Syntactic Annotation for the Spoken Dutch Corpus Project (CGN)

Heleen Hoekstra, Michael Moortgat, Ineke Schuurman, Ton van der Wouden

UiL-OTS, Utrecht University and CCL, KULeuven

Abstract

Of the ten million words of contemporary standard Dutch in the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN), a selection of one million words of natural spoken language will be annotated syntactically. In the present paper we discuss the tag sets and the annotation procedures that are currently being developed and tested. The annotation tags provide information about syntactic constituents and about the semantic relations (dependencies) between these constituents. The annotation graphs allow crossing branches, which makes it possible to represent dependencies independently of surface word order. Moreover, constituents can carry multiple dependency roles, a feature that is exploited in the annotation of non-local dependencies and ellipsis. The annotation process is carried out semi-automatically, using an interactive annotation environment developed within the NEGRA project, a syntactically annotated corpus of German newspaper texts. We illustrate the approach with some real life examples from the CGN corpus, focusing on how some typical spoken language phenomena are dealt with.

1 Introduction: about the CGN

The aim of the Spoken Dutch Corpus project (abbreviated as CGN, from the Dutch name *Corpus Gesproken Nederlands*) is to build an annotated corpus of about one thousand hours of continuous speech, which amounts to 10 million words. The project started in June 1998, and runs for five years. It is a collaborative effort of several Dutch and Flemish universities (Goedertier, Goddijn and Martens 2000, Oostdijk 2000a, Oostdijk 2000b).

The corpus is intended as a major resource both for linguistic research and for language and speech technology. To serve this dual purpose, it contains materials recorded in a variety of communicative settings: spontaneous face-to-face and telephone dialogues, interviews, discussions, debates, lectures, news broadcasts and book passages read aloud. Two-thirds of the material is collected in the Netherlands, one third in the Dutch speaking part of Belgium. Upon completion, the corpus will be the largest and most diverse database of spoken Dutch collected so far.

The project envisages different levels of annotation. The complete corpus is orthographically transcribed; also, every word receives a (contextually disambiguated) part-of-speech (POS) tag (Van Eynde, Zavrel and Daelemans 2000). In addition, broad phonetic transcription and syntactic annotation is provided for a representative selection of 10 percent of the data — the so-called core corpus. One quarter of the core corpus receives a prosodic annotation as well. In this paper, we focus on the syntactic annotation.

2 CGN syntactic annotation

The syntactic annotation structures to be stored with the CGN sound files and the other types of annotations are derived semi-automatically. This annotation process is described in the following sections. Note that the most important aim of the enterprise is the annotation structures rather than the parser used in deriving them.

2.1 Input: POS-tagged orthographic transcription

Input for the syntactic annotation is a POS-tagged orthographic transcription of the primary sound files. The material is segmented in annotation units. POS-tagging is done in a way comparable to the syntactic annotation, viz., semi-automatically: the output of an ensemble of automatic taggers, using some 400 different morphosyntactic tags and with an accuracy around 95%, is checked and corrected by hand. (For details of POS-Tagging and lemmatization within the CGN project we refer to (Van Eynde 2000, Van Eynde et al. 2000).) We give a real life example below in (1). For expository purposes, we have picked a short 14-word unit.[1]

(1) *Ik zal u gaan uitleggen hoe we dat zo'n beetje hebben*
I will you go explain how we that such-a bit have
aangepakt dat probleem .
approached that problem.
'I will explain to you how we more or less approached it, that problem'

The POS-tagged material has a rather straightforward line-oriented format shown below. The leftmost column has the complete sentence in a one word per line manner, the middle column contains the POS-information (main category in caps, features within brackets), the last column has the lexical lemma's.

```
 <au id=1 t=0.000 sp=N00052>
ik          VNW(pers,pron,nomin,vol,1,ev)          ik
zal         WW(pv,tgw,ev)                          zullen
u           VNW(pers,pron,nomin,vol,2b,getal)      u
gaan        WW(inf,vrij,zonder)                    gaan
uitleggen   WW(inf,vrij,zonder)                    uitleggen
hoe         BW()                                   hoe
we          VNW(pers,pron,nomin,red,1,mv)          we
dat         VNW(aanw,pron,stan,vol,3o,ev)          dat
zo'n        VNW(aanw,det,stan,prenom,zonder,agr)   zo'n
beetje      N(soort,ev,basis,onz,stan)             beetje
hebben      WW(pv,tgw,mv)                          hebben
aangepakt   WW(vd,vrij,zonder)                     aanpakken
dat         VNW(aanw,det,stan,prenom,zonder,evon)  dat
probleem    N(soort,ev,basis,onz,stan)             probleem
.           LET()                                  .
```

