LASSY: LARGE SCALE SYNTACTIC ANNOTATION OF WRITTEN DUTCH

Gertjan van Noord

Deliverable 3-4: Report Annotation of Lassy Small

1 Background

Lassy Small is the Lassy corpus in which the syntactic annotations have been manually verified. This part contains one million words. The composition of the corpus is detailed in deliverable 1.1. The annotations include syntactic dependency annotations, as documented in deliverable 3.5 [5], and the annotation of the part-of-speech and lemma of each token, as documented in [3].

2 Annotation Procedures

Both the annotation guidelines manuals and the various tools we used for annotation were initially developed in the STEVIN D-Coi project. The annotation of part-of-speech and lemma proceeded in the same way as in D-Coi: initial assignment of part-of-speech and lemma by TadPole [2]. These automatically assigned annotations were then checked and corrected by students.

The syntactic annotation procedure works in a similar way. The Alpino parser [4] is used to assign initial dependency structures automatically. These automatically assigned annotations were then checked and corrected by students (using an adapted version of TrEd, `http://ufal.mff.cuni.cz/~pajas/tred/`). Unlike the part-of-speech and lemma annotations, most syntactic dependency annotations were checked by students a second time.

Syntactic dependency annotations and part-of-speech and lemma annotations were then integrated into a single XML representation.

For further quality improvement, a large number of heuristics has been defined and implemented to check the resulting annotations both for illegal annotations, as well as for unexpected annotions. In particular, heuristics which confront the syntactic analysis with the part-of-speech tag assignments were very effective. Based on the results of these heuristics, thousands of mistakes have been corrected. This latter step therefore appears to be a very important step, and for this reason we provide an overview of the most important of these heuristics here in the next session.

3 Verification

As indicated above, all annotations were verified systematically using `dtchecks`, a script which applies a large set of rules which check for either illegal annotations or unexpected annotations. The rules are defined as XPATH queries, and any annotation which matches a query is returned. Each of these annotations then undergoes yet another manual inspection to correct the annotation, or to confirm the exceptional annotations.

We now describe the most important rules here as follows.

- each node has atmost a single head

du	dp, sat, nucl, tag, dlink
mwu	mwp
whq	whd, body
oti	cmp,body,mod
ti	cmp,body
pp	hd,obj1,mod,vc,predc,hdf,pobj1,se
ap	hd,mod,vc,pc,obcomp,obj1,predm,predc,me,se,pobj1,obj2
svan	cmp,body,mod
cp	cmp,body,mod
ahi	cmp,body
rel	body,rhd
whrel	body,rhd
whsub	body,whd
conj	cnj,crd
advp	hd,mod,obcomp,me
detp	hd,obcomp,mod,me
inf	hd,su,predc,mod,obj1,vc,ld,pc,svp,predm,obj2,se,me,sup,pobj1
sv1	hd,su,predc,mod,obj1,vc,ld,pc,svp,predm,obj2,se,me,sup,pobj1
smain	hd,su,predc,mod,obj1,vc,ld,pc,svp,predm,obj2,se,me,sup,pobj1
ssub	hd,su,predc,mod,obj1,vc,ld,pc,svp,predm,obj2,se,me,sup,pobj1
ppart	hd,su,predc,mod,obj1,vc,ld,pc,svp,predm,obj2,se,me,sup,pobj1

Table 1: Relations which are expected to occur in nodes of the given category.

- check for expected daughers for a given category. For instance, a noun phrase can have a determiner, but a prepositional phrase does not. Table 1 lists which relations are expected for which category.
- check for expected categories for a given relation. For instance, direct objects are usually NP. The expected categories for each relation name is given in table 2.
- check for expected relations of sister nodes for a given relation. For instance, a node with relation `whd` usually has a `body` sister node. The expected sisters are listed in table 3.
- check co-indexing of nodes: for every node with an index, there should be another node with the same index. Also, a node with a index cannot dominate a node with the same index.
- check that the required attributes `cat, lemma, postag` are present in nodes of the relevant type.
- `rhd`- and `whd`-nodes usually are co-indexed.
- check that nodes of categories such as `np, smain, ...` have a head node.

