



Katri Haverinen

Syntax Annotation Guidelines for the Turku Dependency Treebank

2nd edition, revised for the treebank release of July 2013

TURKU CENTRE *for* COMPUTER SCIENCE

TUCS Technical Report
No 1034, January 2012



Syntax Annotation Guidelines for the Turku Dependency Treebank

2nd edition, revised for the treebank release of July 2013

Katri Haverinen

University of Turku, Department of Information Technology

Joukahaisenkatu 3–5 B, 20520 Turku, Finland

`kahave@utu.fi`

TUCS Technical Report

No 1034, January 2012

Abstract

This document describes the syntax annotation scheme of the Turku Dependency Treebank. The treebank is annotated using a modified version of the well-known Stanford Dependency (SD) scheme, which represents the syntax of a sentence as a tree of labeled, directed dependencies. The SD scheme has originally been designed for English, and thus it has been modified in the annotation process, in order to accommodate the specific features of the Finnish language.

We first give a brief description of the original SD scheme and then proceed to describe the dependency types used in the Finnish specific version. Next, we discuss the most important changes between the original and the Finnish specific schemes, and finally, we give instructions for annotating specific phenomena within the Finnish language.

This document has been revised to reflect the annotation in the July 2013 release of the treebank, as described in the paper of Haverinen et al. [4]. The revisions include, most importantly, describing the second annotation layer of the treebank and related changes, as well as few additional smaller clarifications.

Keywords: syntax, parsing, treebanking, Finnish

TUCS Laboratory
Bioinformatics

1 Turku Dependency Treebank and the Stanford Dependency scheme

This document describes the syntax annotation guidelines of *Turku Dependency Treebank* (TDT), a manually annotated treebank of Finnish. The treebank, its documentation and its current associated publications are available at the address <http://bionlp.utu.fi>. TDT was developed primarily for *natural language processing* purposes, but it can also act as a valuable source of material for other language research.

The Stanford Dependency (SD) scheme was originally developed for English by de Marneffe and Manning [1, 7]. It is a dependency scheme, meaning that the syntactic structure of a sentence is represented as a graph of binary *dependencies* between words. The dependencies are directed: each dependency has a *governor* or *head word* and a *dependent*. Each dependency also has a *dependency type* that describes the syntactic function of the dependent word. The dependency types are arranged in a hierarchy, so that each type is directly or indirectly a subtype of the most general type *dep* (dependent). The most specific type possible in a given situation is always used; the types higher in the hierarchy are meant for situations where choosing a more specific type is impossible.

There are four different *variants* of the SD scheme, each of which includes a different subset of dependency types and gives a different amount of information about the sentence structure. The annotation of TDT is divided into two different layers. The first layer or the *base layer* of annotation is based on the *basic* variant of SD. This means that (with the exception of one dependency type) the analyses in the first layer of TDT are trees, and the dependency types encode mostly syntactic information. The second layer of TDT is termed *conjunct propagation and additional dependencies*, and all the phenomena annotated in it are also covered in extended variants of the SD scheme. The first layer of annotation is described in Section 2 and the *conjunct propagation and additional dependencies* layer in Section 3.

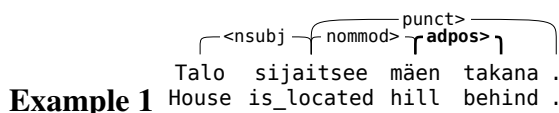
The SD scheme was originally developed for English, and there exists a version for Chinese. Although in general the scheme has been developed partly with language-independence in mind, slight modifications have been made in order for it to suit the specific features of Finnish. The most important differences between the Finnish and English SD schemes are discussed in Section 4, and analyzing specific syntactic structures is described in detail in Section 5.

2 Basic dependency types

This section contains the basic uses of each of the 46 basic dependency types belonging to the Finnish-specific version of the SD scheme.

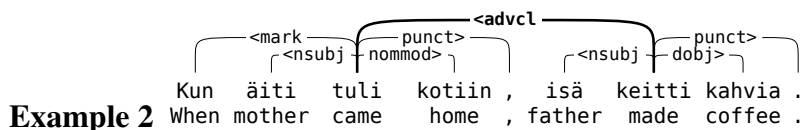
2.1 adpos (adposition)

The dependency type *adpos* is used for the adposition in pre- and postpositional phrases. In the Finnish-specific SD scheme, the head of an adpositional phrase is the nominal, not the adposition, so as to analyze adpositional phrases similarly to nominal modifiers without an adposition. (Such nominal modifiers are frequent in Finnish, as cases are often used for the same purpose as adpositions.) To the same end, the type *adpos* is used in combination with the type *nommod*, which is also used for nominal modifiers when no adposition is present (see Section 2.30).



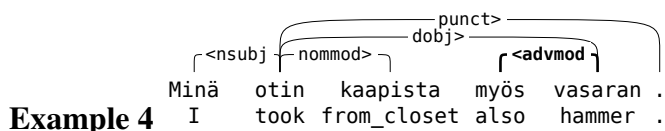
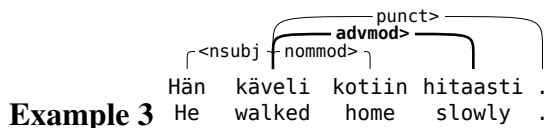
2.2 advcl (adverbial clause modifier)

Adverbial clause modifiers (advcl) are subordinate clauses that are not complements. Also non-complement infinite or temporal clauses¹ are marked as *advcl*. If there is a subordinating conjunction present, it is marked with the dependency type *mark* (see Section 2.26).



2.3 advmod (adverb modifier)

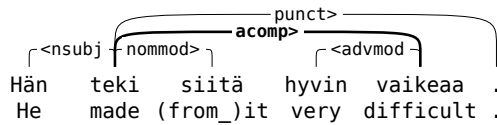
The dependency type *advmod* is used for *adverb modifiers* of verbs, nominals and adverbs alike.



¹*lauseenvastike*, see for instance [3, §876]

2.4 acomp (adjectival complement)

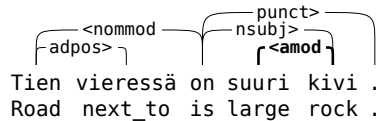
The dependency type *acomp* is used for adjectival complements of verbs, except for predicatives.



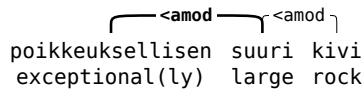
Example 5

2.5 amod (adjectival modifier)

Nouns may take adjectival modifiers, which are marked with the dependency type *amod*. It is also possible for an adjective to take another adjective as a modifier.²



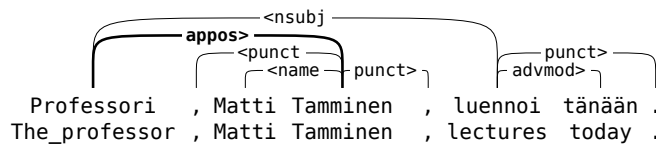
Example 6



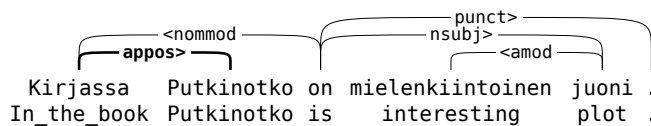
Example 7

2.6 appos (apposition)

An *apposition* (*appos*) is a grammaticalized, paradigmatic addition (usually a noun phrase), which has the same referent as its head word, and the same grammatical function [3, §1059]. Also structures with a *supporting noun* [3, §567] are considered appositional. Appositional structures and the closely related *appellation modifiers* are discussed in detail in Section 5.3.



Example 8



Example 9

²These adjectival modifiers are generally expressed with *-ly* adverbs in English.

2.7 aux (auxiliary)

In TDT, only a closed list of verbs can act as auxiliaries, including the main modal verbs [3, §1562] and in addition the verbs *olla* (to be) and *aikoa* (to be going to). The full list of auxiliaries in TDT is thus as follows:

- täytyä (must)
- pitää (have to)
- tarvita (need)
- joutua (have to)
- voida (be able to, can)
- saattaa (may)
- taitaa (be+probably, may)
- mahtaa (be+probably, may)
- olla (be)
- aikoa (be going to)

Example 10

<nsubj	<aux	<advmod>	punct>
┌──────────┬──────────┬──────────┬──────────┐			
Hän	saattoi	lähteä	jo .
He	may(impf.)	leave	already .

2.8 auxpass (passive auxiliary)

The only *passive auxiliary (auxpass)* in Finnish is *olla* (to be). An auxiliary is only considered a passive auxiliary if the main verb is in passive, not if only the auxiliary is in passive. In the latter case the auxiliary is marked as a non-passive auxiliary, *aux*. The distinction between the passive voice and the zeroth person is discussed in Section 5.15.

Example 11

<nomod	<auxpass>	<obj>	punct>
┌──────────┬──────────┬──────────┬──────────┐			
Suunnitelmaan	on	tehty	muutoksia .
Into_the_plan	have_been	made	changes .

Example 12

<nomod	<aux>	<obj>	punct>
┌──────────┬──────────┬──────────┬──────────┐			
Suunnitelmaan	voidaan	tehdä	muutoksia .
Into_the_plan	can_be	made(1st_inf.)	changes .

2.9 cc (coordinating conjunction)

Coordinating conjunctions are marked as dependents of the first coordinated element, and the dependency type used is *cc*. Coordinating conjunctions are a closed class of words, and the main conjunctions are as follows:

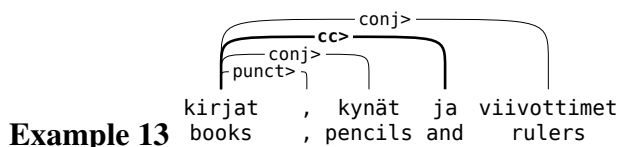
- ja (and)
- sekä (and)
- sekä... että (both... and)
- -kä (attached to negation words, nor)
- eli (a.k.a.)
- tai (or)
- vai (or, in a question context)
- joko... tai (either... or)
- mutta (but)
- vaan (but, in a negative context)

In addition, certain less frequent words or combinations of words are marked as coordinating conjunctions in TDT, namely:

- &
- elikkä (colloquial version of *eli*, a.k.a)
- ja / tai (and / or)
- ja toisaalta (and on the other hand)
- kuin (as/like)
- kuin myös (as also)
- kuten (like also)
- milloin... milloin (when... when)
- mitä... sitä (the... the)
- niin... kuin (as well as)
- niin kuin (like)

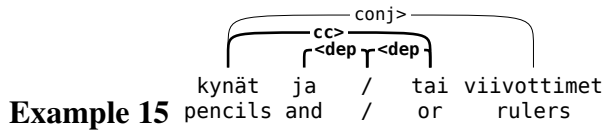
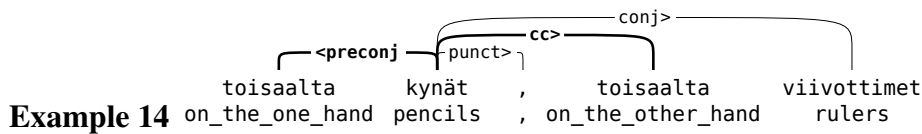
- paitsi että (except that)³
- paitsi... myös (not only... but also)
- saati (let alone)
- saati että (let alone that)
- saatikka (let alone)
- samoin kuin (“the same way as”)
- siinä missä (“as much as”)
- sitä... mitä (the... the)
- sitä mukaa... mitä (a version of *the... the*)
- sun muuta (et cetera)
- toisaalta... ja/mutta toisaalta (on the one hand... and/but on the other hand)
- toisaalta... toisaalta (on the one hand... on the other hand)
- vaikka (although)⁴
- vuoroin... ja vuoroin (in turn... and in turn)
- vuoroin... vuoroin (in turn... in turn)
- yhtä lailla... kuin (+kin) (as well as (also))
- ym. (etc.)

Coordinating conjunctions that consist of parts separated by coordinated elements are marked so that the first part is marked with the type *preconj* (see Section 2.39) and the second part with *cc* in the regular fashion. Adjacent parts of conjunctions are joined together with the most general dependency type *dep* (Section 2.18), the rightmost word being the head.



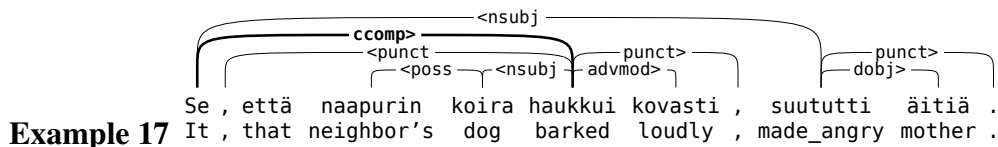
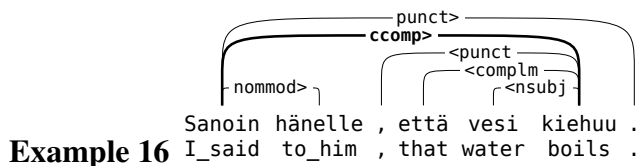
³a two-part *preconjunction*, see Section 2.39

⁴also a subordinating conjunction



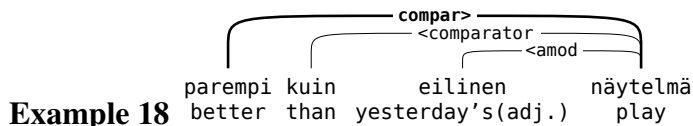
2.10 ccomp (clausal complement)

Clausal complements that have a subject different from that of the main clause⁵ are marked with the dependency type *ccomp*. The governor is most commonly, although not always, the main verb or predicative of the main clause, and the dependent is the main verb or predicative of the dependent clause. The clausal complement can also modify a word other than a verb, most often a noun or pronoun. Most commonly clausal complements are *että*-clauses. Distinguishing different verbal dependents, including different clausal complements, is discussed more closely in Section 5.4.



2.11 compar (comparative)

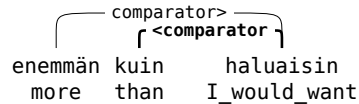
The dependency type *compar* is used in comparative constructions, most often involving adjectives in the comparative form. The head of the *compar* dependency is the comparative wordform, and the dependent is the compared element. Annotating comparative and superlative structures is described in Section 5.9.



⁵Note that a clausal complement need not have a subject present at all; the clause could be, for instance, passive.

2.12 comparator (comparative conjunction)

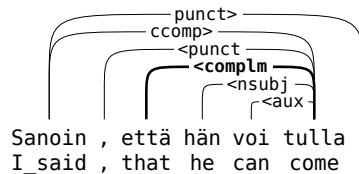
The *comparative conjunction* (most commonly *kuin*, which corresponds roughly to *than* and *as* in English) is marked with the dependency type *comparator*. The head of the dependency is the element being compared.



Example 19 enemmän kuin haluaisin
more than I_would_want

2.13 complm (complementizer)

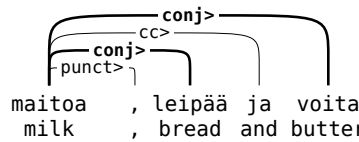
The *complementizer* (*complm*) appears in most clausal complements, and the head of the dependency is the main verb of the subordinate clause. The only complementizer in Finnish is *että* (*that*).



Example 20 Sanoin, että hän voi tulla.
I_said, that he can come.

2.14 conj (coordinated element)

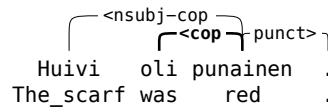
The SD scheme marks coordination so that the first coordinated element acts as the head, and the rest of the elements in the coordination, as well as the coordinating conjunction, depend on it. *Coordinated elements* are marked with the dependency type *conj*.



Example 21 maitoa, leipää ja voita
milk, bread and butter

2.15 cop (copula)

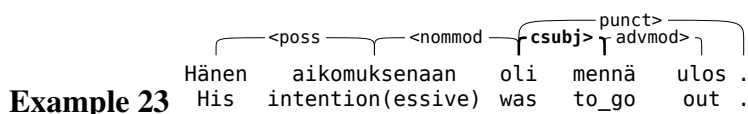
Copular clauses receive a special treatment in the SD scheme. The predicative acts as the head word of the clause, and the copular verb depends on it using a *cop* (*copula*) dependency. The only copular verb in Finnish is *olla* [3, §891]. Distinguishing copular structures from other constructs as well as recognizing the subject and the predicative is discussed in Section 5.2.



Example 22 Huivi oli punainen.
The_scarf was red.

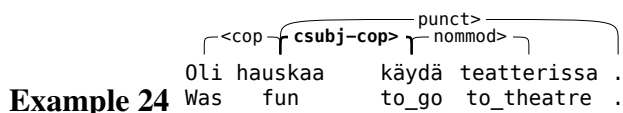
2.16 csubj (clausal subject)

A *clausal subject* (*csubj*) is a clause that acts as the subject of another clause.



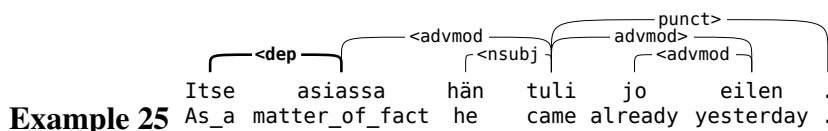
2.17 csubj-cop (clausal copular subject)

A *clausal copular subject* (*csubj-cop*) is a clause that acts as the subject of another, copular clause. As in all copular clauses, the predicative acts as the head of the clause and hence it is also the governor of the copular subject. The distinction between clauses acting as the copular subject of another clause and so called *necessive* structures is discussed in Section 5.14.



