

Towards a multi-layered dependency annotation of Finnish

Alicia Burga¹, Simon Mille¹, Anton Granvik³, and Leo Wanner^{1,2}

¹ Natural Language Processing Group, Pompeu Fabra University, Barcelona, Spain

² Institució Catalana de Recerca i Estudis Avançats (ICREA)

³ HANKEN School of Economics, Centre for Languages and Business Communication

firstname.lastname@upf.edu, anton.granvik@hanken.fi

Abstract

We present a dependency annotation scheme for Finnish which aims at respecting the multilayered nature of language. We first tackle the annotation of surface-syntactic structures (SSyntS) as inspired by the Meaning-Text framework. Exclusively syntactic criteria are used when defining the surface-syntactic relations tagset. Our annotation scheme allows for a direct mapping between surface-syntax and a more semantics-oriented representation, in particular predicate-argument structures. It has been applied to a corpus of Finnish, composed of 2,025 sentences related to weather conditions.

1 Introduction

The increasing prominence of statistical NLP applications calls for creation of syntactic dependency treebanks, i.e., corpora that are annotated with syntactic dependency structures. However, creating a syntactic treebank is an expensive and laborious task—not only because of the annotation itself, but also because a well-defined annotation schema is required. The schema must accurately reflect all syntactic phenomena of the annotated language, and, if the application for which the annotation is made is “deep” (as deep parsing or deep sentence generation), also foresee how each of the syntactic phenomena is reflected at the deeper levels of the linguistic description.

For Finnish, there are two well-known syntactic dependency-based treebanks: the Turku Dependency Treebank (TDT), and the FinnTreeBank. TDT, the most referenced corpus in Finnish (Haverinen et al., 2014), contains 15,126 sentences (204,399 tokens) from general discourse and uses a tagset of 53 relations (although just 46 are used at the syntactic layer), which is an adaptation of the Stanford Dependency (SD) schema for

English (de Marneffe and Manning, 2008). The FinnTreeBank (Voutilainen et al., 2012) contains 19,764 sentences (169,450 tokens), mostly extracted from a descriptive Finnish grammar, which are annotated using a reduced tagset of only 15 relations.¹

In what follows, we present an alternative annotation schema that is embedded in the framework of the Meaning-to-Text Theory (MTT) (Mel’čuk, 1988). This schema is based on the separation of linguistic representations in accordance with their level of abstraction. Subsequently, we distinguish between surface-syntactic (SSynt) and deep-syntactic (DSynt) annotations, and argue that this schema more adequately captures the syntactic annotation of Finnish. We designed our annotation scheme empirically, through various iterations over an air quality-related corpus of 2,025 sentences (35,830 tokens), which we make publicly available. However, since this paper focuses on the principles which underlie our annotation schema, rather than on the quality of the annotated resource itself, we do not provide an evaluation of the annotation quality.

The next section outlines our annotation scheme for Finnish and discusses the main syntactic criteria for the identification of the individual relation tags. Section 3 shows how the presented annotation can be projected onto a deep-syntactic annotation, while Section 4 details the principal differences between the TDT annotation schema and ours, before some conclusions are presented in Section 5.

2 A surface-syntactic annotation of Finnish

Our annotation schema for Finnish follows the methodology adopted for the elaboration of the

¹According to KORP -<https://korp.csc.fi>- the FTB with all its versions joined contains 4,386,152 sentences (76,532,636 tokens). However, the limited number of relations makes an in-depth analysis and/or comparison difficult.

schema of the Spanish AnCora-UPF treebank (Mille et al., 2013). Taking into account a series of clearly cut syntactically-motivated criteria, a tagset of Finnish syntactic dependencies has been established. In what follows, we first present the SSynt relation tagset, and then discuss some of the main criteria applied for the identification of selected tags.

2.1 The SSynt dependency tagset

The SSynt annotation layer is language-dependent, and thus captures the idiosyncrasies of a specific language. An example of a Finnish surface-syntactic structure (SSyntS) is shown in Figure 1.

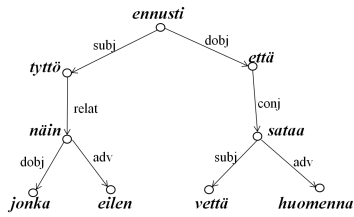


Figure 1: SSyntS of the sentence *Tyttö jonka näin eilen ennusti, että huomenna sataa vettä.* ‘The girl whom I saw yesterday predicted that tomorrow it will rain’.

The Finnish SSynt tagset contains 36 relations, which are presented and described in Table 1 along with their distinctive syntactic properties. For comparison, consider the Spanish tagset, shown in Table 2.