[1] Real life annotation units are anywhere between one and more than 150 words; in the data parsed so far, the average length of an annotation unit is around 15 words.

2.2 The CGN annotation graphs

The CGN syntactic annotation enriches the material with category information and dependency information. We call the annotation graphs *dependency structures*. At the input side, the annotation schemes should be maximally simple in order to minimize the work load involved in annotation and correction. At the output side, the CGN users should be offered annotation information that is maximally rich. We therefore opted for a theory neutral primary annotation level in terms of dependency structures (cf. also (Skut, Krenn and Uzkoreit 1997)). This primary annotation can be enriched with information from the POS tagging and from the CGN lexicon. The combination of these sources of information yields a number of output formats tailored to the wishes of various user groups. The dependency structures used in the CGN syntactic annotation are developing into a *de facto* standard for the computational analysis of Dutch: cf. (Bouma, van Noord and Malouf 2001).[2]

Formally, a CGN dependency structure $D = \langle V, E \rangle$ is a labeled directed acyclic graph. Node labels V and edge labels E are taken from disjunct sets CAT and DEP, respectively.

- Nodes: $\mathsf{CAT} = \mathsf{POSCAT} \cup \mathsf{PHCAT}$: category labels ($c$-labels), the union of lexical (POS) and phrasal labels.
- Edges: DEP: dependency labels (d-labels).

We distinguish between atomic and composite dependency structures. Atomic dependency structures are simply nodes decorated with a c-label from POSCAT. They are the leaves of our annotation graphs. The label set POSCAT is a reduced version of the full set of CGN part-of-speech labels, which contains more than 300 tags. We condense this to a POSCAT set of some 50 labels, retaining distinctions that are relevant for the syntactic annotation procedure.[3] The set of POSCAT labels currently used is given in the Appendix.

The basic building blocks of composite dependency structures we call *local dependency domains*. The mother node of such a domain carries a phrasal label from PHCAT; the daughters have c-labels from CAT. The d-labels for the mother-daughter edges consist of a *head*, together with the *complements* and the *modifiers* of that head.

Head. The head of a dependency domain projects the c-label of the mother node.

Complements. The complementation pattern determines the interpretation of the head in terms of thematic structure. A complement label occurs at most once in a local dependency domain.

[2] Syntactic details of the annotation are spelled out in (Moortgat, Schuurman and van der Wouden 2001).

[3] As the reader will see later, the original POS-labels are preserved in the 'morphology' field of the annotation.

Modifiers. Modifying elements do not change the c-label of the mother node; they can be left out without affecting the thematic structure. There can be multiple occurrences of a given modifier label within a local dependency domain.

The tags sets PHCAT and DEP currently used are given in full in the Appendix. As the reader will notice, the terminology follows traditional grammatical practice rather closely. To keep the DEP set small, we use some overloading. The tag OBJ1, for example, labels the 'direct object' of transitive verbs, but also the 'first complement' of prepositions and adjectival heads.

The tag sets make provision for phenomena that are typical for spoken language: the c-label DU ('discourse unit'), for example, makes it possible to categorize asyndetic constructions; dependency articulation within DU is given in terms of d-labels such as NUCL (nucleus) versus SAT (satellite), TAG, or DLINK (discourse link).

We now draw the reader's attention to some properties of the CGN annotation that follow from the two-dimensional analysis (form/categorial information versus function/dependency information).

