A grammar-book treebank of Turkish

Çağrı Çöltekin

Department of Linguistics
University of Tübingen
E-mail: ccoltekin@sfs.uni-tuebingen.de

Abstract

This paper introduces a new Turkish dependency treebank following the Universal Dependencies annotation scheme. The treebank is built on example sentences from a grammar book, which cover a wide range of the linguistic constructions. Thus, the resulting treebank is a valuable resource for theoretical (linguistic) research as well as testing computational tools for the coverage of the constructions found in the language.

1 Introduction and motivation

Common choices of source material for treebanks include news corpora from a single source [12, 21], random sentences from the Web and other freely available sources [17, 22], or from sentences balanced across a selected set of document categories [2]. Although these treebanks are useful for the purpose they are created for, and they may be representative of the language use to some degree, it is unlikely that they include infrequent grammatical constructions because of the power laws that govern the distribution of linguistic constructions at many levels.

The aim of the present work is to cover a large set of morpho-syntactic constructions with a minimal amount of annotation effort. To this end, comprehensive grammar books provide an excellent source of sentences, since they are selected by the authors to cover all constructions in the language, including infrequent but interesting ones. Such books are also more likely to cover examples of spoken and non-standard language use in comparison to most treebanks that are based on written, and possibly carefully edited, language material.

Our initial motivation for constructing the present treebank has been to set annotation guidelines for Turkish for the Universal Dependencies (UD) project [1]. However, such a treebank can be useful for many other purposes. For example, it is a valuable resource for checking existence of certain features or syntactic constructions in the language. Therefore, it may be useful in (theoretical) linguistic studies, including cross-linguistic comparisons. The rich linguistic descriptions in the source grammar book (e.g., glosses and detailed descriptions that accompany the example sentences) make the use of the treebank even more practical. Researchers can always refer to the original verbal description of the sentence in the grammar. Furthermore, it could be used for testing and qualitative evaluation of parsers, as one can observe type of errors that are difficult to encounter in typical test sets used for parser evaluation. Although the present treebank would not be appropriate as the only training data for parsers, it may improve parser performance by providing the data for infrequent constructions if the treebank is used as additional training data. For both purposes, a well-documented annotation standard is important.

Currently the most prominent treebank of Turkish is the METU-Sabancı treebank [2, 13], which also sets the de facto standard for dependency annotation of Turkish. The treebank contains a selection of sentences from the METU corpus [15] which is built as a balanced corpus across a number of different domains. The METU-Sabancı treebank is relatively small in comparison to the treebanks available for other languages (5 635 sentences and 56 424 tokens, in comparison to approximately 100 000 or more sentences usual in today's treebanks [e.g., 17, 21]). The treebank has not been updated since its first release in 2003, and annotation errors and inconsistencies are frequently reported in the literature [e.g., 6, 9, 16, 19]. Some of these studies also report improvements to the annotation scheme and individual annotations. However, except modifications by Seeker and Çetinoğlu [16] breaking the cycles in the dependency graphs, these improvements have not yet been released. The METU-Sabancı treebank is also converted to UD scheme as part of HamleDT [25], through an automatic process.

Besides METU-Sabancı treebank, other Turkish treebank constructions efforts include automatically or semi-automatically constructed Swedish-Turkish [3] and English-Turkish [24] parallel treebanks, and a small LFG treebank of 32 sentences in the INESS project [14]. The examples of the use of descriptive linguistic information for enriching NLP resources in earlier literature include [4, 23].

The study presented here differs from the earlier work by manually annotating a selection of sentences covering a wide range of constructions in the language. The annotations in the treebank follow the current UD annotation scheme (version 1.2) as closely as possible. In this paper, we focus on introducing the treebank, and discussing some of the issues in the dependency annotation of Turkish. Special attention is paid to divergences from the UD annotation scheme, and differences from the METU-Sabancı treebank.

2 Treebank and the annotation procedure

The treebank consists of 2 803 example sentences or sentence fragments extracted from a recent comprehensive grammar of Turkish by Göksel and Kerslake [10]. 410

of the treebank entries are sentence fragments, e.g., example noun phrases. For the rest of this document, we refer to all entries in the treebank, as 'sentences'.

The average length of the sentences in the treebank is shorter than sentences found in typical treebanks. The treebank consist of 16 516 surface tokens (5.89 per sentence, cf. 10.01 in METU-Sabancı treebank). The number of syntactic tokens, or inflectional groups (see Section 3.1 for details of tokenization), is 18 146, with a ratio of 1.10 syntactic tokens per surface token. This number is lower than the METU-Sabancı treebank (1.20) because of the more conservative approach we took in segmentation of words into syntactic tokens.