hdf	mwu
hd	mwu
cmp	mwu,conj
su	np,conj,cp,ti,oti,whrel,whsub,svan,mwu
obj2	np,pp,conj,mwu,whrel
pc	pp,conj
vc	cp,ti,ppart,inf,oti,conj,whsub,ahi,svan,smain
svp	pp,mwu,ti,ahi,inf
predc	np,ap,ppart,ppres,cp,pp,conj,mwu,whrel,oti
predm	advp,np,ap,ppart,ppres,cp,pp,conj,mwu,whrel
ld	pp,np,conj,mwu,ap,advp,whrel
me	np,ap,conj,mwu
obcomp	cp,oti,ssub,ti,conj
rhd	np,pp,ap,conj,advp,mwu,ppart,ppres
whd	np,pp,ap,conj,advp,cp,mwu,ppart,ppres
mod	rel,cp,np,advp,ap,ppart,pp,mwu,conj,oti,du,smain,whrel,ti,sv1,ppres
body	ssub,ti,sv1,inf,np,conj,pp,cp,mwu,du,ppart,smain,whrel,ap,ppart,ppres
det	detp,mwu,np,conj,ap,pp
app	mwu,np,conj,smain
crd	mwu

Table 2: Categories which are expected to occur in a node with the given relation.

tag	tag,nucl,sat,dlink
rhd	body,rhd
whd	body,whd
dp	dp
sat	sat,nucl,dlink,tag
nucl	tag,sat,nucl,dlink
body	body,cmp,rhd,whd,mod
cmp	body,cmp,mod
crd	cnj,crd
cnj	cnj,crd
mwp	mwp
hdf	hdf,hd,obj1,mod,se

Table 3: Relations of sister nodes which are expected to occur with a node with the given relation.

- if there is a `sup`-node, there should be a `su`-node.
- if there is a `pobj1`-node, there should be a `obj1`- or `vc`-node.
- if there is a `sat`- or `tag`-node, there should be a `nucl`-node.
- a relative should be a modifier
- a comparative should be part of an `obcomp`-node
- nodes cannot have a single daughter
- head nodes cannot have daughters, except for multi-word-units
- nodes with a `cat` attribute should have daughters
- words such as `inclusief, uitgezonderd, voorbehouden, ...` are adjectives, not prepositions
- PPs should have a direct object
- category `smain` does not occur as a `body`.
- check that subjects of infinitives in auxiliary verb and control verb contexts are co-indexed with subject of finite verb
- check that modifiers are attached to main verb in auxiliary verb constructions
- check that multi-word-units are concatenative
- nodes with `cat=rel` cannot modify verbal projections
- in passive construction with a subject there is a coindexed direct object
- subjects without co-indexing do not occur in infinitival VPs
- auxiliaries and copula do not take a direct object
- VP complements should be `vc`, not `obj1`.
- a PP cannot be an obj1, unless it is a complement of another preposition
- check that the value of `postag` is legal
- if there is a subject, there should be a finite verb
- there cannot be a space in lemma, root and word values
- a finite verb should head one of `smain, ssub, sv1`
- the body of a `cp` should not be an infinite vp

- a node with category `ppart` should be headed by a participle
- a node with category `inf` should be headed by an infinitive
- a node with category `oti` should have a body with category `inf`
- a node with category `sv1, smain, ssub` should be headed by a finite verb
- the complementizer `om` heads a node with category `oti`
- the complementizer `te` heads a node with category `ti`
- the complementizers `dat, of` head a node with category `cp`
- a noun phrase with determiner `het, dit, dat` should have a head which is not a `zijd` noun
- a noun phrase with determiner `de, die, deze` should have a head which is not a singular `onz` noun
- a noun phrase with determiner `het, dit, dat, deze` should not have a plural head
- words that are `cmp` should have a complementizer part-of-speech
- words that are `rhd` should have a relative pronoun part-of-speech
- words that have relative pronoun part-of-speech should not be determiner
- `whq` nodes should contain a part-of-speech with `wh` flag
- personal pronouns should not be determiner
- relatives should be `rhd`
- determiners can be nouns only if genitive
- a determiner should be used as determiner
- identical sentences should have identical annotations

In addition, we have adapted the DECCA software to the LASSY Small treebank and corrected some further mistakes [1].