Shallow annotation structures. Taking together complementation and modification within one and the same local dependency domain yields flat annotation structures. Specifically,

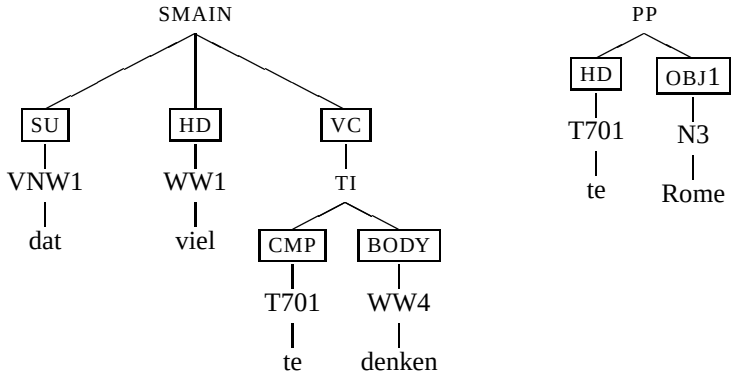
- a new local domain (hierarchical level) will only be opened if there is a new head;
- complementation and modification are *relations* between phrases and a head: if there are no complements or modifiers, there is no reason for (non-branching) projections.

The CGN treatment of verbal projections is a good illustration of this shallow approach. Following the custom in Dutch traditional grammar and elsewhere, we distinguish level between *finite* and *non-finite* verbal projections at the c-label. The inflected verb (=POS tag) is head of the finite clausal types; the infinitive or participle (=POS tag) is head of the non-finite ones. In finite clauses, there is no need then for an intermediate VP level.[4]

Lexical anchoring. A point related to the above is that, in the unmarked case, local dependency domains are lexically rooted: the c-label of the head is a leaf label from POSCAT. As the head projects the c-label of the mother node, we can use the d-label of the head to disambiguate in cases where the information we get from the POS-annotation is underdetermined. For example, in the POS annotation, no distinction is made between *te* as a preposition, i.e. head of a preposition phrase

[4] As a reviewer correctly observes, one would not expect a distinct VP level in a dependency structure either.

(PP), on the one hand, and as head of a non-finite verbal projection, the *te*-infinitive (TI):[5] both are labeled T701, which is an abbreviation for VZ(INIT), i.e., preposition. The syntactic annotation disambiguates the word *te* by means of the *d*-label: the head of the TI is labeled CMP (for complementizer).[6]



Crossing and multiple dependencies. We stress once again that the CGN annotation is a graph, not a tree. Graphs with crossing branches are used to annotate dependency relations that are at odds with surface word order and/or constituency. By assigning constituents multiple dependency roles, we can annotate non-local dependencies as they are found in e.g. relative clauses and constituent questions.[7]

- On the one hand, the elements introducing these kinds of configurations (constituents containing a WH-element or relative pronoun) determine the *c*-label of the mother node; therefore, they are dependency heads.
- On the other hand, we also want to be able to indicate the role these elements play in the rest of the clause; the relevant local dependency domain may be embedded arbitrarily deep.

Examples of crossed and multiple dependencies are given in the next section.

2.3 The annotation process

Our goal is to syntactically annotate a (balanced, representative) sub-corpus of one million words. In order to yield a maximally consistent result in the time allotted, the task is carried out (semi-)automatically. We use the interactive ANNOTATE tool, which was developed in Saarbrücken in the context of the

[5] The jury is still out on the exact status of *te*, but we have chosen to call it a complementizer.

[6] *c*-labels are given in SMALL CAPS, *d*-labels are boxed. The *d*-labels decorate the *arcs* of the annotation graph: they are not *nodes*. POS information is taken directly from the POS annotation and will not always be spelled out.

[7] Specification of other types of non-local dependencies such as the resolution of pronouns and the interpretation of control structures is postponed to a later phase in the project.