The sentences in the treebank include all numbered examples in the grammar book. We have also included some in-text examples. Sentences with optional words or phrases are repeated with all alternatives suggested by the example. If a sentence has multiple, ambiguous interpretations listed in the grammar book, the sentence is repeated and annotated for each alternative analysis (2 sentences with four analyses, 2 with three analyses and 28 with two analyses).

All words are analyzed using TRmorph [7] and disambiguated with a simple morphological analyzer [8]. Morphological analyses are checked and corrected manually. The tokenized and morphologically analyzed sentences were annotated following current specifications of UD, using BRAT [18]. During this process, features or constructions that are not covered by the UD specifications are noted, and treebank-specific annotation guidelines are developed.

All sentences in the treebank are annotated by a single annotator (the author). Pending approval of publisher of the grammar book, we intend to release the treebank (the source sentences and the annotations) with a free/open-source license.

3 Issues in dependency annotation of Turkish

This section discusses some of the major annotation decisions. We focus mainly on the issues that conflict with the current UD specification. Most of these issues relate to morphological complexity of the language. All annotation decisions reflecting the current state of the treebank are documented separately, and it will be proposed as the Turkish-specific UD guidelines after the major issues are resolved.

3.1 Sub-word syntactic units

Turkish exhibits a highly productive derivational morphology. In some cases, the derivational suffixes may be attached late in the affixation process, causing an already inflected word to change its part of speech. This may result in conflicting feature-value assignments within the same word, and parts of a word may participate in different syntactic relations. As a result, taking words as syntactic tokens produces less than satisfactory syntactic analyses of Turkish sentences. The last word in (1) demonstrates a case where both of these problems are present.

(1) *Kaygımız terörün durdurulamamasıydı*
Worry.P3PL terror-GEN stop.CAU.PASS.ABIL.NEG-INF.P3S-COP.PAST.3S
'Our worry was (the fact that) terror could not be stopped.'

The word *durdurulamamasıydı* starts with the verb *dur* 'stop', inflects for *passive* and *causative* voice. The morpheme coded as 'ABIL' modifies the *mood* of the verb, and the verb is also negated. Next, this inflected verb is nominalized by a subordinating suffix and inflected for third person singular possessive agreement.[1] At this point, the clause can approximately be translated to English as 'the fact/case/event that (it/something) cannot be stopped'. Finally, the resulting noun is again verbalized through a copular suffix which carries the third person singular agreement.

The morphological complexity presented in the example above causes both of the problems mentioned above:

1. The same word may contain conflicting lexical/morphological features. For example, in (1) above, although the content verb *dur* is negative, the predicate introduced by the copula is positive.

2. Parts of the word may participate in different, conflicting, syntactic relations. In the example above, the subject of the verb *dur* 'stop' is *terör* 'terror', while the subject of the copular predicate is *kaygımız* 'our worry' (see Figure 1 for the dependency analysis).

These two issues arise with numerous other constructions in the language. We will revisit some of them in this paper.

The solution used for this problem in Turkish NLP literature is to split the words into multiple syntactic tokens, commonly referred to as *inflectional groups* (IG) [11]. In earlier Turkish NLP work, e.g., in METU-Sabancı treebank, words are split at all productive derivational suffixes. Many other suffixes, including the voice and modality suffixes discussed above, also introduce new IGs. For example, the word *durdurulamamasıydı* would be split into six IGs in the METU-Sabancı treebank (*dur-dur-ul-ama-ması-ydı* as opposed to *durdurulama-ması-ydı* in our annotation scheme). We introduce new IGs more conservatively: a word is split into multiple IGs only if (i) the parts of the word may carry the same feature and/or (ii) the parts may participate in different syntactic relations. Following these principles, we explicitly define the morphological contexts in which a new IG is introduced.

Another fundamental difference of our work and the METU-Sabancı annotation scheme is the annotations of relations between the IGs in a word with multiple IGs. The original version of the METU-Sabancı treebank does not specify the dependency relations between the IGs within a word explicitly. The last IG is always assumed to be the head of the other IGs within the word. No explicit or implicit structure is defined for relating the head and the dependent IGs. The version used

[1] The suffix here in fact does not mark for possession, but indicates the subject of the verb.

