(When) do we need inflectional groups?

Çağrı Çöltekin
University of Tübingen
ccoltekin@sfs.uni-tuebingen.de

***Abstract*—Inflectional groups (IGs) are sub-words units that became a de facto standard in Turkish natural language processing (NLP). Despite their prominence in Turkish NLP, similar units are seldom used in other languages; theoretical or psycholinguistic studies on such units are virtually nonexistent; they are typically overused in most existing work; and there are no clear standards defining when a word should or should not be split into IGs. This paper argues for the need for sub-word syntactic units in Turkish NLP, followed by an explicit proposal listing a small set of morphosyntactic contexts in which these units should be introduced.**

I. Introduction

The term *inflectional group* (IG) in Turkish natural language processing literature refers to a sub-word unit. Although it does not seem to stem from (theoretical) linguistics, the unit has been a de facto standard for representing words in Turkish NLP. Representing words as multiple IGs helps dealing with complex interaction between the morphology and syntax in the language. Furthermore, it alleviates the data sparseness problems in machine learning methods that arise due to large (theoretically infinite) number word forms as a result of numerous affixes a word can get. On the other hand, the use of IGs makes it difficult to use well-studied methods from other languages, or common off-the-shelf NLP tools since these methods and tools are designed with the assumption that the word is the basic unit of syntactic processing. While we argue that sub-word syntactic units are necessary for Turkish NLP, the oversegmentation of words into IGs, which is very common in present practice in the field, amplifies these problems, and even defeats its own aim by shifting the data sparseness problem caused by long sequences of potential suffixes per word to one caused by a long sequences of IGs per word. We discuss these issues in detail, and propose a more conservative alternative for segmentation of words into IGs. In this paper, we assume that the IGs are introduced for syntactic reasons, even though the traditional use of the unit seems to link it with derivational morphemes and derivation boundaries. We do not address, or discuss the derivational morphology outside its relation to the IGs.

A. The need for sub-word syntactic units

In many languages, representing a word with a *lemma*, a *POS tag* and a set of (inflectional) *features* is sufficient (and useful) for most NLP tasks. In Turkish, however, this representation is often inadequate. For example, consider the word *arabadakiler* 'the ones in the/a car' in (1) below. The word *araba* 'car' is inflected for *locative case* after which it receives the suffix *-ki* which changes the meaning of the word to 'the one in the/a car'. Finally the word is suffixed with the plural morpheme resulting in *plural number* inflection.

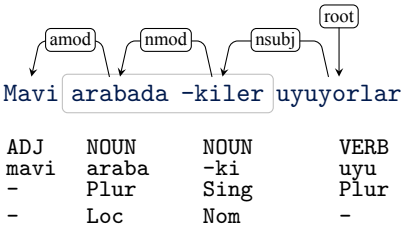


Figure 1. Dependency analysis of the sentence in (1). The dependency and feature labels follow Universal Dependencies (UD, Nivre et al. 2016) conventions. Only the features relevant to our discussion are listed.

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

The conventional representation with a triple ⟨lemma, POS tag, features⟩ fails here, since the word *arabadakiler* refers to two different (sets of) entities, and it carries a separate set of inflections for each. The first part of the word, *arabada* 'in the/a car', is singular and in locative case, while the complete word, *arabadakiler* 'the ones in the/a car', is plural and not marked for case (nominative). Besides the multiple conflicting inflectional features within the word, parts of the word participate in separate syntactic relations. Figure 1 presents a dependency analysis of the sentence in (1).[1] The adjective *mavi* 'blue' modifies the car (not the people in it), while the entities that sleep are the ones in the car (not the car). As a result, in Turkish computational linguistics literature, such words have been represented using multiple sub-word units known as *inflectional groups* (Oflazer 1999).

Although the need for sub-word units is clear in (1), the current practice in the field oversegments the words without any clear linguistic or practical reasons. For example, the subordinated verb *sınırlandırılabilecek* 'that/which can be limited' would be tokenized into six IGs in METU-Sabancı treebank (Say et al. 2002; Oflazer et al. 2003) as in (2).

