ANNO: a Multi-functional Flemish Text Corpus

Ineke Schuurman*

Abstract

In this paper the ANNO Project ("Een Geannoteerde Publieke Gegevensbank voor het Geschreven Nederlands/An Annotated Database for Written Dutch") is reported on[1]. The project aims at laying the foundations for the compilation and linguistic annotation of a large multi-functional Flemish text corpus. The corpus available now consists of language written to be spoken, together with transcribed interviews.
In this paper we present the levels of annotation ANNO comes with at the moment. In general, we will show what can be achieved using taggers, parsers etc. that are currently available for Dutch. A separate issue is whether the tools are as useful for Flemish as they are for Dutch.

Introduction

The ANNO Project is sponsored by the Flemish Research Initiative in Speech and Language Technology. It is a pilot project, aiming at laying the foundations for the compilation and linguistic annotation of a large, multi-functional, standard Flemish text corpus.
Although great efforts have been made in creating machine-readable corpora for English and other major languages, this is only to a lesser degree the case for Dutch. To some extent this is understandable: the market for English NLP products is much larger than that for Dutch NLP products. On the other hand, to safeguard the position of languages like Dutch, Danish etc. both inside the European Union and beyond it is important to develop tools for the automatic processing of these languages as well: taggers, parsers, speech interfaces, etc. Otherwise, these languages are in danger of being pushed aside in our digitized society. For such reasons national governments, the European Union and other bodies promote the development of tools and resources for minor languages as well. As annotated corpora provide an excellent basis for developing NLP tools, corpora of reasonable size should be created for languages like Dutch as well, cf. also Kruyt (1995).

*Centrum voor Computerlinguïstiek, K.U. Leuven

[1] One way or another the following people were also involved in ANNO: Joyce de Booy, Frank van Eynde, Wim Peters and Bruno Tersago.

1 Two variants of standard Dutch

According to the constitution the official language in Flanders is Dutch, just as it is in the Netherlands. So why should there be a corpus of Flemish[2]? Is standard Belgian Dutch different from standard "Dutch" Dutch? Yes, assuming that the language used on radio and television reflects the standard language[3].
Although many speakers of Dutch and Flemish are unaware of this it turns out that there are differences at many levels: phonology, morphology, syntax, semantics, pragmatics).
Some examples:

- Voicing of syllable-initial fricatives

- Stress patterns

- Other past tenses (Flemish *zegden* – Dutch *zeiden* (**said**)) and plurals (Flemish *leraars* – Dutch *leraren* (**teachers**))

- Gender. In Flemish there are three genders (*masculine*, *feminine* and *neuter*), in Dutch only two genders are left (*neuter* and *non-neuter*)

- The behaviour of separable verbs. In Flemish the separable affix often remains with the verb also in cases where this would be 'ungrammatical' for speakers of Dutch, cf. Hoekstra (1987, 35):

  (1) Hij aanhoorde het vonnis onbewogen (Fl)

  (2) Hij hoorde het vonnis onbewogen aan (D and Fl)

  'He listened to the sentence without emotion'

- The occurrence of Verb Projection Raising in Flemish:

  (3) ..., omdat zij wil een appel eten (Fl)

  (4) ..., omdat zij een appel wil eten (D and Fl)

  'because she wants to eat an apple'

- The choice of the auxiliary of the perfect. For a range of verbs in Flemish the choice of the auxiliary of the perfect depends on the main verb:

  (5) Hij heeft haar komen afhalen (Fl)

  (6) Hij is haar komen afhalen (D and Fl)

  'He came to fetch her'

[2] In this paper the notion *Flemish* will be used to refer to standard Belgian Dutch.
[3] See also Hoekstra (1987)

There are also reasons to believe that the distribution of the present perfect and the imperfect past to express that something happened before the moment of speech is not the same in both variants of Dutch (temporal semantics), whereas the same holds for the choice of the personal pronoun *jullie* or *je* vs *u* (**you** pl and sg). And of course there are the differences with respect to the vocabulary.

A sufficiently large corpus of Flemish, especially when contrasted with the same kind of corpus for Dutch, will also tell us more about these and other particularities of the language used.

It will be clear that, although in general both variants have the same properties, there is a whole number of phenomena which are 'out' in one of the variants of Dutch whereas they are perfect in the other variant. Take the role of gender: in Flemish one should use the genders correctly, one should for example refer to a bus with *zij* as it is a feminine noun. In Dutch people will not be aware of its feminine genus, therefore it often will be referred to as *hij*.

Thus far corpus linguistics didn't pay much attention to the variant used in Belgium.

