Annotation differences between CGN and the Alpino Treebank

Leonoor van der Beek, Gosse Bouma, Gertjan van Noord

November 6, 2002

The Alpino Treebank uses CGN Dependency Structures, as defined in the CGN annotation guidelines [1] and exemplified in the Leuven Yellow Pages [2]. However, the Alpino Treebank differs with respect to this annotation convention in a number of ways. The differences are listed here.

We make a distinction between minor and major differences. We regard a difference as minor if it appears to be possible to implement a filter which would transform one representation into the other automatically. Note that such a filter has not actually been implemented.

Major Differences

- In the Alpino Treebank obligatory control relations are represented explicitly. This affects the treatment of auxiliaries, modals, control verbs, passives, etc. Motivation: such control relations are not always predictable from lexical specifications.

- In the Alpino Treebank, modifiers in sentences with auxiliaries are generally annotated such that the modifier is attached to the main verb, not the finite auxiliary verb. Note that in other constructions, if the correct attachment is hard to determine, the modifier is attached high.

- In the Alpino Treebank, partitives are not treated special. Instead, in the Alpino Treebank a noun phrase such as `één van de drie` is analysed by taking `één` as the head, which is modified by a *PP*. Thus, the dependency relation *part* is not used in the Alpino Treebank.

- In the Alpino Treebank, all cases of *prt* are treated as *mod*. They are typically not grouped together, either. If they are grouped together, then a head-modifier structure results. Thus, the dependency relation *part* is not used in the Alpino Treebank.

- As in CGN, leaf nodes contain a part of speech label. However, the inventory of part of speech labels is much smaller in the Alpino Treebank. In addition, the annotation of part of speech labels contains many inconsistencies and should be regarded poor quality.

- Leaf nodes in the Alpino Treebank contain pointers into the string (this is equivalent to CGN), but in addition also contain a canonical form of the word(s) — typically the stem of the word. Note however, that the annotation of canonical word forms still contains many inconsistencies and should be regarded poor quality.

- We never use complex heads. If a word-group has the *hd* relation, then it must be a leaf.

- Sbar complements are never assigned the *obj1* relation; they get the *vc* relation. As a consequence, there is never a need to have multiple *obj1* relations for a given head.

- Idiomatic phrases are assigned the *svp* relation, as in CGN. However, unlike CGN idiomatic phrases are not analyzed. As a consequence, if a word-group has the *svp* relation, it must be a leaf. [1]

Minor differences

- Secondary edges are represented by means of co-indexing in the Alpino Treebank.

- In CGN, multi-word-units are represented by a flat tree (for each word a node), in the Alpino Treebank represented by a single node. NB. There are of course also non-minor differences w.r.t. the decision when something is regarded a multi-word-unit or not. As a consequence, in the Alpino Treebank you can **never** have discontinuous multi-word-units. [2]

- Root sentences introduced by words such as `want, en, maar` get a *dlink, nucl* representation in CGN. In the Alpino Treebank, these are treated as complementizers, so they receive *cmp, body* representation. The *dlink* label is not used in the Alpino Treebank.

- In CGN `te` in a *te-infinitive* is treated as a complementizer. In the Alpino Treebank, a *te-infinitive* is treated as a single (multi-word) unit; `te` is treated as inflection.

- In IPP constructions, we use the *ppart* category, whereas CGN uses *inf*.

[1] This was probably not a good decision.

[2] This seems right: discontinuous multi-word-units are weird.

- In CGN coordination without an explicit coordinator are *lists* where every conjunct gets the *lp* relation. In the Alpino Treebank coordinations with or without explicit coordinator are treated the same. In the Alpino Treebank, the *lp* relation is never used; and the *list* category is never used.
- In CGN interjections (hesitations, disfluencies) are represented as a node without a relation name (`--`). We do not represent these parts of the input at all.
- The category label *compp* is not used in the Alpino Treebank. Instead, we use category labels such as *advp, ap* for word groups that are *obcomp* with respectively an adverbial or adjectival head etc.

This list is incomplete.

References

[1] Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. CGN syntactische annotatie, March 2001. internal report Corpus Gesproken Nederlands.

[2] Bram Renmans and Ineke Schuurman. Yellow pages sa/cgn, June 2001. internal report Corpus Gesproken Nederlands.