Revising the METU-Sabancı Turkish Treebank: An Exercise in Surface-Syntactic Annotation of Agglutinative Languages

Alicia Burga, Alp Öktem
Pompeu Fabra University
Barcelona, Spain
firstname.lastname@upf.edu

Leo Wanner
ICREA and Pompeu Fabra University
Barcelona, Spain
leo.wanner@upf.edu

Abstract

In this paper, we present a revision of the training set of the METU-Sabancı Turkish syntactic dependency treebank composed of 4997 sentences in accordance with the principles of the Meaning-Text Theory (MTT). MTT reflects the multilayered nature of language by a linguistic model in which each linguistic phenomenon is treated at its corresponding level(s). Our analysis of the METU-Sabancı syntactic relation tagset reveals that it encodes deep-morphological and surface-syntactic phenomena, which should be separated according to the MTT model. We propose an alternative surface-syntactic relation annotation schema and show that this schema also allows for a sound projection of the obtained surface annotation onto a deep-syntactic annotation, as needed for the implementation of down-stream language understanding applications.

1 Introduction

Dependency treebanks are crucial for the development of statistical NLP applications, including sentence parsing and generation. To obtain good performance, well-defined and coherent treebank annotation schemas are needed. To provide an outcome that is good not only in quantitative but also in qualitative terms in the sense that it is well-suited for various down-stream applications, the annotation schemas must be equally rigorous from the linguistic viewpoint. Thus, given that different down-stream applications may start from structures of different abstraction or diferent nature, an annotation schema should strive to annotate phenomena of different nature at different layers or focus on just one layer. [1]

A conflation of different types of phenomena in one layer would make the annotation idiosyncratic and thus less appropriate for down-stream applications. In addition, in order to be appropriate for down-stream applications, an annotation schema should differentiate between different phenomena at the same layer. For instance, if a tagset uses just one label for two rather different syntactic relations (e.g., 'adjunct' for both indirect objects and preposition-governed circumstantials), it will not lead to a parse from which, e.g., a semantic role structure can be derived.

The linguistic model of the Meaning–Text Theory (MTT) (Mel'čuk, 1988) accomodates for both of the above needs: it foresees different layers of linguistic representation (each one encoding linguistic descriptions at a specific level of abstraction), and it offers a fine-grained analysis of the phenomena at each of the layers. Furthermore, it provides a theoretically sound framework for the projection of a structure at a given layer to an equivalent structure at the adjacent layer (which is very useful, again, for down-stream applications).

Nearly all available dependency treebanks annotate what in the MTT-model would be the Surface-Syntactic (SSynt) layer. However, given the multi-layer nature of a language model proposed by MTT (Sem ⇔ DSynt ⇔ SSynt ⇔ DMorph ⇔ SMorph ⇔ DPhon ⇔ SPhon), a SSynt annotation schema should accurately reflect all (surface-)syntactic phenomena of the annotated language **and** encode all information that is necessary to derive their equivalents at the DMorph and DSynt layers.

We address the task of the annotation of a Turkish corpus at the SSynt-layer in accordance with the principles of MTT. In order not to start from scratch, we draw upon already available resources.

[1] Note, however, that a specific phenomenon may receive different descriptions at different layers – as, e.g., *grammemes* (discussed in Section 2) are divided into semantic and syntactic grammemes (Mel'čuk, 2012a), and thus described at the semantic and (surface-)syntactic layers.

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 32-41,
Pisa, Italy, September 18-20 2017

For Turkish, two major treebanks are available: the METU-Sabancı treebank (Oflazer et al., 2003) ('MS' from now on), composed of 5635 sentences, and the IMST Turkish Dependency treebank (Sulubacak et al., 2016), which is an adaptation of the first one and contains the same number of sentences. In any case, until now the reference treebank for Turkish has been the MS (see, among others, (Çetinoğlu and Kuhn, 2013; Eryiğit et al., 2008; Eryiğit et al., 2011), etc.).[2]

The remainder of the paper is structured as follows. In Section 2, we discuss the separation of deep-morphological and surface-syntactic phenomena in agglutinative languages such as Turkish in general and analyze to what extent the annotation schema of the MS treebank complies with this separation. In Section 3, we present an alternative annotation schema, which respects the multilayered nature of language established by the MTT framework and allows subsequent transitions from surface to deeper layers. Section 4 outlines how this transition can be realized between the surface and deep-syntactic layers. Section 5, finally, draws some conclusions and sketches the plans for continuation of our work on MTT-based corpus annotation.