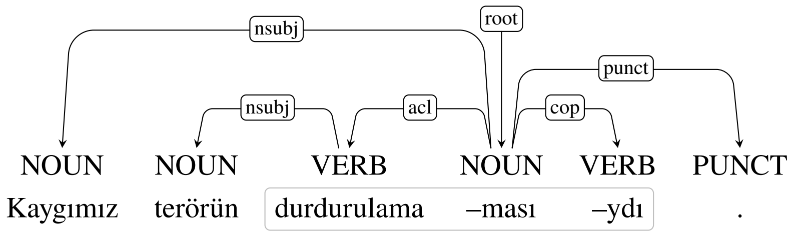

Figure 1: Analysis of (1), which includes a verbal noun. The details of the analysis are discussed in 3.5.

<table>
<tr><th>Token</th><th>Form</th><th>Lemma</th><th>UPOS</th><th>Feats</th><th>Head</th><th>Deprel</th></tr>
<tr><td>1</td><td>Kaygımız</td><td>kaygı</td><td>NOUN</td><td>Number=Sing|Number[psor]=Plur|Person[psor]=1</td><td>4</td><td>nmod</td></tr>
<tr><td>2</td><td>terörün</td><td>terör</td><td>NOUN</td><td>Case=Gen|Number=Sing</td><td>3</td><td>nsubj</td></tr>
<tr><td>3-5</td><td>durdurulamamasıydı</td><td>_</td><td>_</td><td>_</td><td>_</td><td>_</td></tr>
<tr><td>3</td><td>durdurulama</td><td>durmak</td><td>VERB</td><td>Mood=Abil|Negative=Neg|Person=3|Voice=Cau-Pass</td><td>4</td><td>acl</td></tr>
<tr><td>4</td><td>-ması</td><td>-me</td><td>NOUN</td><td>Number=Sing</td><td>0</td><td>root</td></tr>
<tr><td>5</td><td>-ydı</td><td>-0</td><td>VERB</td><td>Mood=Ind|Negative=Pos|Person=3|Tense=Past</td><td>4</td><td>cop</td></tr>
<tr><td>6</td><td>.</td><td>.</td><td>PUNCT</td><td>_</td><td>4</td><td>punct</td></tr>
</table>

Figure 2: The analysis of (1) in CoNLL-U format. All language specific columns (including the XPOS column which normally is the fifth column) and some features with default values (e.g., `Tense=Pres` from token 3, and `Number=Sing` from both predicate tokens) are left out for readability. The forms of the morphemes on column 2 are added for demonstration. Currently, the forms of the suffixes are left unspecified (annotated as '_').

during CoNLL-X shared task [5] introduces an explicit/dummy dependency label, `DERIV`, that relates the last IG (the head) to the other IGs in the word by a chain-like structure. In the present work, we always use dependency labels from UD dependency inventory to reflect the relations between the IGs. Furthermore, following the UD preference for marking the content words as heads, we do not always mark the last IG as the head of the other IGs in the word.

Figure 1 demonstrates the dependency analysis of the example sentence in (1) graphically, and Figure 2 presents the same analysis in CoNLL-U format. Since some suffixes are altered (and sometimes deleted) based on morpho-phonological context, determining surface forms of IGs is sometimes non-trivial. Current version of the treebank leaves the surface forms for non-root IGs unspecified. The lemma field is always filled consistently for both root and non-root IGs.

3.2 Morphological features

The morphological complexity of the language requires special attention to the morphological features assigned to each syntactic unit. Many linguistic functions that are expressed through word order or function words in English are expressed using inflectional suffixes in Turkish. For example, a verbal root may receive over

Table 1: The features used in the treebank. The features or values not in the current UD specification are *emphasized*. For definitions of the existing features and values, the reader is referred to UD specification at `http://universaldependencies.github.io/docs/`.

Feature	Possible values	POS
Aspect	Perf, Prog, *Hab*, *Rapid*, *Dur*, Pro	VERB
Case	Acc, Dat, Gen, Ins, Loc, Nom	NOUN, PRON, PROP
Definite	Def, Ind	DET
Degree	Cmp, Sup	ADV
Evidential	*Fh*, *Nfh*	VERB
Mood	*Abil*, Cnd, Des, *Gen*, Imp, Ind, Nec, *Prs*	VERB
Negative	Neg, Pos	VERB
Number	Plur, Sing	NOUN, PRON, PROP, VERB
Number[psor]	Plur, Sing	NOUN, PRON, PROP
NumType	Card, Dist, Ord	NUM
Person	1, 2, 3	NOUN, PRON, PROP, VERB
Person[psor]	1, 2, 3	NOUN, PRON, PROP
PronType	Dem, Int, Loc, Prs	PRON
Reflex	Yes	PRON
Tense	Fut, Past, Pres, Pqp	VERB
VerbForm	Part, Trans	VERB
Voice	Cau, Pass, Rcp, *Rfl*	VERB

10 inflectional suffixes, some of which may repeat multiple times. All IGs in the treebank are annotated with the lexical and inflectional features. We used features from the UD feature inventory as much as possible, and introduced new feature labels and/or values when necessary. Table 1 lists the features and their values. Here we will discuss the features and/or values that diverge from their traditional interpretation or from the current UD specification.