No corpus of reasonable size at all was available in machine-readable format. The only completely Flemish, i.e. standard Belgian Dutch, corpus we are aware of is the one collected by Willy Martin (Martin (1967), cf. also Dutilh-Ruitenberg (1992)).

2 The objective of the project

The objective of the ANNO Project was twofold:

- the inventory of corpora, taggers, parsers, etc. that are available, especially for Dutch and Flemish;
- the compilation of a multi-functional database for Flemish, containing a corpus with a series of annotation schemes representing various levels of linguistic analysis

With respect to the second task: at this moment texts are annotated for their part-of-speech, morphological, syntactic and phonological information, and discourse information.

The tools to be used are preferably freely available for research purposes and have a good performance: correction of output is very time-consuming.

Another initial requirement was platform independence, i.e. the ANNO database should be usable in both DOS and UNIX environments[4].

3 Inventory

Our inventory, cf. the first objective (reported on in Peters and Tersago (1996)), showed that there are quite a number of corpora for Dutch, and the same holds for

[4] During the project we learned about JAVA, therefore the new objective is to make ANNO available on the Web.

tools to treat them. But, as we expected, there was almost nothing available for Flemish.

Peters and Tersago (1996) contains chapters (in Dutch) on the design and compilation of corpora, on annotations, existing corpora, tools and recent initiatives. Several of these are made available on the Web[5].

The outcome of the inventory also to a large extent determined the choice of our tools.

4 Corpus

4.1 Composition of the corpus

As is clear from the full project title "Een Geannoteerde Publieke Gegevensbank voor het Geschreven Nederlands", ANNO[6] is an annotated corpus for *written* Dutch. Still the texts it contains are transcriptions of radio news and current affairs broadcasts, i.e. *spoken* language[7].

More specifically, ANNO contains texts

- with a wide circulation,
- intended for a broad population,
- treating non-specialist topics, and
- as recent as possible (Kruyt and Putter (1992), Martin, Platteau, and Heymans (1985))

The text material the ANNO corpus consists of has been derived from BRTN (Belgian Radio and Television) radio news broadcasts and the current affairs programme Actueel[8]: language written to be spoken together with transcribed interviews. The latter contain spontaneous speech.

4.2 Some obstacles

The BRTN-texts are not available in electronic format, so we had to scan several thousands of sheets of paper as every item is written on a separate sheet. A very time-consuming job by which also a considerable amount of structural (scanning) errors is introduced. These were corrected in a semi-automatic way.
The texts we received were not meant to be made public: the texts contain many

[5] http://www.ccl.kuleuven.ac.be/about/ANNO/inleiding.html.

[6] In what follows the notion ANNO is used to refer to the whole project as well as to the corpus and/or the resulting database.

[7] A database of spoken Flemish as such is taken care of by another project within the Flemish programme for speech- and language processing, FONILEX.

[8] News: 21 - 26 March 1995, 17 - 30 April 1995, 1 - 30 May 1995 and 12 - 30 June 1995 , always the 08.00, 13.00, 18.00 and 24.00 broadcasts; Actueel: 20 - 29 March 1995, 1 - 31 July 1995, 1 - 31 August 1995, the 13.00 and 18.00 broadcasts (no broadcasts on Sundays and on holidays). A quite similar corpus for Dutch is described in Sterkenburg (1989).

typing errors and the spelling is very inconsequent (both preferred and alternative spelling within one item, many inaccuracies, even the names of the reporters themselves are written in three, four ways). Whenever the spelling didn't influence pronunciation we normalized the texts (preferred spelling) in order to simplify consultation of the corpus by future users[9].

07mei13u: binenland ⟶ binnenland

However, sometimes a word was 'misspelled' deliberately as a pronunciation help for the newsreader: *biezonder*, *honderste* and *Andaloesisch* instead of *bijzonder*, *honderdste* and *Andalusisch*. Such 'mistakes' are preserved as the newsreaders apparently tried to avoid a spelling pronunciation of these words: their pronunciation had to sound natural.
Abbreviations are always spelled out, as they will be used in their full form in the broadcasts themselves:

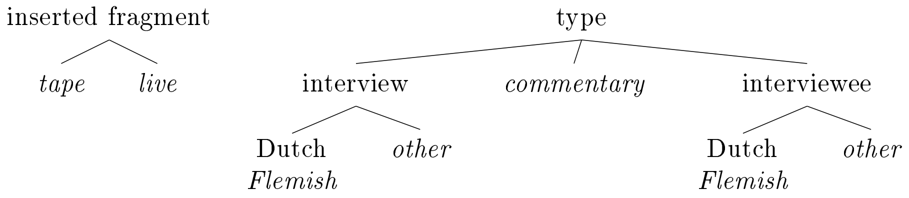
m/s ⟶ meter per seconde

One may question our approach with respect to these phenomena: we tried to come as close as possible to what was actually said (and how it was said), although we didn't have the tapes. Of course the original texts (without interventions from us) will be made available as well, whereas all our interventions (or lack of interventions, cf. the pronunciation help) will be motivated in the documentation that comes with the database. And all interventions are recorded in SGML-annotations.