2 Annotation of agglutinative languages

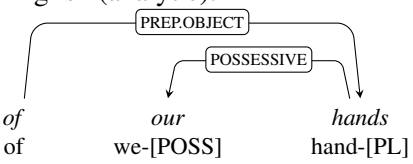
As an agglutinative morphologically rich language (MRL), Turkish poses challenges to tools and annotation schemas broadly used for non-agglutinative languages with a simpler morphology. As Eryiğit et al. (2008, p. 2) point out, agglutinative languages such as Turkish raise the question about "to what extent our models and algorithms are tailored to properties of specific languages or language groups". In order to assess how and to what extent the common models and algorithms should be modified and adapted, we need to spell out the phenomena in agglutinative languages that are, in contrast to non-agglutinative languages, intertwined. In our task, these phenomena concern deep morphology and surface syntax.

2.1 Agglutination: SSynt vs. DMorph

Agglutinative languages are synthetic languages in which words consist of a base and a set of *agglutinated* morphemes that modify the meaning of the base, each one separately in a predefined sense. In other words, each morpheme (whose boundaries are explicit) encodes a specific meaning, without undergoing context-dependent modifications.[3]

Thus, whereas analytical languages construct, as a rule, meaning through the grouping of words into phrases with a clear syntactic structure, agglutinative languages handle a good share of this process through the agglutination of morphemes; cf. a contrastive example in (1) and (2).[4]

(1) English (analytic):



(2) Turkish (agglutinative):

el	*-ler*	*-imiz*	*-in*
hand	PL	POSS-1PL	CASE-gen

'of our hands'

From the viewpoint of a grammatical theory, a morpheme is the realization (or instantiation) of a specific *grammeme* or a specific *derivateme*, each as a separate element.[5] In Turkish, grammemes capture noun inflection (number, possession, case, and clause-type) and verb inflection (person, number, tense-aspect, voice, reflexivity, reciprocality/collectivity, causativity, negation, impossibility, auxiliarity); derivatemes encode noun derivation (from other nouns, adjectives or verbs), adjective derivation (from other adjectives, nouns or verbs), verb derivation (from other verbs, nouns or adjectives), and adverb derivation (from other adverbs, nouns, adjectives or verbs); see (Oflazer et al., 1994) for details. Instantiation of grammemes and derivatemes is a purely morphological procedure, which in the MTT-model is modeled at the DMorph-layer. Thus, in the syntactic structure, grammemes should be already attached to lexemes, with the information encoded by each

[2]Most of the reported work has been done prior to the release of the IMST corpus. Note also that in the meantime some modifications of the original MS treebank have been made; cf. (Atalay et al., 2003). However, we use the original version.

[3]Morphemes can be ambiguous in the sense that two different meanings can be encoded by the same form, but individual morphemes do not carry combined meanings.

[4]The names of the SSynt relations in the example do not belong to any SSynt tagset; we have just chosen them for the sake of the transparency of the principal characteristics of the corresponding relations.

[5]Since we deal here with syntactic and morphological phenomena only, we can define a grammeme as "an element of an inflectional category" and a derivateme "as an element that is formally expressed by the same linguistic means as a grammeme, but that is not obligatory and not necessarily regular" (Melčuk and Wanner, 2008).

of them stored as a feature-value pair assigned to the lexeme in question (e.g., *table* [number = PL]).

In the next subsection, we analyze the MS treebank annotation schema from the perspective of this phenomenon separation as well as from the perspective of the coverage of the individual syntactic phenomena.

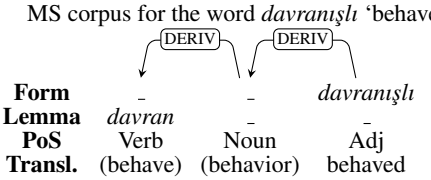
2.2 Analysis of the MS tagset

Let us assess the MS tagset first with respect to its uniform treatment of morphological and syntactic phenomena and then with respect to its treatment of syntactic phenomena as such.

2.2.1 Uniform treatment of morphological and syntactic phenomena

The MS syntactic relation tagset has been designed to cover both (surface) syntax and derivational morphology, such that no separation in the spirit of an MTT model is given. To conciliate the inclusion of both derivational morphology and surface-syntactic phenomena at the same level of annotation, derivatemes are treated as independent nodes in the structure. The annotation thus contains the derivative and the base lexeme as two different nodes; consider, for illustration the codification of *davranışlı* 'behaved' in (3).

(3) An example of the use of the relation DERIV in the MS corpus for the word *davranışlı* 'behaved':



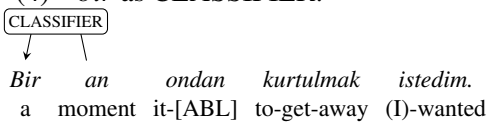
This practice leads to the appearance of extra lexical items in the annotation (the base lexemes do not materialize in the corresponding sentence(s) of the corpus), which are not present in the original corpus and which duplicate (or even multiply) specific meanings in the sentence; see also (Çetinoğlu and Kuhn, 2013). Such "artificial" lexical items that are introduced as auxiliary nodes to model a morphological phenomenon may even become the head of a syntactic relation (and thus also the root of a syntactic tree). As a consequence, the derivation of, for instance, a genuine semantic structure in the course of further analysis becomes a very tedious and unnecessarily complex task.