In Turkish, `Case` is an inflectional feature of nouns (POS tags `NOUN`, `PROPN` and `PRON`). Besides the five cases accepted in traditional grammars, we also use the case label `Ins` for instrumental or comitative marker *-(y)lA*.[2] We also use the same label when the suffix is not used in this case-like function but as a coordinating conjunction. The treatment of *-(y)lA* is similar to the METU-Sabancı treebank. Besides the suffix *-(y)lA*, there are a few productive suffixes (most notably *-lI* 'with', *-sIz* 'without') with case-like functions. Like the case-marked nouns, the derived word often functions like adverbs or adjectives. In this usage, it is possible to introduce non-standard case labels, or specific inflectional features for annotating these forms.

[2] In describing variable suffixes we use capital letter 'A' to denote alternative letters 'e' or 'a', capital letter 'I' for 'ı', 'i', 'u', 'ü', capital letter 'C' is used for 'c' or 'ç'. Buffer consonants or vowels are written in parentheses. According to this notation, the forms *-(y)lA* can take based on the morphophonological context are *-la*, *-le*, *-yla* and *-yle*.

However, we split these suffixes, and treat them like postpositions. The suffix is attached to the noun with the `case` relation. See Section 3.3 for more discussion on splitting productive suffixes.

The most challenging aspects of the inflectional features are related to verbal features. One aspect that currently does not fit well into the UD framework is the `Voice` feature. Turkish verbs can be inflected for reciprocal (`Rcp`), reflexive (`Rfl`), causative (`Cau`) and passive (`Pass`) voice. Current UD specification does not list `Rfl` as a possible voice value.[3] Additionally, current UD specification does not allow combination of voice values, e.g., for verbs that are inflected for both passive and causative voices as in (1) above, which occurs often in Turkish. A further complication is caused by the fact that the causative suffix is recursive. Even though it is very rare to see more than two iterations, a verb can be made causative multiple times, without a principled limit. For lack of an agreed solution, we currently annotate multiple `Voice` values as a list (see annotation of token 3 in Figure 2).

Despite the fact that the voice suffixes are considered as inflectional suffixes by descriptive grammars, METU-Sabancı treebank introduces a new IG for each voice feature. Since none of the IGs but the last one can be inflected, this creates 'inflectional groups' without any potential inflections. In other words, feature conflicts are not possible. The intermediate IGs cannot be modified by syntactic relations either.[4] As a result, the voice suffixes fail on both criteria set in Section 3.1 for introducing new syntactic tokens.

Turkish has a complex tense/aspect/modality (TAM) system. A single TAM suffix often marks a combination of tense, aspect and modality. Similar to [20], we annotate *evidentiality* as another feature dimension alongside tense, aspect and modality. We introduce a new feature, `Evidential` with two possible values `Nfh` (non-first hand) and `Fh` (first hand). We also use the following `Aspect` and `Mood` values that are not defined in the current UD specification.

- `Aspect=Hab` (habitual): *Güneş doğudan doğar* 'The sun rises from east'
- `Aspect=Dur` (durative): *bakakaldı* 'he/she looked (for a while, she was frozen while looking)' (durative stative) or *yapagelmiştir* 'he/she has gone on doing (something)' (durative progressive)
- `Aspect=Rapid` (for rapid or sudden action): *eve gidiver* 'quickly go home!'
- `Mood=Pers` (persuasive): *eve gitsene* 'go home (please)'
- `Mood=Abil` (abilitative or potentiality): *eve gidebilir* 'he/she may go home' or 'he/she is permitted to go home'. A negative verb may be inflected twice with this morpheme *eve gidemeyebilir* 'he/she may not be able to go home'

[3] The term *reflexive* here means that the subject of the predicate is also the direct object, i.e., the subjects acts on him/her/itself. This should not be confused with 'reflexive' verbs in some languages, e.g., German, which require an obligatory reflexive pronoun.

[4] One potential exception is that the subject of the non-causative predicate, i.e., content verb, may also be indicated by a noun phrase within the clause. In this case, the noun phrase acts as an argument or modifier of the complete (causative) predicate as well. Hence, we do not use another subject relation, but use language-specific subtypes of `dobj` and `nmod` relations (`dobj:cau` and `nmod:cau` respectively).