2.18 dep (dependent)

The name of the dependency type *dep* stands for *dependent*. It is the most general dependency type in SD, and it is meant to be used when no other, more specific dependency type applies. In TDT, the type *dep* is mostly used in idiomatic two-word expressions. There are also some fixed three-word expressions in the treebank.



The following expressions are considered idiomatic, and their parts are to be combined with the dependency type *dep* in TDT. Note that this is not intended to be a closed list, but rather a list of examples encountered while annotating TDT. The two-part expressions that fall into the categories of coordinating and subordinating conjunctions are omitted here, and instead listed in Sections 2.9 and 2.26, respectively. Due to the idiomatic nature of these two-part expressions, the translations may on occasion not be very natural.

2.18.1 Adverbs:

- aika lailla (quite some)
- aina vain (forever and ever)

- aivan kuin (just like)
- alun alkaen (from the beginning, originally)
- alun perin (originally)
- ennen aikojaan (prematurely)
- ennen kaikkea (first and foremost)
- ennen muuta (first and foremost)
- ennen pitkää (before long)
- entä jos (what if)
- heti perään (right after)
- hyvissä ajoin (on time, in good time)
- ihan vaan (only)
- ikään kuin (kind of)
- ilman muuta (of course)
- itse asiassa (as a matter of fact, in fact)
- ja niin edelleen (and so on)
- jonkin verran (some, to some extent)
- jossain määrin, siinä määrin, missä määrin (some, to some extent, to that extent)
- kaiken aikaa (all the time)
- kaiken kaikkiaan (all in all)
- kaikin puolin (in all ways)
- kerta kaikkiaan (completely, once and for all)
- loppujen lopuksi (in the end)
- muun muassa (among others)
- miten niin (how so)
- missä sattuu, mistä sattuu, minne sattuu (wherever)

- mitä jos (what if)
- niin ikään (also)
- niin kuin (like)
- niin sanotusti (so to say)
- noin vain (just like that)
- no kun (well)
- no niin (alright)
- näillä näkymin (with the current knowledge)
- näin ollen (this being so)
- pikku hiljaa (little by little)
- pilvin pimein (plenty of)
- piri pintaan (full)
- päällisin puolin (from the surface of it)
- saman tien (at once)
- saman verran (the same amount)
- sen koom(m)in (since then)
- sen suuremmin (any more than that)
- sen kun vaan (go ahead)
- sen verran (that amount)
- siellä täällä (here and there)
- siinä sivussa (on the side)
- silloin tällöin (every now and then)
- sillä aikaa (meanwhile)
- sitä mukaa (“accordingly”)
- sitä paitsi (besides)
- sivumennen sanoen (by the way)

- summa summarum (all in all)
- suuna päänä (headfirst)
- suurin piirtein (just about)
- ties vaikka (who knows)
- toisin sanoen (in other words)
- tuon tuosta (all the time)
- tuosta vain (just like that)
- tämän tästä (all the time)
- vähän kuin (a bit like)
- yhtä aikaa (at the same time)
- yhtä kaikki (all the same)
- yhtä paljon (the same amount, as much)
- yleisesti ottaen (generally speaking)

2.18.2 Adjectives:

- niin kutsuttu (so called)
- niin sanottu (so called)

2.18.3 Adpositions:

- lukuun ottamatta (disregarding)

2.18.4 Determiners:

- itse kukin (each)
- joka ainoa (each and every one)

2.18.5 Interjections:

- ai ai (oh oh, tut tut)
- ai niin (oh yeah)
- ei jumalauta (goddammit)
- ei vitsit (oh dear)
- hei hei (hey hey, bye bye)
- hip hip hurraa (hip hip hooray)
- hitto vie (dammit)
- jep jep (yep yep)
- kas kummaa (surprise surprise)
- mitä vittua (what the fuck)
- no joo (well yeah)
- piru vie (dammit)
- totta kai (of course)
- voi että (oh dear)
- voi po(i)jat (oh boy)

2.18.6 Nominals:

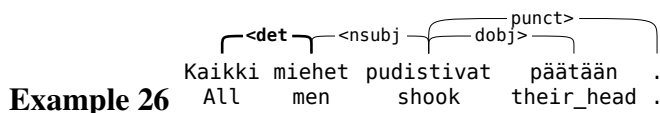
- missä ikinä (wherever)

2.18.7 Other: (the POS may vary)

- mikä tahansa (whichever, whatever)
- mikä vain (whichever, whatever)

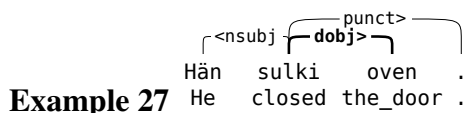
2.19 det (determiner)

There are no definite or indefinite articles in the Finnish language, but there are other determiners (see for instance [3, §1409]). In TDT, mostly pronouns are marked as *determiners* (*det*), because numerals, which can also be analyzed as determiner-like, are marked as *numeral modifiers* (*num*, see Section 2.34), and genitive modifiers, also determiner-like, are marked with *poss* (Section 2.38).

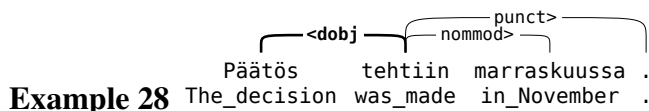


2.20 dobj (direct object)

The dependency type *dobj* is used for (nominal) direct objects of the verb.



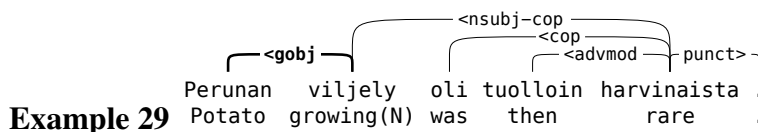
As in Finnish a passive clause does not have a subject, and what in English would be considered the passive subject, is in Finnish the direct object, the type *nsubjpass* is not used in TDT, but the type *dobj* is used instead.



Fine distinctions in special cases of subjects, objects and object-cased amount adverbials are discussed in Section 5.1.

2.21 gobj (genitive object)

Certain nouns, those which have been directly derived from a verb or otherwise have a verb counterpart, can take an object in Finnish. These objects closely resemble more general genitive modifiers (*poss*, see Section 2.38).



2.22 gsubj

Genitive subjects are subject-like arguments taken by a noun. This is in parallel to genitive objects (*gobj*, see Section 2.21). For further discussion on subjects and objects of nouns, see Section 5.11.

Example 30

	{<gsubj>	
maljakon		särkyminen
vase(gen.)		breaking

2.23 iccomp (infinite clausal complement)

The dependency type *icomp*, which stands for *infinite clausal complement*, is a subtype of *ccomp* (*clausal complement*). It is used for clausal complements where the complement clause has a different subject from that of the governing clause and is infinite, i.e. where the verb is an infinitive or a participle. The differences between types of verbal dependents, such as *icomp*, are thoroughly discussed in Section 5.4.

Example 31

	{<icomp>	{punct>	
Sain	hänet	itkemään	.
I_made	him	cry	.

2.24 infmod (infinitive modifier)

The dependency type *infmod* is used for infinitives that modify a nominal or a noun phrase.

Example 32

	{<nommod>	{<nsbj>	{<infmod>	{<advmod>	
Minulla	oli	lupa	mennä	ulos	.
I	had	permission	to_go	out	.

2.25 intj (interjection)

Interjections are typically exclamations or wordlike entities. They are attached to the main verb or predicative of the sentence with the *intj* dependency type.

Example 33

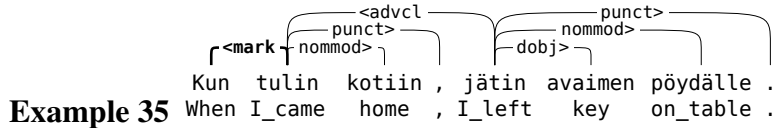
	{<intj>	{<punct>	{<xcomp>	{<punct>
Hei	,	tule	auttamaan	!
Hey	,	come	to_help	!

Example 34

	{<intj>	{<doobj>	{<nommod>	{<punct>	
Hmm	...	Mitähän	tuohon	sanoisi	?
Umm	...	What	to_that	to_say	?

2.26 mark (marker)

A *marker* (*mark*) is the subordinating conjunction in a non-complement subordinate clause.



The main subordinating conjunctions in TDT are:

- että (that)
- jotta (so that)
- koska (because)
- kun (when)
- jos (if)
- vaikka (even though)
- kunnes (until)
- kuin (as, than)

Note that the conjunction *että* (usually) starts a complement clause, in which case it is marked as a complementizer (*complm*, see Section 2.13). On a similar note, the conjunction *kuin* also has several uses. In addition to a subordinating conjunction, it can also serve as an adverb modifier (see Section 2.3) as well as a comparative conjunction (Section 2.12).

In addition to the basic subordinating conjunctions, the following words or word combinations can be considered subordinating conjunctions in TDT. Not all of these expressions have a direct counterpart in English, and thus the translations are approximate.

- ennenkuin (before)
- jahka (as soon as)
- jos kohta (even if)
- kun taas (whereas)
- kuten (like, as)
- mikäli (if)

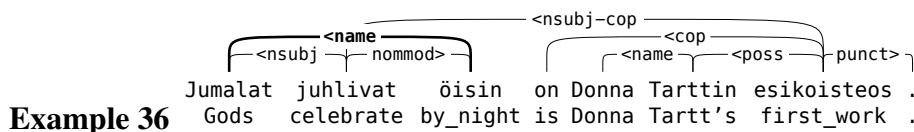
- mitä nyt (only)
- muuten (otherwise)
- niin (so)
- niinkuin/niinku (like, standard and colloquial version)
- paitsi (except)
- paitsi että (except that)
- paitsi jos (except if)
- sikäli kuin (if)
- sillä (because)
- sitten kun (then when)
- vähän kuin (a bit like)

2.27 name (multi-word named entity)

Multi-word named entities are marked using the dependency type *name*. The rightmost word of the named entity is considered the head, and the leftmost word is the dependent. If there are more than two words, these are not marked in any way, as the *name* dependency can be expanded automatically if needed.

There are two different cases in which the dependency type appears. If the multi-word named entity does not have an obvious internal syntactic structure, as is the case with for instance names of people (*Matti Virtanen*) or cities (*New York City*), only the *name* dependency is used.

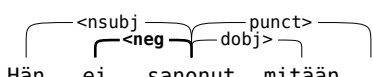
If the named entity has an obvious internal structure, as is often the case in names of books and movies for instance, this structure is marked as well, and the *name* dependency is placed on top of this structure, despite it being extraneous in the tree. In these cases, the head of the internal structure, not the rightmost word of the named entity, is considered to be the true syntactic head. It is possible for the user of the treebank to choose their preferred analysis for these cases according to need, and automatically discard the alternative analysis.



2.28 neg (negation marker)


In Finnish, negation is marked using the verb *ei*, which is used as an auxiliary. This means that the *negation marker (neg)* is a subtype of *aux* (see Section 2.7). The most commonly negated elements are verbs and verb phrases, but occasional exceptions in verbless constructions are allowed.

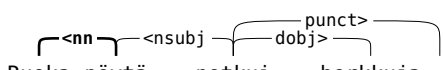
If a conjunction or adverb has been merged together with *ei*, as in for instance *ettei (että+ei, that+not)* or *miksei (miksi+ei, why+not)*, then the word is marked as a conjunction or an adverb rather than a negation verb. However, *eikä (and+not)*, when it appears alone and not coordinating another clause or phrase, is still marked as *neg*.


Example 37 Hän ei sanonut mitään .
He didn't say anything .

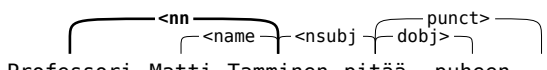
2.29 nn (noun compound modifier)

The dependency type *nn*, which stands for *noun compound modifier* has two basic uses in the Finnish SD scheme. The first use involves actual compounds. In Finnish, compounds are generally written as a single word, but for instance some compounds involving foreign words or proper names are written separately using a dash, and in written Finnish, erroneously writing compounds as two words is a common mistake. Both of these cases are marked using *nn*.


Example 38 Da Vinci -merkkinen luomiväri
Da Vinci -make eyeshadow


Example 39 Ruoka pöytä notkui herkkuja .
Food table was_full_of goodies .

The second use of the type *nn* is to mark *appellation modifiers*, which are modifying, non-inflecting noun phrases that generally express profession, rank, position, assignment or other such classifiable property [3, §1062]. The phenomenon is closely related to that of *apposition*, and the distinction between the two is described in Section 5.3.


Example 40 Professori Matti Tamminen pitää puheen .
Professor Matti Tamminen gives a_speech .

2.30 nommod (nominal modifier)

Nominal modifiers are inflected nominals which modify most commonly a verb or a noun phrase. They can occur alone or together with an adposition in an adpositional phrase. Both cases are analyzed similarly, as semantically nominal modifiers and adpositional phrases are similar.

Example 41

```
Maljakko oli pöydällä .
The_vase was on_the_table .
```

Example 42

```
Maljakko oli pöydän päällä .
The_vase was table on_top_of .
```

2.31 nommod-own

In Finnish, there is no equivalent for the verb *have*. Rather, *having* is expressed using the verb *olla*, *to be*. For instance, the meaning of the sentence *I have a pen* would be expressed in Finnish by *Minulla on kynä*, literally “*At me is a pen*”. In TDT, these so called *possessive clauses*⁶ are analyzed as a subtype to *existential clauses*,⁷ making the thing had (*kynä* in the previous example) the subject. For more information on special cases of subjects, see Section 5.1.

This kind of an analysis would naturally result in the *haver* being marked as a nominal modifier, *nommod*. However, as *nommod* is a very frequent dependency type that encodes many different meanings, the information that the clause is about having or owning would be lost. Therefore, the Finnish-specific SD scheme introduces a separate dependency type for nominal modifiers that encode owning, *nommod-own*. The governor of the dependency is the verb *olla*, and the dependent is the haver or owner, which is required to be in the *adessive* case. The haver must also be an animate being or a group of animate beings.

Example 43

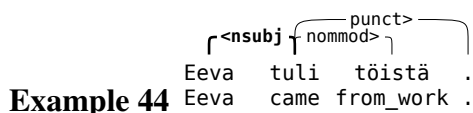
```
Matilla on uusi auto .
At_Matti is new car .
```

2.32 nsubj (nominal subject)

The dependency type *nsubj* marks nominal subjects of the non-copular clause. For thorough discussion of different types of subjects in Finnish, see Section 5.1.

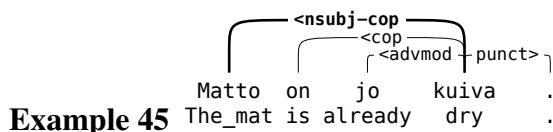
⁶omistuslause

⁷eksistentiaalilause



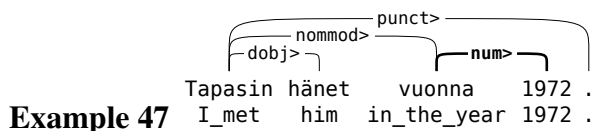
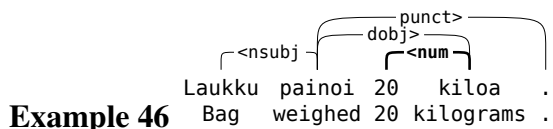
2.33 nsubj-cop (nominal copular subject)

The dependency type *nsubj-cop* is used for the nominal subject of a copular clause. The predicative is the head of the copular clause, and also the governor of the *nsubj-cop* dependency. Annotating copular clauses is discussed in Section 5.2.



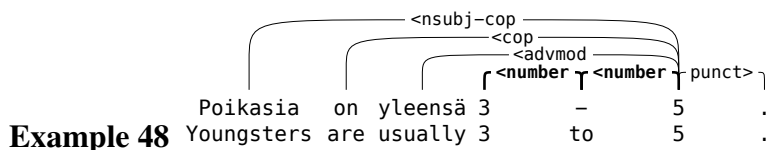
2.34 num (numeral modifier)

Numeral modifiers of a noun or NP, including both cardinal and ordinal numbers, are marked with the *num* dependency type. This dependency type is used also with for instance years and program versions.



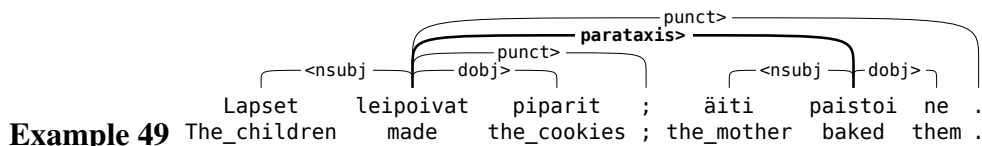
2.35 number (numerical expression)

Numerical expressions consisting of multiple tokens are annotated using the *number* dependency type. The last word of the numerical expression is the governor, and the number dependencies are chained. Special cases of numerical expressions are discussed in Section 5.12.