2.2.2 Treatment of syntactic phenomena

Apart from the problem resulting from the merge of DMorph and SSynt layers of annotation, the MS annotation reveals some issues that originate mainly from the underlying annotation guidelines and that affect directly the syntactic annotation. Let us go over these issues in what follows.

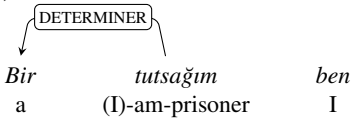
Vagueness in syntactic relation delimitation. The MS guidelines for the annotation of specific syntactic relations seem to be not sufficiently precise to ensure an unambiguous choice. Inconsistencies in the annotation are recurrent. For instance, from a total of 829 relations that take as dependent *bir* (unit that works either as an indefinite article or as a cardinal number), 738 are DETERMINER (not in all of them the unit acts, actually, as a determiner), 83 are MODIFIER, 4 are CLASSIFIER and 9 are SUBJECT (from these 96 cases, not always the unit has a cardinal number status). The remaining five cases are labeled as COORDINATION and S.MODIFIER. For illustration, compare (4) with (5).

(4) *bir* as CLASSIFIER:



'I wanted to get away from it for a moment.'

(5) *bir* as DETERMINER:



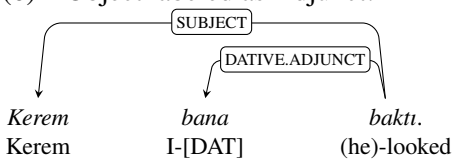
'I am a prisoner.'

This vagueness also affects the distinction between specific relations (cf. ADJUNCT vs. OBJECT) or the overuse of some relations (cf. MODIFIER) as "default" relations. Thus, hardly any criteria are given to decide whether a verbal dependent is to be annotated as ADJUNCT (potentially further detailed by the case; cf., DATIVE.ADJUNCT) or as OBJECT. In the guidelines it is only stated that adjuncts are optional elements related to a verb,[6] and that objects are either nouns or pronouns. In cases like DATIVE.ADJUNCT, the only criterion to consider the relation to be ADJUNCT seems to be that the

[6] In the annotation, this condition is not always followed either: some elements related to a verb as ADJUNCT are obligatory, and in some cases, the head is a noun rather than a verb.

element must be in dative; consider, for illustration the relation between *baktı* '(he-)looked' and *bana* 'I' in (6). Obviously, the decision whether a relation is annotated as ADJUNCT or as OBJECT has important consequences for the projection of the annotated SSynt structures onto more abstract structures (such as DSynt).

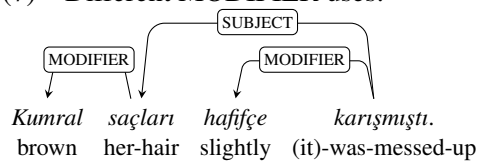
(6) Object labeled as Adjunct:



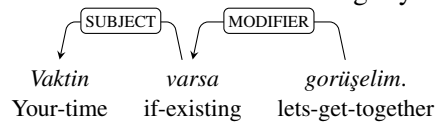
'Kerem looked at me.'

MODIFIER is defined only with respect to the possible PoS combinations of the head and the dependent, which makes it impossible to understand or systematize the behaviour of the relation. Therefore, as mentioned above, it becomes a "default relation", overused across the corpus with very different morphosyntactic behavior among its instances, as can be observed in (7).

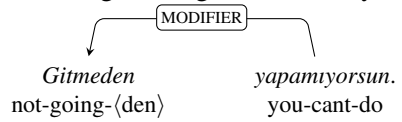
(7) Different MODIFIER uses:



'Her brown hair was slightly messed up.'



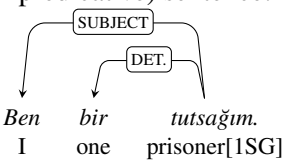
'Let's get together if you have time.'



'You can't do without going.'

Vagueness in copulative construction annotation. To express what is known as a copulative construction of the type 'A is B', in Turkish special predicative forms of nouns and adjectives are common, in which the subject is directly linked to the predicate.[7] The predicate takes (beyond its own PoS and internal structure) verbal inflectional suffixes (person, number, tense) and becomes thus the syntactic head of the sentence; cf. (8).[8]

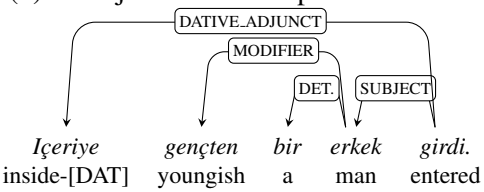
(8) Subject in a "copulative" (nominal predicative) sentence:



'I am a prisoner.'