NEGRA project (cf. (Plaehn 1998), and see www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html). The functionality of this tool is very well adapted to the two-dimensional annotation philosophy we have adopted in the CGN project:

> Annotate is a tool for the efficient semi-automatic annotation of corpus data. It facilitates the generation of context-free structures and additionally allows crossing edges. Functions for the manipulation of such structures are provided. Terminal nodes, non-terminal nodes, and edges are labeled. In the NEGRA project, these labels are used for parts-of-speech and morphology (terminal nodes), phrase categories (non-terminal nodes), and grammatical functions (edges). Type and number of labels are defined by the user. Annotated corpora are stored in an SQL database. Annotate has a specified interface for communication with external taggers and parsers. (Plaehn 1998)

For an illustration, we return to our real life example (1). With the tag sets given in the Appendix, ANNOTATE allows us to produce the graph of Figure 1 (on the next page).
The example illustrates some salient features of the annotation we have discussed in the previous section.

- The question word *hoe* ('how') has two parent nodes: it is the head of the subordinate interrogative WHSUB, and at the same time it plays the role of modifier within the PPART (past participle phrase) embedded in the body of that interrogative. The two edge labels WHD and MOD connecting the question word to the parent nodes WHSUB and PPART respectively, encode this double dependency role.[8]

- The dependency articulation is independent of surface order and constituency: the temporal auxiliary verb *hebben* ('have') selects the past participle phrase PPART as a complement, but it occurs within that phrase (between the direct object and the participle head) in surface order, leading to crossing dependecies in the annotation graph.

- Phenomena such as "right dislocation" are not seen as part of clausal syntax proper, but rather as belonging to discourse. The discourse coherence between the "main clause" and the "dislocated constituent" (in this case the noun phrase *dat probleem* 'that problem') is expressed by grouping these consituents under the label DU (for Discourse Unit) where they are assigned the dependency roles of NUCL and SAT respectively. If, in a later phase of the annotation process, anaphoric relations are going to be marked as well, a link may be made between the pronominal element *dat* in the nucleus component, and the satellite full noun phrase *dat probleem*.

[8] Nothing precludes, in principle, the possibility of an element having more than two dependency roles.