The lack of tapes also complicated the encoding of the corpus in SGML as it was not always clear whether a paragraph belonged to an interview or was part of the text of the newsreader.

4.3 Typical properties of the texts involved

Typical for the Flemish news broadcasts as they are incorporated in our corpus is that they are read by two newsreaders and that they contain (live) interviews and commentaries. These inserted news fragments can be in a foreign language. Some of the inserted fragments are live, others are taped. Within both types *interviews* show an interaction between interviewer and interviewee, *commentaries* often contain prepared speech whereas speech fragments containing only statements by *interviewees* are often spontaneous.



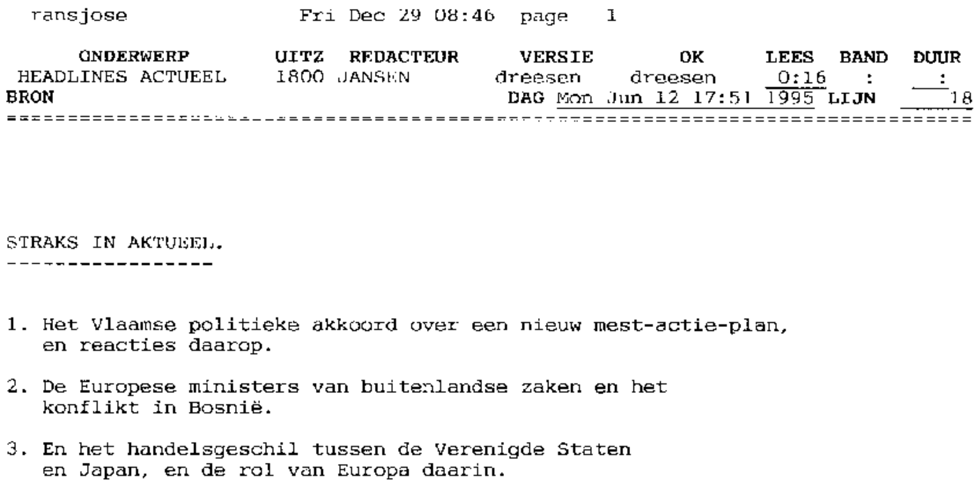
[9] When in doubt nothing was changed. Note that we couldn't consult the tapes as the BRTN did not want us to have them.

The same distinctions hold for Actueel, be it that the items are much longer and have a larger share of spontaneous speech.

Next to what you hear when listening to the radio, the corpus also contains a considerable amount of text-not-to-be-read-aloud: directions for the newsreader, administration, etc., see the following fragment (LW means "last words of the tape").

> 01mei08u: ..., maar als je wil kampioen worden dan moet je dat gewoon presteren, drie keer winnen.
>
> LW gewoon presteren, drie keer winnen.

Another example is the header in Figure 1 .



```
  ransjose              Fri Dec 29 08:46   page    1

      ONDERWERP         UITZ   REDACTEUR      VERSIE        OK       LEES   BAND   DUUR
 HEADLINES ACTUEEL      1800  JANSEN        dreesen     dreesen      0:16     :      :
BRON                                        DAG Mon Jun 12 17:51 1995 LIJN          18
===================================================================================




STRAKS IN AKTUEEL.
------------------


1. Het Vlaamse politieke akkoord over een nieuw mest-actie-plan,
   en reacties daarop.

2. De Europese ministers van buitenlandse zaken en het
   konflikt in Bosnië.

3. En het handelsgeschil tussen de Verenigde Staten
   en Japan, en de rol van Europa daarin.
```

Figure 1: Part of an original text: 12 June 1995, 18h: the headlines of Actueel

5 Annotation

In this section the various types of annotation will be discussed. Often tools require their input to be in a well-defined format (without accents, without ASCII-codes, etc.), each tool having its own desiderata. Several small AWK-programmes had to be written to convert the corpus into the desired formats.