As can be observed, many labels in the Finnish and Spanish tagsets are identical (e.g., *clitic*, *modif*, *relat*). This uniformity of labels across languages is one of the major motivations behind the Universal Stanford Dependencies (de Marneffe et al., 2014). We also think that using the same labels across languages facilitates the understanding of the annotations but, unlike in the USD proposal, we make the different syntactic characteristics encoded by identical relations in different languages explicit. Some prominent examples of relations with the same label in both tagsets, but with different definitions are *subj*, *obl_obj* and *copul*. The relation *subj* refers in both tagsets to the element that agrees with the verb in person and number, but in Finnish the relation is also defined with respect to the case: the dependent of this relation takes the case assigned by the verb. In Spanish, given that nominal phrases do not carry case (or, at least, they do not show any case marker), the case assignment is not used for the definition of the relation.

DepRel	Distinctive properties
adjunct	mobile sentential adverbial
adv	mobile verbal adverbial
appos	right-sided apposed element
attr	genitive complement of nouns
aux	non finite V governed by auxiliary verbs
aux_phras	multi-word marker
bin_junct	relates binary constructions
clitic	non-independent adjacent morpheme attached to its syntactic governor
compar	complement of a comparative element
conj	complement of a non-coordinating Conj (right-sided)
compl	non-removable adjectival object agreeing with another verbal actant
compos	relates a nominal head with prefixed modifiers in compound nouns
copul	non-locative complement of the copula <i>olla</i> ; agrees with subject in number; its canonical order is to the right
coord	relates the first element of a coordination with the coord. conjunction (recursive)
coord_conj	complement of a coordinating Conj (right-sided)
det	non-repeatable first left-side modifier of noun
dobj	verbal dependent with case partitive, genitive, nominative or accusative (for pronouns); no agreement with verb
hyphen	reflects the orthographic necessity of hyphenating compounds
juxtapos	for linking two unrelated groups
modal	relates modal auxiliaries (which require genitive subjects) and main verb
modif	element modifying a noun; agrees in case and number
noun_compl	non-genitive complement of nouns
obj_copred	relates the main verb with a predicative adjective that modifies an object
obl_obj	verbal dependent with locative case (adessive, ablative, elative, illative, allative)
postpos	left-sided complement of an adposition or of an adverb acting as such
prepos	right-sided complement of an adposition or of an adverb acting as such
punc	for punctuation signs
quasi_coord	for coordinated elements with no connector; (e.g. specifications)
relat	right-sided finite verb modifying a noun
relat_expl	adjunct-like finite clause
restr	invariable & non-mobile adverbial unit
sequent	for numerical or formulaic elements belonging together (right-side)
subj	verbal dependent that controls number agreement on its governing verb; acquires the case assigned by the verb
subj_obj	subject-like element governed by passive, existential-possessive and impersonal verbs, with some object properties
subj_copred	relates the main verb with a predicative adjective that modifies the subject
verb_junct	right-sided verbal particle that gives the expression a particular meaning

Table 1: Dependency relations used at the Finnish surface-syntactic layer.

DepRel	Distinctive properties
abbrev	abbreviated apposition
abs_pred	non-removable dependent of an N making the latter act as an adverb
adv	mobile adverbial
agent	promotable dependent of a participle
analyt_fut	Prep <i>a</i> governed by future Aux
analyt_pass	non-finite V governed by passive Aux
analyt_perf	non-finite V governed by perfect Aux
analyt_progr	non-finite V governed by progressive Aux
appos	right-sided apposed element
attr	right-side modifier of an N
aux_phras	multi-word marker
aux_refl	reflexive Pro depending on a V
bin_junct	for binary constructions
compar	complement of a comparative Adj/Adv
compl1	non-removable adjectival object agreeing with subject
compl2	non-removable adjectival object agreeing with direct object
compl_adnom	prepositional dependent of a stranded Det
conj	complement of a non-coordinating Conj
coord	between a conjunct and the element acting as coordination conjunction
coord_conj	complement of a coordinating Conj
copul	cliticizable dependent of a copula agrees with subject in number and gender
copul_clitic	cliticized dependent of a copula;
det	non-repeatable left-side modifier of an N
dobj	verbal dependent that can be promoted or cliticized with an accusative Pro
dobj_clitic	accusative clitic Pro depending on a V
elect	non-argumental right-side dependent of a comparative Adj/Adv or a number
iobj	dependent replaceable by a dative Pro
iobj_clitic	dative clitic Pro depending on a V
juxtapos	for linking two unrelated groups
modal	non-removable, non-cliticizable infinitive verbal dependent
modif	for Adj agreeing with their governing N
num_junct	numerical dependent of another number
obj_copred	adverbial dependent of a V, which agrees with the direct object
obl_compl	right-side dependent of a non-V element introduced by a governed Prep
obl_obj	prepositional object that cannot be demoted, promoted or cliticized
prepos	complement of a preposition
prolep	for clause-initial accumulation of elements with no connectors
punc	for non-sentence-initial punctuations
punc_init	for sentence-initial punctuation
quant	numerical dependent which controls the number of its governing N
quasi_coord	for coordinated elements with the no connector
quasi_subj	a subject next to a grammatical subject
relat	right-sided finite V that modifies an N
relat_expl	adverbial finite clause
sequent	right-side coordinated adjacent element
subj	dependent that controls agreement on its governing V
subj_copred	adverbial dependent of a V agreeing with the subject

Table 2: Dependency relations used at the Spanish surface-syntactic layer.

obl_obj refers in Spanish to those verbal objects that are introduced by a preposition and cannot be demoted, promoted or cliticized. In Finnish, due to its case-inflected nouns, *obl_obj* is defined as the relation that links verbs with objects containing locative cases. Finally, *copul* is defined in both tagsets as the complement of copular verbs, which agrees with the subject in number. However, in the case of Spanish this element can cliticize, but in Finnish it cannot.

