Using Grammatical Description as a Metalanguage Resource

F. de Vriend*, P.A. Coppen[‡], W. Haeseryn*

* Department of Dutch, University of Nijmegen
Erasmusplein 1, Nijmegen, The Netherlands
[‡]Department of Language and Speech, University of Nijmegen
Erasmusplein 1, Nijmegen, The Netherlands
{F.deVriend, P.A.Coppen, W.Haeseryn}@let.kun.nl

Abstract

The present paper is concerned with the advantages of a digitised descriptive grammar over its traditional print version. First we discuss the process of up-conversion of the ANS material and the main advantages the E-ANS has for the editorial staff. Then from the perspective of language resources, we discuss different applications of the grammatical descriptions for both human and machine users. The discussion is based on our experiences during the project ‘Elektronisering van de ANS’, a project in progress that is aimed at developing a digital version of the Dutch reference grammar *Algemene Nederlandse Spraakkunst* (ANS).

1. Introduction

One of the key resources of the Dutch language is the *Algemene Nederlandse Spraakkunst* (ANS) (Haeseryn et al. 1997), a descriptive grammar of Dutch for general purposes. The ANS is meant as a reference grammar, comparable in scale and scope to grammars such as *A Comprehensive Grammar of the English Language* (Quirk et al. 1985), *Le bon usage* (Grevisse/Goosse 1986, 12th edition) for French and *Duden. Grammatik der deutschen Gegenwartssprache* (Drosdowski et al. 1995, 5th edition) for German. The ANS is aimed at different user groups, academic as well as non-academic, both Dutch speaking users and advanced learners of Dutch.

At this moment a digital version of the ANS (called the ‘Electronic ANS’ or E-ANS) is being developed at the University of Nijmegen in the Netherlands. The project has been made possible by financial support of the Faculty of Arts and the Nederlandse Taalunie (‘Dutch Language Union’, a Dutch-Flemish intergovernmental organisation responsible for Dutch language).

In terms of language resources, the ANS can be considered metalanguage. Although it contains many primary language elements in the form of example sentences provided with qualitative labels specifying their grammaticality or acceptability, the main part of the ANS is devoted to a discussion of the grammatical properties of the Dutch language. This discussion is elaborated by means of case lists, in which words and phrases with certain properties are enumerated, often exhaustively.

The difference between example sentences and case lists is that the former can be considered as a biased corpus of Dutch sentences, whereas the latter can be seen as a special kind of lexical information.

2. Electronic ANS

2.1. Aims of the Project

The two main aims of the project ‘Elektronisering van de ANS’ are:

1. To make the grammatical information of the ANS more accessible to different user groups (including machine users);
2. To create a format and file structure for efficient content management (i.e. storing, editing and extending) of the grammatical information.

The project thus has clear advantages for both the users and the editorial staff of the ANS. In this paper we will be focusing on the first aim of the project by looking at the advantages the E-ANS has for human and machine users (see sections 3 and 4 respectively). First we will briefly discuss the process of up-conversion of the material and the main advantages the E-ANS has for the editorial staff.

2.2. Up-conversion

We have chosen eXtensible Markup Language (XML) as the encoding standard to be employed for the files of the Electronic ANS (De Vriend, 1999). Since the original files were in Word Perfect format, an up-conversion was needed. Moreover, the original coding was minimal, and aimed at superficial lay-out matters only. For example, index lists and enumerations were not automatically generated. This is why conversion to a richer, content oriented code (XML) had to be done in two separate steps.

First the WP files were automatically converted to HTML. Next the HTML encoded files were semi-automatically converted to the intended XML encoding. The conversion strategy is depicted in figure 1.