5.1 Standard Generalized Markup Language

By means of SGML-codes all information in the corpus is captured unambiguously, cf. Sperberg-McQueen and Burnard (1994), Ide and Véronis (1994). When scanning texts and/or transferring the corpus to another platform the lay-out of the texts may change. The SGML-codes will tell you exactly how the original texts looked like. In the following example part of the news broadcast of 21mei08u is reproduced without and with codes. In this case only *representative* information is involved:

> In de Burundese hoofdstad Bujumbura loopt de etnische spanning op. Bij nieuwe gevechten vannacht zijn er opnieuw doden gevallen.
>
> In Tokio zijn nu al acht doden geteld na de aanval met sarin-gas in de metro. Volgens een Japanse ochtendkrant zou één verdachte zijn geïdentificeerd; de politie gaat ervan uit dat er een georganiseerde bende aan het werk is geweest.
>
> De Franstalige socialisten willen dat premier Dehaene bemiddelt in het dispuut rond de uitbouw van communicatie-netwerken in ons land.

```
<div1 ID=210508.2><HEAD>Headlines<HEAD>
<p>
<list type=simple>

<item> In de Burundese hoofdstad Bujumbura loopt de etnische spanning
op. Bij nieuwe gevechten vannacht zijn er opnieuw doden gevallen. </item>

<item> In Tokio zijn nu al acht doden geteld na de aanval met sarin-gas in de
metro. Volgens een Japanse ochtendkrant zou &eacute;&eacute;n verdachte
zijn ge&iuml;dentificeerd ; de politie gaat ervan uit dat er een georganiseerde
bende aan het werk is geweest. </item>

<item> De Franstalige socialisten willen dat premier Dehaene bemiddelt
in het dispuut rond de uitbouw van communicatie-netwerken in ons land.
</item>
</list>
</p>
</div1>
```

Interpretative information is to be coded as well. In the following fragment the dots indicate that the newsreader has to wait a few moments before he completes the sentence (the listener is informed that this time Ireland didn't win the European Song Contest)(14mei13u):

> Of toch niet helemaal. Het winnende nummer, Nocturne van de groep Secret Garden, heeft maar een tekst van 24 Noorse woorden. De rest van het nummer is een vioolsolo, gespeeld door ... een Ierse violiste.
>
> Of toch niet helemaal. Het winnende nummer, Nocturne van de groep Secret Garden, heeft maar een tekst van 24 Noorse woorden. De rest van het nummer is een vioolsolo, gespeeld door <pause>...</pause> een Ierse violiste.

A series of dots may also mean that the transcriber didn't understand what was said. In such cases a correct sentence was constructed for linguistic annotation as the original construction will have been correct:

> - Met Swissair hebben we meer bepaald beslist dat onze streefdoelen competitiviteit, kwaliteit en winst zullen zijn. ... zullen zo snel mogelijk en maximaal verwezenlijkt worden.

becomes

> - Met Swissair hebben we meer bepaald beslist dat onze streefdoelen competitiviteit, kwaliteit en winst zullen zijn. **Deze** zullen zo snel mogelijk en maximaal verwezenlijkt worden.

In the SGML-coded original the gap is respected:

> <int> <speaker> – </speaker><p>Met Swissair hebben we meer bepaald beslist dat onze streefdoelen competitiviteit, kwaliteit en winst zullen zijn.<gap reason="inaudible" resp="transcriber"><completion>Deze</completion> zullen zo snel mogelijk en maximaal verwezenlijkt worden.< /p>< /int>

It will be clear that coding texts in SGML the way described above will always involve human interference. Our decisions in this matter may be questioned, especially with respect to our treatment of gaps. We have opted for this solution in order to give our tools a fair chance. The completions are always as neutral as possible. And of course the original texts are available as well.

As remarked before the whole corpus was tagged with SGML, including the parts in a foreign language. These parts, however, have been taken out of the corpus when it comes to linguistic annotations as we didn't have the means to treat these.
This means that of a fragment like the following only the first and the last paragraph are annotated for part of speech, phonology etc.

> De uitslag van de verkiezingen die vandaag beginnen zal bijzondere aandacht krijgen op de verschillende politieke hoofdkwartieren.
>
> Das Oberkommando der Wehrmacht gibt bekannt: Seit mitternacht schweigen nun an allen Fronten die Waffen auf Befehl des Grossadmirals ...
>
> I only wish that Franklin Lee Roosefelt[10] had lived to witness this day. General Eisenhower informs me that the forces of Germany have surrendered to the United Nations. The flags of freedom fly all over Europe.
>
> U hoorde eerst een Duitse omroeper, en daarna de Amerikaanse president Truman, die elk op hun manier het officiële einde afkondigden van de Tweede Wereldoorlog in Europa. Dat is vandaag precies vijftig jaar geleden.

With SGML-annotation this looks like:

[10] Cf. note 9 about misspellings.