In contrast, such relation labels as *appos*, *coord* or *relat* share exactly the same properties across the two languages.

2.2 Syntactic criteria

The syntactically-motivated criteria described in (Burga et al., 2014) were used for creating the Finnish SSynt tagset. In this section, some remarks about Finnish idiosyncrasies related to these criteria are detailed.

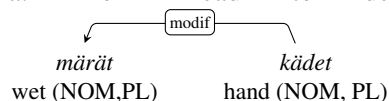
- **Agreement:** Two elements are involved in agreement if they share some morphological features, such as number, person or case. If such agreement arises because one element transmits those features to the other, we conclude that those elements are syntactically related. On the other hand, if an element that admits morphological variation does not vary according to its governor/dependent, we can conclude that no agreement is involved in the dependency relation between the two. However, as already pointed out for Spanish (Burga et al., 2014), one has to be careful when analyzing agreement, because it depends not only on the licensing from the syntactic relation, but also on the Part-of-Speech (PoS) of each element. Thus, if the element to which the morphological feature(s) is (are) transmitted from another has a PoS that does not allow any morphological variation –or is lexically invariable, despite having a PoS that admits variability–, the agreement will not be visible. Then, to evaluate if agreement actually exists, one needs to use the prototypical head and dependent for each relation.² When applying this criterion, it is also important to keep in mind that different syntactic relations allow different types of agreement, namely: i) head transmits features to dependent (e.g., *modif*) (1a); ii) dependent transmits features to head (e.g., *subj*) (1b); and iii) dependent transmits features to a sibling

²This point is important because the non-visibility of agreement can cause a wrong division of relations, as happens in the TDT annotation scheme (see Section 4).

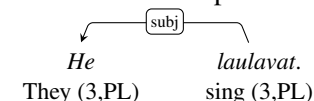
(e.g., *copul*) (1c).

(1) Possible agreement transmissions:

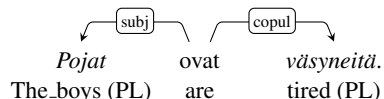
a. from head to dependent:



b. from dependent to head:

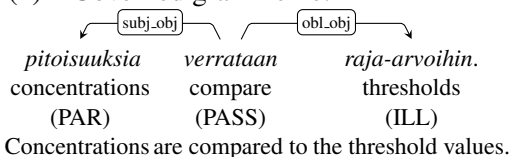


c. between two siblings:

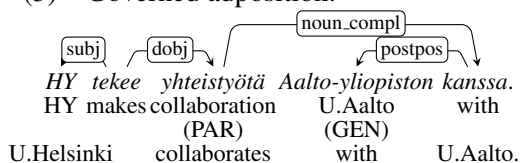


• **Governed Adposition / Conjunction / Grammeme:** Some relations require the presence of a preposition, a subordinating conjunction, or a grammeme (as, e.g., verbal finiteness or case). In Finnish, differently from English or Spanish, adpositions and inflected nouns are both admitted as alternative ways of expressing the same meaning.³ However, beyond the way the meaning is conveyed at the surface, some units (namely the functional elements) are governed and some units (namely the content elements) are not. The governed elements in Finnish are mostly grammemes (case features), although it is also possible to find specific examples with governed adpositions. In the annotation scheme presented in this paper, this criterion is used for establishing the tagset (e.g., the relation *subj* does not require a particular case – the acquired case depends on the verbal head – whereas the relation *attr* requires genitive in the dependent), but does not imply a different analysis of configurations with governed and non-governed elements.

(2) Governed grammeme:

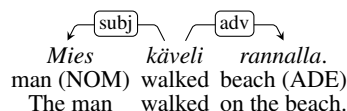


(3) Governed adposition:

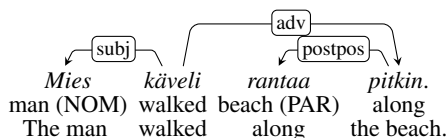


³This is the reason behind the TDT treating both kinds of configurations in the same way (see Section 4).