(2) *sınır -lan -dır -ıl -abil -ecek*
NOUN VERB.Deriv VERB.Caus VERB.Pass VERB.Abil ADJ

In this annotation scheme, as well as the derivational morpheme *-lan*, the causative (*-dır*) and the passive (*-ul*) *voice* suffixes, the *mood* suffix *-abil* expressing ability or possibility and the subordinating suffix *-ecek* which forms

[1] We present example analyses using dependency annotations, since this is where the IGs were first introduced, and due to popularity of dependency parsing and annotation in the NLP community. However, the parallel examples can easily be constructed for other grammar formalisms.

a verbal adjective introduce new IGs. The segmentation in (2) does not have the same grounding as the one introduced by the suffix *-ki* in (1). All suffixes except the first one are considered part of inflectional morphology by modern grammars of Turkish (e.g., Kornfilt 1997; Göksel and Kerslake 2005). Even if we consider first three inflectional suffixes as verb–verb derivations, none of the intermediate forms can carry any separate inflections, and there is no possibility of conflicting features. The case for verbal adjective suffix is slightly more complicated (discussed in Section II-C). However, the verbal adjective forms in Turkish are not much different than participle forms in other languages where an additional inflectional feature is sufficient to indicate that the word carries properties of both adjectives and verbs. That is, the word acts similar to verbs within the subordinate clause, while acting like an adjectival outside the subordinate clause.

The current paper proposes tokenizing a surface word into multiple IGs only in case one of the following is true.[2]

(3) a. Parts of the word may have potentially conflicting inflectional features.
    b. Parts of the word may participate in different syntactic relations.

These guidelines also imply that the syntactic units should have clearly defined syntactic functions, unlike, for example, the relation *deriv* introduced in the CoNLL-X version of the METU-Sabancı treebank (Buchholz and Marsi 2006). Under our guidelines, the word in (2) would not be segmented at all.

The next section presents a critical summary of the use of IGs to date, mainly pointing out when segmentation of words are *not* necessary. Section III lists the cases where we need to introduce IGs after which we provide a brief discussion followed by a summary and outlook.

II. Inflectional groups

The term *inflectional group* first appeared in work related to Turkish dependency parsing and annotation (Oflazer 1999), and used in later studies with similar aims (Say et al. 2002; Oflazer et al. 2003; Sulubacak and Eryiğit 2013; Çöltekin 2015). It is also used in work on Turkish syntax with different grammar formalisms (Çetinoğlu and Oflazer 2006; Çakıcı 2008), and in pre- or non-syntactic analysis such as morphological analysis and disambiguation (e.g., Hakkani-Tür, Oflazer, and Tür 2002; Çöltekin 2014). The similar units are also used by NLP work on other Turkic languages (Tyers and Washington 2015). Although we are not aware of a precise definition of the term, both the use in the literature so far and the name *inflectional* group indicates that the unit was introduced based on morphosyntactic concerns. More precisely, we assume inflectional groups are sub-word units required by syntax. The remainder of this section outlines the earlier use of IGs, and discusses the morphological constructions where the current practice oversegments words according to the guidelines defined in (3).

[2]The conditions 'conflicting features' and 'separate syntactic relations' depend on the annotation scheme. Ideally, the tagsets should avoid spurious conflicts. However, the guidelines are useful even if the tagset choice is not free, and causes spurious conflicts.

A. Earlier use in the literature

Following Oflazer (1999), almost all Turkish NLP tools and resources annotate a word as a sequence of IGs as shown in (4) below.

(4) `root+Infl`$_1$`^DB+Infl`$_2$`+…+^DB+Infl`$_n$

where `root` is the root of the word, `Infl`$_i$ are a group (presumably a set) of inflections and `^DB` is a special symbol indicating a derivation boundary. According to this annotation scheme, the word *sınırlandırılabilecek* in (2) is represented as (5) below.[3]

(5)
```
sınır+Noun+A3sg+Pnon+Nom
      ^DB+Verb+Acquire
      ^DB+Verb+Caus
      ^DB+Verb+Pass
      ^DB+Verb+Able+Pos
      ^DB+Adj+AFuttPart
```

The same annotation scheme is used in most of the Turkish computational linguistics literature to date. Below we discuss the differences between the current practice and the scheme suggested in this paper.