However, significant syntactic differences remain between such nominal (and adjectival) predicative constructions and non-copulative constructions. Despite these differences, MS uses the same tag, SUBJECT, to mark the subjectival relation in both of them; cf. (9).

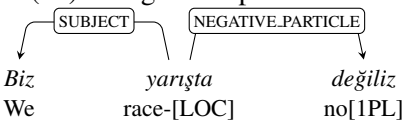
(9) Subject in a non-copulative sentence:



'A fairly young man entered inside.'

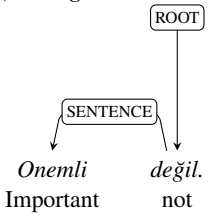
When negation comes into play (i.e., when we have a construction 'A is not B'[9]) the annotation is very inconsistent in the MS. Sometimes, the predicative element is considered head of the sentence, as in (10), and sometimes the negation element, as in (11).

(10) Negated copulative sentences:



'We are not in a race.'

(11) *değil* as ROOT:



'It is not important.'

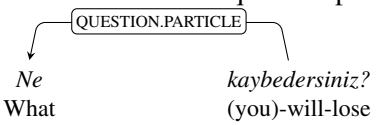
[7] According to the traditional grammar, the copula is expressed through the suffix *-dir*. However, the suffix is not really productive in modern Turkish.

[8] In order to keep the terminology simple, we continue to call those predicative constructions "copulative" (in quotes), although, strictly speaking, they are not copulative (Mel'čuk, 2012b).

[9] In the case of negation in a "copulative" construction, the verbal inflectional suffixes are taken by the negation element *değil*; cf. (10).

Indiscriminate annotation of *WH*-words. Many times, just because they are included in a question, the *WH*-words (*ne* 'what'/ *hangi* 'which'/ *kim* 'who'/ *kimin* 'whose', etc.) are linked to the verb through the relation QUESTION.PARTICLE, as in (12), which is the relation used to link the verb with particles that take verbal inflectional suffixes (when the questioned element is the verb) and mark *yes-no* questions. At least two important problems arise from this annotation: (i) syntactic differences between the links 'verb–question particle' and 'verb–*wh*-words' are ignored, given that *wh*-words can take case suffixes (governed by the head) and question particles cannot, and the first ones can only take verbal inflectional suffixes in copulative sentences (as any other noun), whereas particles take them in any *yes–no* question, if the questioned element is the verb; (ii) the mapping to deeper levels becomes truncated, given that the real syntactic function of the *wh*-words (e.g., OBJECT, SUBJECT, etc.) is not annotated at the SSynt layer.

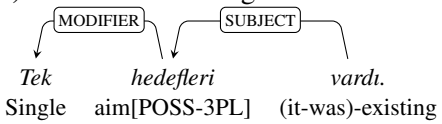
(12) *Wh*-word treated as question particle:



'What will you lose?'

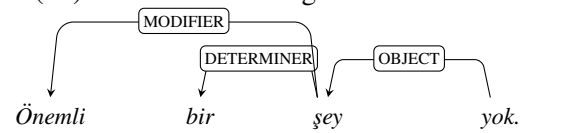
Inconsistent annotation of existential sentences. Annotation of existential sentences (which are expressed in Turkish through attributive configurations with *var* and *yok*, as in (13) and (14)) is not unified: the attributee is linked to the existential attributes either via the relation SUBJECT (13) or via the relation OBJECT (14). Given that the syntactic characteristics of the relation between existential attributes and the attributee are always the same, only one relation should be consistently chosen. The chosen relation should depend on whether the existing element shares its syntactic behavior with other subjects or objects. If its syntactic characteristics are unique and exclusive, a new relation should be created.

(13) Existential taking SUBJECT:



'They had a single aim.'

(14) Existential taking OBJECT:



'There is nothing important.'

3 Revising the surface-syntactic annotation of Turkish

The problems discussed in the previous section and some others that were not touched upon due to the lack of space made us revise the SSynt annotation schema followed in the MS treebank, with the MS' tagset as basis. The design of our annotation schema follows the principles of the MTT framework (Mel'čuk, 1988) and the methodology adopted for the elaboration of the annotation schema of the Spanish AnCora-UPF treebank (Mille et al., 2013) and the Finnish weather corpus (Burga et al., 2015).

The mapping of the original MS annotation into our annotation was carried out in two stages; its result is henceforth referred to as "UPF-METU SSynt". In the first stage, general transformations have been made. These transformations targeted, first of all, the removal of the relation DERIV (encoding the corresponding morphological information in terms of morphological feature-values assigned to the corresponding nodes) and conversion of the relation SUBJECT into SUBJNOUN in nominal and adjectival predicative ("copulative") constructions. In the second stage, the outcome of the transformation has been revised manually and modifications discussed in Section 2.2.2 have been implemented.