(4) Non-governed grammeme:



(5) Non-governed adposition:



In (2–5), we display examples that illustrate governed and non-governed cases and adpositions. In (2), the case ILL of *raja-arvo* ‘threshold values’ is governed by the verb *vertaa* ‘compare’, and this requirement is what defines the type of relation holding between the verb and the inflected noun (*obl_obj*). In (3), the postposition *kanssa* is required by the predicate *tehdä yhteistyötä* ‘collaborate’, which motivates the relation *noun_compl*.⁴ On the other hand, the adessive case in *ranta* ‘beach’ in (4) and the adposition *pitkin* ‘along’ in (5) are not required by any element. As a consequence, they contribute by themselves to the semantics of the sentences – which should be reflected at the deep-syntactic layer.

• **Linearization / Canonical order:**⁵ By linearization/canonical order we make reference to the required (or preferred) direction between governor and dependent within a specific dependency relation. Although Finnish is a language with a quite flexible word order, there are certain syntactic relations that require a rigid linearization (e.g., *appos*) or, at least, prefer a certain order between head and dependent (e.g., *dobj*, *copul*).

As these criteria contribute to the definition of SSynt relations, they also serve, along with some features of the elements involved, to distinguish different syntactic configurations. For instance, the verb *olla* ‘to be’ is used in copulative, locative, and existential configurations. Therefore, we need some criteria to identify each of these uses.

In a copulative sentence, the subject is the element that agrees in person and number with the

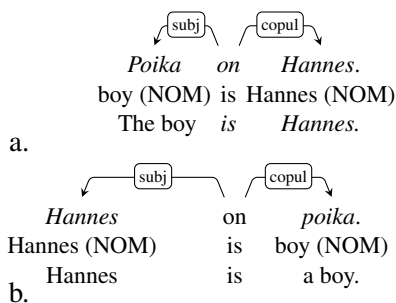
⁴As the predicate comprises two elements, and the predicate itself is a noun, the relation is *noun_compl*. However, if the predicate were composed by just one verbal element, the relation received by the adposition would be the same as in (2), *obl_obj*.

⁵Thanks to a reviewer for providing some important Finnish judgments that have contributed to clarify this section.

verb and carries nominative case. The complement of the copula, on the other hand, is “the element that says something about the subject”. It can be of four different types: i) a non-nominal element (such as an adjective), ii) a nominal element in a case different from nominative, iii) a nominal element in nominative that does not agree with the verb in person and/or number, and iv) a nominal element in nominative that also agrees with the verb in person and/or number.

In cases i–iii), the two previous criteria – agreement and governed grammeme – are enough for detecting subjects and complements of the copula. However, in cases where the two elements related to the verb are nominal elements that agree with the copula and are in nominative case, as in (6), linearization helps to determine which element is the subject (i.e., the element appearing before the copula) and which one is the complement of the copula (i.e., the element appearing after the copula).⁶ Thus, as observed, (6a) and (6b) do not carry the same meaning: they are not exchangeable and (6b) is not the result of exchanging directions over the relations of (6a).

(6) Copulative:

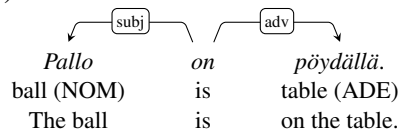


The *copul* relation, thus, conveys a rigid linearization when combined with certain morphological features, and therefore this criterion should explicitly intervene in the definition of the relation.

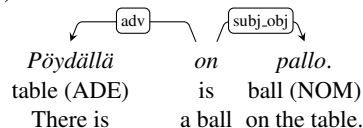
In the same way, locative sentences containing *olla* require the relation *adv* to be right-sided (7), opposite to existential sentences, which require it to be left-sided (8). Again, this distinction only applies in cases where the non-locative element is non-definite. If it is definite (e.g., a definite modifier is explicitly added), no existential interpretation is possible and therefore the distinction between locative and existential vanishes.

⁶Even if it is possible to find sentences with the two nominal elements at the same side of the copula, they are not interpreted as neutral copulative sentences, but are communicatively marked.

(7) Locative:



(8) Existential:



3 Towards a deep-syntactic annotation

Since we approach linguistic description in a multilayered way, our annotation scheme aims at obtaining not only the Surface-Syntactic layer, but also a shallow semantics-oriented layer, referred to as *Deep-Syntactic* (DSynt) layer in the Meaning-Text Theory. An example of a DSynt structure for Finnish is shown in Figure 2.

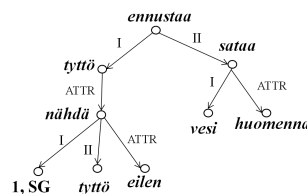


Figure 2: DSyntS of the sentence *Tyttö jonka näin eilen ennustaa, että huomenna sataa vettä.* ‘The girl whom I saw yesterday predicted that tomorrow it will rain’.

The main differences between a Surface-Syntactic structure (SSyntS) and a Deep-Syntactic structure (DSyntS) are the following:

- (i) a SSyntS contains all the words of a sentence, while in a DSyntS all functional elements (such as governed adpositions or auxiliaries) are removed, so that only meaning-bearing (content) elements are left; Figure 2, for instance, does not contain the subordinating conjunction *että* present in Figure 1;
- (ii) the SSynt tagset is language-idiosyncratic whereas in the DSyntS relations between the content elements are generic and predicate-argument oriented (thus, language-independent); for instance, *subj* and *dojb* in Figure 1 map to argumental relations in Figure 2 (respectively *I* and *II*), while *relat* and *adv* are mapped to the non-argumental relation *ATTR*.