B. Derivation boundaries are not necessarily syntactic-token boundaries

In the current literature, it is common to see inflectional group boundaries inserted before some derivational morphemes, such as *-lan* in (2). However, not every derivation warrants introducing a new syntactic unit. In the noun–verb derivation example, *sınır-lan* 'border-*lan* (= to restrict)', the noun *sınır* cannot be inflected. Hence, it cannot have an inflectional group of its own. It is also not accessible from syntax: neither it can be modified by another syntactic word, nor is it possible for it to modify another one. Although keeping the derivational history may be helpful for some applications, it is not related to determining syntactic units. For the purpose of determining syntactic units, the (derivational) morphemes of interest are typically those that modify an already inflected word, like the suffix *-ki* in (1) in Section I. However, attaching to an already inflected verb is not sufficient for forming a new syntactic token. Also, the condition we are seeking here is more strict than morphemes that scope over the phrases. Some productive derivational suffixes may attach to already inflected forms, and scope over whole phrases, as exemplified by the suffix *-sIz* 'without' in (6) below.

(6) [*Takım arkadaşlarım*]*sız* *yapamam*
Team friend.PL.POSS1S.*without* do.AOR.NEG.1P
'I cannot do without my team mates'

It may be tempting to segment the word *arkadaşlarımsız* into two IGs, since the noun *takım* modifies the stem *arkadaş*, and the suffix *-sız* scopes over the complete phrase. Furthermore, the suffix *-sız* attaches to an already inflected noun and derives an adverbial. However, according to our criteria, these do not warrant introduction of a new syntactic token. A large number of inflections scope over the phrases headed by

[3]The analysis here follows the annotation scheme in METU-Sabancı treebank (Oflazer et al. 2003) which is a typical example of other resources and tools for Turkish NLP with respect to representation of words.

the words carrying the inflection. For example, the possessive suffix attached to the same noun also scopes over the whole phrase (it is 'my [team mates]', not '*team [my mates]'). The word *arkadaşlarımsız* in this example cannot have conflicting features either (adverbs are not inflected in Turkish). Hence, there are no strong reasons for segmenting words at derivation boundaries introduced by the suffixes similar to *-sIz*. The suffixes in this category include *-lI*, *-lIk*, *-(n)CA*, *-CI*, and also *-ki* when it derives an adjectival. These suffixes should be represented with adequate morphological features, rather than separate syntactic units. Note that we make a distinction between the cases where these suffixes derive adjectivals or adverbials and the cases that some these suffixes derive nominals. Nominal case is discussed in Section III-B.

C. Inflectional morphemes should not introduce IGs

In the current literature, a large number of inflections introduce new IGs. The majority of these inflections are verbal inflections including *voice* suffixes, as well as some *mood* and *aspect* modifiers. The passive and causative suffixes and the modal suffix glossed as *Abil* in (2) are examples of such inflectional suffixes.

One of the motivations for segmenting at these inflectional morphemes may be the fact that some of them can attach repeatedly to the same verbal stem. In this respect, the *causative* morpheme is particularly interesting, since, similar to *-ki* described in Section I, it can repeat multiple times with no principled limit on the number of consecutive causative suffixes. In practice, however, the use of multiple causative suffixes is rare, and it often indicates emphasis rather than multiple levels of causation. Example (7) demonstrates a verb with two causative suffixes which, indeed, can be interpreted as having two levels of causation.[4]

(7) *Ders* *bütün* *okullarda*
Subject all school.PL.LOC
oku-t-tur-ulacak.
study.CAU.CAU.PASS.FUT.3SG
'The subject will be caused to be caused to be studied all schools.' (literal)
'The subject will be taught in all schools.'

Besides the causative suffix, the *passive* suffix, and forms of the modal suffix *-Abil* may attach to the same verb multiple times. The double passive (on a transitive verb) creates impersonal (passive) expressions (Göksel and Kerslake 2005, p.136). The double use of *-Abil* modifies the modality of the verb for both of its senses (ability and possibility). In all of these cases, these suffixes do not create a new predicate with potentially different inflections than the verbal stem they are attached to. For example, in the multiple levels of causatives above, all actions have to share the same tense, aspect and modality. As a result, if these suffixes form inflectional groups, the resulting inflectional groups will not have any independent inflections. A set of features that allow marking multiple levels of causation and distinguishing the effects of single or double passive or *-Abil* suffixes is sufficient for avoiding additional syntactic tokens.

Another aspect of the voice inflections that may have affected the current practice of oversegmentation is the fact that they change the valency of the verb, and modify the meanings of the arguments of the verb. For example, a causative or passive verb will assign different roles to its arguments. However, even if the verb valency is changed, there will still be a single grammatical subject and/or object, and their roles can be inferred from the transitivity of the verb and the voice inflections it carries. As a result, none of the suffixes discussed above meet the criteria set in (3). With a proper morphological tag set, we do not need to introduce new IGs for voice suffixes as well as other aspect or modality modifiers.