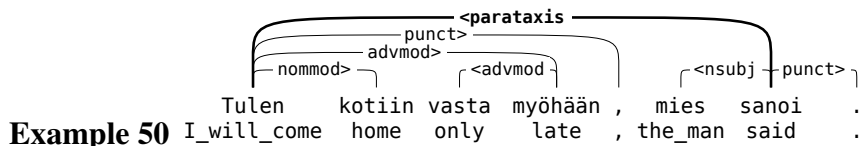


2.36 parataxis (parataxis)

Parataxis dependencies mark two different phenomena. Firstly, they are used with certain implicit coordinations. These coordinations are recognized by two factors: there is no coordinating conjunction, and the independent clauses are separated by a colon, semicolon or a dash. As with explicit coordinations, the first element is the governor. Also parenthetical clauses can receive the *parataxis* dependency. If there is a coordinating conjunction present (regardless of punctuation) or if the clauses are separated by merely a comma, the coordination type *conj* is used.

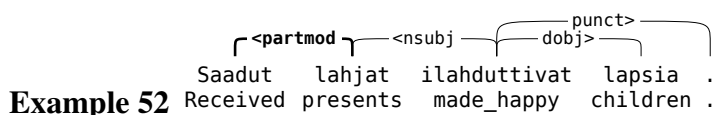
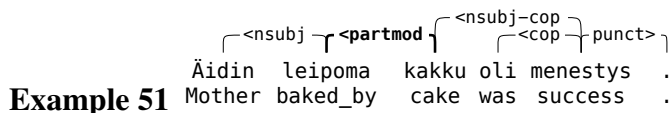


Parataxis is also used for direct speech. The verb of saying⁸ acts as the governor, and the main verb or predicative of the utterance is the dependent.

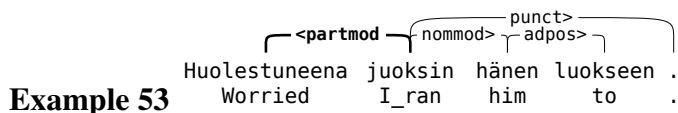


2.37 partmod (participial modifier)

The *participial modifier* (*partmod*) most commonly modifies a noun phrase. Note that the participle⁹ can take arguments, for instance a subject, just as any verb.

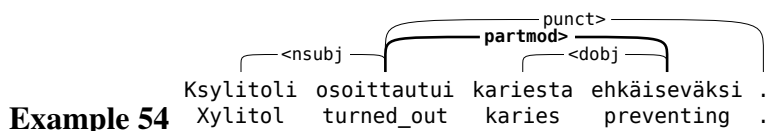


Occasionally, participles can modify a verb as well. These uses include cases that are clearly modifiers, as well as some more complement-like situations. Note that in the complement-like cases of *partmod*, the complement is not a clause; otherwise it would be marked as an *infinite clausal complement* (see Section 2.23).



⁸or thinking, etc.

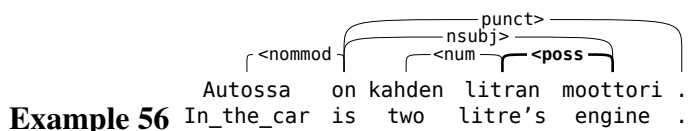
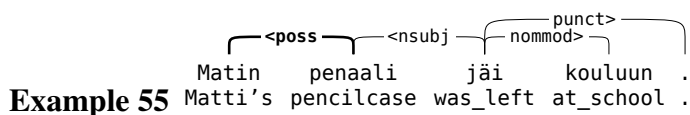
⁹Also the MA-derivation is treated as a participle in TDT.



For more information on different verb-headed constructions as dependents, see Section 5.4.

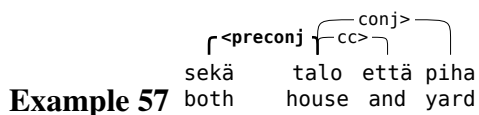
2.38 poss (genitive modifier)

The dependency type *poss* stands for *possessive* in the original SD scheme, but in TDT, it is used for genitive modifiers in general, which in Finnish often but not nearly always imply possession. There are two kinds of genitive modifiers that are not annotated using the general genitive modifier type: *the genitive object*, *gobj* (see Section 2.21) and *the genitive subject*, *gsubj* (Section 2.22).



2.39 preconj (preconjunction)

Preconjunction (*preconj*) marks the first part of those two-part coordinating conjunctions where the two parts are separated by coordinated elements.



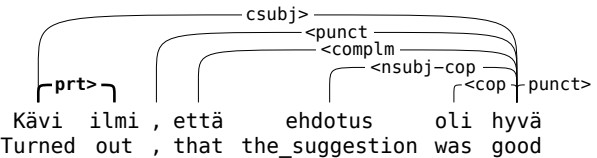
Such two-part coordinating conjunctions in TDT are:

- joko... tai (either... or)
- milloin... milloin (when... when)
- mitä... sitä (the... the)
- niin... kuin (as well as)
- paitsi... myös (not only... but also)
- sekä... että (both... and)

- sitä... mitä (the... the)
- sitä mukaa... mitä (a version of *the... the*)
- toisaalta... ja toisaalta (on the one hand... and on the other hand)
- toisaalta... mutta toisaalta (on the one hand... but on the other hand)
- toisaalta... toisaalta (on the one hand... on the other hand)
- vuoroin... vuoroin (in turn... in turn)
- yhtä lailla... kuin (+kin) (as much... as also)

2.40 prt (phrasal particle)

Phrasal particles (prt) are used in connection with *phrasal verbs*,¹⁰ where the particle is considered an integral part of the verb expression. The governor of the dependency is the verb, and the dependent is the phrasal particle.



Example 58

Turned out, that the_suggestion was good .

Verb particles (see [3, §455, §680]) are the only case where particles are distinguished from adverbs in TDT. This distinction can be made by the following rough rules. A word is a verbal particle if it, together with the verb, forms an expression that has a meaning that differs from the meaning of the verb alone, and if the word cannot be modified by an adverb.

For instance, *laittaa kiinni* (*make closed, close*) is not a phrasal verb, as *kiinni* can be modified.

Example 59 *Laitoin oven kokonaan kiinni.* (*I closed the door entirely.*)

In contrast, *ottaa kiinni* (*catch*) is a phrasal verb, as it has a meaning distinct from the verb *ottaa* (*take*), and *kiinni* cannot be modified.

Example 60 **Poliisi otti rosvon kokonaan kiinni.* (**The police caught the robber entirely.*)

The following verb expressions are considered phrasal verbs in TDT:¹¹

¹⁰partikkeliverbi, “particle verb” in Finnish grammar

¹¹The list is not closed, but includes the phrasal verbs encountered in the corpus text. Also, due to the figurative meanings of many of these expressions, the English translations are approximate.

- ajaa takaa (chase)
- antaa periksi (give up)
- astua voimaan (become valid)
- jäädä jälkeen (be left behind)
- jäädä kiinni¹² (be caught)
- jäädä käteen¹³ (“be left in one’s hand”, one is left with something)
- jäädä väliin (be passed¹⁴)
- kiriä kiinni (close the distance)
- kuroa kiinni (close the distance)
- kutsua kokoon (summon)
- kutsua koolle (summon)
- käydä ilmi (come up)
- käydä kateeksi (make jealous)
- käydä läpi (go through)
- käydä sääliksi (be pitied)
- laskea alleen (wet one’s pants)
- lyödä laimin (neglect)
- lyödä läpi (strike through)
- nukkua pommiin (oversleep)
- olla kaupan (be for sale)
- olla meneillään (be happening)
- olla tarpeen (be necessary)
- olla tarvis (be necessary)
- olla voimassa (be valid)

¹²only in the sense “be caught”, not in the sense “be stuck into something”

¹³The figurative reading only.

¹⁴In the sense “I’ll pass.”

- ottaa irti¹⁵ (“take sth out”, make the most of)
- ottaa kiinni (catch)
- ottaa lukuun (take into account)
- ottaa mukaan (take along)
- ottaa selvää (find out)
- ottaa vaari(n) (take advice)
- ottaa vastaan (receive)
- painaa päälle (push, stress on)
- panna merkille (take note)
- panna täytäntöön (put into action)
- pidellä kiinni (hold on)
- pitää kaupan (keep for sale)
- pitää kiinni (hold on)
- pitää voimassa (keep valid)
- pitää yllä (maintain)
- päästä käsiksi (get one’s hands on)
- päästä läpi (get through)
- päästää irti (let go)
- saada aikaan (get sth done)
- saada aikaiseksi (get sth done)
- saada kiinni (catch)
- saada läpi (get sth through)
- saada vireille (get sth started)
- tulla mukaan (come along¹⁶)

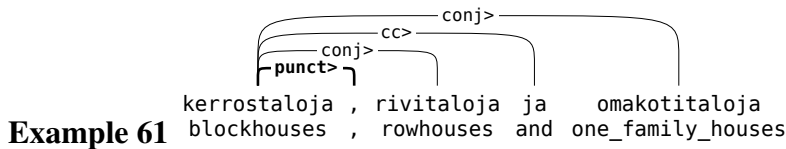
¹⁵in expressions such as “ottaa ilo irti”

¹⁶In the sense of “follow”, not the social sense.

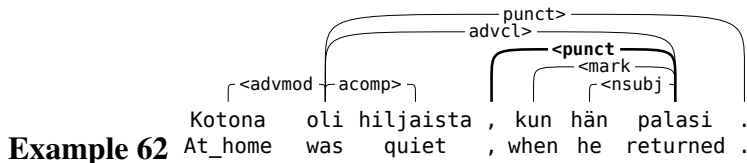
- tulla tarpeen (become necessary)
- tulla vastaan (come across)
- tulla voimaan (become valid)

2.41 punct (punctuation)

The dependency type *punct* is used to mark punctuation. The dependent is the punctuation symbol, and the governor is the element which the punctuation symbol delimits. For instance, with coordination, the first coordinated element is the head of all *punct* dependencies in the coordination, and with subordinate clauses, the head of the subordinate clause is the governor of the *punct*.



Example 61

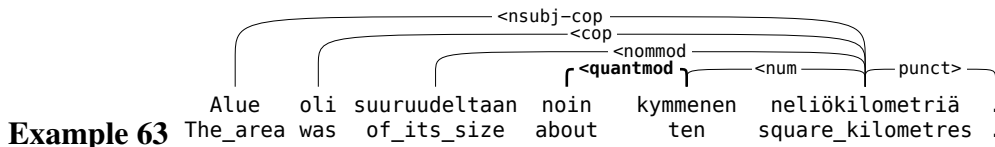


Example 62

Attaching punctuation correctly is described more closely in Section 5.17.

2.42 quantmod (quantification modifier)

Quantification modifiers (*quantmod*) are quantifiers that modify a numerical expression. Certain adverbs¹⁷ and few adjectives are allowed as quantifiers. Note that adverbs that describe the writer's attitude towards the quantity, such as *vain* (*only*), are not considered quantification modifiers, but regular adverb modifiers, although they modify the number. Some examples of words that can act as quantification modifiers include *noin* (*about*), *vähintään* (*at least*), *lähes* (*almost*) and *yli* (*over*).

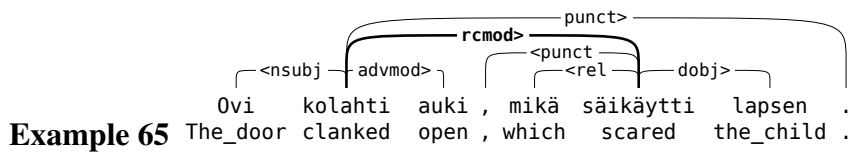
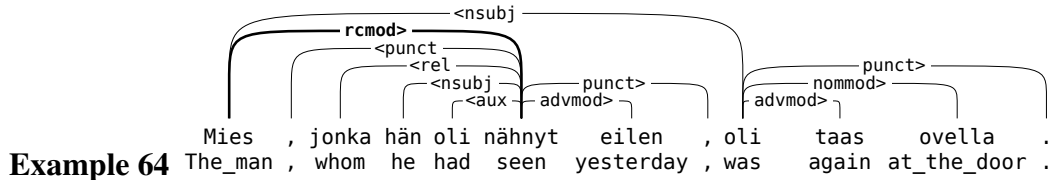


Example 63

¹⁷and ad-adjectives, which are sometimes regarded a separate category from adverbs but treated identically in TDT

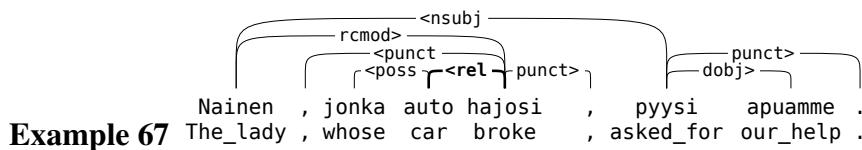
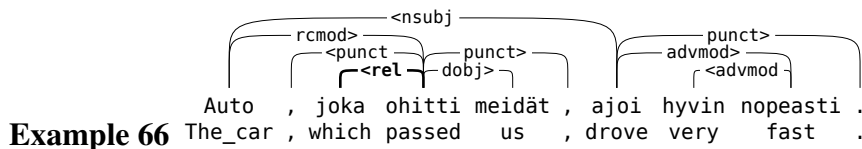
2.43 rcmod (relative clause modifier)

A *relative clause modifier* (*rcmod*) marks relative clauses. The governor is the phrase or clause modified, most often a noun phrase but occasionally a full clause as well. The dependent is the main predicate of the relative clause.



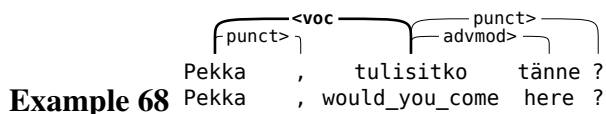
2.44 rel (relativizer)

The *relativizer* (*rel*) is the head of the phrase containing the relative pronoun (or other relative word). Most often, but not always, this is the relative word itself. The governor of the dependency is the main predicate of the relative clause. Annotation of relative clauses is more closely examined in Section 5.6.



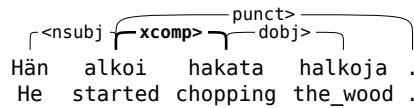
2.45 voc (vocative)

The dependency type *voc* is used for *vocatives*, that is, expressions where someone is being addressed. The governor of the dependency is the main predicate of the clause where the addressing occurs.



2.46 xcomp (open clausal complement)

The dependency type *xcomp* is reserved for clausal complements which have an external subject, that is, whose subject is shared with the complemented verb (phenomenon also known as subject control). Note that the subject of the complementing clause must be the *subject* of the complemented verb, not any other sentence element.



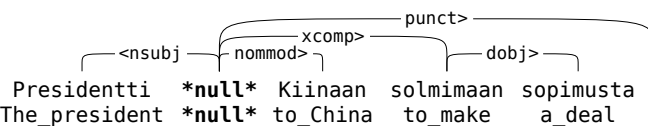
Example 69 Hän alkoi hakata halkoja .
He started chopping the_wood .

Many of the complements with an external subject resemble cases that are analyzed as main verbs with auxiliaries. Both auxiliaries and *xcomp* complements share their subject with another verb, but only a closed list of verbs are analyzed as auxiliaries in TDT (see Section 2.7). Note also that in auxiliary cases the second verb is the governor, whereas with *xcomp* the first verb becomes governor (unless the word order is inverse).

2.47 The null token

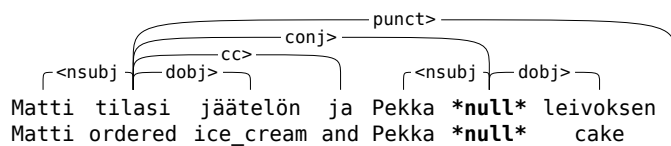
The *null token* is not a dependency type, but an extra token that is added into the sentence to represent a missing token. A null token is only added when the missing token is required in order to construct an analysis, that is, when it governs another token that is present in the sentence. Thus, for instance copulas and auxiliaries are not represented by null tokens when absent, because if they are absent, their dependents are as well. The null token is most commonly, but not always, a verb.

There are two basic uses for the null token in TDT. First, it is used in *fragments*: sentences or clauses with an omitted main predicate.



Example 70 Presidentti *null* Kiinaan solmimaan sopimusta .
The_president *null* to_China to_make a_deal .

Second, the null token is used in *gapping*, a type of *ellipsis* where a head word has been omitted to avoid repetition. Gapping is the only type of ellipsis marked with null tokens, as according to the definition of a null token, only words required for constructing an analysis should be represented by one.



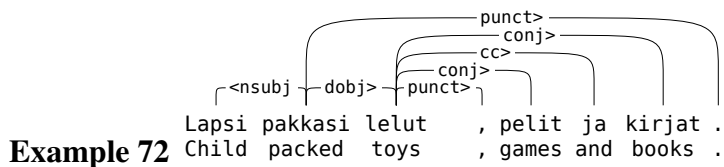
Example 71 Matti tilasi jäätelön ja Pekka *null* leivoksen .
Matti ordered ice_cream and Pekka *null* cake .