In other words, during the mapping between surface- and deep-syntax, functional elements and

predicate-argument relations have to be identified. Thanks to the existence of dedicated tools such as the graph-transducer MATE (Bohnet et al., 2000), the mapping of the SSynt-annotation onto the DSynt-annotation is facilitated. For instance, Mille et al. (2013) describe how they obtain the DSynt annotation of a Spanish treebank. To make the mapping straightforward, predicate-argument information is included in the tags of surface-syntactic annotation, enriching surface-syntactic relations with semantic information. Thus, for instance, instead of simply annotating the relation *obl_obj* when this relation is identified, specifying the argument number in the label is also required: *obl_obj0* corresponds to the first argument, *obl_obj1* to the second argument, *obl_obj2* to the third argument, etc. Then, their mapping grammar simply converted the labels and removed functional elements, before removing the predicate-argument information from the superficial annotation. For Finnish, instead, we followed another approach: we included a valency dictionary in which we store subcategorization information, i.e., the distribution of the arguments of a lemma and required functional elements associated with each of the arguments⁷. For illustration, see a sample entry of such a lexicon in Figure 3.

```
ennustaa {
  POS = V
  gp = {
    I = {rel= subj|dpos = N|case = NOM}
    II = {rel = dobj|dpos = N|case = GEN}
    III = {rel = comp1|dpos = A|case = GEN}
  }
  gp = {
    I = {rel= subj|dpos = N|case = NOM}
    II = {rel = dobj|dpos = V|case = GEN
        conj = että|finiteness = FIN}
  }
}}
```

Figure 3: Sample lexicon entry for *ennustaa* ‘to predict’.

The entry for *ennustaa* ‘to predict’ states that this word is a verb (*Pos* = *V*) and that it has two possible government patterns (*gp*): one with three arguments and one with two arguments. Consider *HSY ennustaa pölyämisen jatkuvan* ‘HSY predicts the dust to continue’ for the first and *Metla ennustaa, että koivu kukkii ...* ‘Metla predicts that the birch will be in bloom ...’ for the latter.

Thanks to this lexicon, rules can check in the input SSyntS if a word has a dependent of the type described in its entry, and perform the adequate mapping. For instance, if a dependent of *ennustaa* is a noun in the nominative case with the depen-

dency *subj*, the latter will be mapped to *I* in the DSyntS. A nominal dependent in the genitive case with a dependency *dobj* would be mapped to the second argument (*II*), while a nominalized verb in genitive receiving the dependency *comp1* would be mapped to its third argument (*III*). In the lexicon, governed conjunctions are also described, as in the description of the second argument of the second governed pattern: in this case, if *ennustaa* has a dependent *dobj* which is the conjunction *että*, which itself introduces a finite verb, not only will *dobj* be mapped to second argument (*II*), but the governed (functional) element will be removed, so that *II* will link both content words of the substructure, i.e., *ennustaa* and the dependent verb.

The lexicon currently contains more than 1400 entries, including about 300 verbs, 750 nouns, 220 adjectives, 50 adverbs and 100 prepositions, postpositions and conjunctions.⁸

One great advantage of this method is that this resource is not only useful for obtaining lexical valency information from syntactic structures, but also in the framework of rule-based text generation, that is, for the exact opposite mapping (producing syntactic relations and functional elements from abstract predicate-argument structures (Wanner et al., 2014)).⁹

4 Comparison with the TDT annotation scheme

In this section, we present a contrastive analysis of the TDT annotation scheme, the most referenced scheme for Finnish, with respect to its treatment of certain phenomena.

The last version of TDT (Haverinen et al., 2014) contains two layers of annotation. The first layer (the base-syntactic layer) contains 46 relations and

⁸The lexicon furthermore contains additional information about the entries which is not related to subcategorization, such as morphological invariability, as well as the values for some lexical functions.

⁹A number of other annotations have resemblance with DSyntSs; cf. (Ivanova et al., 2012) for an overview of deep dependency structures. In particular, DSyntSs show some resemblance, but also some important differences, with PropBank structures, mainly due to the fact that the latter concern phrasal chunks and not individual nodes. The degree of “semanticity” of DSyntSs can be directly compared to Prague’s tectogrammatical structures (Hajič et al., 2006), which contain *autosemantic* words only, leaving out *synsemantic* elements such as determiners, auxiliaries, (all) prepositions and conjunctions. Collapsed SDs (de Marneffe et al., 2006) differ from the DSyntSs in that they collapse only (but all) prepositions, conjunctions and possessive clitics, they do not involve any removal of (syntactic) information, and they do not add semantic information compared to the surface annotation.