Finally, we have created various frequency lists (some of these are part of the Lassy Small release) which we have inspected for outliers (for instance, words that were annotated many times with a given part-of-speech and in a small number of times with another part-of-speech were manually checked, etc.

4 Limitations

Given the huge amounts of mistakes that have been corrected after the initial manual corrections, it is clear that more mistakes are present in the released version. In addition to obvious mistakes, there are also many cases in which more than a single analysis appears to be reasonable. In such cases we have aimed at a consistent approach in which the same analysis is chosen in similar cases, but we are aware that this objective has not been established in all cases.

We now list a number of issues in which we are aware of inconsistencies. The following issues are relevant for the assignment of part-of-speech labels.

- The distinction between the part-of-speech labels `SPEC(deeleigen)` and `N(eigen,...)` is unreliable
- The distinction between the part-of-speech labels `SPEC(symb)` and `TW(hoofd,vrij)` is unreliable
- The distinction between part-of-speech labels `SPEC(vreemd)` and `SPEC(eigen)` is unreliable
- For names `N(eigen,...)`, it is often arbitrary what value is assigned for gender.
- The part-of-speech assignment of words that are part of a multi-word-unit are unreliable

The following issues are relevant for the syntactic dependency annotations:

- Verbal modifiers with relation `mod` are almost always assigned to the main verb, even in cases where the modifier clearly modifies a higher verb (such as the modal `kunnen`). For verbal modifiers that receive the relation `predm` the situation is reversed. Predicate modifiers are almost always attached to the finite verb, even in cases where a lower attachment is more reasonable.
- For noun-phrases such as *een zak friet, een berg kleren, een verzameling boeken* it is often unclear to which word a following modifier should attach. A minimal pair is *een berg kleren in de voorkamer* versus *een berg kleren met vieze vlekken*. In most cases, however, the correct assignment is unclear or irrelevant (*een berg kleren van mijn dochter*).
- Lists. In some cases, lists are represented as a conjunction without a coordinator. In other cases, a `dp` category is assigned where each element receives the `du` role.
- Incomplete coordinations. In coordinations in which one conjunct contains material that appears to be elided in the other conjunct, this material is sometimes (but not always consistently) present in the analysis of both conjuncts using co-indexing.

- 'Plaatsonderwerp' `er`. In the Dutch linguistic tradition, some usages of the word `er` are called 'plaatsonderwerp'. In contrast and for practical reasons, these cases are all assigned a modifier role.

5 Tools

The tools that were used during the annotation process are all available free of charge.

The TadPole part-of-speech tagger and lemmatizer is developed at the Tilburg University, and is available free of charge under the Gnu General Public License, `http://ilk.uvt.nl/tadpole/`.

The Alpino parser, as well as the various tools to browse, search, and check dependency structures encoded in XML, is developed at the University of Groningen. Alpino is available free of charge under the Gnu Lesser General Public License.

For editing the syntactic annotations, we used TrEd. TrEd is developed at Charles University in Prague. TrEd is free software distributed under the Gnu General Public Licence. In order to use TrEd with Lassy files, there are a number of files which define the interface and various settings specific for Lassy. These files are part of the Alpino distribution.

References

[1] Jasper Hoenderken. Inconsistencies in dependency treebanks. Master's thesis, University of Groningen, 2009.

[2] Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. An efficient memory-based morphosyntactic tagger and parser for Dutch. In Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, and Frank van Eynde, editors, *Computational Linguistics in the Netherlands 2006. Selected papers from the seventeenth CLIN meeting*, LOT Occassional Series, pages 99–114. LOT Netherlands Graduate School of Linguistics, 2007.

[3] Frank van Eynde. Part of speech tagging en lemmatisering van het D-COI corpus, 2005.

[4] Gertjan van Noord. **A**t **L**ast **P**arsing **I**s **N**ow **O**perational. In *TALN 2006 Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues naturelles*, pages 20–42, Leuven, 2006.

[5] Gertjan van Noord, Ineke Schuurman, and Gosse Bouma. Lassy syntactische annotatie, revision 19053. http://www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf, 2010.