3 Conjunct propagation and additional dependencies

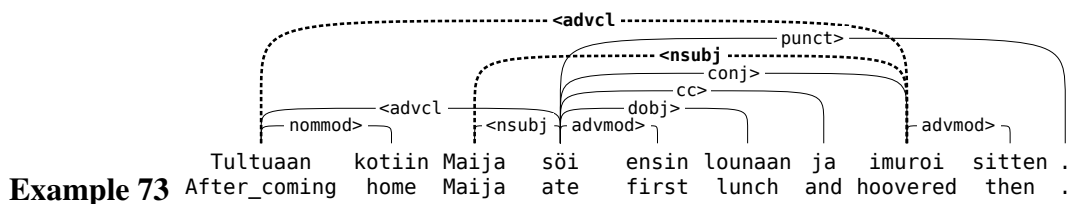
This section describes four phenomena that are annotated in the second annotation layer of TDT, termed the *conjunct propagation and additional dependencies* layer. These phenomena are *the propagation of conjunct dependencies*, *external subjects*, *syntactic functions of relativizers* and *gapping*. The annotation of this layer is added on top of the first layer, meaning that the analyses are no longer trees.

3.1 Conjunct propagation

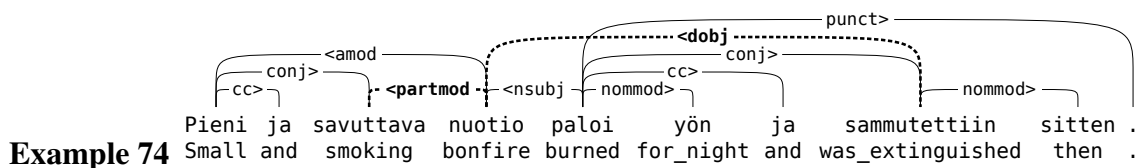
The SD scheme analyzes coordinations so that the first coordinated element is the head of the whole coordination.



In this analysis alone, it is not possible to distinguish dependents and governors of the first coordinated element from those of the whole coordination, nor from elements that depend on or govern some but not all conjuncts. Therefore in the extended variants of the SD scheme, this distinction is made explicit with additional dependencies on top of the tree structure. That is, if an element modifies or governs multiple coordinated elements, it should be *propagated* to them. In principle, any dependency type introduced in Sections 2 and 3, with the exceptions of *punct*, *conj*, *cc* and *ellipsis*, can propagate in the second layer of annotation. Note especially that the dependencies introduced in the second layer, that is, external subjects and syntactic functions of relativizers, are also allowed to propagate.

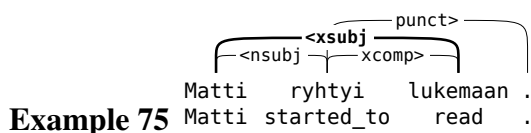


In addition to simply propagating to other coordinated elements, it is possible for a dependency to change its type while doing so. This may happen in coordinations of elements with differing parts-of-speech, and cases where a sentence element acts in one syntactic role for the first conjunct and in an another role for some other conjunct.

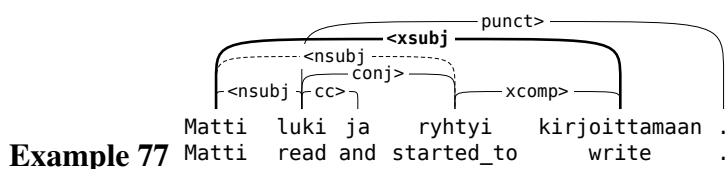
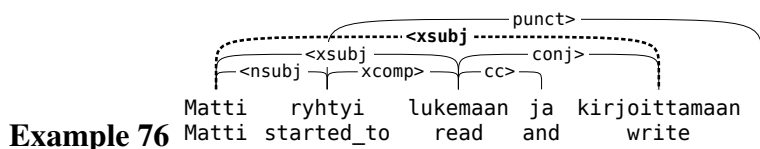


3.2 External subjects (xsubj)

Open clausal complements, as discussed in Section 2.46, share their subject with another verb. The fact that the subject of the main verb is also the subject of the complement cannot be annotated on the first layer of TDT, as this would violate the treeness restriction. Therefore, these subjects are marked on the second layer of annotation using the dependency types *xsubj* (*external subject*) and *xsubj-cop* (*external copular subject*). Note also that an open clausal complement may not always have a subject, in for instance passive constructions.

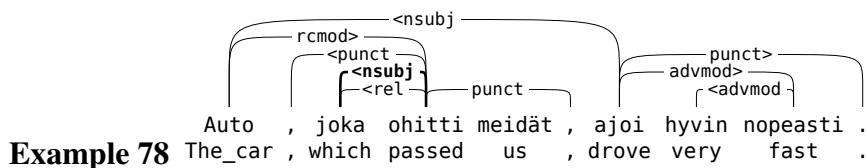


External subjects interact with conjunct propagation in two ways: an external subject may propagate, and also a propagated *nsubj* dependency may be the source of a new *xsubj* dependency.

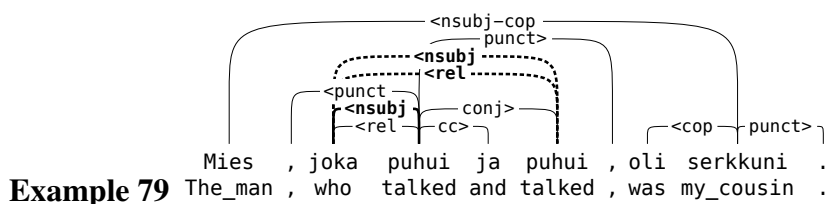


3.3 Syntactic functions of relativizers

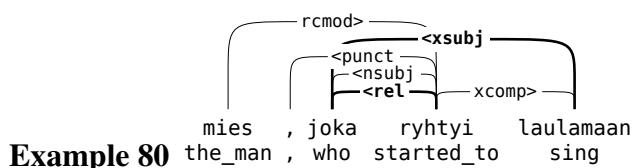
Relativizers, that is, the phrases containing the relative word are marked as such in the base layer of annotation, using the dependency type *rel* (see Section 2.44). However, the relativizers also always have a secondary syntactic function, such as a subject, which cannot be marked on the base layer of annotation due to the treeness restriction. Therefore these functions are marked on the *conjunct propagation and additional dependencies* layer on top of the tree structure. In principle any dependency type from Section 2 may represent the syntactic function of a relativizer, although in practice certain types (such as *punct*) will not do so.



Relativizers and their secondary functions may propagate in coordinations, and if the dependencies are between the same tokens (see Section 5.6 for discussion of cases where they are not), they will propagate together.



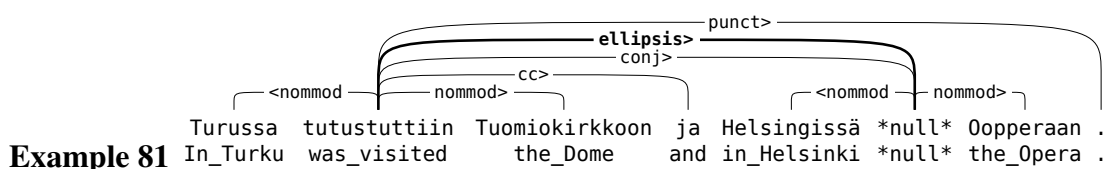
If the syntactic function of a relativizer is a subject, the relativizer may also act as an external subject to another verb.



3.4 Gapping (ellipsis of a head word)

As described in Section 2.47, gapping in TDT is marked by inserting a so called *null token* to represent the omitted token. In addition, the second layer annotation uses the dependency type *ellipsis* to mark the elided word so that the null token acts as the dependent, and the governor is the non-elided occurrence of the word.¹⁸

Note that according to the policy of only inserting a null token where necessary for constructing an analysis (see Section 2.47), gapping is the only form of ellipsis marked using null tokens and *ellipsis* dependencies in TDT. Some other elliptical structures may be less explicitly marked as *conjunct propagation* (see Section 3.1).



¹⁸Note that the elided word can also precede the non-elided occurrence.

4 Differences between the Finnish and English versions of the SD scheme

The original SD scheme by de Marneffe and Manning [1, 7] includes in total 55 dependency types arranged in a hierarchy; excluding six intermediate types that are not meant to be used if a more specific type can be selected, the total number of dependency types is 49. The Finnish-specific scheme version used in this work includes 46 dependency types in the *base layer* and 3 additional types in the *conjunct propagation and additional dependencies* layer. This section discusses the differences between the two scheme versions. Figure 1 shows the original SD type hierarchy as described in the SD scheme manual [1], and Figure 2 the hierarchy of the Finnish-specific version.

To maintain a hierarchy similar to the original one, Figure 2 includes four intermediate types which have not been introduced above and are not used in TDT: *arg* (argument), *comp* (complement), *subj* (subject) and *mod* (modifier). This makes the overall number of types in the Finnish SD scheme 53.

4.1 Dependency types not used in Finnish-specific SD

There are several reasons why the Finnish-specific SD scheme differs from the original scheme. First, some dependency types from the original scheme have been removed, as the corresponding phenomenon does not occur in Finnish. Types omitted for this reason include *expl* (expletive *there*), *csubjpass* (clausal passive subject), *nsubjpass* (nominal passive subject), *agent*, *possessive* (the possessive *'s*) and *iobj* (indirect object). Finnish existential clauses do not contain an expletive *there*, nor do passive clauses have a subject. What in English is considered the passive subject is the direct object in Finnish, and thus the corresponding type, *dobj* is used instead, or in the case of a clause acting as the direct object, it is marked as a clausal complement (*ccomp*). Similarly, there are no agents in Finnish passive clauses, and constructions resembling the English agent can be analyzed according to their syntactic structure rather than semantically as agents. This makes the type *agent* unnecessary. Also the possessive *'s* does not occur in Finnish, and thus the dependency type *possessive* is not needed. Finally, indirect objects do not occur in Finnish, as regardless of word order, the corresponding argument is expressed by a nominal modifier.

Second, adpositional phrases are handled differently from the original SD scheme, so as to analyze them similarly to nominal modifiers without a pre- or postposition present. Thus, the original SD types *prep* and *pobj* are not used. Third, the type *ref* (*referent*) is not included in the current TDT annotation. When used, it causes the structures to not be trees, meaning that it would be part of an additional layer of annotation.

Fourth, three dependency types are considered semantic in nature, and thus not included in the first layer of annotation in TDT. These types include *purpcl*

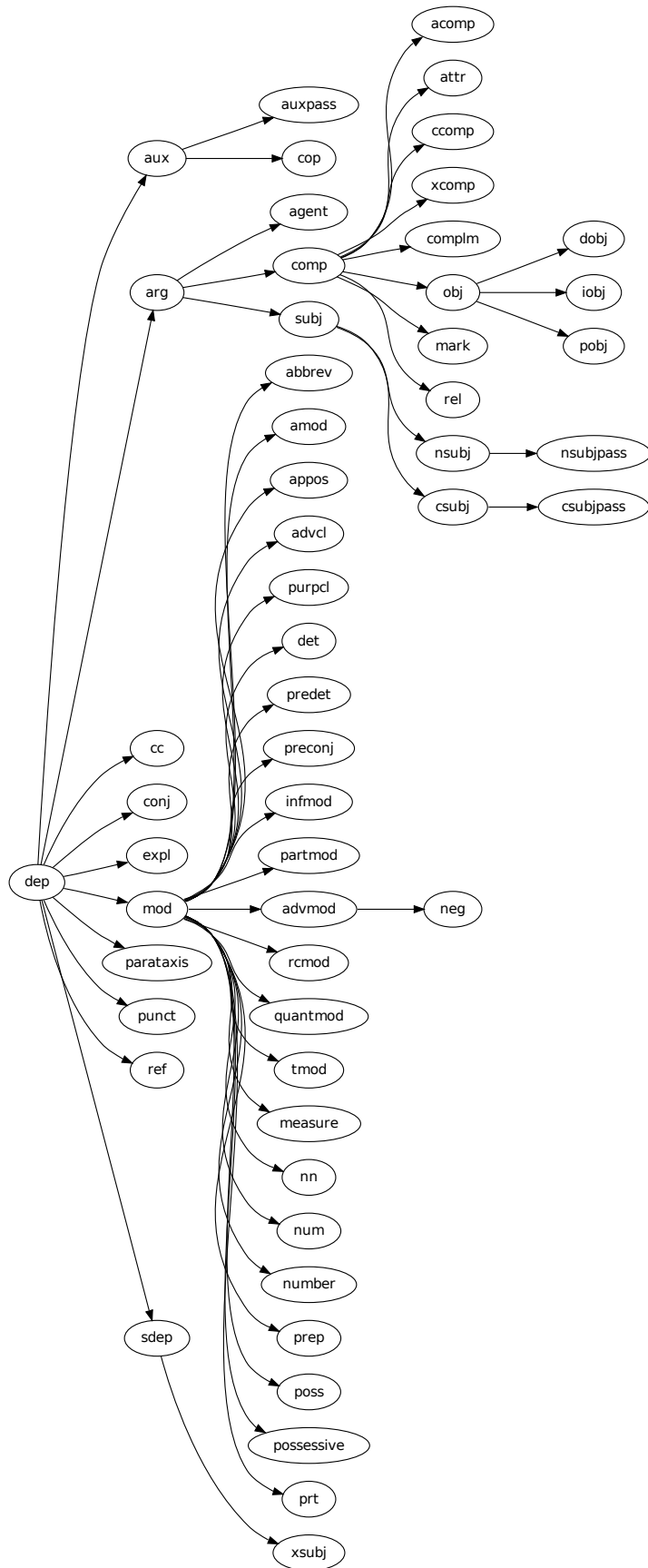


Figure 1: The original SD scheme for English. Figure adapted from de Marneffe and Manning [1].

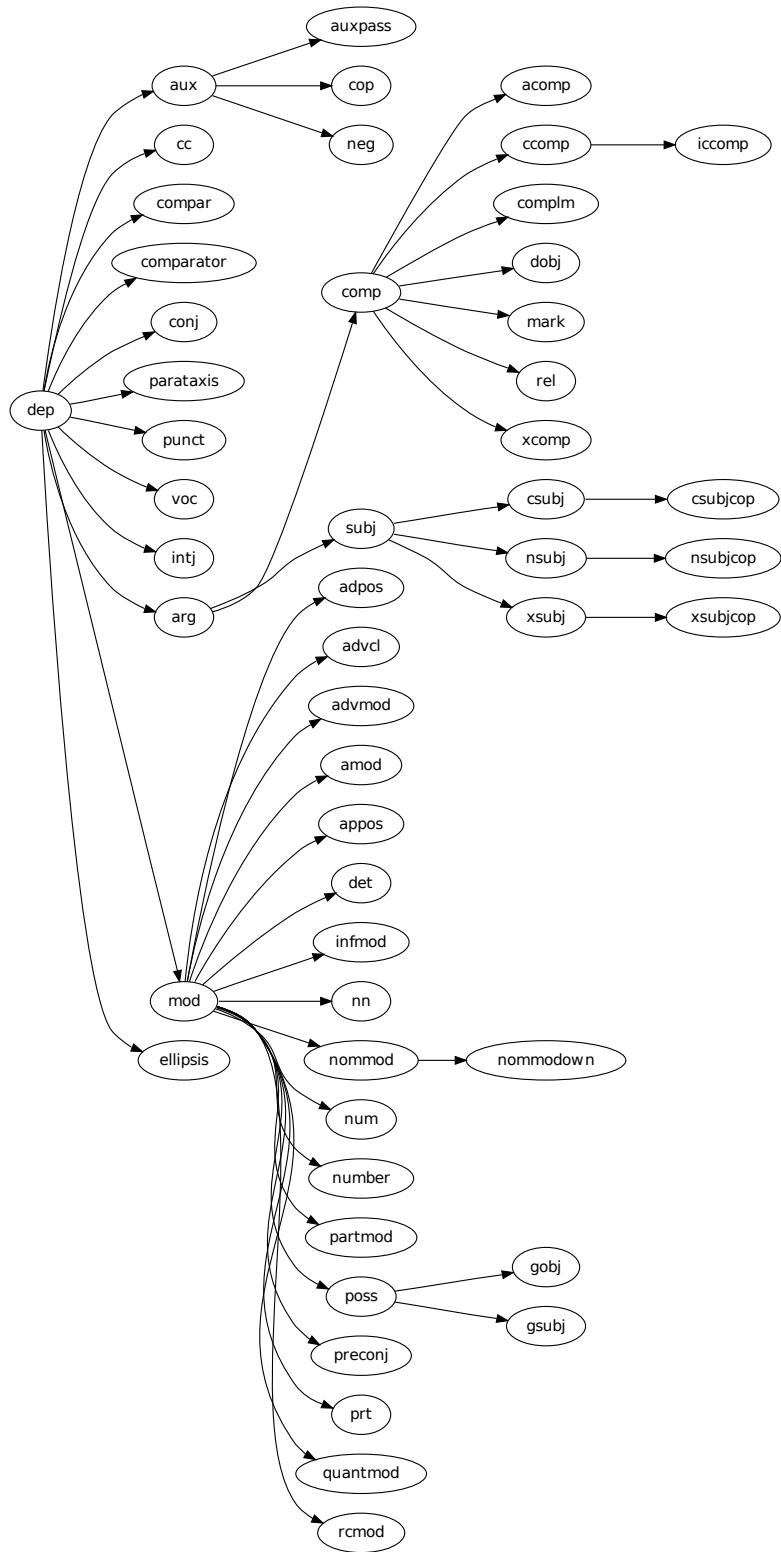


Figure 2: The Finnish SD scheme as used in TDT.