⁷As, e.g., in (Gross, 1984), and the *Explanatory Combinatorial Dictionary* (Mel’čuk, 1988).

uses the SD scheme adapted to Finnish. The second layer inserts additional dependencies over the first layer. This second layer tries, on the one hand, to cover more semantic phenomena (conjunct propagation for coordinations, and external subjects), but, on the other hand, it aims at covering some syntactic phenomena—gaps resulting from the first layer annotation—such as describing the function of relative pronouns.¹⁰

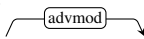
In the following, we present the principal characteristics of the pure-syntactic first layer annotation of TDT, focusing on the most relevant differences between TDT and the annotation scheme presented in this paper.

- Many relations in the TDT annotation scheme are based on the PoS and internal morphological processes of the dependent and/or the governor, rather than on particular syntactic properties of the relations themselves. Even if it cannot be denied that some PoS carry restrictions that others do not, it is important to recognize when those restrictions are imposed by morpho-syntactic factors and, therefore, should not be confused with pure syntactic restrictions. Thus, the TDT annotation scheme distinguishes between two different relations *advmod* and *nommod* for verbal modifiers (9), but the distinction is based only on the PoS of the dependent.¹¹

(9) Distinguishing relations using PoS:


- a. The dependent is an adverb:

Hän käveli kotiin hitaasti.
He walked home slowly.



- b. The dependent is a noun:

Maljiakko oli pöydällä.
The_vase was on_the_table.



Not only is the PoS information duplicated in the annotation, but in those cases in which it is difficult to decide if a word is a noun or an adverb (e.g., *pääasiassa* ‘mainly’ (adverb) / ‘main thing’ (noun)), if a wrong PoS tag is chosen, the annotation error directly propagates to the syntactic annotation, as Haveri-

¹⁰The authors explain that this information is omitted in the first layer because of treeness restriction (Haverinen et al., 2014, p.505).

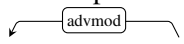
¹¹In this section, we have tried to use the examples presented in (Haverinen, 2012), but in some cases these examples have been shortened/adapted according to format restrictions.

nen et al. (2013) point out. If the syntactic behavior is not different when a dependent is an adverb or a noun, only one syntactic relation should be needed.

Given that the TDT tagset sub-specifies some dependency tags according to the PoS of the elements involved, it is perfectly possible to choose an annotation that links heads and dependents that belong to different clauses (without being a relative configuration), as in (10). Such analysis is not syntactically accurate, given that it completely ignores the syntactic independence of each clause.

(10) Edge between independent clauses:

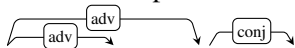
Tulen heti, kun pääsen.
I_will_come right_away, when I.can.



In contrast, we keep the syntactic independence of each clause, and relate one to each other through the relation *adv* (11).¹²

(11) Clause independence respected:

Tulen heti, kun pääsen.
I_will_come right_away, when I.can.

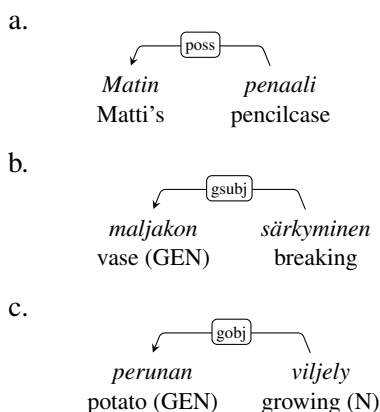


- When adapting the SD scheme to Finnish, some relations in the TDT annotation were ruled out for being considered “semantic in nature” (Haverinen et al., 2014, p.504). Nevertheless, the analysis of some other phenomena – and the consequent definition of dependencies related to them – still has a more semantic justification than a syntactic one. A first example of this observation, also related to the previous point, is the division of the genitive modifiers of nouns into three different relations: *poss* (12a), *gsubj* (12b) and *gobj* (12c). Although it is argued that such a division responds to the desire of obtaining a higher granularity of the scheme (Haverinen et al., 2014, p.507), the relation division actually depends on the semantics of the governor and not on the syntactic properties of these constructions. Thus, in (12a), *Matin* is a genitive modifier of the noun *penaali* ‘pencilcase’; in (12b), due to the semantics of the head, *maljakon* ‘vase’ is considered a “subject-like” modifier of *särkyminen*

¹²Another way to analyze this sentence is considering a relative configuration, the subordinating clause being a specification of *heti* ‘right away’ / ‘this moment’.

‘breaking’; and in (12c), *perunan* ‘potato’ is considered a nominal modifier of *viljely* ‘growing’, but it is actually analyzed as a genitive object of the verb *viljellä* ‘to grow’. The annotation scheme assumes, as (12b) and (12c) show, that the nominalization process undergone by the verb makes it transmit not only its semantics, but also its syntactic properties. As expected, when the annotation concerns genitive modifiers of nouns, the annotation errors propagate (Haverinen et al., 2013).