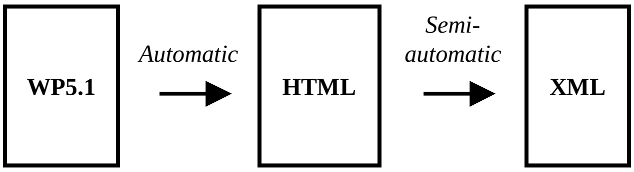


Figure 1: The up-conversion

We have chosen for this strategy with an intermediate conversion to HTML because an immediate and completely automatic conversion from WP to XML was impossible, since the WP lay-out code could not be translated one-to-one to a content-oriented code. The conversion to XML was based on a Document Type Definition (DTD) constructed beforehand. This DTD can be thought of as a document describing the lexicon and grammar of the ANS-specific XML encoding system, i.e. the collection of distinct tags and the rules on how to use them. For the construction of a DTD the structure of the text that is to be encoded has to be determined. In particular, it needs to be established what different kinds of content will have to be encoded (titles, bodies, examples, etc.) and in what types of structures they are related to each other (a file must contain a title and a body, a body may contain examples, etc.). For these structure types generic rules were defined in the DTD. The conversion software involved a DTD guided interpretation of the HTML code which was automatically generated in the first step.

2.3. Advantages for the Editorial Staff

For the editorial staff the two most important advantages of the XML sources are the possibilities for validating edited documents and easily generating different end products from the same source. These advantages enable efficient content management. Figure 2 describes these editorial processes.

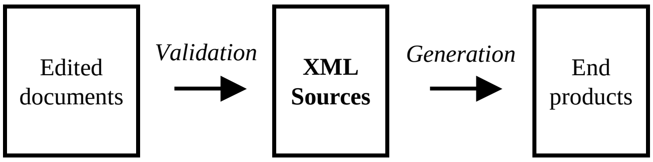


Figure 2: Editorial processes

2.3.1. Validation

The texts of the ANS are produced by several authors working in different locations, mostly universities in the Netherlands and Belgium. Because of the relatively complex structure of the XML encoded files this requires clear protocols by which the authors' work in progress can be efficiently monitored. XML offers excellent possibilities in this respect for controlling the structure of documents. In particular, a document can be validated against the DTD it is associated with. The structures in the documents must obey the rules declared in the DTD. Proper validation ensures that coding variations between authors no longer occur. Although ideally validation takes place during the editing process, so that mistakes can be corrected immediately, validation by the editorial staff after receiving the edited material is also very well possible.

2.3.2. Generation

Because of the content-oriented nature of XML, the ANS material is encoded in a media and product neutral way. This allows for different products based on the ANS material to be produced faster and at a lower cost. From the source a hypertext structure can be generated for electronic publications but also a linear structure for a possible future edition of the book. Producing for different electronic publishing media (webbrowser, eBook, pda) in the ideal situation will only mean changing the style sheet that is associated with the data. Whenever updates or upgrades of the sources have been made the altered material can be made available to the public immediately via the Internet. Another important improvement is the inclusion of the index terms into the files they point to as this enables automatic generation of the index.

Next to the electronic publishing products based on the ANS, products based on the ANS sources that are intended for machine use can also be generated. In the next two paragraphs we will take a closer look at the advantages of the E-ANS for the human and the machine user respectively.

3. The Human User

As an electronic publication the ANS has several advantages over the original print version. The most important ones for its users are the improved accessibility and new ways of navigating through the text.

3.1. Navigation

Because of its modular text organization the ANS has already the nature of a hypertext document. At various places, the reader is referred to other paragraphs containing further information or alternative examples. This makes the conversion to a real hypertext document a very natural one. Hypertext links offer a quick access to the many references in the text, which are typical for a reference work like the ANS. After all, it is uncommon to read long stretches of text in a reference grammar. The hypertext links not only make it possible to follow the numerous references that are present in the text very quickly, but they also offer the possibility to direct users much more precisely to referenced points in the ANS. In the printed version, reference pointers were limited to pages. Now hyperlinks can point to specific words or phrases in the text.

In the hypertext structure the modularity of the text is further extended by making a division between types of information units that are of primary and of secondary importance. An example of a type that is considered of secondary importance is the numerous examples that are given throughout the ANS. These will be hidden by default, which will make the text better organised and ultimately easier to read. Only when a user wants to read the examples and clicks on a button are the examples shown. The advantage of hiding certain types of information units becomes especially evident when considering lists of examples that are meant to be exhaustive. These lists can be very long which was a problem in the print version of the ANS. In the electronic version however quantity does not matter anymore because computer storage capacity is not a problem and more importantly huge quantities of data can easily be hidden.