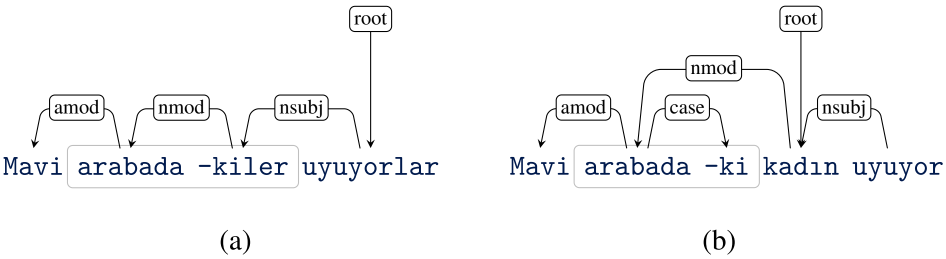

Figure 3: Dependency analyses of sentences in (2), demonstrating a nominal (a) and adjectival (b) derived with the suffix *-ki*.

- Mood=Gen (generalized modality): this marks statements with a more general or theoretical nature as opposed to statements of direct experience [10, p.295]. For example, *hastadır* '(I hypothesize/deduce that) she must be sick' or *iki, iki daha dört eder* 'two plus two is four'

Similar to voice suffixes, a verb may be inflected for multiple Aspect or Mood values. For example, *eve gidiverdim* 'I went home (quickly)' includes a completed (Perf) action that is performed quickly (Rapid). For multiple values, we follow the same strategy with multi-valued Voice features. Features in METU-Sabancı treebank are one-to-one mappings from the morphemes. As a result, a verb like *gitmiş* 'he/she (evidently) left' would be assigned a single +Narr (for narrative) feature. In our annotation scheme, the same verb receives tense, aspect, mood and evidentiality features Tense=Past|Aspect=Perf|Evidentiality=Nfh|Mood=Ind. Detailed documentation of these features and further examples can be found in the annotation guidelines document.

3.3 Productive derivational suffixes

As described in Section 3.1, some derivational suffixes cause an inflectional feature to be assigned multiple times, potentially with conflicting values. Example sentences in (2) demonstrate this with the suffix *-ki*. In (2a), the word *arabadakiler* refers to multiple people in the car. In the situation described, there are multiple people, but only a single car. Hence, *araba* 'car' carries the feature assignment Number=Sing, but *arabadakiler* 'the ones in the car' has the feature assignment Number=Plu. Furthermore, the adjective *mavi* 'blue' clearly refers to the car (not to the people), and the entity that is/are sleeping is the people, not the car. As a result, the suffix fulfils both criteria defined in Section 3.1 for introducing a new syntactic token.

(2) a. *Mavi arabadakiler uyuyorlar*
Blue car.LOC-ki.PL sleep.PROG.1P
'The ones in the blue car are sleeping.'

b. *Mavi arabadaki kadın uyuyor*
Blue car.LOC-ki woman sleep.PROG.1S
‘The woman in the blue car is sleeping.’

If the suffix *-ki* derives an adjective as in (2b), admitting multiple units is not equally justified. We still observe that the adjective modifies *araba* ‘the car’, not the resulting adjective. This, however, is not unlike the case suffixes that often scope over the phrase headed by the noun they are attached to. A possible way to annotate the adverbial and adjectival forms could be introducing features for these suffixes. However, we currently split the word into multiple IGs in both uses of the suffix *-ki*.

Besides the suffix *-ki*, the suffixes *-lI*, *-sIz*, *-lIk*, *-sI* deriving (pro)nouns from adjectives and determiners and *-dIr* and *-lArI* that derive time adverbials introduce new syntactic units. In case the derivation results in an adjective or adverb, we mark the content word as the head, and attach the suffix using the dependency relation `case`. In case the derivation results in a noun, we mark the final (noun) IG as the head of the word. Figure 3 shows the dependency analyses for examples in (2). As a general rule, however, we do not split a derivational suffix if the word as a whole is lexicalized. For example, the word *kitaplık* (3a) is annotated as a single syntactic token, while it is annotated as two tokens in (3b).

(3) a. *Kitaplık dolu*
Bookshelf full
‘The *bookshelf* is full.’

b. *Çantamda üç kitaplık yer var*
Bag-P1S-LOC three book-lIk space exist
‘I have *space for* three *books* in my bag.’

3.4 Copular constructions and the null copula

The copular constructions in Turkish include the verb *ol-* ‘be / become’, the suffix *-(y)* attached to the subject complement or, with a much lower frequency, its clitic counterpart *i-*. We split the copular suffix and its inflections since the IG introduced by the copula carries features that conflict with the features of the subject complement. Figure 4 shows example analyses. In both analyses, the subject complement, *spor arabalar* ‘sports cars’, is plural. However, the in both examples the copula does not carry explicit inflections for `Number`, defaulting to the singular agreement. Furthermore, if the copular suffix is attached to a verbal noun, as shown in Figure 1, it may cause further feature conflicts. Whether they are suffixes, or free morphemes, copulas are always annotated as dependents (not as the head).