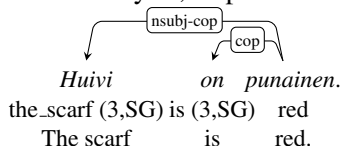
(12) Distinguishing modifiers of nouns:



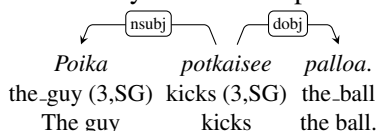
In the annotation schema presented in this paper, the three constructions are parallel and use the relation *attr*.

Another clear example of the prevalence of semantics over syntax in TDT is the treatment of copular verbs. They are treated in a specific way (13), different from any other verb (14), due to the semantic link between the subject and the complement of the copular verb.¹³

(13) TDT analysis, copulative sentences:



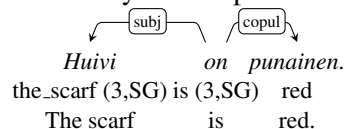
(14) TDT analysis of non-copulative:



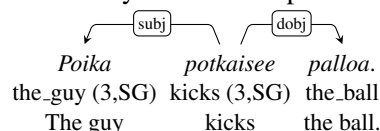
¹³The TDT annotation faces a problem of not resulting in a tree when, instead of a subject noun, a participial modifier appears. Thus, in those cases, they treat a copulative configuration as any other verbal construction, which weakens their original analysis (Haverinen, 2012, Section 5.13).

In both sentences, the verb agrees with the preverbal element in person and number, which is the morphological marker of the syntactic phenomenon of being a subject. However, the analysis assigned to each sentence does not capture such parallelism. The difference between both sentences concerns the second verbal complement: in copulative sentences, if its PoS licenses agreement, this element agrees with the subject in number; in non-copulative sentences, such an agreement does not happen. Therefore, two different relations hold between the verb and this complement, as (15) and (16) show.

(15) Our analysis of copulative sentences:

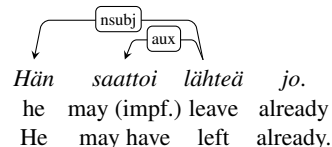


(16) Our analysis of non-copulative:



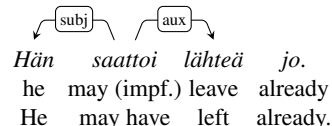
Finally, the prevalence of semantics over syntax in TDT is exemplified through the treatment of subjects, auxiliaries and content verbs. The TDT annotation schema takes the content verb as head of the sentence, and makes the subject hold on it (17).

(17) TDT treatment of auxiliaries:



If syntactic properties are prioritized in the course of the definition of the annotation schema, the subject relation should link the subject and the auxiliary (18), given that agreement holds between these two elements. Consequently, the auxiliary should head the relation between the two verbs. In the same way, the negative auxiliary should be also treated as the element heading the subject and the content verb.

(18) Our treatment of auxiliaries:



- Given the semantic motivation for annotating differently similar syntactic phenomena (or vice versa), we would expect the TDT annotation schema to allow for a direct mapping from surface-syntax to deeper linguistic levels (or, in more concrete terms, to a predicate-argument structure, which we refer to as “semantics”). However, this is not the case.

As detailed in Section 2.2, case markers and adpositions can be either functional or meaning-bearing, and each of them should be treated differently. TDT, however, treats as the same, on one hand, case markers and adpositions (Haverinen, 2012, p.2) and, on the other hand, elements that are purely functional and those ones that do convey a content. The examples in (19) show TDT’s parallel treatment of case markers and adpositions (compare (19a) to (19b)), and of governed and non-governed elements (compare (19b) to (19c)). As can be observed, the same syntactic analysis is offered to sentences that differ in syntax: in (19a), the adessive case of *pöytä* ‘table’ is required for expressing a locative meaning with the verb *olla*, whereas in (19b), the genitive case is required by the adposition and not by the verb or the configuration itself. On the other hand, non-governed elements (such as *päällä* ‘on_top_of’ in (19b)) are treated in the same way as governed elements (such as *kanssa* ‘with’ in (19c)).

(19) TDT treatment of adpositions:

- a.
-
- Maljiakko oli pöydällä.*
The_vase was on_the_table
- b.
-
- Maljiakko oli pöydän päällä.*
The_vase was table on_top_of
- c.
-
- HY tekee yhteistyötä Aalto-yliopiston kanssa.*
U.H. collaborates U.Aalto. with

One problem of treating functional and content elements in the same way is the difficulty in reaching an actual abstract structure which contains only content words. (20) is an expansion of (19c) where, apart from the governed adposition, there is a translative case

(*-ksi*), expressing purpose, which is not required by the predicate. In an abstract structure corresponding to (20), the governed adposition should not appear, unlike the non-governed case.

- (20) *HY tekee yhteistyötä Aalto-yliopiston kanssa uudenlaisen digitaalisen oppimisen tukemiseksi.*
‘The university of Helsinki collaborated with the University Aalto to promote a new way of digital learning.’