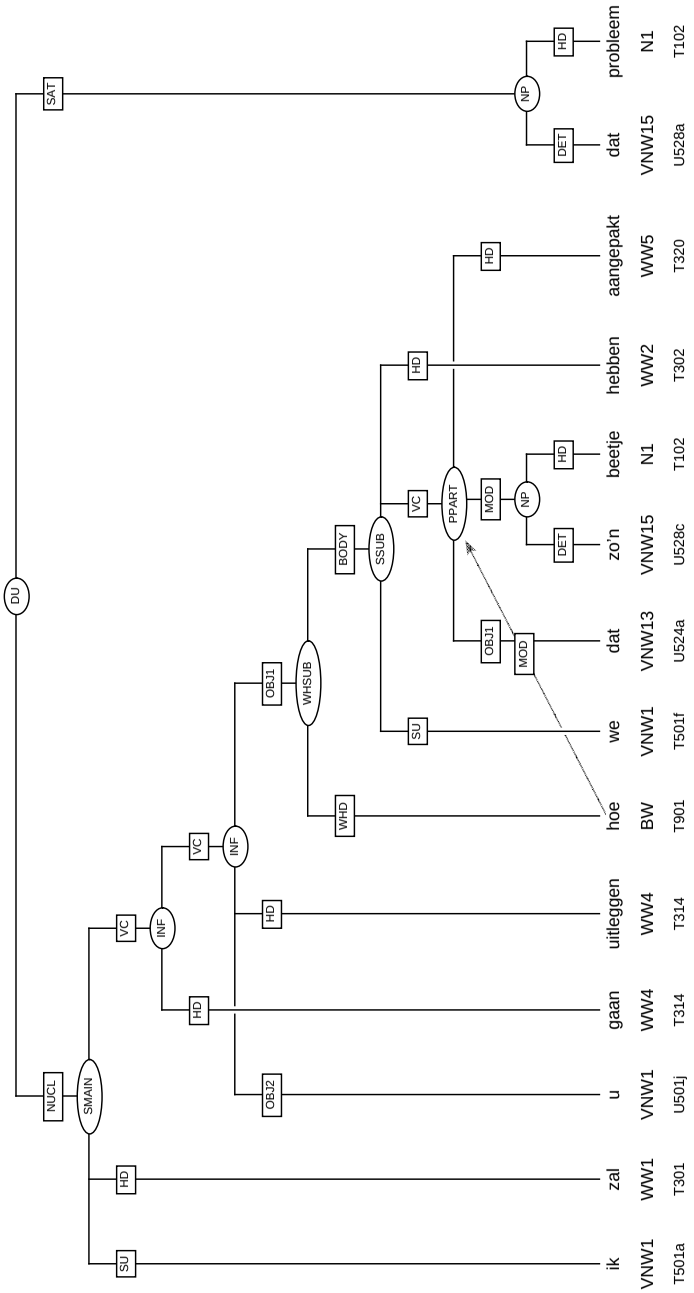


Figure 1: Example annotation

The ANNOTATE tools are designed to work together with parsers supporting the manual annotation and running in the background via a defined interface. In this phase of the project, we work with Thorsten Brant's (Brants 1999) Cascaded Markov Models (CMMs) approach which supports learning on the basis of an existing annotated corpus (a tree bank). The CMM approach implements a bootstrapping strategy: starting off with a small corpus, using the hypotheses of the parser to gain speed and quality in manually annotating the next part, add this part to the corpus and let the program refine its hypotheses, and so forth. In later phases of the project, the CMM approach will be used in combination with other parsers, so that we can integrate the rich information of the CGN lexicon with the statistical approach.

Currently, the parser is parsing the first subset of the CGN. The Spring 2001 CGN release already contained annotations for some 50.000 words, two thirds from the Netherlands, one third from Belgium, all checked by hand.

Given the fact, however, that even the most successful statistical POS-tagging algorithms reach only accuracy rates between 96 and 97% for new, unseen texts (Brants 1999), it may not be expected that the output of CMM parsing, which probably is a much more difficult task than mere POS-tagging, will be completely trustworthy. Therefore, all output will be checked and, whenever necessary, corrected by humans, who will often return to the sound signal for disambiguation clues. Tools have been developed to check for inconsistencies in the corrected output, which will again be fed into the parser's database in order to further increase the quality.

2.4 Customized export formats

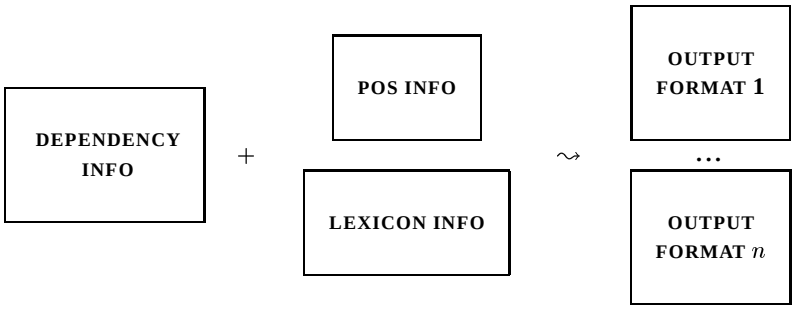
The ANNOTATE environment has a line-oriented export format, which makes it possible to interface with other applications. Our running example is exported as follows:

```
%% word     tag     morph   edge   parent   secedge   secparent
#BOS 41 6 984562209 0
ik          VNW1    T501a   SU     507
zal         WW1     T301    HD     507
u           VNW1    U501j   OBJ2   505
gaan        WW4     T314    HD     506
uitleggen   WW4     T314    HD     505
hoe         BW      T901    WHD    504      MOD       502
we          VNW1    T501f   SU     503
dat         VNW13   U524a   OBJ1   502
zo'n        VNW15   U528c   DET    500
beetje      N1      T102    HD     500
hebben      WW2     T302    HD     503
aangepakt   WW5     T320    HD     502
dat         VNW15   U528a   DET    501
probleem    N1      T102    HD     501
.           LET     T007    --     0
#500        NP      --      MOD    502
#501        NP      --      SAT    508
#502        PPART   --      VC     503
#503        SSUB    --      BODY   504
#504        WHSUB   --      OBJ1   505
#505        INF     --      VC     506
#506        INF     --      VC     507
#507        SMAIN   --      NUCL   508
#508        DU      --      --     0
#EOS        41
```