The analyses in Figure 4b shows a case where the copular suffix is not present in the sentence because of the morpho-phonological process. Since the suffix version of the copula is just a buffer consonant, with third person singular agreement combined with present tense, it is not realized on the surface. Although there is no

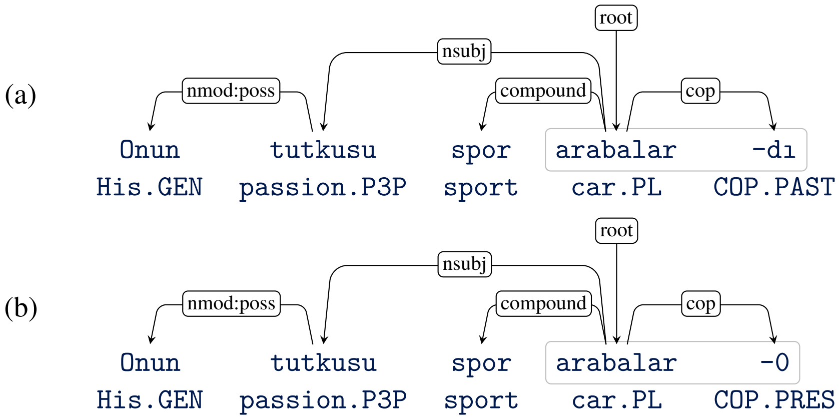

Figure 4: Example copula analyses (a) with overt past copula and (b) present copula with no overt suffix.

overt copular suffix, the predicate in Figure 4b still carries the third person singular agreement features, which conflicts with the plural number feature on the subject complement. As a result, we introduce an empty syntactic unit for the missing copula, despite UD's stand against null or missing elements. Besides the potential feature conflicts demonstrated above, failing to introduce the empty copular suffix results in analyses with different number of syntactic units for the same syntactic structure with trivial differences in their inflectional features. For example, the example sentences in Figure 5 differ only in the person agreement of the copular predicate. If we do not admit a null unit, as demonstrated in Figure 5, we assign different structures to these sentences.

The only exception where we do not introduce a null copula is in secondary predicates like *soğuk* 'cold' in *Ali çayını soğuk içer* 'Ali drinks his tea cold', or *arkadaş* 'friend' in *Ali'yi arkadaş sayarız* 'We consider Ali a friend'. The adjectives or nouns in these constructions are annotated with predicative relations without a copula.

3.5 Non-finite subordinate clauses

The main means of subordination in Turkish is through a set of subordinating suffixes. Resulting subordinate clauses may function as *adjective*s, *adverb*s or *noun*s. Adjectival and adverbial constructions behave like the simple words with the same functions, and they do not receive further suffixes. As a result, we do not introduce a new IG in these cases, but assign a feature that indicates the verb form as *participle* and *converb* respectively [10, p.84].[5]

[5] For converbs we use the label Trans since it has already been defined in the UD feature inventory, and the definition covers the converbs in Turkish.

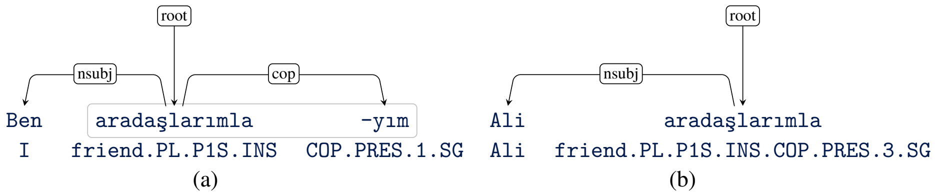

Figure 5: Inconsistent analyses of copula in case an empty syntactic unit is not introduced. (a) Overt copula: *Ben arkadaşlarımlayım* 'I am with my friends'. (b) No surface copula: *Ali arkadaşlarımla* 'Ali is with my friends'. Besides the conflicting number features (PL and SG) in (b), the same structure is analyzed differently.