Another example of the difficulty of getting an appropriate mapping between syntax and semantics is the treatment of relative pronouns: in the first layer of annotation, all relative pronouns receive the same relation from the subordinate verb (i.e., *rel*), without taking into account the syntactic function of the pronoun within the subordinate clause (21).

(21) TDT treatment of relative pronouns:

- a.
-
- auto, joka ohitti meidät*
the_car that (NOM) passed us
- b.
-
- mies, jonka näin eilen*
the_man that (GEN) I_saw yesterday

Even though a case can indicate the function occupied by the element to which it is attached, it is not enough for obtaining a direct mapping to semantics. First of all, many times, cases themselves are not enough for indicating such function, but their combinability with the involved verbs is also needed. Secondly, and more importantly, the same cases are used by elements that occupy different semantic slots. Thus, for instance, both subjects and objects accept the same set of cases (nominative, partitive and genitive), which clearly blurs a direct mapping to predicate-argument structures. In our syntactic annotation scheme, *rel* would be annotated as a subject in (21a), and as object in (21b).

5 Conclusions

In this paper, we presented an annotation schema for Finnish that can be considered an alternative

to the SD-oriented schema used in the TDT treebank. We justify and present a syntactically motivated tagset for Finnish, and the creation of a lexicon which facilitates the annotation of a deep syntactic (semantics-oriented) representation which captures lexical valency relations between content lexical items. Having two distinct levels for capturing syntactic and semantic information, has been shown to allow for developing different NLP applications in the parsing and the natural language generation fields (Ballesteros et al., 2014; Ballesteros et al., 2015).

The corpus annotated following the SSynt and DSynt annotation schemata described in this paper are made available upon request.

Acknowledgements

The work described in this paper has been carried out in the framework of the project *Personalized Environmental Service Configuration and Delivery Orchestration* (PESCaDO), supported by the European Commission under the contract number FP7-ICT-248594.

References

- M. Ballesteros, B. Bohnet, S. Mille, and L. Wanner. 2014. Deep-syntactic parsing. In *Proceedings of COLING*, Dublin, Ireland.
- M. Ballesteros, B. Bohnet, S. Mille, and L. Wanner. 2015. Data-driven sentence generation with non-isomorphic trees. In *Proceedings of NAACL-HLT*, Denver, CO, USA.
- B. Bohnet, A. Langjahr, and L. Wanner. 2000. A development environment for an MTT-based sentence generator. In *Proceedings of INLG*.
- A. Burga, S. Mille, and L. Wanner. 2014. Looking behind the scenes of syntactic dependency corpus annotation: Towards a motivated annotation schema of surface-syntax in Spanish. In *Computational Dependency Theory. Frontiers in Artificial Intelligence and Applications Series*, volume 258. Amsterdam:IOS Press.
- M.C. de Marneffe and Ch. Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation (COLING)*, Manchester, UK.
- M.C de Marneffe, B. MacCartney, and Ch. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 449–454, Genoa, Italy.
- M.C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and Ch. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4585–4592, Reykjavik, Iceland.
- M. Gross. 1984. Lexicon-grammar and the syntactic analysis of French. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING) and the 22nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 275–282, Stanford, CA, USA.
- J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, and Z. Žabokrtský. 2006. Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia.
- K. Haverinen, F. Ginter, V. Laippala, S. Kohonen, T. Viljanen, J. Nyblom, and T. Salakoski. 2013. A dependency-based analysis of treebank annotation errors. In K. Gerdes, E. Hajičová, and L. Wanner, editors, *Computational Dependency Theory*. IOS Press.
- K. Haverinen, J. Nyblom, T. Viljanen, V. Laippala, S. Kohonen, A. Missilä, T. Salakoski, and F. Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. In *Proceedings of LREC*, Reykjavik, Iceland, September.
- K. Haverinen. 2012. Syntax Annotation Guidelines for the Turku Dependency Treebank. *Technical Report 1034*, Turku Centre for Computer Science, Turku, Finland.
- A. Ivanova, S. Oepen, L. Øvrelid, and D. Flickinger. 2012. Who did what to whom? A contrastive study of syntacto-semantic dependencies. In *Proceedings of the 6th Linguistic Annotation Workshop*, pages 2–11, Jeju, Republic of Korea.
- I. Mel'čuk. 1988. Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, 1(3):165–188.
- I. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.
- S. Mille, A. Burga, and L. Wanner. 2013. AnCora-UPF: A multi-level annotation of Spanish. In *Proceedings of DepLing*, Prague, Czech Republic.
- A. Voutilainen, K. Muhonen, T. Purtonen, and K. Lindén. 2012. Specifying treebanks, outsourcing parsebanks: Finntreebank 3. In *Proceedings of LREC*, Istanbul, Turkey.
- L. Wanner, H. Bosch, N. Bouayad-Agha, G. Casamayor, Th. Ertl, D. Hilbring, L. Johansson, K. Karatzas, A. Karppinen, I. Kompatsiaris, et al. 2014. Getting the environmental information across: from the web to the user. *Expert Systems*.