395.
- Oflazer, K. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9: 2.
- Oflazer, K. & Say, B. & Hakkani-Tür, D. Z. & Tür, G. 2003. Building a Turkish treebank. In: Abeillé, A. (ed.) *Treebanks: Building and using parsed corpora*. Kluwer. 261–277.
- Sak, H. & Güngör, T. & Saraçlar, M. 2011. Resources for Turkish morphological processing. *Language Resources and Evaluation* 45: 2, 249–261.
- Sulubacak, U. 2015. *ITU Treebank annotation guide v2.7*. Istanbul Technical University.
- Sulubacak, U. & Eryigit, G. (to be published). A redefined Turkish Dependency Grammar and its implementations: The revised Turkish treebank and a new Turkish web treebank.
- Taylan, E. E. 2014. The structure of Modern Turkish. Morphology. [Class Notes]. Boğaziçi University.
- Universal Dependencies documentation. <http://universaldependencies.org/> (accessed 19 March, 2016).
- Tiedemann, J. 2003. Recycling translations. Extraction of lexical data from parallel corpora and their applications in Natural Language Processing. PhD Thesis. Uppsala University.
- Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.
- Uplug documentation and download: <https://bitbucket.org/tiedemann/uplug> (accessed April 29, 2016).
- Uyghur, D. 2009. Semantiken av turkiska verbet *ver-* ‘att ge’ och dess motsvarighet i svenska. [Term thesis.] (In Swedish.) Department of Linguistics and Philology, Uppsala University.
- Yuret, D. & Türe, F. 2006. Learning morphological disambiguation rules for Turkish. In: *Proceedings of HLT NAACL'06*, New York. 328–334.