We stress that this output is fully equivalent (modulo the interpunction, which is not a genuine part of the of the transcription *per se*, but rather an artifact of it) to the annotation graph in (1). Some clarification, however, may be appropriate. The structure of this table is as follows: each line describes one element in the data structure. The first line, for example, tells us that there is an end node valued *ik* with POS-label VNW1 (a personal pronoun) and morphological information T501 (which is a shorthand for the original POS-tag, cf. above) which fulfills the SU (subject) role with respect to some mother node 507. In one of the last lines we see that this node 507 itself is of category SMAIN (main clause), that it has no morphological information (what could it possibly be?) which functions as the NUCL (nucleus) of a DU (discourse unit).

Another element, *hoe* in line 8, has a POS-label BW (adverb) and functions as WHD (head of an interrogative clause) whose mother node carries the label 504; additionally (this information is in the final columns), it functions as a MOD (modifier) in the structure headed by node 502.

The primary annotation can be enriched with information from the POS tagging and from the CGN lexicon. The combination of these three information sources can lead towards a number of customized output formats for various user groups:



As regards the derived output formats mentioned, one may think of

- enriching category labels (c-labels) with morphosyntactic feature information;
- enriching dependency labels (d-labels) with 'deep' dependencies (e.g.: semantic control information);
- present surface constituent trees in a user friendly notation (with or without 'empty elements' etc.);
- presentation matters: choices as regards the 'language' of the label sets (Dutch, English, ...) and of the output (HTML, MSWORD, LaTeX, Postscript, XML, ...).
- (Moortgat and Moot 2001) discuss the feasability of deriving type-logical grammatical representations from the CGN output;
- etc.

Of course, the possibility of realizing such customized output formats depends heavily on other annotation levels, such as lemmatization (i.e. the linking up of all words with a rich lexicon), POS-tagging and prosodic annotation on the one hand, and the planned CGN exploitation software module on the other. In this paper, however, we have concentrated on the primary dependency annotation.

3 Concluding remarks

In this paper, we have given an overview of the CGN approach to syntactic annotation of a subset of the 10 million word spoken Dutch corpus. We stress once again that the goal of the annotation enterprise is not to build a maximally efficient, ninety odd percent trustworthy real time annotating tool, but rather to produce a ninety nine odd percent trustworthy output which is maximally useful for the maximal number of users. Given the current state of affairs in computational linguistics, this is only possible by means of human intervention. The output is a

set of 'theory neutral' dependency trees, which are input to other modules producing data structures useful for users from various theoretical backgrounds and with various practical aims.[9]

References

Bouma, G., van Noord, G. and Malouf, R.(2001), Alpino: Wide-coverage computational analysis of Dutch. available via http://odur.let.rug.nl/alfa/papers/papers/.

Brants, T.(1999), Cascaded Markov Models. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics EACL-99*, Bergen, Norway, 1999.

Goedertier, W., Goddijn, S. and Martens, J.-P.(2000), Orthographic transcription of the Spoken Dutch Corpus. Proceedings LREC 2000.

Moortgat, M. and Moot, R.(2001), CGN to Grail. extracting a type-logical lexicon from the CGN annotation, this volume.

Moortgat, M., Schuurman, I. and van der Wouden, T.(2001), Syntactische annotatie. Internal working document CGN, Utrecht, May 2001.

Oostdijk, N.(2000a), Building a corpus of spoken Dutch, *in* P. Monachesi (ed.), *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, Utrecht University, Utrecht Institute of Linguistics OTS, Utrecht, pp. 147–157.

Oostdijk, N.(2000b), The Spoken Dutch Corpus. Overview and first evaluation. Proceedings LREC 2000.

Plaehn, O.(1998), Annotate: Bedienungsanleitung. Document Projekt C3 Nebenläufige Grammatische Verarbeitung. Universität des Saarlandes, FR 8.7 Computerlinguistik.

Skut, W., Krenn, B. and Uzkoreit, H.(1997), An annotation scheme for free word order languages, *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C. available via http://arxiv.org/format/cmp-lg/9702004.

Van Eynde, F.(2000), Part of speech tagging en lemmatisering. Internal working document CGN, Centrum voor Computerlinguïstiek K.U. Leuven, May 2000.