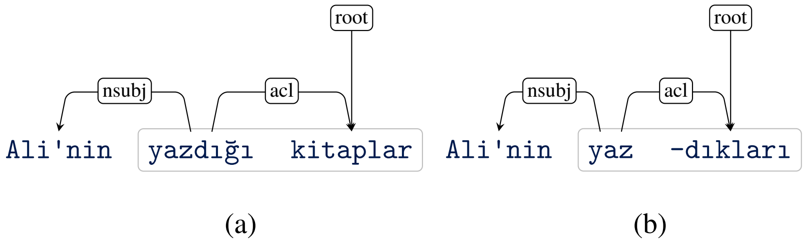

Figure 6: (a) A normal relative clause headed by a noun 'the books Ali has written'. (b) A headless relative clause 'the ones Ali has written'.

The verbal nouns, on the other hand, can be followed by most of the noun inflections. Furthermore, they can also be followed by POS-changing suffixes, most notably by the copular suffixes. An example of such a construction is given earlier in (1) and Figure 1. Figure 6 provides a simpler example with so-called *headless relative clauses* [10, p.389]. In this structure the head noun of a relative clause is omitted, and the relative clause is promoted to a (pro)noun referring to the missing noun phrase, and it can be inflected with all noun inflections. Note that in Figure 6b the predicate requires `Number=Sing`, while the resulting headless relative clause refers to multiple 'things', hence, having the feature assignment `Number=Plur`. Introducing a new syntactic token avoids this conflict. Although there are other conceivable solutions,[6] all other solutions would require major changes in the UD feature scheme. Besides solving potential feature conflicts, introducing a new IG makes the analysis similar to the 'headed' case shown in Figure 6a, and UD analysis of the corresponding English sentence where the pronoun 'one' would be analyzed as the head.

The conflict demonstrated in Figure 6 is very common for the headless relative clauses. With limited productivity, it also occurs with verbal nouns which denote entities of more abstract nature. This is demonstrated in (4) below, where the verb

[6] For example, by specifying all `Number` features as pertaining to *predicate* or the *noun phrase*.

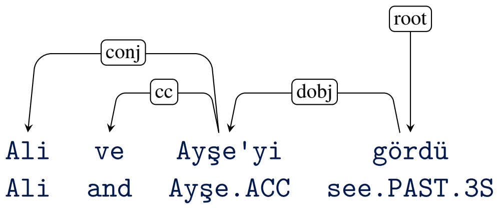

Figure 7: The analysis of sentence ‘She/he saw Ali and Ayşe’. Note that annotating *Ali* as the head would make it difficult to search for accusative subjects, or mislead a parser to assign a subject relation rather than object, since the relevant feature is not immediately available on the head as it is in majority of the other cases. The problem becomes more severe when there are more than two conjuncts, and in case of covert coordination where no explicit conjunction or punctuation exists.

kaç ‘run away’ carries singular predicate-subject agreement feature, while the verbal noun *kaçmaları* formed by suffix *-mA* is plural.

(4)	*Ali’nin*	*dersten*	*kaçmaları*	*annesini*	*kaygılandırıyor*
	Ali.GEN	class.ABL	run away-VN.PL.P3S	mother.P3S.ACC	worry.PROG.3S

‘(The events of) Ali skipping classes worries his mother.’

3.6 Issues related to the dependency labels

Once the morphology of Turkish is represented well through the sub-word syntactic units and the additional features described above, annotating the syntax with existing UD dependency relationships is relatively straightforward. The only major divergence from the current UD scheme is related to the head direction in some of the constructions where the choice of head seems arbitrary (e.g., `conj` and `name`). For these relations, the UD specification requires a head-initial analysis. This results in suffixes that scope over the whole constituent to be attached to a non-head word, making it difficult to locate morphological features during a treebank search or during feature extraction for the statistical tools. Figure 7 presents an example. Currently, we annotate `conj` and `name` in a head-final fashion, otherwise following the UD guidelines where all the dependents are directly attached to the head.

Except the head-direction difference above, the only other noteworthy difference is additional dependency labels which are subtypes of the UD dependencies. Some of these subtypes are also used in other languages. Due to lack of space, we provide a list with brief descriptions. The reader is referred to the annotation guidelines for detailed descriptions of the dependency subtypes used. The additional dependencies currently in use are: `nmod:cau` and `dobj:cau` (‘causee’ of a causative predicate, see Section 3.2); `nmod:comp` (for comparatives); `nmod:pass` (actor of a passive predicate); `nmod:tmod` (temporal modifier); `nmod:own` (owner in a possessive existential sentence); nmod:poss (possessor in genitive-possessive construction); nmod:part (whole in a partitive construction); compound:redup (compounds formed by reduplication); aux:q (question particle).

4 Concluding remarks

This document introduced a Turkish grammar-book treebank following the UD annotation scheme. We believe that the current treebank could be a valuable resource for a number of purposes including (theoretical) linguistic research and testing NLP tools. We also see this effort as a first step towards constructing larger and better documented treebanks for Turkish that conform with the latest standards in dependency parsing and annotation.