(purpose clause), *tmod* (temporal modifier) and *measure*. Instead of using these semantic types, the dependency types appropriate for the syntactic structure of each phenomenon are used in TDT. Finally, no distinction between apposition-like abbreviations (*abbrev*)¹⁹ and appositions (*appos*) is made, meaning both dependency types have been merged under the *appos* type, and instead of marking predicatives in some situations as attributives (*attr*), they are always analyzed as predicatives.

4.2 New dependency types in Finnish-specific SD

In addition to types not used in the Finnish-specific SD scheme, there are also dependency types that are new to this scheme version. First, two new dependency types were needed to accommodate the similar handling of nominal modifiers and adpositional phrases: *nommod* and *adpos*, both under the *modifier* subcategory. It should be noted that although *nommod* is considered a modifier type, many nominal modifiers in the treebank are at least borderline complements, due to the fact that many Finnish verbs take inflected nominals as their arguments.

Second, the dependency types *gsubj* and *gobj* were also added under the *modifier* subcategory, to accommodate the frequent constructions of nouns that take a subject- or object-like argument. The genitive subject and object take the form of a genitive modifier, and thus they are direct subtypes of *poss*.

Third, under the *subject* category, both the nominal and clausal subject types have received a new subtype, *nsubj-cop* and *csubj-cop*, respectively, to be used for subjects of copular clauses, which have their own special treatment in the SD scheme. These two new types come in place of the passive subject types that were, as explained above, removed as unnecessary. Also we have moved the existing *xsubj* type from under the *sdep* category to under the *subject* category, and added a new subtype for *xsubj*, *xsubj-cop*. The external subject types are part of the *conjunct propagation and additional dependencies* layer of the treebank.

Fourth, in the *complement* category, we have introduced one new subtype for clausal complements (*ccomp*): that of *infinite clausal complement*, *icomp*. This is due to the fact that clausal complements in Finnish often involve an infinite main verb.

Fifth, we have added five other new dependency types. The types *compar* and *comparator* are to be used in structures involving comparisons of adjectives.²⁰ The type *voc* is introduced to be able to analyze *vocatives*, and the type *intj* is for *interjections*. The treebank contains only written Finnish, but both vocatives and interjections are fairly common in more informal genres, such as blog text. The type *ellipsis* is part of the *conjunct propagation and additional dependencies* layer and used to mark the elided word in gapping.

¹⁹such as *United States of America (USA)*

²⁰and occasionally other parts-of-speech

4.3 Hierarchy changes

Finally, there are two minor changes made in the SD hierarchy. First, as prepositional objects are no longer needed in the Finnish-specific scheme due to the changes made to handling adpositional phrases, and as indirect objects do not occur in Finnish, the type *doobj* was the sole subtype of the intermediate, unused type *obj*, we have removed this intermediate type, and made *doobj* a direct subtype of *complement*.

Second, the *neg* dependency type, for marking negations, has been moved from under adverbial modifiers to under auxiliaries in the hierarchy. This is because in Finnish, the negation word *ei* is in most contexts a verb and acts in an auxiliary-like manner. It should be noted, however, that in TDT there are few cases where it is considered that for instance a noun phrase has been negated or where *ei* functions as the counterpart of *kyllä* (*yes*), and is thus an adverb.

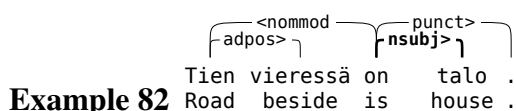
5 Annotating phenomena of Finnish

This section gives detailed instructions on annotating certain common phenomena that require detailed decision rules.

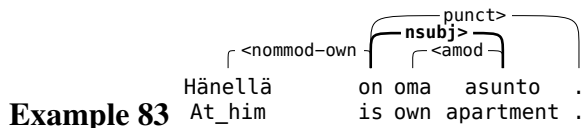
5.1 Subjects and objects

Both subjects and objects are straightforward to recognize in their prototypical cases, but both phenomena also have some difficult cases, which are discussed here.

The subject is the primary complement of the verb, usually denoting the entity doing something. In addition to the *basic subject* (see [3, §910]), also *existential subjects*²¹ are considered subjects in TDT.



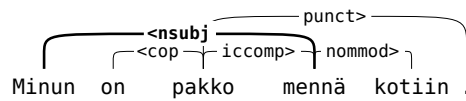
*Possessive clauses*²² are considered a subtype of existential clauses, and analyzed similarly. As explained in Section 2.31, the owner in possessive clauses is marked using the type *nommod-own*.



²¹eksistentialisubjektiksi, e-subjektiksi

²²omistuslause

Also the *genitive subject*²³ in for instance *necessive* structures (see Section 5.14) is annotated as an *nsubj*.



Example 84 I(gen.) is obligation go home .

In TDT, subjects are allowed to be in the *nominative*, *genitive* and *partitive* cases, and in addition, also an *accusative*²⁴ subject is possible. Two notable situations where a complement in the accusative form is analyzed as the subject are:

1. Infinite clausal complements (Sain *hänet* itkemään. I made *him* cry.)
2. Possessive clauses (Minulla on *sinut*. I have *you*.)

The same cases are allowed for objects as for subjects: the nominative, the partitive, the genitive and the accusative. Complements in other cases are analyzed as *nominal modifiers (nommod)*, despite their complement status.

Object cased amount adverbials,²⁵ which, as the name implies, use the same cases as objects, are analyzed as nominal modifiers. However, certain verbs are considered such that they can take as their object an expression that would otherwise be considered an amount adverbial. Examples where an amount is considered the object are for instance:

Example 85 *Juoksin kilometrin.* (I ran a kilometer.)

Example 86 *Moottori pyöri kymmenen kierrosta.* (The motor ran ten rounds.)

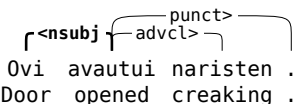
Example 87 *Maitotölkki painaa kilon.* (A milk jar weighs a kilogram.)

Passive verbforms take a direct object and not a passive subject, like in for instance English.



Example 88 Lesson was_prepared carefully .

However, there are certain verbs, so called *derived passives* [3, §336], which may resemble passive verbforms in meaning, but which in fact take a subject, not an object.²⁶



Example 89 Door opened creaking .

²³not to be confused with the genitive subject of a noun, discussed in Section 2.22

²⁴The accusative case only exists for certain pronouns.

²⁵objektin sijainen määrän adverbiaali, OSMA [3, §972]

²⁶In English, the Finnish derived passives generally correspond to intransitive uses of a verb, such as *the door opens*, sometimes termed *inchoative*.

5.2 Copulas

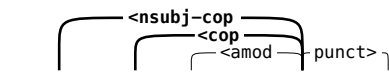
The copular clause deserves some attention, in part due to the special treatment given to it in the SD scheme. This section discusses first defining copular verbs and predicatives, then copulas in combination with auxiliaries, and finally the distinction between the subject and the predicative in copular clauses.

5.2.1 What can be a predicative?

In the SD scheme, the head of a copular clause is the predicative, not the verb (copula), unlike in other clauses. The Finnish language only has one copular verb, *olla* [3, §891], and in order to avoid marking other verbs as copular and to prevent copular clauses from having multiple head words, strict rules are needed to define what is accepted as a predicative.

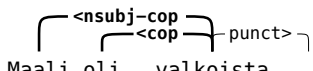
The basic alternatives for predicatives are nominals (nouns, adjectives, pronouns and numerals). Words of these parts-of-speech are required to be in *nominative*, *partitive* or *genitive* to be accepted as predicatives.

Example 90



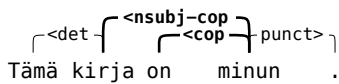
Varpunen on pieni lintu .
Sparrow is small bird(nom.) .

Example 91



Maali oli valkoista .
Paint was white(part.) .

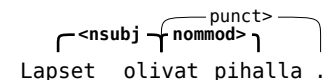
Example 92



Tämä kirja on minun .
This book is mine(gen.) .

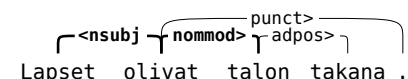
Nominals in any other case are not marked as predicatives, even if they are associated with the verb *olla*. They, similarly to adpositional phrases, are marked as *nominal modifiers (nommod)*, and the verb is marked as the head of the clause, even if it is *olla*.

Example 93



Lapset olivat pihalla .
Children were on_yard .

Example 94



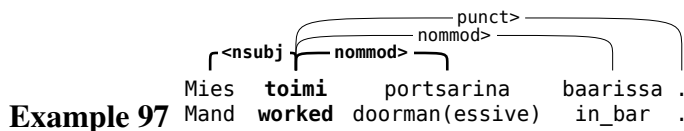
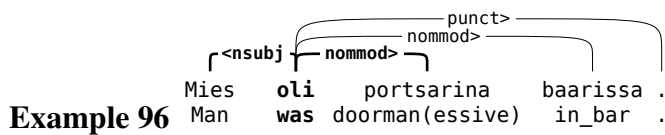
Lapset olivat talon takana .
Children were behind house .

This restriction is to prevent a clause from having two predicatives and hence two heads, which would be the case in a sentence such as the following:

Example 95 *Paketti on Oulusta ystävältäni.* (The package is from Oulu from my friend.)

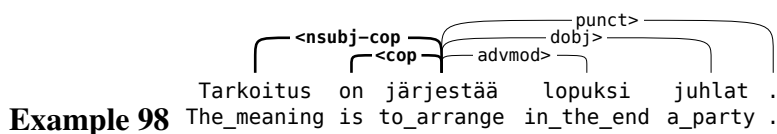
Here both *Oulusta* and *ystävältäni* could be interpreted as predicatives, resulting in a clause with two heads, or alternatively, a decision between two as likely head-candidates. Therefore, only nominative, genitive and partitive are allowed as cases for predicatives.

Note that cases *not* allowed for predicatives include the *essive* case; this is to avoid marking verbs other than *olla* as copulas.



In addition to nominals, also adverbs can act as predicatives, given that they do not express location or time. Note that with adverbs, there is no restriction with regard to case, only that they are not locational or temporal. As a result, adverbs such as *täällä* (*here*) or *huomenna* (*tomorrow*) can not act as predicatives, but others, such as *naimisissa* (*married, inessive adverb*) and *raskaana* (*pregnant, essive adverb*) can, regardless of their case.

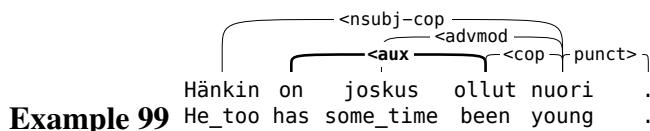
In TDT, also a full clause can act as a predicative, in addition to nominals and adverbs. In these cases, the head of the clause acting as the predicative becomes also the head of the main clause.²⁷



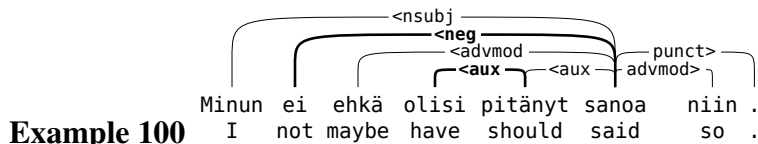
5.2.2 Copulas and auxiliaries

In the Finnish-specific version of the SD scheme, copular verbs and auxiliaries take no dependents of their own, with one exception. An auxiliary of a copular verb is attached to the copula, and not the main predicative as is the usual case. Note that this is the case even if the resulting analysis becomes non-projective.

²⁷If the clause acting as the predicative is also a copular clause, this results in the predicative clause seemingly having two copula subjects and copulas. However, this is not how the analysis should be interpreted.



The same rule is applied to the auxiliary of another auxiliary as well. All other dependents are attached to the main verb or predicative. (Note that this includes negation as well, even though negation verbs are generally considered auxiliaries.)



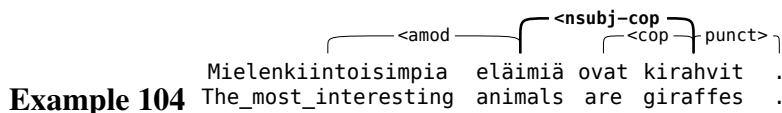
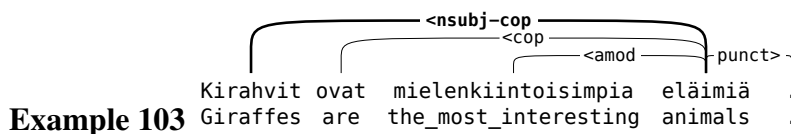
5.2.3 The distinction between the predicative and the subject

Distinguishing the subject from the predicative in copular clauses can be difficult, as it would often be possible to invert the word-order and thus swap the positions of the two elements. For instance in the following sentences, either *kirahvit* or *eläimiä* could be the subject and the other the predicative.

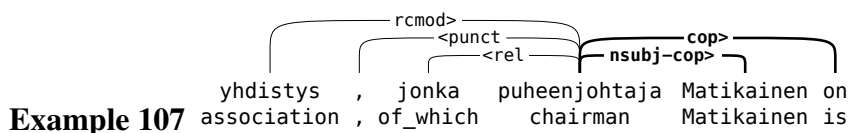
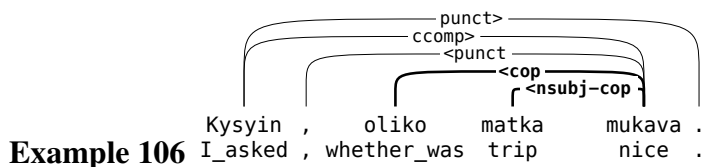
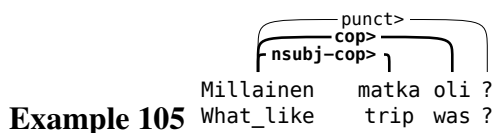
Example 101 *Kirahvit ovat mielenkiintoisimpia eläimiä.* (*Giraffes are the most interesting animals.*)

Example 102 *Mielenkiintoisimpia eläimiä ovat kirahvit.* (*The most interesting animals are the giraffes.*)

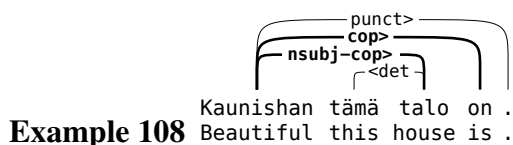
In TDT, the main rule in annotating copular structures is that the leftmost element is the subject and the rightmost one the predicative. Hence, the above sentences would be annotated in the following manner:



Semantic considerations such as which concept is a subconcept of the other are not taken into account in the annotation. However, it is possible to mark the leftmost element the predicative in cases where the word order is clearly inverted. This occurs for instance in (indirect) questions and sometimes relative clauses. Note that especially in questions, several different word orders are possible.



Also, if the leftmost element of the copular clause is an adjective rather than a noun or pronoun, it is considered that the word order is inverted, and thus the adjective is marked as the predicative, not the subject.



5.3 Appositions and appellation modifiers

The Finnish Grammar [3, §1059, §1062] distinguishes between three similar phenomena: the apposition, the appellation modifier²⁸ and the supporting noun.²⁹ Out of these, the apposition (see Section 2.6) and the appellation modifier (Section 2.29) are distinguished in TDT, and supporting noun structures are considered appositions.

All of these structures have in common that they all include two (usually adjacent) elements, most often noun phrases, which refer to the same entity or entities and have the same function in the sentence. Thus, in order to be considered an apposition, an appellation modifier or a supporting noun structure, a structure has to fulfill the following criteria (the same as in the Finnish grammar [3, §1059]):

1. Both elements of the structure must refer to the same entity or group of entities.
2. Both elements of the structure must have the same function in the sentence (for instance, the subject).

²⁸nimikemäärite

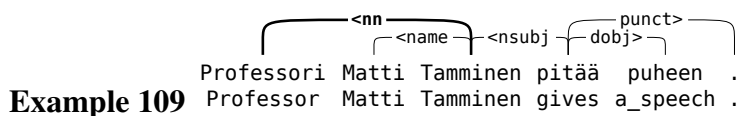
²⁹tukisubstantiivi

These criteria are interpreted rather loosely, and there are no restrictions on the part of speech of the elements involved. Most appositions (and appellation modifiers) in TDT consist of noun phrases, but there are occurrences of different parts of speech as appositions; notably the fiction section of the treebank contains few examples of verbal appositions.

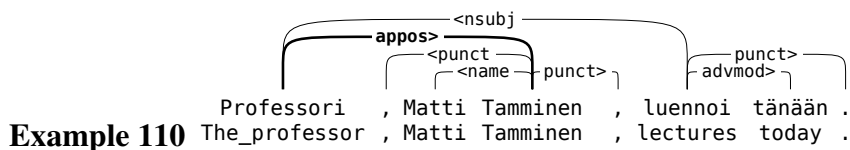
Among the expressions that fulfill criteria 1 and 2, six common cases can be distinguished according to inflection and punctuation.