<p>De uitslag van de verkiezingen die vandaag beginnen zal bijzondere aandacht krijgen op de verschillende politieke hoofdkwartieren.< /p>

<int><lang=german><p>

Das Oberkommando der Wehrmacht gibt bekannt: Seit mitternacht schweigen nun an allen Fronten die Waffen auf Befehl des Grossadmirals <gap reason="inaudible" resp="transcriber">< /p>

< /lang><lang=english><p>

I only wish that Franklin Lee Roosefelt had lived to witness this day. General Eisenhower informs me that the forces of Germany have surrendered to the United Nations. The flags of freedom fly all over Europe.< /p>< /lang>< /int>

<p> U hoorde eerst een Duitse omroeper, en daarna de Amerikaanse president Truman, die elk op hun manier het offici&euml;le einde afkondigden van de Tweede Wereldoorlog in Europa. Dat is vandaag precies vijftig jaar geleden.< /p>

5.2 Part-of-speech annotation

WOTAN (WOordklasse TAgger voor het Nederlands), cf. Berghmans (1994), is a POS-tagger developed at the University of Nijmegen on basis of the TOSCA-tagger for English. The tagset is based on Geerts, Haeseryn, de Rooij, and van den Toorn (1984) and satisfies the EAGLES-standard[11] for corpus annotation, also with respect to their *recommended* tagset. Next to its quite reasonable performance for Dutch, these features made WOTAN an attractive candidate for us.

The tagset of WOTAN distinguishes 10 main word classes (plus 2 additional ones):(**N**oun, **V**erb, **Art**icle, **Adj**ective, **Adv**erb, **Num**eral, **Prep**osition, **Pron**omen, **Conj**unction, and **Int**erjection (plus **Punc**tuation and **Misc**ellaneous). They all come with further specifications (person, number, gender, valency, case, etc.). One of these further specifications concerns the way the element is used: attributive, substantive, or adverbial. As many mistakes are due to this distinction, the developers of WOTAN suggest to leave this feature out in future. As this distinction is not recommended by EAGLES either, it is not included in the reduced WOTAN tagset with which the complete corpus is tagged (see also Schuurman and Tersago (1996), and the ANNO webpages). An example with both tagsets:

In de Burundese hoofdstad Bujumbura loopt de etnische spanning op. (21mrt08u.txt, sentence 6)

[11] EAGLES: Expert Advisory Group on Language Engineering Standards. EAGLES is part of the LRE programme of the EU (DG-XIII). The EAGLES recommendations are to be found at http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

	full tagset	reduced tagset
∧		
In	Prep(voor)	Prep(voor)
de	Art(bep,zijd_of_mv,neut)	Art(bep,zijd_of_mv,neut)
Burundese	Adj(attr,stell,verv_neut)	Adj(stell,verv_neut)
hoofdstad	N(soort,ev,neut)	N(soort,ev,neut)
Bujumbura	N(eigen,ev,neut)	N(eigen,ev,neut)
loopt	V(intrans,ott,3,ev)	V(intrans,ott,3,ev)
de	Art(bep,zijd_of_mv,neut)	Art(bep,zijd_of_mv,neut)
etnische	Adj(attr,stell,verv_neut)	Adj(stell,verv_neut)
spanning	N(soort,ev,neut)	N(soort,ev,neut)
op	Adv(deel_v)	Prep(op)
.	Punc(punt)	Punc(punt)

Note that in the reduced version of WOTAN the separable verbal particle **op** is considered to be a preposition, a simplification suggested by the developers because too many mistakes were made. This is to be corrected by hand if so desired. Within the ANNO project this was corrected indeed.

In both tagsets WOTAN makes use of so-called **portmanteau** tags like **zijd_of_mv** (non-neuter or plural) or **hulp_of_kopp** (auxiliary or copula).

For Dutch the performance when using the full tagset is claimed to be 90 % at the level of the tags, and 95 % at the level of the word class for the extended tagset (for the reduced tagset the performance comes close to 94 % for the tags). Post-editing is therefore necessary.

The scores (full tagset) for our Flemish corpus were not that good: 86 % at the level of the tags and 94 % at the level of the word class[12]. Analysis of the mistakes showed us that many mistakes are made in constructions with typical Flemish properties (order of verbs, verb projection raising, colloquial speech). Ideally the tagger should be adapted to Flemish.

5.3 Phonological annotation

The complete corpus comes with phonological annotations by means of TreeTalk (beta version), a grapheme-to-phoneme conversion tool developed at the Universities of Antwerp and Tilburg.

Its output is in YAPA (Yet Another Phonetic Alphabet) which is IPA in 7-bits ASCII. It is developed at the K.U.Leuven and will be used by all projects within the programme "Spraak- en Taaltechnologie". It is to reflect the Flemish pronunciation.