In parallel, a rule-based projection of the SSynt annotation onto the deeper DSynt annotation has been implemented. Both the original syntactic annotation of the MS treebank and the UPF-METU SSynt annotation have been mapped onto DSynt to validate the conversion of the MS treebank annotation into the UPF-METU annotation.

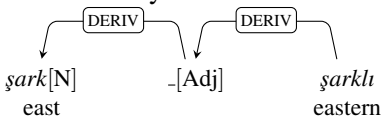
In what follows, we discuss first the initial transformation and then the modifications applied to it in the second stage.

3.1 Removal of DMorph traces

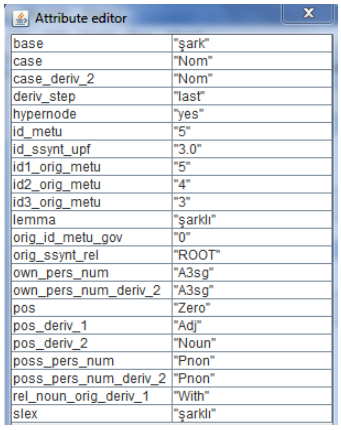
As argued in Section 2.1, it is convinient and cleaner from the theoretical point of view to separate the different levels of linguistic representation. Since the relation DERIV, included in the MS

SSynt tagset, relates the inflectional groups between each other and thus encodes a phenomenon that belongs to the DMorph layer (see Section 2.2), it needs to be removed from the SSynt tagset. For this purpose, the nodes related through DERIV are merged into one and the information of each node is stored in terms of feature-value pairs of the resulting node using a MATE graph transduction grammar (Bohnet and Wanner, 2010). As a consequence, an MS subtree such as shown in (15) is converted into a single node with many morphological features (as in (16)).

(15) DERIV as SSynt relation:



(16) Morphological information related to DERIV: *şarklı*



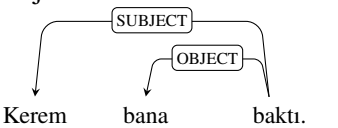
A consequence of this transformation is that the resulting single node becomes the head of the SSyntRels that before were defined between the different nodes related through DERIV, which inevitably results in a relaxation of the head restrictions for each relation (in that relations that prototypically were headed by nouns can after the merge be headed by a lexeme with anothr PoS). In this regard, the second stage of the conversion (manual revision of relations) needs to put special attention to sentences in which automatic transformations applied, and the annotator decisions need to take into account the nature of the originally encoded derivations.

3.2 Making changes to syntactic annotations

In this subsection, we outline how the MS SSynt tagset has been revised in order to account for the issues identified in Section 2.2.2. The updated tagset contains 21 relations summarized in Table 1 at the end of this subsection.

Addressing the vagueness in syntactic relation delimitation. According to the MTT principles, it is crucial to distinguish between adjuncts and objects in SSynt, given that each of them maps to different relations in deeper layers. Therefore, in order to distinguish between the relations ADJUNCT and OBJECT in the case of a verbal head, we consult the case suffix added to the dependent and the analysis of the meaning of the verb. In MTT, the case of objects is governed by the verbal head, while the case of adjuncts is determined by the type of information these adjunts convey. The adjuncts in Turkish can take dative,[10] locative, ablative, instrumental, or equative suffixes. Objects, on the other hand, most of the times take either accusative or nominative,[11] and they can promote (become subjects in passive sentences). Although dative, ablative or instrumental case is also possible, it is more seldom. Which case it actually is depends on lexical restrictions of each verb, which are assumed as intuitively known by native speakers of Turkish. Also, those verbs that require "non-standard" objects cannot passivize through promotion of their objects, and do not admit adjuncts carrying the same case. Thus, our analysis of (6) would be as shown in (17).

(17) Object with dative case:



'Kerem looked at me.'

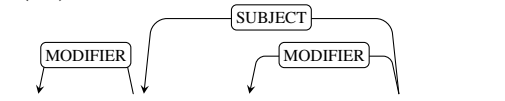
In order to sharpen the definition of MODIFIER, we draw upon the conditions established in MTT for the presence of a SSynt relation. According to these conditions a SSynt relation between two lexical items is present if (i) the position of one of the items in the sentence is established

[10] Even though adjuncts taking dative are uncommon – as one of the reviewers pointed out, and which is confirmed by the fact that in traditional Turkish grammar, nominal phrases in dative are always considered objects – we argue that they exist.