3.2. Accessibility

Access to the text can be gained in traditional ways but also in ways unique to an electronic publication. The traditional ways are by means of the table of contents and the index. For the table of contents the advantage is the ease with which the user can jump to a specific paragraph by just clicking on a hypertext link. When clicking on a link to a term in the index, that same term can be highlighted in the text when arriving there. This directing of users to referenced points is even more precise than the direction to lines as discussed before. Also users can now be offered several specialised indices. For instance an index with frequently asked questions or an index containing canonical forms of grammatical constructions.

Horton says that access is essential for online documents. Either make information twice as accessible as on paper or leave the information on paper (Horton, 1994). Therefore in the electronic publication we offer the following new ways in which the text can be accessed:

- By traversing a Yahoo-like directory structure offered immediately on the homepage.
- Via the bibliography. This way people can jump from within the bibliography directly to the parts of the ANS where a particular publication has been used.
- Via a special page called "new in the ANS" containing information on the latest changes made to the content.
- By using a search engine. Here again the modularity of the ANS material can be used to the user's advantage: since the ANS material is now coded in various types of information, these can all be separate input sources to the search engine. For example, the user may want to look for the word *vandaan* (the postposition meaning approximately "from") but only when it occurs in focus in an ANS example sentence. Or he may want to look for the word *diegene* ("the one"), but only where it occurs in a case list. These search possibilities were not present in the print version of the ANS.

4. The Machine User

Although emphasis is put on the use of the ANS for the human user, the XML-encoded sources also enable several ways of access to the information contained in the ANS by machine users.

4.1. Extraction of Corpus Material

Being primarily a metalanguage resource, the ANS grammar, and especially the digitally encoded version, is well suited to serve as a data mining source for computational grammars. Three types of information can be extracted from the ANS very easily.

First, the ANS contains some 15,000 example sentences illustrating various (the ultimate goal is: *all*) syntactic constructions in the Dutch language. Although the frequency of constructions will not reflect everyday Dutch language, the ANS example collection is a syntactically rich collection, suited to serve as a bench mark for computational grammars. It may not be suited to do corpus research on, but it is likely to contain all problematic constructions. Moreover, the example sentences have been judged as to their acceptability and possibly regional character (not only Dutch versus Flemish variants, but also the four main Dutch regions are distinguished).

Secondly, the ANS contains many so-called *case lists*, in which words and phrases with certain grammatical properties are enumerated, often exhaustively. For example, the ANS lists all verbs with different prepositional objects, or all nouns with two different plural forms (e.g. *aardappel* "potato", which has both a plural *aardappelen* and *aardappels*). In fact, the print version of the ANS suffered from some restrictions at this point since some case lists could not be completed due to space considerations. The electronic version obviously does not suffer from this restriction. These case lists offer a welcome extension to current sources of lexical information, like e.g. Wordnet. Since the case lists are coded as such, they can very easily be extracted for use in computational grammars or NLP software in general.

Finally, the ANS contains many tables with lexical information, aiming at completeness. For example, the ANS contains a table with a complete listing of the irregular verbs and their conjugations. As all of these tables are very precisely coded, they too can be easily extracted.

4.2. Meta-information

The ANS offers a complete syntactic description of the Dutch language, that may very well serve as a "golden standard" for syntactically annotated corpora. The ANS is meant to be theory independent, and it can be characterised as moderately shallow: although words and phrases are functionally labelled, no deeper derivations from arcane sources are investigated. This makes the ANS not only theory independent, but also robust against theoretic developments. The ANS description method has survived more than a century. Current syntactic coding as in e.g. the CGN corpus (cf. http://lands.let.kun.nl/CGN) can be (and in fact, *is*) based on ANS codes. It is even possible to link from the corpus annotation directly to the relevant ANS paragraphs for more information and alternative examples.