The conversions by TreeTalk are not corrected. At the moment the idea is just to give the user an indication of the kind of phonological annotation we have in mind for the future. TreeTalk is first to be improved (for example on basis of the outcome of the aforementioned FONILEX project). Correction by hand was infeasible within the current project.

As far as we are aware TreeTalk is the only tool available to get phonological

[12] Note that one can not just compare the scores as the composition of the corpora involved is different. The WOTAN corpus consists of newspapers.

annotation for Flemish (The CELEX database, for example, reflects the Dutch pronunciation! And especially for phonological annotation one can not work with tools for Dutch-in-general. The relation between grapheme and phoneme in both language variants is not the same.)

> De Verenigde Naties zijn er niet in geslaagd om in Bosnië het bestand te verlengen dat vanmiddag afloopt.

de	d@	bosnie	bOsniE
verenigde	v@ren@Gd@	het	@t
naties	nasis	bestand	b@stAnt
zijn	zE^n	te	t@
er	@r	verlengen	v@rlEN@n
niet	nit	dat	dAt
in	In	vanmiddag	vAnmIdAx
geslaagd	G@slaxt	afloopt	Aflopt
om	Om	.	@
in	In	∧	@

5.4 Morphological annotation

It was quite difficult to find a morphological tagger for Dutch. Asking around on the net resulted in two candidates XSoft (Xerox) and KEPER (Polderland). XSoft turned out not yet to be available at the moment we needed it, therefore we only considered KEPER. It soon turned out that its functionality was not what we were looking for. We just needed in three fields 1) the item itself, 2) the lemma and 3) its internal structure (with special features, cf. below).

Therefore it became rather unappealing to tag the whole corpus with KEPER. Instead we developed our own tagset (AnnoMorf), which was applied to a very small part of the corpus (as tagging by hand is very time-consuming). This exercise gave us the possibility to adjust the tagset. AnnoMorf makes use of both the CELEX-database and the outcome of WOTAN.

In the third field for verbs not the 'neutral' stem should be given (that is already contained in the second field) but the past stem (like *zou*) or the participle stem (like *bombardeer*), TENSE meaning present tense affix, PTENSE past tense affix, PASTP past participle affix, etc. (cf. Schuurman (1997)):

```
zouden\zal\zou+PTENSE\
kunnen\kan\kan+TENSE\
gebombardeerd\bombardeer\bombardeer+PASTP\
gestegen\stijg\steeg+PASTP\
```

A tool with this functionality is under construction. In a later version another functionality should be added as well: of complex words it should be made clear what is the status of the boundary when no connective sound (as in “voorjaarS-buien”) is involved:

voorjaarsbuien\voorjaarsbui\voorjaar+S+bui+EN\
aardbeving\aardbeving\aarde+beving
regelgeving\regelgeving\regel+geef+ing
media\medium\medium+PL\

Note that in "regelgeving" (issuing of rules) the part "geving" is not a word in Dutch, whereas in "aardbeving" (earthquake) both "aarde" and "beving" are existing words. In "voorjaar" (spring) both parts do exist as separate words, but still the word "voorjaar" is to be considered a simplex word.

5.5 Syntactic annotation

The syntactic annotation should add two further clues:

- constituents
- functions fulfilled by the constituents

In ANNO part of the METAL-parser developed by Siemens-Nixdorf was used in order to obtain a flat, bracketed structure (cf. the recommendations by EAGLES, section 1.3.3.2
(URL: http://www.ilc.pi.cnr.it/EAGLES96/browse.html.)), enriched with syntactic functions like *Subject*, *SCOMP*, etc[13]. METAL was chosen because it is the only syntactic parser for Dutch we are aware of yielding a flat, bracketed structure. As the results were not what we expected them to be[14] we will move over to another syntactic parser, probably one based on AGFL[15] or on ALEP[16]. In parallel a tool taking care of so-called *partial parsing* should be taken care of.

Below an example parsed with METAL: 21mrt08, sentences 2 and 6. Note that in sentence 2 some words (16/19) are not included in any constituent, nor are they considered constituents themselves. METAL is robust enough not to fail when it cannot handle part of the input. On the other hand there were too many sentences not receiving any constituent structure at all. Of course, everything can be corrected by hand. But as soon as there are too many 'mistakes' this is not feasible from a practical point of view.

> Het KMI verwacht vooral in het westen van het land mooie opklaringen, elders af en toe ook bewolking.
>
> In de Burundese hoofdstad Bujumbura loopt de etnische spanning op.

[13] At the moment METAL is distributed by LANT and it is called LanTmark.

[14] In fact we made an improper use of the METAL technology: the rules in the METAL parser were written with other applications and other types of sentences in mind. It turned out not to be possible to adapt the parser to our needs, at least not during the project. This appears to be one of the drawbacks of working with a commercial product.