[11] Objects in nominative are also unusual, but they also exist, as in *Çiçek aldım*, lit. flower [nom] buy[1SG, past] 'I bought a flower.' In any case, we take the information about cases as it is included in MS. If this information is incorrect, we do not correct it.

with respect to the other item; (ii) the two lexical items have a prosodic link that connects them; or (iii) one item imposes agreement on the other item. The new relation MODIFIER that shall substitute the original MS MODIFIER has been defined as a repeatable relation, in which the dependent is not verbal, there is no agreement between the head and the dependent, the dependent always appears to the left of the head, and the head and the dependent are adjacent.[12] Thus, from the examples of MODIF in 7, the only ones that are kept as MODIFIER are those in (7a), repeated here as (18).

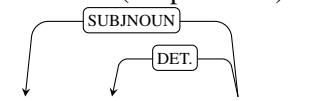
(18) Restricted MODIF:



"Her brown hair was slightly messed up."

Addressing the vagueness in copulative construction annotation. Given that subjects in predicative nominal and adjectival (what we called "copulative") and non-copulative constructions have different properties regarding agreement, and agreement is one of the criteria used for differentiating SSynt relations in the MTT model, we have decided to distinguish between "typical" subjects (in which the head is a conjugated verb) from subjects in "copulative" sentences (in which the head is, strictly speaking, not a conjugated verb), we have created the relation SUBJNOUN; cf. (19) for illustration. Whereas SUBJECT implies agreement with the head in both person and number, SUBJNOUN does it obligatorily with person and optionally with number.

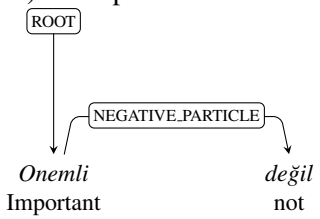
(19) Treatment of subjects in nominal predicative ("copulative") constructions:



'I am a prisoner.'

"Copulative" constructions that contain negation are treated in the same way as those without negation, but the particle *değil* is linked to the negated element through the relation NEGATIVE_PARTICLE, even if, in "copulative" constructions, it takes the inflectional suffixes; cf. (20).[13]

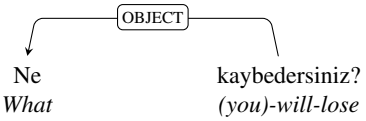
(20) "Copulative" constructions with negation:



'It is not important.'

Addressing the indiscriminate annotation of *WH*-words. Regarding the treatment of *wh*-words, the adapted SSynt tagset restricts the relation QUESTION.PARTICLE to those cases in which the dependent is the particle *mA*, which indicates *yes-no* questions (taking into account the prosodic link between elements involved in the relation). The governor is the element that is questioned and always appears to the right before the particle. This relation, then, always goes from left to right and its members are adjacent. If the questioned element is the verb (as the head of QUESTION.PARTICLE), the particle is conjugated. On the other hand, *wh*-words are labeled according to their syntactic similarity with other relations, without taking into account their PoS. Thus, the suggested annotation of (12) is as shown in (21).

(21) Treatment of *wh*-words:



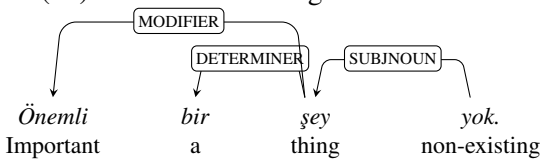
"What will you lose?"

Addressing the inconsistent annotation of existential sentences. Existential sentences are treated as a subset of copulative sentences in which the attributive element is either the adjective *var* 'existing' or the adjective *yok* 'non-existing' Thus, the relation connecting these elements with the existing element is SUBJNOUN, as illustrated in (22).

[12] This adjacency is broken in those cases in which the same head governs more than one MODIFIER relation.

[13] One of the reviewers questioned the correctness of this analysis, given that Turkish is a strong head-final language. Although we have kept our initial proposal, in the near future, it will be necessary to evaluate which analysis (the one that prioritizes the head-final property, or the one in which the parallel treatment of affirmative and negative copulative sentences is followed) prevails.

(22) Existential taking SUBJNOUN:



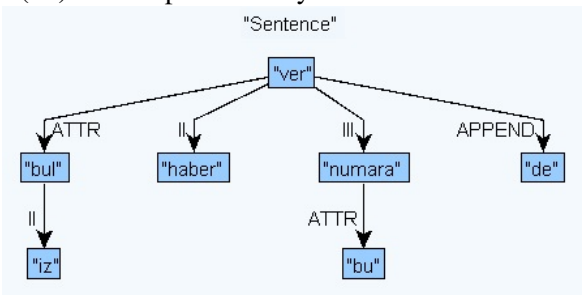
'There is nothing important.'