Van Eynde, F., Zavrel, J. and Daelemans, W.(2000), Lemmatisation and morphosyntactic annotation for the spoken dutch corpus, *in* P. Monachesi (ed.), *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, Utrecht University, Utrecht Institute of Linguistics OTS, Utrecht, pp. 53–62.

[9] See the CGN web site at http://lands.let.kun.nl/cgn/ehome.htm for general information on the CGN corpus and how to receive the current corpus materials. A free viewer for ANNOTATE analyses can be downloaded from the Utrecht CGN site http://cgn.let.uu.nl/.

Appendix: The tag sets

POS-labels

The next table contains all part of speech labels: presently, four types of nouns are distinguished (common nouns and proper nouns, singular and plural), twelve types of adjectives (differentiating between attributive and predicative usage, and various morphological variants, including comparative and superlative forms), six (morphologically defined) verb forms, ordinals and cardinals, and many types of pronouns (morphologically defined). The first column shows the abbreviation we use, the second column gives a more verbose variant, the third column offers some information in English.

POS-labels

N1	N(soort,ev)	common noun, singular
N2	N(soort,mv)	common noun, plural
N3	N(eigen,ev)	proper noun, singular
N4	N(eigen,mv)	common noun, singular
ADJ1	ADJ(prenom,basis)	prenominal adjective, base form
ADJ2	ADJ(prenom,comp)	prenominal adjective, comparative form
ADJ3	ADJ(prenom,sup)	prenominal adjective, superlative form
ADJ4	ADJ(nom,basis)	nominalized adjective, base form
ADJ5	ADJ(nom,comp)	nominalized adjective, comparative
ADJ6	ADJ(nom,sup)	nominalized adjective, superlative
ADJ7	ADJ(postnom,basis)	postnominal adjective, base form
ADJ8	ADJ(postnom,comp)	postnominal adjective, comparative form
ADJ9	ADJ(vrij,basis)	adjective used predicatively, base form
ADJ10	ADJ(vrij,comp)	
ADJ11	ADJ(vrij,sup)	
ADJ12	ADJ(vrij,dim)	adjective, diminutive form
WW1	WW(pv,ev)	inflected verb form, singular
WW2	WW(pv,mv)	inflected verb form, plural
WW3	WW(pv,met-t)	inflected verb form with *-t*
WW4	WW(inf)	infinitive
WW5	WW(vd)	past participle
WW6	WW(od)	present participle
TW1	TW(hoofd)	ordinal number
TW2	TW(rang)	cardinal number
VNW1	VNW(pers,pron)	personal pronoun
VNW2	VNW(pr,pron)	
VNW3	VNW(refl,pron)	reflexive pronoun
VNW4	VNW(recip,pron)	reciprocal pronoun
VNW5	VNW(bez,det)	possessive pronoun
VNW6	VNW(vrag,pron)	question word
VNW7	VNW(betr,pron)	relative pronoun
VNW8	VNW(vb,pron)	
VNW9	VNW(vb,adv-pron)	
VNW10	VNW(excl,pron)	exclamative pronoun
VNW11	VNW(vb,det)	
VNW12	VNW(excl,det)	
VNW13	VNW(aanw,pron)	demonstrative pronoun

VNW14	VNW(aanw,adv-pron)	
VNW15	VNW(aanw,det)	
VNW16	VNW(onbep,pron)	indefinite pronoun
VNW17	VNW(onbep,adv-pron)	
VNW18	VNW(onbep,det)	
VNW19	VNW(onbep,grad)	
LID	LID	determiner
VZ	VZ	preposition
VG1	VG(neven)	coordinating element
VG2	VG(onder)	subordinating element
BW	BW	adverb
TSW	TSW	interjection
SPEC	SPEC	rest category
LET	LET	interpunction

Category labels

The next table contains all category labels currently in use. The first column shows the label, the second one an explanation in Dutch, with one in English below it.