1. singular, both elements in nominative, no punctuation: *professori Matti Tamminen* (*professor Matti Tamminen*)
2. singular, first element in nominative, second element inflected: *professori Matti Tammissen mukaan* (*according to professor Matti Tamminen*)
3. singular, both elements in nominative, punctuation in between: *professori, Matti Tamminen* (*the professor, Matti Tamminen*)
4. singular, first element inflected, second element in nominative: *romaanissa Putkinotko* (*in the novel Putkinotko*)
5. singular, both elements inflected: *professorin, Matti Tammissen, mukaan* (*according to the professor, Matti Tamminen*)
6. plural, elements either in nominative or inflected: *professorit Matti Tamminen ja Erkki Koivunen* (*the professors Matti Tamminen and Erkki Koivunen*) or *professoreiden, Matti Tammissen ja Erkki Koivusen, mukaan* (*according to the professors, Matti Tamminen and Erkki Koivunen*) or *professoreiden Matti Tamminen and Erkki Koivunen mukaan* (*according to the professors Matti Tamminen and Erkki Koivunen*)

Out of these six cases, the first two are considered appellation modifiers, and thus marked with the dependency type *nn*. Note that the governor of the dependency in appellation modifiers is the latter of the two words.



The remaining four cases are all considered appositions and marked with the type *appos*. Contrary to appellation modifiers, in apposition structures the first word is considered the governor.



It should be noted that case number 4 is in fact an example of a supporting noun structure, but in TDT, these are marked as appositions. In plural (case number 6), all possible case combinations are considered appositions.

The only difference between the cases 1 and 3 is the presence or absence of punctuation. Often, said punctuation is a comma, but also parentheses, a dash or a colon are possible. As can be seen from the examples above, the punctuation produces a semantic difference, which is taken into account in the annotation. Punctuation variations of the cases 2, 4, and 5 need not be considered, as these variations are ungrammatical. (Naturally, ungrammatical phenomena can and do occur in a corpus of actual language, but these cases are resolved on a case-by-case basis.)

Example 111 **professori, Matti Tammisen mukaan*

Example 112 **romaanissa, Putkinotko*

Example 113 **professorin Matti Tammisen mukaan*³⁰

5.4 Verbal dependents: Clauses, non-clauses, complements and modifiers

One particularly difficult task in annotating in the SD scheme is selecting the correct dependency type for dependents that are verbal. Verbal dependents include different kinds of subordinate clauses, as well as infinitive and participial complements and modifiers. A simplified description of the decision procedure for verbal dependents is given in Table 1, and the full details are given below.

Some basic cases are relatively easy to decide. If the dependent is a regular subordinate clause, the choices are clear. For relative clauses the type to be used is *rcmod* and as indirect questions are clausal complements, the correct type for them is *ccomp*.

If the subordinate clause is an conjunction clause, it can be either a complement or a modifier. In the majority of cases, conjunction clauses starting with the conjunction *että* are complements and clauses starting with any other conjunction are modifiers. However, it should be noted that the conjunction *että* can be used instead of the conjunction *jotta*, and respectively, also *jotta* can (especially in spoken language) be used instead of *että*.

Example 114 *Minun täytyy nyt mennä, että en myöhästy. ~jotta en myöhästy.*

Example 115 *Hän sanoi, jotta tulee vasta illalla. ~että tulee vasta illalla.*

³⁰unless a possessive reading, *the professor's Matti Tamminen*, is intended

subordinate clause?	type?								
yes	relative clause	<i>rmod</i>							
	indirect question	<i>ccomp</i>							
	conjunction clause	complement?							
		yes		<i>ccomp</i>					
		no		<i>advcl</i>					
no	governor?								
	noun	dependent?							
		participle		<i>partmod</i>					
		infinitive		<i>infmod</i>					
	verb	complement/modifier?							
		complement		clausal?					
				yes		subject?			
						shared		<i>xcomp</i>	
						not shared		dependent?	
								infinitive/participle	<i>icomp</i>
								referative/temporal	<i>ccomp</i>
		modifier		no		<i>partmod</i>			
				dependent?		<i>advcl</i>			
				infinitive/temporal		<i>partmod</i>			
				participle					

Table 1: Table guide for selecting a dependency type for verbal dependents.

In these cases, a clause starting with *että* is a modifier, and a clause starting with *jotta* is a complement. Complement conjunction clauses are marked with *ccomp* and modifier ones with *advcl*.

If the dependent is not a subordinate clause, the next deciding factor is the POS of the governor. If the governor is a noun, the dependent can be an infinitive modifier (*infmod*) or a participle modifier (*partmod*).

If, in turn, the governor is a verb, then the dependent can be either a complement or a modifier. A complement can be either clausal or non-clausal. With clausal complements, there are three alternative dependency types available: *xcomp*, *icomp* and *ccomp*.

If the subject of the dependent is shared with the governor (subject control), the correct type to use is *xcomp*. If not, the decision is made by the morphology of the dependent. If the form of the verb is an infinitive or a participle, the correct type is *icomp*; also participles are considered infinitival verb forms in TDT. If, in turn, the verb is in a finite form,³¹ the correct type is *ccomp*.

If the dependent is a non-clausal complement, it is a participial complement that resembles adjectival complements. Some of these complements can be modified, but all the same they do not form clauses. These participial complements do not have their own dependency type, but the type *partmod* is used.

Example 116 *Poika vei kotitehtävän opettajan tarkastettavaksi.* (The boy took the homework to be inspected by the teacher.)

If the dependent is not a complement but a modifier, again the morphology of the dependent decides the dependency type. If the dependent is either an infinitive or a temporal form, then the correct dependency type is *advcl*. These cases are usually easily recognized as *lauseenvastike* (“substitute of a clause”).

Example 117 *Pyyhittyään pölyt hän imuroi.* (After dusting, he hoovered.)

If the dependent is a participle, the correct type is *partmod*. These participial modifiers of a verb are often in the essive case.

Example 118 *Huolestuneena seurasin tilanteen kehittymistä.* (Worried, I followed the development of the situation.)

5.5 Attachment issues: word-order-dependent structures and ambiguity

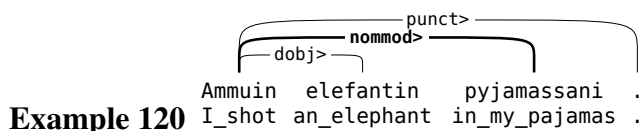
Occasionally determining the correct head word for a dependency may be difficult. Some structures are inherently ambiguous, and with some structures, often

³¹For instance, the verb form *juoksevan* can, in addition to a participle, be a finite form, as in *näin miehen juoksevan*. See for instance [3, §938, §1452] about referative and temporal structures, which are considered finite.

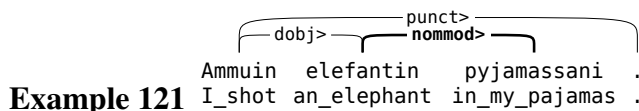
ones involving nominal modifiers, the dependent is most naturally seen to modify different sentence elements depending on the word-order. The following classic example is ambiguous:

Example 119 *Ammuin elefantin pyjamassani.* (*I shot an elephant in my pajamas.*)

In this example, it is possible that the shooting happened while wearing the pajamas, in which case the correct syntax tree would be as follows:



On the other hand, it is also possible that the elephant wore the pajamas, in which case the correct analysis is:



In TDT, ambiguities such as this one are resolved as far as possible, and also context is used to determine the correct reading where applicable. That is, if in the context there exists another sentence which makes it clear whether the shooter or the elephant wore the pajamas, then that sentence is used to disambiguate the structure.

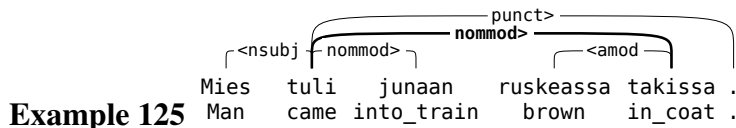
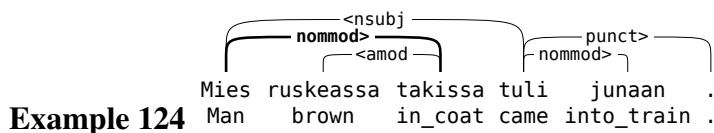
If, however, the ambiguity cannot be resolved even given context, or if an element seems to modify two or more elements simultaneously, then the attachment higher in the tree is chosen. In the case of the previous example, this would be the reading in which the shooting happens wearing the pajamas.

In some structures, the most natural analysis may be word order dependent. Consider the following two examples.

Example 122 *Mies ruskeassa takissa tuli junaan.* (*A man in a brown coat came into the train.*)

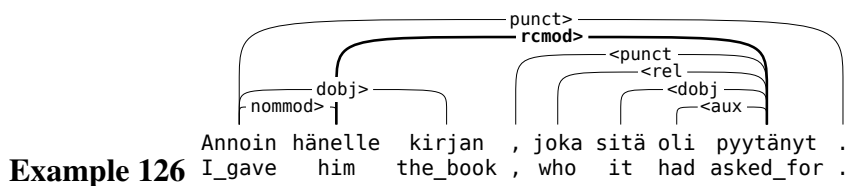
Example 123 *Mies tuli junaan ruskeassa takissa.* (*A man came into the train in a brown coat.*)

In the former example, there is clearly *a man in a brown coat*, whereas in the latter case, the coming into the train happened *while wearing a brown coat*. Therefore, the correct TDT analyses for these examples differ in their attachment of the phrase *in a brown coat*. These attachment rules are akin to those used in the Prague Dependency Treebank [2].

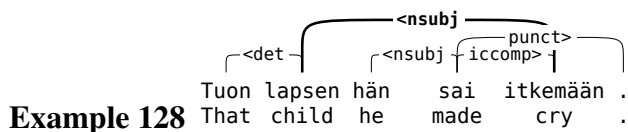
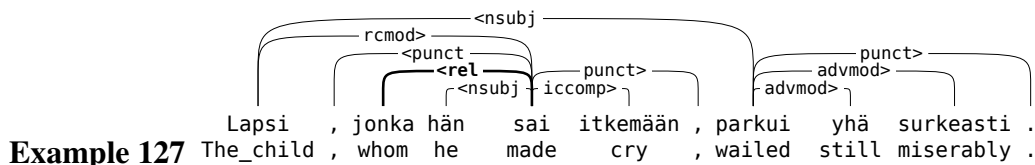


5.6 Relative clauses

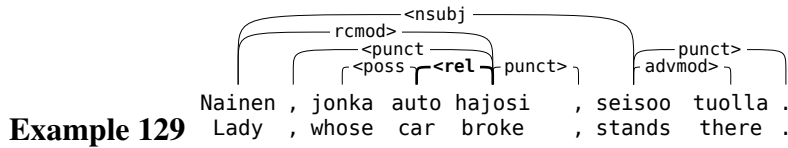
Relative clauses most often modify noun phrases, but it is also possible for them to modify a whole clause. From a prescriptive perspective, the relativizer that should be used in relative clauses that modify noun phrases is *joka*, and the relative clause should always modify the word directly before it. The relativizer that should be used in relative clauses modifying full clauses is *mikä*. However, in real, especially spoken, language, the use of the two relativizers is mixed, and not every *joka* clause actually refers to the word adjacent to it. In TDT, the actual reference for the relative clause is chosen as the head of the *rcmod* dependency wherever possible.



As the analyses of the *base* layer of TDT are trees, the relativizer is always marked using the dependency type *rel*, and its secondary syntactic function is marked in the separate *conjunct propagation and additional dependencies* layer (see Section 3.3). In most cases the *rel* dependency and its corresponding second layer dependency are between the exact same tokens. However, because the governor of the *rel* dependency is always the head of the relative clause, this does not hold for all cases.

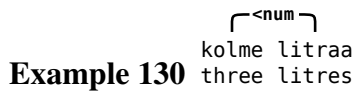


Note also that the dependent of the *rel* dependency is always the head of the relative phrase, which may or may not be the relative word itself.



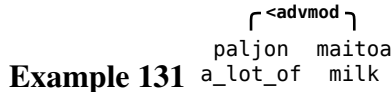
5.7 Units, measures and amounts

There are several ways to express amounts. The most simple case is expressing amount with numbers: *three apples, sixteen litres*.

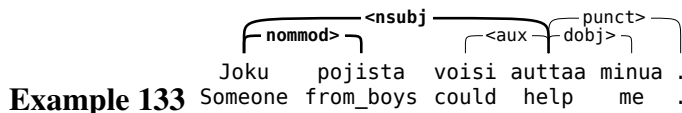
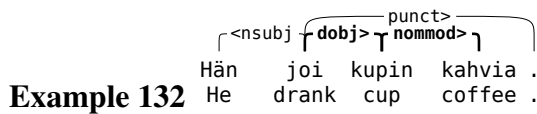


As in the English SD scheme, also in the Finnish-specific scheme version the semantic head, *litraa* in the above example, is selected as the head, and the number is marked as a numeral modifier, *num*.³² For more information on the internal structure of numerical expressions, see Section 5.12.

Amount can also be expressed with adverbs. This, too, is handled by selecting the semantic head as the head of the structure, that is, the noun.



In addition, amount can be expressed using a nominal, often in expressions such as *kuppi kahvia* (*a cup of coffee*) or *joku pojista* (*one of the boys*, “*someone from the boys*”). In these cases, the first nominal is marked as the head.



These structures are considered different from the amount expressions with numerals or adverbs, as their inflection behaves differently. Consider the following examples.

³²Morpho-syntactically, the number *kolme* could also be considered the head, as it determines the case used for the word *litra*.

Example 134 *Kieltäydyin kolmesta donitsista.* (I refused three doughnuts.)

Example 135 *Kieltäydyin kupista kahvia.* (I refused a cup of coffee.)

In the first example, both parts of the amount expression inflect as required by the verb *kieltäytyä* (to refuse), whereas in the latter case, only the first nominal inflects, signaling that the head, the thing refused in this expression, is the cup. The structure *Joku pojista* behaves and is annotated similarly.

Two things should be noted about the above analysis of *joku pojista*. First, this analysis leads to *yksi pojista* (one of the boys) being analyzed similarly to *joku pojista* rather than *yksi poika* (one boy).

Example 136

$\left. \begin{array}{c} \text{<nommod>} \\ \text{Yksi} \end{array} \right\} \text{<nsubj>} \left. \begin{array}{c} \text{pojista} \\ \text{ran} \end{array} \right\} \text{<advmod>} \left. \begin{array}{c} \text{juoksi} \\ \text{ulos} \end{array} \right\} \text{<punct>} \\ \text{One} \quad \text{from_boys} \quad \text{ran} \quad \text{out} \quad .$

Second, this analysis allows a structure like *joku pojista* to act as a predicative, as the head of the expression is in nominative.

Example 137

$\left. \begin{array}{c} \text{<nsubj-cop>} \\ \text{Se} \end{array} \right\} \left. \begin{array}{c} \text{oli} \\ \text{was} \end{array} \right\} \text{<cop>} \left. \begin{array}{c} \text{joku} \\ \text{someone} \end{array} \right\} \text{<nommod>} \left. \begin{array}{c} \text{pojista} \\ \text{from_boys} \end{array} \right\} \text{<punct>} \\ \text{It} \quad \text{was} \quad \text{someone} \quad \text{from_boys} \quad .$

5.8 Noun phrases without nouns

In TDT, it is considered that it is possible for a phrase with a head word other than a noun (or pronoun) to act as a noun phrase. Typical cases of this include adjective-headed and participle-headed noun phrases.

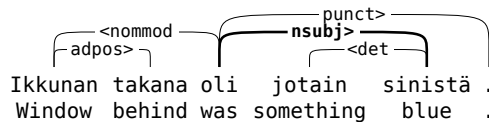
Example 138 *Ikkunan takana oli jotain sinistä.* (There was something blue behind the window).

Example 139 *Kukista kaunein oli punainen ruusu.* (The most beautiful of the flowers was a red rose.)

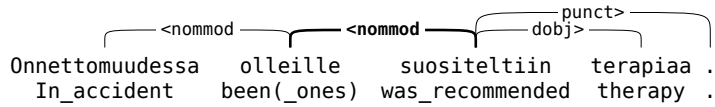
Example 140 *Kirjaa kirjoittavat sanoivat samaa.* (The (ones) writing a book said the same.)

Example 141 *Onnettomuudessa olleille suositeltiin terapiaa.* (Therapy was recommended for the (ones) been in the accident.)

These structures are analyzed as standard noun phrases. For instance, they can be marked as the subject of a clause, or a nominal modifier, regardless of the part-of-speech of the head word.



Example 142 Ikkunan takana oli jotain sinistä .
Window behind was something blue .



Example 143 Onnettomuudessa olleille suositeltiin terapiaa .
In_accident been(_ones) was_recommended therapy .

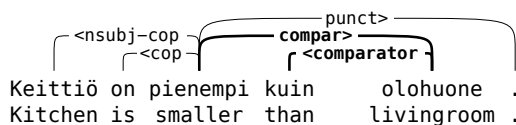
5.9 Comparatives and superlatives

This section describes annotating comparative and superlative structures, which, in TDT, are considered to include also certain similar structures that do not contain a comparative or superlative wordform.

5.9.1 Comparatives

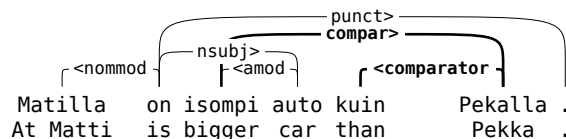
Structures with comparative adjectives and adverbs may be difficult to annotate: they are often elliptical, and it may be difficult to tell what is being compared with what. In the Finnish-specific version of the SD scheme, there are two dependency types that are reserved for comparative structures, *compar* and *comparator*. Both of these types are new types not present in the original SD scheme.