4 Projecting SSynt Structures onto Deeper Levels of Annotation

The challenge of the SSynt annotation schema design is not only to cover the syntactic phenomena of a specific language, but also to facilitate an appropriate projection to deeper levels, in our case DSynt. In contrast to the SSynt tagset, the DSynt tagset is language-independent. It is composed of the argumental relations I, II, III, IV, V, VI, and the non-argumental relations ATTR, APPEND and COORD(INATION); cf. also (Mel'čuk, 1988) An example of a DSynt tree of a sample from MS corpus is shown in 23.[14]

In total, 122 rules that map specific SSynt relations in specific configurations onto DSynt relations were created. The mapping resulted in well-formed DSynt trees, whose relations (participants as well as labels) are being manually corrected, in parallel to SSynt structures.

(23) Example of a DSynt structure:



İzini	*bulursanız,*	*bu*	*numaraya*	*haber*	*verirsiniz,*	*dedi.*
iz	*bul*	*bu*	*numara*	*haber*	*ver*	*de*
his-trace	if-you-find	this	number	notice	you'd-give	he-said

'If you find his trace, you'll notify this number, he said.'

In what follows, we discuss how the issues that we identified with the original MS treebank inevitably have negative consequences for the projection of SSynt structures to DSynt structures, and how the revision offered in our proposal helps obtain a better SSynt-DSynt mapping.

First of all, the relation DERIV (that should be encoded within DMorph, as discussed in Section

[14]For details about the differences between SSynt and DSynt structures, see, for instance, (Burga et al., 2015).

Table 1: Dependency relations used after adaptation of the Turkish surface-syntactic layer

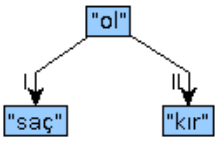
DepRel	Distinctive properties
adjunct	non-required element; non NOM/ACC case
apposition	for clarification; right-sided for nouns, left-sided for statements
classifier	noun modifying another noun; case NOM; left-sided relation
collocation	relates base and collocation
coordination	links coordinated elements or the 1st coordination member with the coord. Conj
coord_conj	complement of a coord Conj
determiner	non-repeatable left-side modifier of an N
intensifier	particles emphasizing the head; right-side relation
juxtaposition	for linking unrelated groups
modifier	non-required modifying element; no case taken left-sided relation
neg_particle	right-sided relation between the negated element and the particle *değil*
object	required element. It takes NOM and ACC most times, but can take DAT, ABL, INSTR
possessor	links possessed thing (in genitive case) and possessor (with possessive suffix)
punc	for punctuation signs
quasi_subj	relates object and subject of an ommited verb
ques_particle	links questioned element and question particle *mI*
relativizer	links a verb-based element to the subordinating elements *de/da* and *ki*
s_modifier	acts as a sentential adjunct; left-sided relation
subject	unrepeatable verbal dependent that controls number and person; takes NOM case
subjnoun	subjects in copulative sentences;agreement only in person
vocative	element marking the addressee; always in NOM; at the beginning or end of sentence

3.1) would lead to spurious nodes in the DSynt structure, which have absolutely no theoretical or practical justification. Obviously, auxiliary measures during the projection can be implemented in order to avoid the introduction of such spurious nodes, but this would mean a cumbersome and unnecessarily complex projection. Second, even if it is not always possible to map a SSynt relation to just one DSynt relation, the SSynt tagset should at least drastically limit the mapping options. This is why the lack of syntactic criteria when defining a tagset also generates problems for the projection of a SSynt structure to a DSynt structure.

The inconsistency in annotation, as well as the use of the same relation for pairs that behave syntactically different (see below), decreases the quality of DSynt structures (e.g., the above-discussed argument–adjunct confusion). In this sense, our attempt to restrict the syntactic characteristics of each SSyntRel serves not only the SSynt layer itself, but also to the corresponding DSynt layer.

As far as the structures of nominal / adjectival predicative (what we called "copulative") and non-copulative sentences are concerned, at the DSynt layer, their structures become homogenized since both receive a verbal root; in the case of the "copulative" construction, the subject is the first argument of the root and the predicative element its second; see (24) for illustration.

(24) DSynt tree of a adjectival predicative ("copulative") construction:



Saçları kır.
saç ol kır.
hair be gray
"Her hair is gray."

Given that Turkish is a pro-drop language, the mapping of SSynt structures to DSynt structures introduces a subject node when it is absent in SSynt (acting as the first argument of the verbal root). This node contains the morphological features that allow agreement.