Besides the verbal suffixes discussed above, existing work also segments the words at subordinating suffixes (suffixes that cause phrases headed by the verbs to function as adjectives, adverbs or nouns). These suffixes change the function of the word they are attached to. However, there is no principled reason for not representing their status by setting a feature, e.g., *verb form* to an appropriate value, e.g., *verbal adjective* (participle), *verbal adverb* (converb) or *verbal noun* (gerund/infinitive). This avoids segmentation by indicating that the word functions as a verb within the subordinate clause, while acting like a noun, adjective or adverb outside the subordinate clause. Note that even the subordinate clauses that function as nouns (verbal nouns and headless relative clauses, Göksel and Kerslake 2005, p.84) do not require segmentation since nominal predicates cannot be subordinated without an auxiliary verb and inflectional features, and syntactic relations of verbs can easily be distinguished from that of nouns, adjectives and adverbs (the copula attached to the subordinate verbs is discussed in Section IV). In many ways, the subordinating suffixes are similar to the productive derivational suffixes discussed in Section II-B, and do not need to introduce a new syntactic tokens.

D. Uniform representation of all syntactic units

Another issue with the present use of IGs as represented in (4) is the asymmetry between the first IG and the ones that follow. In this representation, the only IG with a lemma is the first one. This hinders the uniform treatment of the syntactic tokens since some of the tokens are not represented as ⟨lemma, POS tag, features⟩ triples, and introduces difficulties with using existing NLP tools like parsers.

The current proposal requires a syntactic token to always be associated with a lemma. For non-root IGs, the lemma should be a canonical representation of the (derivational) morpheme that introduces the IG. For example, for the proposed tokenization of *arabada-kiler* in (1), the suffix *-ki* should be treated as the lemma rather than an inflection. This also serves as a test for introducing new IGs. If the segmentation of a word results in IGs that cannot have any inflections of their own (except for the lemma), the segmentation is not justified.

[4] A bit of context may be useful for non-native speakers to understand the double causative in this example. The example, taken from a news text about a new educational regulation, expresses that (the authorities who made) the regulation will cause schools or teachers to cause the students to study the subject.

III. Inflectional group boundaries

So far, our focus in this paper has been on where or when *not* to segment a word to sub-word syntactic units. In this section, we list the cases where sub-word units are necessary.

A. The relativizer -ki

The suffix *-ki* has two main functions (Hankamer 2004). It either forms either adjectivals or pronominal expressions from nouns. We already argued in Section II-B that when the suffix *-ki* derives adjectivals, there is no need for introducing a new syntactic unit. However, as the example in Section I demonstrates, if it derives a pronominal a new IG is necessary.

If the suffix *-ki* is attached to a noun in genitive case, the resulting pronominal expression refers to an entity that belongs to the object or person the original noun refers to. If it is attached to a locative noun, the resulting expression refers to an entity in/on/at the object the original noun refers to. The parts of the word referring to these two entities may have their own set of inflections, and may participate in different syntactic relations. The example (1) and the corresponding dependency analysis in Figure 1 demonstrate the need for separate syntactic units. Without segmenting the word into multiple syntactic tokens, we cannot tell whether the expression refers to multiple cars or a single car, and we cannot tell whether the car or the objects in the car are blue, or even whether the car is sleeping or the people/objects inside are sleeping. Both problems can be solved by introducing a new syntactic token as in the analysis presented in Figure 1.

Furthermore, the nominals derived with *-ki* may be suffixed with genitive or locative suffixes again, and in turn, with another *-ki* suffix. Although multiple *-ki* suffixes are rare in real language use, the process is recursive, and there is no principled limit that one can place on number of *-ki* suffixes in a word form. This fact also underlines the need for introducing new IGs in pronominal usage of suffix *-ki*.

B. Other productive noun–noun derivations

Like the suffix *-ki* discussed above, some productive noun derivations result in word forms that refer to multiple entities. This is demonstrated using the derivational suffix *-CI* in (8).