Acknowledgments

The author gratefully acknowledges the financial support by the German Ministry for Education and Research (BMBF) as part of the CLARIN-D research infrastructure grant given to the University of Tübingen.

References

[1] Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Lindén, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Héctor Alonso Martínez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. *Universal Dependencies 1.1*. 2015. URL: http://hdl.handle.net/11234/LRT-1478.

[2] Nart B. Atalay, Kemal Oflazer, and Bilge Say. "The annotation process in the Turkish treebank". In: *Proceedings of the 4th International Workshop on Linguistically Interpreteted Corpora (LINC)*. 2003.

[3] Eva Pettersson Beáta Megyesi Bengt Dahlqvist and Joakim Nivre. "Swedish-Turkish Parallel Treebank". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008, pp. 470–473.

[4] Emily M Bender, Sumukh Ghodke, Timothy Baldwin, and Rebecca Dridan. “From database to treebank: On enhancing hypertext grammars with grammar engineering and treebank search”. In: *Electronic Grammaticography*. Honolulu: University of Hawai’i Press, 2012.

[5] Sabine Buchholz and Erwin Marsi. “CoNLL-X shared task on multilingual dependency parsing”. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning*. 2006, pp. 149–164.

[6] Ruket Çakıcı. “Wide-Coverage Parsing for Turkish”. PhD thesis. University of Edinburgh, 2008.

[7] Çağrı Çöltekin. “A freely available morphological analyzer for Turkish”. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Valetta, Malta, 2010, pp. 820–827.

[8] Çağrı Çöltekin. “A set of open source tools for Turkish natural language processing”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014.

[9] Çağrı Çöltekin. “Turkish NLP web services in the WebLicht environment”. In: *Proceedings of the CLARIN Annual Conference*. 2015.

[10] Aslı Göksel and Celia Kerslake. *Turkish: A Comprehensive Grammar*. London: Routledge, 2005.

[11] Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. “Statistical Morphological Disambiguation for Agglutinative Languages”. In: *Computers and the Humanities* 36.4 (2002), pp. 381–410.

[12] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. “Building a large annotated corpus of English: The Penn Treebank”. In: *Computational linguistics* 19.2 (1993), pp. 313–330.

[13] Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. “Building a Turkish treebank”. In: *Treebanks: Building and Using Parsed Corpora*. Ed. by Anne Abeillé. 2003. Chap. 15, pp. 261–277.

[14] Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. “An open infrastructure for advanced treebanking”. In: *META-RESEARCH Workshop on Advanced Treebanking at LREC2012, Istanbul*. Ed. by Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco. 2012, pp. 22–29.

[15] Bilge Say, Deniz Zeyrek, Kemal Oflazer, and Umut Özge. “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, 2002.

[16] Wolfgang Seeker and Özlem Çetinoğlu. “A Graph-based Lattice Dependency Parser for Joint Morphological Segmentation and Syntactic Analysis”. In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 359–373.

[17] Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. “A Gold Standard Dependency Corpus for English”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik, Iceland, 2014, pp. 2897–2904.

[18] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. “BRAT: a web-based tool for NLP-assisted text annotation”. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 102–107.

[19] Umut Sulubacak and Gülsen Eryiğit. “Representation of Morphosyntactic Units and Coordination Structures in the Turkish Dependency Treebank”. In: *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*. 2013, pp. 129–134.

[20] John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. “A Universal Feature Schema for Rich Morphological Annotation and Fine-Grained Cross-Lingual Part-of-Speech Tagging”. In: *Systems and Frameworks for Computational Morphology*. Ed. by Cerstin Mahlow and Michael Piotrowski. Springer, 2015, pp. 72–93.

[21] Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. “The TüBa-D/Z treebank: Annotating German with a context-free backbone”. In: *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. 2004, pp. 2229–2232.

[22] Francis M. Tyers and Jonathan Washington. “Towards a free/open-source universal-dependency treebank for Kazakh”. In: *3rd International Conference on Computer Processing in Turkic Languages (TURKLANG 2015)*. 2015.

[23] Atro Voutilainen and Krister Lindén. “Specifying a linguistic representation with a grammar definition corpus”. In: *Proceedings of corpus linguistics 2011*. 2011.

[24] Olcay Taner Yıldız, Ercan Solak, Onur Görgün, and Razieh Ehsani. “Constructing a Turkish-English Parallel TreeBank”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 112–117.

[25] Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. “HamleDT: Harmonized Multi-Language Dependency Treebank”. In: *Language Resources and Evaluation* 48.4 (2014), pp. 601–637.