Node labels (category information)

SMAIN	declaratieve zin (V2)
	main clause (V2)
SSUB	bijzin (V-finaal)
	subordinate clause (verb-final)
SV1	zin met V op de eerste plaats
	any sentence with a sentence-initial inflected verb
INF	kale-infinitiefgroep
	short infinitive group
PPART	voltooid/passief-deelwoordgroep
	past/passive participle group
PPRES	tegenwoordig-deelwoordgroep
	present participle group
CP	zinsdeel ingeleid door onderschikkend vw. of vw./vz. v. verg.
	clause headed by any kind of complementizer
MWU	merged-word-unit (‘drie en twintig’, ‘Jan van den Berg’)
	merged-word-unit (used for complex numbers and names)
TI	te-infinitiefgroep
	long infinitive group
OTI	om-te-infinitiefgroep
	long infinitive group headed by *om* ($\sim$ *for to*)
AHI	aan-het-infinitiefgroep
	long infinitive group headed by *aan het* (‘a Dutch progressive form’)
ADVP	bijwoordgroep (alleen voor echte bijwoorden)
	adverbial phrases
DETP	determinatorgroep (‘bijna alle’ in ‘bijna alle boeken’)
	determiner group
AP	adjectiefgroep (ook voor adverbiaal gebruikte adjectieven)
	adjectival group

PP	prepositiegroep prepositional group
NP	nominale groep nominal group
SVAN	van-zin (complement in directe rede) subordinate clause headed by *van*
REL	relatiefzin relative clause
WHREL	hoofdloze relatiefzin headless relative
WHQ	constituentvraag: hoofdzin WH-question, V2
WHSUB	constituentvraag: bijzin embedded WH-question
CONJ	conjunctie conjunction
DU	discourse-unit (asyndetische constructie) discourse-unit
LIST	asyndetische conjunctie asyndetic conjunction
COMPP	zinsdeel met 'meer' of 'even' als hoofd en CP als complement various comparative constructions

Edge labels

The next table contains all edge labels currently in use. The first column shows the label, the second one an explanation in Dutch, with one in English below it.

Edge labels (dependency information)

HD	hoofd head
HDF	staart (scheidbaar deel) van circumpositie second part of a circumposition (*tot hier **toe***
DET	determinator determiner
PART	partitief partitive
SU	subject, onderwerp subject
SUP	voorlopig subject provisional subject
OBJ1	direct object van V, (eerste) complement van P, A, N direct or first object
POBJ1	voorlopig OBJ1 provisional direct or first object
OBJ2	secundair object (IO, EO, BO) secondary object
SE	verplicht reflexief object obligatory reflexive object

SVP	scheidbaar deel van werkwoord verbal particle
PREDC	predicatief complement predicative complement
PC	voorzetselvoorwerp prepositional complement
VC	verbaal complement, beknopte bijzin verbal complement
LD	locatief of directioneel complement locational or directional complement
ME	maat(/duur/gewicht)-complement measure complement
CMP	complementeerder/hoofd van CP, SVAN, TI, OTI of AHI grammatical complementizer
RHD	complementeerder/hoofd van (hoofdloze) relatiefzin complementizer heading (headless) relative
WHD	complementeerder/hoofd van WHQ of WHSUB complementizer heading WH question
BODY	romp van CP, SVAN, TI, OTI, AHI, REL, WHQ of WHSUB body of subordinate clause
PREDM	bepaling v. gesteldheid 'tijdens de handeling' secondary predicate
MOD	algemeen label voor bepaling/modificeerder modifier
CRD	nevenschikker coordinator
CNJ	lid van nevenschikking member of conjunction
NUCL	kernzin (in DU) nuclear clause
SAT	satelliet: aan- of uitloop (in DU) met binding in NUCL satelite
TAG	aanhangsel, voor- of tussenvoegsel tag
DP	elk der delen van een DU any part of a DU
PRT	elk der delen van een partikelgroep any part of a particle group
OBCOMP	vergelijkingscomplement (compl. van 'meer'/'even') comparative complement
APPOS	bijstelling apposition
LP	elk der delen van een LIST any part of a LIST
DLINK	"en", "maar", "want" o.i.d. aan het begin van een uiting discourse particles joining discourse fragments
MWP	elk der delen van een MWU any part of a MWU