The basic usage of these two types is as follows. The comparative adjective or adverb acts as the head for a *compar* dependency, and *the element being compared* is its dependent. The element being compared also acts as the head for a *comparator* dependency, the dependent of which is a comparative conjunction, nearly always *kuin*.



Example 144 Keittiö on pienempi kuin olohuone .
Kitchen is smaller than livingroom .

Note that the comparative adjective or adverb remains the head of the *compar* dependency even if the word order is such that the dependency becomes non-projective.

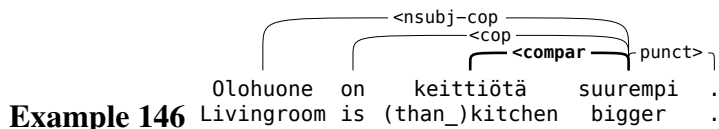


Example 145 Matilla on isompi auto kuin Pekalla .
At_Matti is bigger car than Pekka .

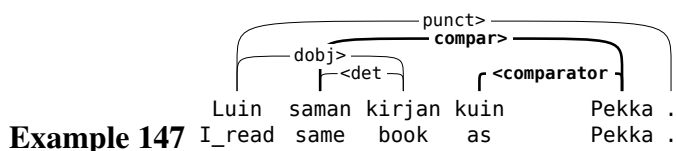
From the previous example it can also be seen that comparative structures are often elliptical in some way. Strictly speaking, the example does not compare Matti and Pekka, but rather their cars, and the car owned by Pekka is not explicitly present in the sentence. As a general rule of thumb, the different kinds of ellipsis

present in comparative structures are not marked with null tokens, but rather the available elements are used wherever possible.

It is also possible to make comparisons without the comparative conjunction *kuin*. In these cases, only the dependency type *compar* is used, marking the comparative adjective or adverb as the head, and the element compared as the dependent, just as in the case with the comparative conjunction present.



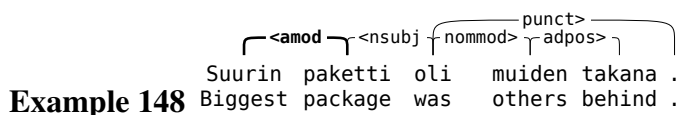
In TDT, also some structures not involving a comparative adjective or adverb can be marked as comparatives. In order to qualify as a comparative construction, a structure has to contain either a comparative word form or a word form that otherwise semantically entails comparison, such as *samanlainen* (*similar*), *sama* (*same*), *erilainen* (*different*) or *eri* (*differing, separate*).³³



An additional difficulty is posed by the fact that in Finnish, the comparative conjunction *kuin* can also appear as a subordinating conjunction as well as an adverb. Borderline situations are resolved on a case-by-case basis, considering whether or not there is a comparison involved in the structure and, secondarily, whether the dependent structure is a clause.³⁴

5.9.2 Superlatives

Superlatives are less problematic than comparatives but deserve some attention nevertheless. The basic case with superlatives is simple: a lone superlative modifying a noun. The superlative form in this case is not marked in any particular way in the syntax annotation, but the structure is annotated similarly to any adjective modifying a noun. The same strategy of not marking the superlative in any particular way is also used in cases where the superlative acts as a predicative.



³³Note that for example the word *sama* is in fact a pronoun in Finnish.

³⁴Comparative structures can also occasionally be full clauses.

Often a superlative is modified by nominal in some manner. A very common phenomenon is a genitive modifier modifying a superlative. For instance, in an expression such as

$\left\{ \begin{array}{l} <poss> \\ <amod> \end{array} \right.$
 Suomen paras kokki
 Finland's best cook

Example 149

the cook is the best *of those in/of Finland* and thus the correct head word for the genitive modifier is *paras*. Similarly, an ordinal number can act as the head of a genitive modifier. For example, in

$\left\{ \begin{array}{l} <poss> \\ <num> \end{array} \right.$
 Virtasen kuudes mestaruus
 Virtanen's sixth championship

Example 150

the championship is the sixth out of *those of Virtanen*, and thus the genitive modifier should modify the ordinal number.

However, it is still possible for the noun to act as the head word in some cases. For instance, in

$\left\{ \begin{array}{l} <poss> \\ <amod> \end{array} \right.$
 Rusakon pahin vihollinen
 The_hare's worst enemy

Example 151

the enemy is not the worst *of the hare*, but rather it is an enemy of the hare, and it is the worst enemy. Thus, the head word should be *hare*.

As a rule of thumb, if the noun phrase containing the genitive modifier can be turned into a copular clause in the following fashion, then the genitive modifier should modify the superlative or ordinal number.

Example 152 *Kokki on Suomen paras.* (*The cook is the best in Finland.*)

Example 153 *Mestaruus on Virtasen kuudes.* (*The championship is the sixth for Virtanen.*)

are perfectly valid, but

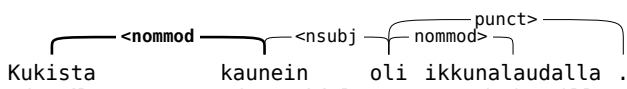
Example 154 *?Vihollinen on rusakon pahin.* (*?The enemy is the worst of the hare.*)

is questionable at best. Thus, in *Suomen paras kokki* and *Virtasen kuudes mestaruus*, the genitive modifier is considered to modify the superlative adjective, but in *rusakon pahin vihollinen*, it is considered to modify the noun directly.

In this context, it should also be noted that in addition to superlatives, also certain other adjectives can also act as the head of a genitive modifier. These

adjectives can be semantically superlative-like (*viimeinen* (*last*)), but there are also many others, such as *oma* (*own*), *kaltainen* (*-like*), *välinen* (*between* (*adj.*)), and *vastainen* (*against* (*adj.*)).

Also other nominal modifiers are possible, expressing the set of beings from which the objects are drawn when making the comparison. These are treated similarly to the genitive modifiers, making the superlative wordform the head of the modifier if the modifier expresses the set of beings to draw from.

Example 155 
 Kukista kaunein oli ikkunalaudalla .
 From_the_flowers most_beautiful was on_windowsill .

Note how in the previous example the phrase *kukista kaunein* can act as a noun phrase (it is the subject of the clause), even though its head word is an adjective. See Section 5.8 on nounless noun phrases.

5.10 Subordinate clauses and expressions of time

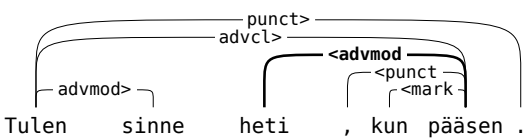
Many subordinate clauses, especially ones starting with the conjunction *kun* (*when*), come with an adverbial, usually expressing time. Consider the following examples.

Example 156 *Tulen sinne heti, kun olen imuroinut.* (*I'll come there right away, when I have hoovered.*)

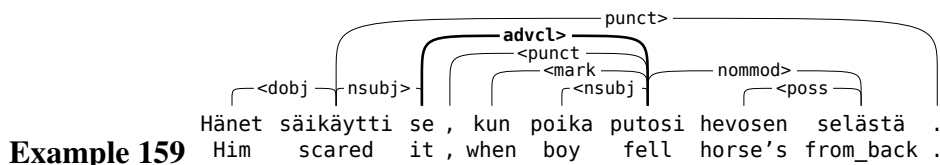
Example 157 *Tapasin hänet sen jälkeen kun olin tullut kaupasta.* (*I met him after I had come from the store.*)

It is often unclear where these time adverbials should be attached. On the one hand, they seem to modify the main clause, expressing when the action of the main clause takes place. On the other hand, they could also modify the subordinate clause, being a part of the time condition given in the subordinate clause. A third option would be to make the time adverbial depend on the subordinating conjunction, to make the whole expression a two-part conjunction. The third option has some intuitive appeal, but this would make the number of subordinating conjunctions excessively large.

In TDT, a very limited number of these cases are considered especially tightly bound with the subordinating conjunction. These cases are considered multi-part subordinating conjunctions and listed as such in Section 2.26. Otherwise, these adverbials are consistently made dependents of the subordinate clause.

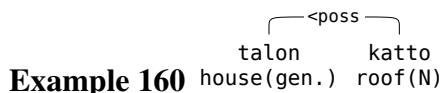
Example 158 
 Tulen sinne heti , kun pääsen .
 I_will_come there right_away , when I_can .

However, it should be noted that all subordinate clauses themselves are not dependents of the main verb. As discussed in Section 2.10, clausal complements can depend on nouns, pronouns or adverbs. Similar situations can occur with subordinate clauses that are modifiers, and they are also analyzed similarly. Most commonly this occurs with the pronoun *se* (*it*).

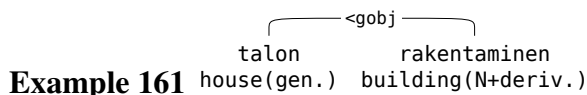


5.11 Subjects and objects of a noun

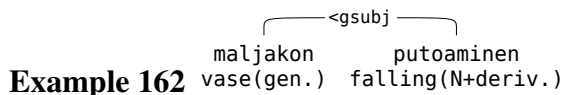
In Finnish, it is possible for certain nouns which either are direct derivations of a verb or otherwise have a verb counterpart³⁵ to take a subject- or object-like complement. Both of these are identical in form to more general genitive modifiers of a noun, marked with the dependency type *poss* in the SD scheme.



Genitive objects of a noun are marked the *gobj*, which is a subtype for the more general genitive-modifier type *poss*. Both nominal derivations and other nouns with verb counterparts can take a genitive object, with the exception of JA-derivations, the genitive modifier of which is never considered an object in TDT (*talon rakentaja, the builder of the house*).



Genitive subjects, in turn, are marked using the *gsubj* dependency type, also a subtype of *poss*. Only nouns that are marked as derivations of a verb in the morphological tagging present in TDT receive a *gsubj* dependent.³⁶

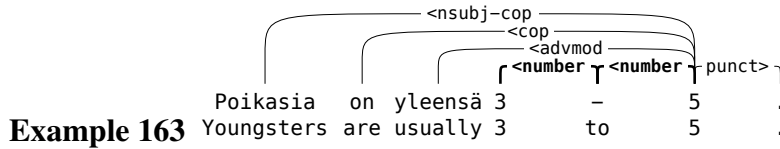


³⁵verbivastineellinen substantiivi [3, §560]

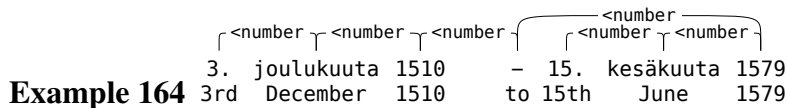
³⁶These dependencies were added in a separate annotation phase, and finding verb derivations based on the morphological tagging was feasible, while finding other nouns with a verb counterpart was not.

5.12 Numerical expressions

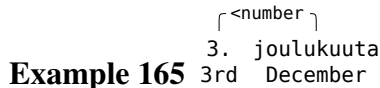
The dependency type *number* in the SD scheme is reserved for numerical expressions. Generally, with multi-token numerical expressions, the rightmost token of the expression is considered the head and the dependencies are chained.



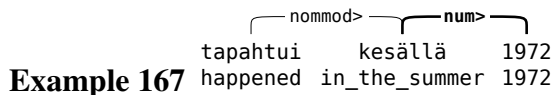
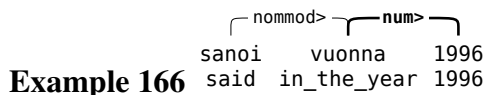
However, it is possible that rather complex expressions are considered numerical, and in these cases the structure of the expression is also marked, showing the parts of which the expression consists. Often these complex expressions involve dates, which are also considered numerical expressions in TDT.



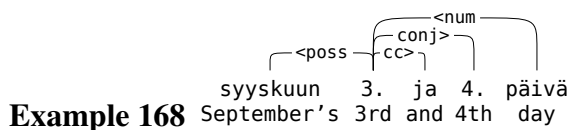
Dates can be expressed using many different forms, and all full dates are considered numerical expressions in TDT, also those where some or all parts of the date are written with characters. Even partial dates such as



are considered numerical expressions. However, year expressions such as the following are not considered dates in TDT, and thus not complex numerical expressions.



If a date expression has a clear internal syntactic structure, this structure is annotated instead of the default chain of *number* dependencies.



If a date has a more specific time (such as *kello kuudelta*, *at six o'clock*) attached to it, the date is considered the head of the expression, and the more specific time depends on it. Clock expressions, alone or in conjunction with a date, are not considered dates or numerical expressions in TDT.

$\left[\begin{array}{c} \langle \text{number} \rangle \langle \text{nommod} \rangle \langle \text{num} \rangle \\ \text{6. joulukuuta kello 18} \\ \text{6th December o'clock 18} \end{array} \right]$

Example 169

In addition to dates, there is one more case of numerical expressions that deserves attention: numerical expressions with multiple units. If a single amount expression involves multiple units, the units are considered a compound unit so to say, and combined using the dependency type *nn*.

$\left[\begin{array}{c} \langle \text{num} \rangle \langle \text{nn} \rangle \langle \text{num} \rangle \\ \text{2 kg 315 g} \end{array} \right]$

Example 170

In rare cases, however, the previous situation may occur with the rightmost part of the expression lacking the unit. These cases are annotated flatly as numerical expressions, with no compound units.

$\left[\begin{array}{c} \langle \text{number} \rangle \langle \text{number} \rangle \\ \text{2 kg 315} \end{array} \right]$

Example 171

5.13 Participial modifiers and predicatives

In connection with participial modifiers, predicatives are given a slightly different treatment than in other contexts. In a regular copular clause, the analysis is as follows.

$\left[\begin{array}{c} \langle \text{nsubj-cop} \rangle \langle \text{cop} \rangle \langle \text{punct} \rangle \\ \text{Eeva on raskaana .} \\ \text{Eeva is pregnant .} \end{array} \right]$

Example 172

However, if the same analysis were applied in a situation where *olla* acts as a participial modifier, this would result in a non-tree structure:

$\left[\begin{array}{c} \langle \text{nsubj-cop} \rangle \langle \text{cop} \rangle \langle \text{nsubj-cop} \rangle \langle \text{cop} \rangle \langle \text{punct} \rangle \\ \text{Raskaana oleva nainen on nälkäinen .} \\ \text{Pregnant being woman is hungry .} \end{array} \right]$

Example 173

Therefore, in conjunction with participial modifiers, copular verbs are analyzed similarly to regular verbs, in order to avoid non-tree structures.

$\left[\begin{array}{c} \langle \text{advmod} \rangle \langle \text{partmod} \rangle \langle \text{nsubj-cop} \rangle \langle \text{cop} \rangle \langle \text{punct} \rangle \\ \text{Raskaana oleva nainen on nälkäinen .} \\ \text{Pregnant being woman is hungry .} \end{array} \right]$

Example 174

The same rule is applied to certain special constructions that are normally considered passive structures but can also appear in conjunction with participial modifiers. Here the application of the rule results in two chained participial modifiers.

Example 175

<dobj	<auxpass	punct>
┌──────────┴──────────┐		
Resurssit	ovat	käytettävissä .
Resources	are	usable .

Example 176

<partmod	<partmod	<nsbj-cop	<cop	punct>
┌──────────┴──────────┐		┌──────────┴──────────┐		
┌──────────┴──────────┐		┌──────────┴──────────┐		
Käytettävissä	olevat	resurssit	ovat	rajalliset .
Usable	being	resources	are	limited .

5.14 Necessive structures and clausal subjects

A clause can act as a subject to another clause,³⁷ in which case it should be marked as a clausal subject, *csbj*, or, if the main clause is copular, a clausal copular subject, *csbj-cop*. However, in the case of clausal-copular subject, it may be difficult to determine whether a clause is, in fact, the subject of another clause, as the construct is similar to that of a *necessive structure*. Consider the following example.

Example 177 *On tärkeää syödä hyvin. (It is important to eat well.)*

At first glance, it seems that the clause *syödä hyvin* is the subject of *on tärkeää*. However, in TDT, this is not considered a clausal subject. Instead, it is considered a *necessive structure*, as *on tärkeää* can be given a subject in the genitive form:

Example 178 *Hänen on tärkeää syödä hyvin. (It is important for him to eat well.)*

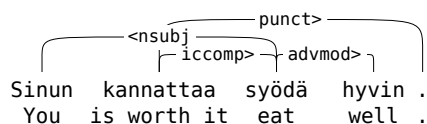
The whole structure is considered a single unit, and the genitive subject is considered the subject of the latter verb (which expresses what it is that is necessary).

Example 179

<nsbj	<cop	iccomp>	nommod>	punct>
┌──────────┴──────────┐		┌──────────┴──────────┐		
┌──────────┴──────────┐		┌──────────┴──────────┐		
Hänen	on	pakko	mennä	kotiin .
He	has	to	go	home .