[15] "Affix grammars over a Finite Lattice" (AGFL) is developed in Nijmegen, at the Department of Software Engineering. For more information, cf. http://www.cs.kun.nl/agfl/

[16] The "Advanced Language Engineering Platform" (ALEP) is an initiative of the European Commission. For more information, cf. http://www.iai.uni-sb.de/alep/

One problem concerns verbs with separable affixes as in “oplopen” (increase). In sentence 6 the affix is left out, other times it is considered a preposition used in postposition. Discontinuous structures in general present problems for the parser.

```
(2  [CLS [CLS [NP $SUBJ (”Het” 1) (”KMI” 2) ] [PRED
    (”verwacht” 3) ]
    [PP (”vooral” 4) (”in” 5) (”het” 6) (”westen” 7) ] [PP $POBJ
    (”van” 8) (”het” 9) (”land” 10) ]
    [NP $DOBJ (”mooie” 11) (”opklaringen” 12) ] ] (”,” 13) [CLS
    (”elders” 14)
    [PRED (”is” 15) ] (”er” 16) (”af” 17) (”en” 18) (”toe” 19) [NP
    $SUBJ (”ook” 20) (”bewolking” 21) ]
    [PP (”met” 23) (”vooral” 24) (”in” 25) (”de” 26) (”Ardennen”
    27)
    [PP (”op” 30) (”nog” 28) (”kans” 29) ] (”lichte” 31)
    (”voorjaarsbuien” 32) ] ] ] (”.” 33) )
(6  [CLS [PP $MOV (”In” 1) (”de” 2) (”Burundese” 3) (”hoofdstad”
    4) (”Bujumbura” 5) ]
    [PRED (”loopt” 6) ] [NP $SUBJ (”de” 7) (”etnische” 8)
    (”spanning” 9) ] ] (”.” 11) )
```

5.6 Discourse annotation

In a last annotation round semantic information concerning Tense and Aspect is added. At the moment this is done by hand. Within the NFWO-project Linguaduct this approach will be worked out and implemented in ALEP.
Per sentence six types of information are given in just as many fields, cf. Booij (1996).

Field 1: TEMPORAL ANAPHORA
Does the *point of reference* of the sentence under consideration coincide with the point of reference in the previous sentence? **g** says that both points of reference are *simultaneous*, **n** that they are *not simultaneous.*

Field 2: TENSE
What is the relation between the *point of reference R* and the *point of perspective P*? **v** describes the relation as being *anterior*, **g** as *simultaneous* and **n** as *posterior.*

Field 3: TEMPORAL ADJUNCTS
In case the sentence contains a temporal adjunct this adjunct is qualified as being **l** (*locational*) or **r** (*relational*). If it is relational there is a further distinction in *deictic* (**d**) and *anaphoric* (**a**) ones. A third value tells whether the adjunct expresses *anteriority* (**v**), *simultaneity* (**g**) and *posteriority* (**n**) or whether it is to be considered a *general adjunct* (**a**).

Field 4: ASPECT
What is the relation between the *time of event E* and the *point of reference R*? **p** says it is *perfective*, **d** *durative*, **r** *retrospective*, **t** *terminative*, **i** *inchoative* and **pr** *prospective.*

Field 5: ASPECTUAL ADJUNCTS

Are the aspectual adjuncts to be classified as *durative adjuncts* (**d**) or as *frame adjuncts* (**g**)? Durative adjuncts are subdivided in *in*-adjuncts (**i**) and *for*-adjuncts (**f**), frame adverbials in adjuncts marking the *beginning* (**b**) or the *end* (**e**) .

Field 6: AKTIONSART
Is the basic proposition *bounded* (**b**) or *unbounded* (**o**)?

For a sentence containing several finite clauses the information is expressed for all of these clauses. In such a case the values for the clauses is separated by a “+” (as shown in the second example). Note that in the fields 3 and 5 the values will be complex ones. On the other hand, they may remain empty since adjuncts are optional.

> In de Burundese hoofdstad Bujumbura loopt de etnische spanning op.
>
> \ n \ g \ \ d \ \ o
>
> Morgen blijft het nog aan de frisse kant, vanaf donderdag wordt het overdag heel wat zachter.
>
> \ n”+”n \ n”+”n \ rdn \ d”+”t \ gb \ o”+”o