(8) a. [*eski kitap*]*çı*
old book.CI
'[old book] shop/seller'

b. *eski* [*kitapçı*]
old book.CI
'old [book shop]'

If the word *kitapçı* in (8) is not segmented, we do not have a way to represent the ambiguity between 8a and 8b. The same issue surfaces in case of other noun–noun derivations or noun–adjective derivations when the derived adjectival is nominalized, referring to an object with the property described by the derived adjective. In such cases, similar to *-ki*, the parts of the word refer to entities which may have their own set of inflections, and may participate in different syntactic relations. The other suffixes with similar behavior are *-sIz*, *-lI* and *-lIk* (which overlap with the ones listed in Section II-B). We present an example for each of the cases in (9).

(9) a. *Kayıt belgesizlere 2 bin TL ceza*
Registation doucment-SIZ.PL.DAT 2 thousand TL fine
kesilecek
cut.PASS.FUT
'Those without a registration document will be fined 2000 TL.'

b. *2-3 metrelikleri adamdan saymıyor*
2-3 meter-LIK.PL.ACC man.ABL count.NEG.PROG
musun?
QuesP.2SG
'Are you not considering 2 to 3 meter long ones worthy? (referring to boats)

c. *1.5 crdi motorlusuyla 170 tl'lik*
1.5 CRDI engine-LI.POS3S.INS 170 TL.LIK
dizelle Istanbul-Sivas mesafesini yaptım.
diesel.INS Istanbul-Sivas distance do.PAST.1SG
'I rode the Istanbul-Sivas distance with the one with 1.5 CRDI engine using 170 TL worth of diesel fuel.'

In (9a), without segmenting the word *belgesizlere* 'the ones without documents', we cannot represent the fact that the noun *kayıt* 'registration' modifies the word *belge* 'document', not the people who do not have the document. This is unlike the earlier example (6) where the relation is unambiguous since the attributive noun can only modify the noun, not the resulting adjectival. Similarly, in (9b), the numeral modifies the *metre* 'meter(s)', not the pronominal expression derived by the suffix *-lik*. In other words, the expression refers to (unknown number of) 2 to 3 meter boats, not 2 or 3 boats of one meter long. In (9c), too, the numeral and the abbreviation modifies the *motor* 'engine', not the car with that particular engine. Also note that the suffix *-lık* in this example does not have to be segmented, since it derives an adjectival. The preceding number here can only modify the noun, not the adjectival.

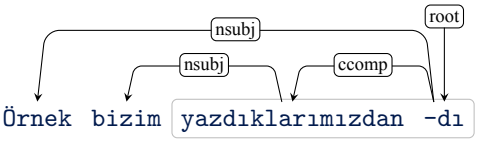
The suffixes listed in (9) are a lot less productive than *-ki* discussed Section III-A, and they attach to already inflected words with a varying but lower degree than *-ki*. Nevertheless, the cases exemplified in (8) exist. For a uniform treatment, our proposal is to segment words into multiple tokens when these suffixes derive a (pro)nominal expression.

Although the suffixes discussed here require segmentation of words, this is not true if the same suffix is part of a lexicalized derivation. For example, in contrast to the use of suffix *-siz* in (9b), the lexicalized word *ev-siz* 'homeless' should not be segmented since the root here cannot be inflected, and it cannot participate in separate syntactic relations.

C. Copular suffixes and the suffix -lAş

In Turkish, main means of forming copular predicates is through suffixation. In most cases, copular suffixes attach to a simple noun or adjective, where one may avoid segmenting the word by setting a feature that indicates the copular nature of the word. However, if the copula is attached to a verbal noun or a headless relative clause, as in (10) below, segmentation is unavoidable.

(10) *Örnek bizim*
Example we.GEN
yazdıklarımızdandı.
write.PART.PAST.PL.POSS1P.ABL-COP.PAST.3SG
'The example was from the ones we wrote'



POS:	NOUN	PRON	VERB	VERB
Lemma:	örnek	biz	yaz	-y-
Number:	Plur	Plur	Plur	Sing
Case:	Nom	Gen	Abl	-
Number[psor]:	-	-	Plur	-
Person[psor]:	-	-	1	-
VerbForm:	-	-	Part	-
Tense:	-	-	Past	Past
Person:	3	1	3	3

Figure 2. Dependency analysis of the sentence in (10). The dependency and feature labels follow the Universal Dependencies conventions (marking copula as the head is against one of the UD principles which is violated frequently). Only the features relevant to our discussion are listed. The features `Person[psor]` and `Number[psor]` mark the person and number of the possessor in a noun. The same suffixes also indicate the person and number of the subject on a subordinate verb.