The name *necessive structure* comes from the fact that these structures often express the necessity of doing something, but it does not mean that all of these structures would have such a meaning; for example, *on vaikea(a)* (*it is difficult*) is a *necessive structure* the meaning of which does not express necessity. Common *necessive structures* include expressions such as *on pakko*, *on tärkeää*, *on oleellista* and *on välttämätöntä*. They usually, but not always, involve the verb *olla* and an adjective. There are also some verbs, such as *kannattaa* (*be worth it*) and *kuulua* (*be supposed to*), that are analyzed in a *necessive manner*.

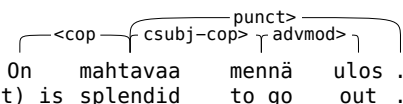
³⁷as well as an object, but these are marked as clausal complements (*ccomp*)



Example 180 You is_worth_it eat well .

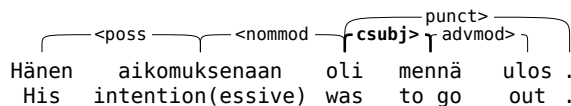
If it is not possible to insert a genitive subject into the clause, then the structure is considered a clausal subject case.³⁸

Example 181 **Hänen on mahtavaa käydä ulkona.* (It is splendid for him to go out.)³⁹



Example 182 (it)_is splendid to_go out .

Note that due to the copular nature of the main clause, the clausal subjects in these clauses which resemble necessive structures are in fact clausal copular subjects. There are also other clausal subjects which cannot be confused with necessive structures.



Example 183 His intention(essive) was to_go out .

5.15 Passive structures and zeroth person constructions

The Finnish language has two notable cases of subjectless expressions: the passive voice and the zeroth person. In most cases, distinguishing these two is rather simple, as the zeroth person uses the same verb forms as the third person, whereas there is a morphological passive form that is used in constructions considered passive. However, there are at least two particular phenomena that deserve special attention. First, the *on tehtävä* -structure is worth examining:

Example 184 *Tämä työ on tehtävä tänään.* (This work has to be done today.)

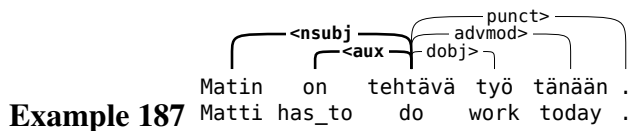
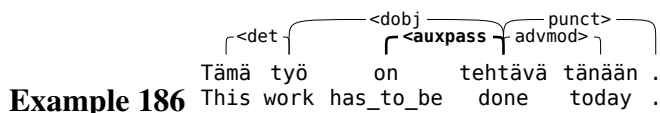
The form *tehtävä* is morphologically a passive participle of the verb *tehdä* (to do). Still, *on tehtävä* can take a subject, which could perhaps point towards to the subjectless version being zeroth person after all.

Example 185 *Matin on tehtävä työ tänään.* (Matti has to do the work today.)

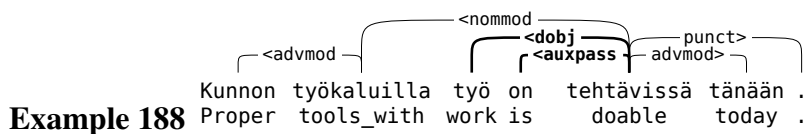
In TDT, we use the presence or absence of a subject as a cue to whether the structure is passive or not. If a subject is present, the structure is marked as an active construction, and if not, it is assumed to be passive.

³⁸This is an area where language intuitions differed between annotators, and these decisions were made on a case-by-case basis in TDT.

³⁹The Finnish sentence is ungrammatical, whereas the translation may be grammatical.



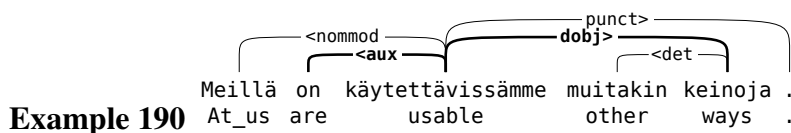
Second, the *on tehtävissä* structure deserves a mention. Similarly to *tehtävä*, *tehtävissä* is a passive verb participle — in fact, the difference between the two forms is only that *tehtävissä* is the plural inessive form of the base participle *tehtävä*. The annotation of *on tehtävissä* follows a strategy similar to the previous one. In general, it is assumed that the structure is passive.



Unlike *on tehtävä*, *on tehtävissä* cannot take a genitive form subject:

Example 189 **Minun on tehtävissä tämä.* (“*I this is doable.”)

However, in some cases it is possible to attach a possessive suffix to the participle and use a corresponding personal pronoun as a nominal modifier.⁴⁰ This case is analyzed as an active structure.



However, as can be seen from the example, no subject is marked, but rather an object. It is still understood that *means* are the object of *using* in this example.

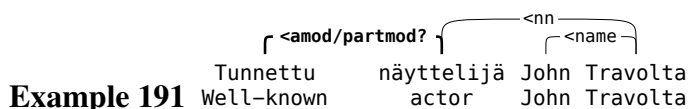
5.16 Morphological distinctions

Distinctions between certain dependency types, most commonly between participial modifiers (*partmod*) and adjectival modifiers (*amod*) as well as adverbial modifiers (*advmod*) and nominal modifiers (*nommod*), are based on the corresponding morphological distinction, which can sometimes be rather difficult. This section describes heuristics used in TDT to make these two most common morphology-based distinctions. Some of these heuristics resemble those used in the Penn Treebank [6].

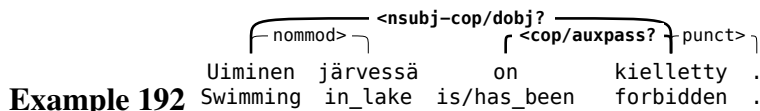
⁴⁰This is a rare phenomenon and not seen with many verbs.

5.16.1 Participles versus adjectives

The distinction between verb participles and adjectives is difficult in several languages, and Finnish is no exception. In TDT, this distinction affects the syntax annotation of mainly two kinds of structures. First, it affects the choice between the dependency types *partmod* (participial modifier) and *amod* (adjectival modifier).



Second, it affects whether certain structures should be marked as copular clauses, or alternatively, as passive clauses in the present or past perfect form.⁴¹ The same structure can be considered copular if the head word is an adjective, or a passive clause if the head word is considered a passive participle.



The syntax in TDT has been annotated using the output of a Finnish morphology tool, FinTWOL,⁴² and the July 2013 release includes morphological information based on the open source tool OMorFi [5, 8]. Thus the first source of information for annotators in cases of morphological ambiguity are the analyses given by FinTWOL and OMorFi. However, some words receive several readings, and it is fairly common that a word receives both a participial reading and an adjectival one. In addition, it is also possible that the most natural reading for the word in the current context has been omitted. Thus, the following heuristics are used when deciding whether a word is an adjective or a participle.

If a word can receive comparative and superlative forms, it is likely to be an adjective. For instance, the word *tunnettu* (*well-known*), which has both an adjectival and a participial reading, inflects in these forms: *tunnettu*, *tunnetumpi*, *tunnetuin*.

If, on the other hand, the word is modified by for instance a nominal or adverbial modifier, it is likely to be a verb participle. For instance, with the word *tunnettu*, the following contexts would be possible:

Example 193 *laajalti tunnettu näyttelijä* (*widely known actor*)

Example 194 *kalliista autoistaan tunnettu näyttelijä* (*actor known for his expensive cars*)

⁴¹*perfekti* and *pluskvamperfekti* in Finnish grammar

⁴²<http://www.lingsoft.fi/>

Thus, it is the case that the same word can act both as an adjective and as a verbal participle, depending on context, and the decisions are made on a case-by-case basis. As a third heuristic used in the decision, the annotators are asked to consider whether someone is actively doing something in the example under consideration. If so, then the word is likely a verbal participle, otherwise it is an adjective. Consider the following examples:

Example 195 *Maijan tuleva aviomies* (*Maija's future husband*, “*Maija's coming husband*”)

Example 196 *Maijan Turusta tuleva aviomies* (*Maija's husband coming from Turku*)

In the first example, the husband is not actively doing anything, he simply is going to be Maija's husband in the future. Thus *tuleva* in this example would be considered an adjective. In the second example, he is actively coming from the direction of Turku, and thus *tuleva* here would be a verbal participle.

As a rule of thumb, if an adjectival reading is possible in a given context, it is generally preferred. For instance, in *tunnettu näyttelijä*, if it was not specified by whom or for what the actor is known, it would be assumed that the adjectival reading is intended. Similarly, in *uiminen on kielletty*, if the context does not reveal that there has been active forbidding of the swimming (the example is genuinely ambiguous), then it is assumed that it is a property of the swimming that it is forbidden.

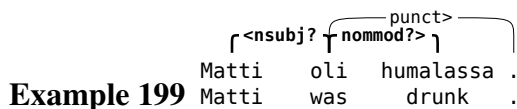
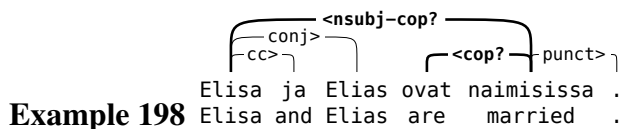
5.16.2 Adverbs versus nouns

Due to the fact that certain Finnish adverbs have a partial case inflection, it is sometimes difficult to decide whether a word is an inflected form of a noun (or adjective), or rather an adverb. For instance, the word *pääasiassa* (*mainly*) could be analyzed as an adverb, or alternatively, as an inflected form of the noun *pääasia* (*the main thing*).

This distinction affects the choice between the dependency types *advmod* (adverb modifier) and *nommod* (nominal modifier). Additionally, it can affect the choice of whether a word can be marked as a predicative (if it is an adverb) and thus head of the clause, or if it should be marked as a nominal modifier for the verb *olla*. In the latter case, the structure of the whole clause is affected by the decision.

Example 197

<advmod/nommod?	<nsbj	punct>	nommod>	<poss	
Pääasiassa	tämä	vaikuttaa	koron	suuruuteen	.
Mainly	this	affects	interest's	level	.



Again, the main source of information while annotating is the morphological analysis of the word, but occasionally it is possible that the syntactic annotation uses a reading that has been omitted. It is less common that both an adverb and noun reading would be available. Decision heuristics are needed here as well.

The main deciding factor between a noun and an adverb reading is whether there exists a corresponding noun in its baseform and whether and to what degree the word under question is related to that noun. For example, in the case of *pääasiassa* (*mainly*) there exists a corresponding noun *pääasia* (*main thing*), but in the case of *naimisissa* (*married*) the only candidate for such a noun would be *naiminen*, which could technically be translated as *marrying*, but is in fact more often used (usually in spoken language) in the meaning *having sex*. As for *humalassa* (*drunk*), there is a candidate noun, *humala*, which can be used to refer to the state of being drunk.

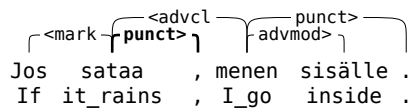
As a test used to see whether the possible candidate noun is closely (enough) related to the word under question, annotators are asked to reflect on the hypothetical baseform of the noun reading and on whether it could be imagined to be involved in the current sentence. For instance, is there a *main thing* (*pääasia*) in which the interest rate is affected? Is there a state of being married (“*naimiset*”) in which Elisa and Elias are? Is there a state of being drunk (*humala*) in which Matti is? The answer to the first two questions is no, and thus *pääasiassa* and *naimisissa* are considered adverbs. The answer to the third question, however, is yes, and therefore the word *humalassa* is analyzed as an inflected form of the noun *humala* in TDT.

5.17 Attaching punctuation

Dependencies signaling punctuation are labeled with the dependency type *punct*, and the main rule is that the dependency should be attached to *that element which it delimits*. Thus, sentence-delimiting punctuation, such as “.”, “!” or “?” should be attached to the main verb (or predicative) of the sentence.

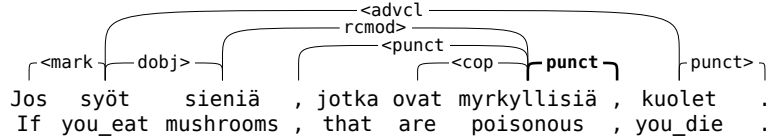


According to the same rule, the comma delimiting a subordinate clause should be attached to the head word of said clause.



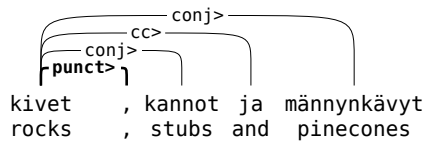
Example 201 If it_rains , I_go inside .

If there are several subordinate clauses within each other and the punctuation could delimit any of them, the shortest-spanning (closest) clause is selected.



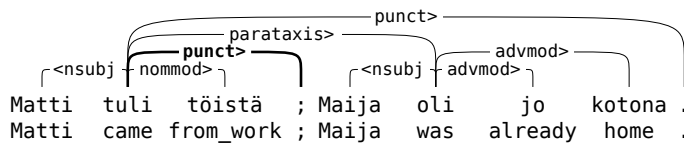
Example 202 If you_eat mushrooms , that are poisonous , you_die .

In coordinations, the punctuation symbols (usually commas) are treated similarly to the coordinating conjunction and attached to the head of the coordination, which is the first coordinated element.



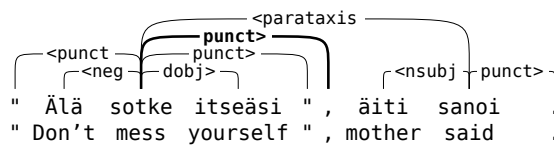
Example 203 rocks , stubs and pinecones

Punctuation related to coordination-like parataxis, that is, parataxis used in connection with a semicolon, colon or dash, is attached as in coordinations.



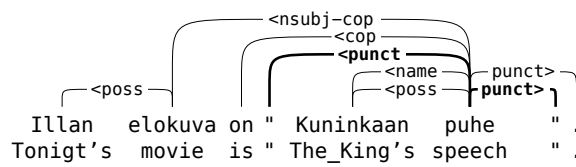
Example 204 Matti came from_work ; Maija was already home .

Punctuation with direct speech -type parataxis, however, is seen to delimit the utterance of the speaker.



Example 205 " Don't mess yourself " , mother said .

Single and double quotes as well as parentheses are attached to the head of the quoted/parenthetical clause or phrase. Dashes signifying quotes are also attached to the head of the quote.



Example 206 Tonight's movie is " The_King's speech " .

Example 207 Matikainen (s. 1943) on ammatiltaan kirjailija .
 Matikainen (born 1943) is by_profession author .

Example 208 - Älä sotke itseäsi , sanoi äiti .
 - Don't mess yourself , said mother .

If the quotes or parentheses contain two or more items, such as parts of a coordination, then the punctuation is attached to the closest enclosed element, so as to avoid unnecessary non-projectivity.

Example 209 Hän pitää kirjoista (ja näytelmistä) .
 He likes books (and plays) .

Punctuation can also delimit short additions, such as nominal modifiers or appositions, and in such cases, the punctuation should be attached to the head of the addition.

Example 210 Matti Tamminen , professori
 Matti Tamminen , the_professor

Example 211 Lähden matkalle , ainakin viikoksi .
 I_am_going to_trip , at_least for_a_week .

Finally, list item markers such as bullets of a bulleted list are marked as punctuation attached to the head of the list item.⁴³

Example 212 * Käy kaupassa .
 * Visit store .

Acknowledgements

Heartfelt gratitude goes to (in alphabetical order) Filip Ginter, Samuel Kohonen, Veronika Laippala, Anna Missilä, Jenna Nyblom, Stina Ojala, Tapio Salakoski and Timo Viljanen for their various direct and indirect contributions to the treebank and its annotation scheme. We would also like to thank Lingsoft Ltd. for making the morphology tool FinTWOL available to us.

⁴³It should be noted that in TDT, when selecting text for annotation, certain items with no structure, such as bulleted lists of single words, have been discarded as non-annotatable material.

References

- [1] Marie-Catherine de Marneffe and Christopher Manning. Stanford typed dependencies manual. Technical report, Stanford University, September 2008.
- [2] Jan Hajič. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová, editor, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Karolinum, Charles University Press, Prague, Czech Republic, 1998.
- [3] Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja-Riitta Heinonen, and Irja Alho. *Iso suomen kielioppi / Grammar of Finnish*. Suomalaisen kirjallisuuden seura, 2004.
- [4] Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missil, Stina Ojala, Tapio Salakoski, and Filip Ginter. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 2013. In Press. Available online. DOI: 10.1007/s10579-013-9244-1.
- [5] Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. HFST tools for morphology — an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, volume 41 of *Communications in Computer and Information Science*, pages 28–47. 2009.
- [6] Mitchell Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.
- [7] Marie-Catherine de Marneffe and Christopher Manning. Stanford typed dependencies representation. In *Proceedings of COLING'08, Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, 2008.
- [8] Tommi Pirinen. Suomen kielen äärellistilainen automaattinen morfologinen jäsennin avoimen lähdekoodin resurssien. Master's thesis, University of Helsinki, 2008.

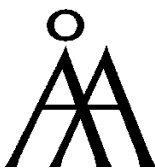
TURKU
CENTRE *for*
COMPUTER
SCIENCE

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www.tucs.fi



University of Turku

- Department of Information Technology
- Department of Mathematics



Åbo Akademi University

- Department of Information Technologies



Turku School of Economics

- Institute of Information Systems Sciences

ISBN ISBN 978-952-12-2936-7
ISSN 1239-1891