5 Summary and Future Work

In this paper, we first briefly analyzed the manifestation of morphological and (surface) syntactic phenomena in agglunitative languages such as Turkish, arguing (in accordance with the Meaning–Text Theory) that both should be described separately at different layers of the linguistic model, namely at the D(eep)Morp(logical) and S(urface)Synt(actic) layers. With the MTT model in mind, we studied the annotation schema of the MS Turkish treebank, which does not make this separation, and identified some issues that result from the uniform treatment of morphological and syntactic phenomena or from the MS-specific treatment of some syntactic phenomena. Then, we presented an MTT-based schema annotation for the SSynt of Turkish. This schema has been followed to convert the original MS annotation of the training set of the MS treebank (4997 sentences) into an MTT-affine annotation. The conversion has been carried out in two stages. In the first stage, a number of regular transformations was applied via graph transducer rules (Bohnet and Wanner, 2010). In the second stage, the automatically obtained annotation in the first stage was revised manually. Tests show that the MTT-affine annotation allows us not only to get higher quality SSynt structures, but also to derive from these SSynt structures an additional more abstract level of annotation, namely that of DSynt. As a result, downstream NLP applications that must rely upon more semantically-oriented linguistic representations can use different levels of the same annotated treebank.

The goal is to offer the MTT-oriented annotation of the MS treebank to the community. Depending on the legal constraints, which still need to be clarified, we count on being able to provide it shortly either on the webpage of the authors of the original MS treebank (`https://web.itu.edu.tr/gulsenc/treebanks.html`) or on our webpage `https://www.upf.edu/web/taln/resources`.

In the future, we plan to carry out an evaluation of parser performance when trained on the original MS-annotated treebank and on the revised treebank. Even if the size of the training treebanks is small, we expect to see clear differences. We also plan to explore how the morphological information that corresponds to the eliminated relation DERIV and the nodal feature values that specify the type of derivation should be structured, stored in DMorph structures and exploited in sentence analysis and generation tasks. In this context, it is to be noted that morphological analysis in Turkish

is a real challenge due to the ambiguity of derivational suffixes themselves and also due to the ambiguity of their combination. Thus, for instance, the morphological analysis of *yarının* using the TRMorph (Çöltekin, 2010) gives us 40 possibilities of analysis, the first three having different roots (25):[15]

(25) Morphological analysis of *yarının*:
yarı<Adj><0><N><gen> 'of the half'
yarın<N><gen> 'of tomorrow'
yar<N><p3s><gen> 'his lover's'

According to one of the reviewers, in the original MS treebank the morphological disambiguation has been done manually.

Acknowledgements

The presented work has been funded by the European Commission as part of the H2020 Programme, under the contract numbers 645012–RIA and 700024-RIA. Many thanks to the three anonymous reviewers for their detailed comments that helped us improve the paper considerably.

References

Nart B. Atalay, Kemal Oflazer, Bilge Say, and Informatics Inst. 2003. The Annotation Process in the Turkish Treebank. In *Proc. of the 4th Intern. Workshop on Linguistically Interpreteted Corpora (LINC)*.

B. Bohnet and L. Wanner. 2010. Open Source Graph Transducer Interpreter and Grammar Development Environment. In *Proceedings of the International Conference on Linguistic Resources and Evaluation (LREC)*.

Alicia Burga, Simon Mille, Anton Granvik, and Leo Wanner. 2015. Towards a multi-layered dependency annotation of Finnish. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 48–57, August.

Özlem Çetinoğlu and Jonas Kuhn. 2013. Towards Joint Morphological Analysis and Dependency Parsing of Turkish. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 23–32, August.

Çağrı Çöltekin. 2010. A Freely Available Morphological Analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 820–827.

Gülşen Eryiğit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Gülşen Eryiğit, Tugay İlbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, SPMRL '11, pages 45–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Igor Melčuk and Leo Wanner. 2008. Morphological mismatches in machine translation. *Machine translation*, 22(3):101–152.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.

Igor Mel'čuk. 2012a. *Semantics, Volume 1*. John Benjamins Publishing Company, Amsterdam.

Igor Mel'čuk. 2012b. Syntax. Bi-nominative sentences in Russian. In V. Makarova, editor, *Russian Language Studies in North America: New Perspectives from Theoretical and Applied Linguistics*, pages 86–105. Anthem Press, London.

Simon Mille, Alicia Burga, and Leo Wanner. 2013. AnCora-UPF: A multi-level annotation of Spanish. In *Proceedings of DepLing*, Prague, Czech Republic.

Kemal Oflazer, Elvan Gmen, and Cem Bozsahin. 1994. An Outline of Turkish Morphology.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In *Treebanks: Building and Using Parsed Corpora*, pages 261–277. Springer.

Umut Sulubacak, Tuğba Pamay, and Gülşen Eryiğit. 2016. IMST: A Revisited Turkish Dependency Treebank. In *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*, pages 1–6.

[15] Each morphological analysis is composed by the base lexeme, its PoS, and the associated grammemes and derivatemes; as soon as a derivateme appears (as <0> in the first line), a new PoS is assigned (<N> in the mentioned example).