In (10), the word *yazdıklarımızdandı* includes two predicates (*yaz* 'write' and the past copula). As it is also presented in Figure 2, both predicates have their own subjects in the sentence. Furthermore, these two predicates have their own feature sets which may conflict. For example, the subordinate verb carries the first person plural subject–verb agreement (indicated by the feature labels `Person[psor]` and `Number[psor]` in Figure 2), while the inflections on the copula indicate a third-person singular subject (marked by feature labels `Person` and `Number`). This example also demonstrates that the potential conflict of *person* and *number* features between the predicate and resulting nominal is avoided by using different labels for these features (although the labels may be confusing in this particular tagset).

The morpheme *-lAş* 'to become' presents a slightly different case. *-lAş* forms verbs from nouns and adjectives, often leaving the possibility of modifying the stem. The sentence in (11) presents an example where the adjective *pembe* within the verb derived by *-lAş* is modified by an adverb.

(11) *Koyu pembeleşinceye kadar kavurun.*
Dark pink-lAş.CONV until fry
'Fry until it it becomes dark pink.'

IV. Discussion and further issues

This paper argues for limiting the segmentation of words into sub-word syntactic tokens based on two principles listed in (3). Based on these principles, the same affix may or may not introduce an new IG depending on whether it derives a nominal or an adjectival. In general, the need for tokenization arises when the same word contains multiple (pro)nouns or predicates. Furthermore, if a derived word with an otherwise transparent and productive suffix is fully lexicalized, there is no need for segmenting the word, as the stem cannot be inflected or modified by other words in the sentence.

Our proposal introduces a new IG in case a suffix derives a (pro)nominal from a noun in a way that allows modification of both nouns in the word, but not when the same suffix derives an adjective or adverb. A potential disadvantage of this approach is that it requires tokenization decisions to be made based on morphosyntactic information, which may cause difficulties for a pipeline approach to NLP.

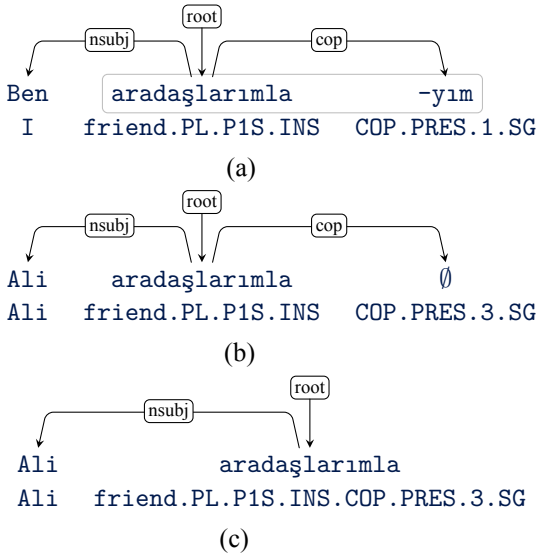


Figure 3. Inconsistent analyses of copula in case an empty syntactic unit is not introduced. (a) Overt copula: *Ben arkadaşlarımlayım* 'I am with my friends'. (b) No surface copula: *Ali arkadaşlarımla* 'Ali is with my friends', a null syntactic element is introduced. (c) same sentence as in (b) analyzed without a null element.

A second issue, we left unspecified in Section III-C is the use of null-copula, which surfaces (pun intended) in case of copular constructions with present tense and third person singular subject. Failing to introduce a null syntactic token will result in inconsistent analyses of copular expressions that differ only in trivial future assignments, e.g., first person or third person subject–verb agreement. Figure 3 demonstrates this inconsistency. In Section III-C we demonstrated that the copular suffixes should be segmented to be able to properly analyze sentences like (10). For the same reasons, we need to segment the copula in the sentence analyzed in Figure 3a. However, unless we introduce a null-copula as in Figure 3b, the tokenization and syntactic analysis of these two sentences will be different (as presented in Figure 3c), despite the fact that two sentences differ only in the person/number features of the copular predicates. It seems, introducing null copula becomes a necessity, unless one wants to introduce an inconsistency in the analyses of these two similar structures. Note, however, the null element introduced here is unlike the null units introduced in certain grammar formalisms as a result of syntactic processes (e.g., movement). Nevertheless, null elements will typically not be allowed in a wide range of grammatical frameworks, where an alternative method may be needed to avoid this inconsistency.