5.7 Some figures

The full corpus, i.e. the corpus as it was scanned, contains approximately 646.500 words (± 4.2 MB), of which 340.000 words (2.2 MB) news broadcasts and 306.500 words (2 MB) Actueel.
The whole corpus is corrected for errors which may result from scanning. Of these 4.2 MB 2.65 MB is edited as described in section 4.2 (1.85 MB news, 0.8 MB Actueel).
SGML-codes have been added for all corrected texts, i.e. 2.65 MB.
Everything (± 4.1 MB as foreign text fragments were excluded) was tagged for part-of-speech with the reduced WOTAN-tagset, 2.65 MB was also tagged with the extended tagset (cf. section 5.2). Of this 2.65 MB 1.3 MB has already been corrected by hand.
0.5 MB is annotated for syntactic information with METAL (section 5.5) and 0.2 MB for morphological information. The latter was done by hand, cf. section 5.4.
The whole corpus is provided with a phonetic annotation (cf. section 5.3), the outcome is not corrected.
A small part of the corpus (0.07 MB) is also annotated for discourse information, more specifically for temporal information (Tense and Aspect). This was done by hand.

6 Conclusion

Creating a multi-functional, annotated linguistic database from scratch is quite a job. There is still a long way to go: tools should be adapted for Flemish (WOTAN),

others should be improved (TreeTalk) and further developed (AnnoMorf, the discourse tool). The whole corpus is to be parsed once more with another parser. We have the feeling that this duplication of work does pay off when we find a parser giving a better result. In that case the correction phase will be far less time-consuming. Remember that such a correction phase will return time after time! So it is worth the effort.
More text genres are to be added as well. At the moment we are collecting a subcorpus with texts from Flemish newspapers.

It will be clear that especially for phonological annotation one cannot work with tools for Dutch-in-general, we didn't even give such a tool a try. The relation between grapheme and phoneme is different in both language variants. Phonological information out of the CELEX database can not be used.
For other annotation tools the situation is less clear: the from our point of view unsatisfying performance of both METAL and KEPER is not to be attributed to the fact that Flemish texts were involved. They just don't satisfy our needs. On the other hand we have the impression, based on an error analysis, that the performance of WOTAN will be better when it is tuned for Flemish.

A last task will be to make everything available via the Web, making use of JAVA and *Abundantia Verborum* (see Speelman (1997)). A complication, however, is that the BRTN doesn't allow us to distribute their texts freely, at least not for commercial purposes. We will have to find a means to make as much as possible of the corpus public.

References

Berghmans, J. (1994). Wotan, een automatische grammatikale tagger voor het Nederlands. Master's thesis, Katholieke Universiteit Nijmegen.

Booij, J. d. (1996). Tense en Aspect in het Nederlands. Master's thesis.

Dutilh-Ruitenberg, W. (1992). Corpus Annotation Schemes in the Netherlands. INL Working Papers 92-03.

Geerts, G., W. Haeseryn, J. de Rooij, and M. van den Toorn (Eds.) (1984). *Algemene Nederlandse Spraakkunst.* Groningen/Leuven: Wolters-Noordhoff.

Hoekstra, E. (1987). Verb Raising and Verb Projection Raising in Flemish and Dutch. A report prepared for the Ministerie van de Vlaamse Gemeenschap.

Ide, N. and J. Véronis (1994). Corpus Encoding. Eagles Document EAG-CSG/IR-T2.1, EAGLES.

Kruyt, J. (1995). Nationale tekstcorpora in internationaal perspectief. *Forum der Letteren 36*(1), 47–58.

Kruyt, J. and E. Putter (1992). Corpus Design Criteria. INL Working Papers 92-11.

Martin, W. (1967). *De inhoud van krant en roman. Een frequentieonderzoek.* Antwerpen: Plantyn.

Martin, W., F. Platteau, and R. Heymans (1985). Naar een corpus voor een woordenboek hedendaags Nederlands. Mogelijkheden en beperkingen van het gebruik van corpora in lexicografisch onderzoek. UIA.

Peters, W. and B. Tersago (1996). Tekstcorpora. de stand van zaken. ANNO-project, Centrum voor Computerlinguïstiek, K.U.Leuven.

Schuurman, I. (1997). AnnoMorf. Centrum voor Computerlinguïstiek, K.U.Leuven.

Schuurman, I. and B. Tersago (1996). ANNO – IWT 940048. Wetenschappelijk Verslag, Centrum voor Computerlinguïstiek, K.U.Leuven.

Speelman, D. (1997). *Abundantia Verborum. A Tool for representing and presenting data of lexicological and lexicographic studies.* Ph. D. thesis, Katholieke Universiteit Leuven.

Sperberg-McQueen, C. and L. Burnard (1994). *Guidelines for Electronic Text Encoding and Interchange* (TEI P3 ed.). Chicago, Oxford: Text Encoding Initiative.

Sterkenburg, P. v. (1989). *Taal van het Journaal. Een momentopname van hedendaags Nederlands.* 's-Gravenhage: SDU-Uitgeverij.