As noted earlier, the criteria we set in (3) depends on the choice of the feature set. For example, many tag sets, e.g., UD, use the same feature label for the *number* feature of predicates and nominals. This causes either feature conflicts or inconsistent labels for morphological and/or syntactic tags in representation of participles and verbal nouns, which should not be tokenized according to our proposal. For example, the word *yazdıkları* 'the ones he/she wrote' in (10) requires two number features, the nominal is plural, but the predicate has

a singular subject. The analysis in Figure 2 avoids conflicting feature values within the word *yazdıklarımızdan*, by indicating the number and person of the subject of the predicate *yaz* using a different tag than the person and number of the subject of the copula. As a result, this word cannot be represented as a single syntactic token by assigning separate labels for these two different roles. Similar issues may also arise because of overloaded use of some syntactic relations.

V. Summary and outlook

This paper presented an analysis of the current use of sub-word syntactic units, IGs, and proposed a more conservative alternative than the current practice while segmenting words into multiple IGs. We show that sub-word syntactic units are necessary even under such a conservative approach. However, the number of sub-word units can be dramatically reduced with appropriate choice of tagset for morphological features and syntactic relations. Our concrete proposal is that introduction of IGs should be motivated by syntactic analysis, and a word should be tokenized into multiple IGs when (1) it cannot be represented as a simple triple ⟨lemma, POS tag, features⟩ and/or (2) the part of the word participates in different separate syntactic relations.

The principles set in this paper for (not) segmenting a word into multiple units, depend on the tagset in use. A logical next step is to complemented this proposal with a tagset that is useful for a wide range of NLP applications. Although defining a proper tagset for morphological features is out of scope of this paper, the guidelines above are useful in design of such a tag set. We note that the efforts like Universal Dependencies project (Nivre et al. 2016) may facilitate constructing such tag sets through the consensus of the broad community of Turkish/Turkic NLP researchers.

Our motivation in this paper has been identifying syntactic units for computational processing of the language. However, the sort of units discussed in this paper are interesting from the perspective of (general/theoretical) linguistics as well. At present, the problems discussed here are underexplored in all subfields of linguistics including computational linguistics (with the notable exception of Bozşahin 2002). This discussion may motivate further research with more theoretical flavor, which in turn may benefit the computational methods.

In closing, we also note that even though our discussion in this paper covers only Turkish, the same approach is likely to be relevant for other Turkic languages.

References

Bozşahin, Cem (2002). "The Combinatory Morphemic Lexicon." In: *Computational Linguistics* 28.2, pp. 145–186.

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

Çakıcı, Ruket (2008). "Wide-Coverage Parsing for Turkish." PhD thesis. University of Edinburgh.

Çetinoğlu, Özlem and Kemal Oflazer (2006). "Morphology-Syntax Interface for Turkish LFG." In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, pp. 153–160.

Çöltekin, Çağrı (2014). "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).

Çöltekin, Çağrı (2015). "A grammar-book treebank of Turkish." In: *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*. Ed. by Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski. Warsaw, Poland, pp. 35–49.

Göksel, Aslı and Celia Kerslake (2005). *Turkish: A Comprehensive Grammar*. London: Routledge.

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

Hankamer, Jorge (2004). "Why there are two ki's in Turkish." In: *Current Research in Turkish Linguistics*. Ed. by Kamile Imer and Gürkan Dogan. Eastern Mediterranean University, pp. 13–25.

Kornfilt, Jaklin (1997). *Turkish*. London and New York: Routledge.

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

Oflazer, Kemal (1999). "Dependency Parsing with an Extended Finite State Approach." In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, pp. 254–260.

Oflazer, Kemal, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür (2003). "Building a Turkish treebank." In: *Treebanks: Building and Using Parsed Corpora*. Ed. by Anne Abeillé. Springer. Chap. 15, pp. 261–277.

Say, Bilge, Deniz Zeyrek, Kemal Oflazer, and Umut Özge (2002). "Development of a Corpus and a TreeBank for Present-day Written Turkish." In: *Proceedings of the Eleventh International Conference of Turkish Linguistics*. Eastern Mediterranean University, Cyprus.

Sulubacak, Umut and Gülsen Eryiğit (2013). "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*, pp. 129–134.

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