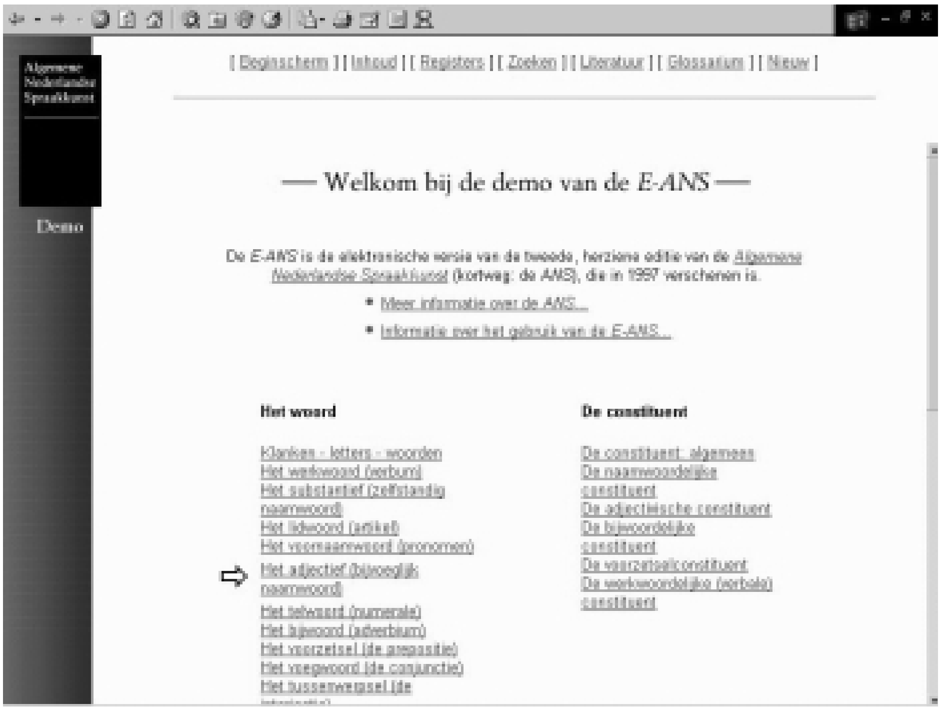
Since the ANS aims at a general public, the grammatical discussion often touches upon normative questions of “correct language use”. Although the ANS mainly takes the linguistic point of view, experience has learned that the ANS has gained a certain authoritative status: language users tend to rely on the ANS as they do on standard dictionaries. In practice, the ANS is often used by professional writers, language teachers and language consultants. This use can be further developed with the electronic ANS: now, relevant passages in the ANS can be referred to from concrete language advice. In stead of referring to the ANS pages, hyperlinks now enable the user to jump directly to the relevant ANS paragraphs.

An exciting application of these possibilities is at the moment being explored in the form of a research project aiming at enriching the grammar checker for Microsoft Word with links to various metalinguistic sources. In this project links to relevant ANS paragraphs are also considered. This way, if the checker detects an error, the user is offered a specific explanation of the possible source of the error in the form of the ANS meta-information. Of course this will eventually mean that some paragraphs will have to be rewritten, so as to meet the requirements of this new application, but the possibilities for this are being created by the digitising of the ANS material.

5. Future Developments

In the last year of the project special attention will be paid to the editorial side of the electronic publishing of the ANS. Especially matters concerning efficient content management will be considered.

At the end of the year 2002 the E-ANS will be fully operational. The ANS material will be published on the Internet free of charge. The reader is referred to the project website www.kun.nl/e-ans for further details and up to date information on the development of the E-ANS. On the website a link to a demo version of the E-ANS can be found as well.



[Beginscherm] [Inhoud] [Registers] [Zoeken] [Literatuur] [Glossarium] [Nieuw]

— Welkom bij de demo van de *E-ANS* —

De *E-ANS* is de elektronische versie van de tweede, herziene editie van de *Algemene Nederlandse Spraakkunst* (kortweg: de *ANS*), die in 1997 verschenen is.

- Meer informatie over de *ANS*...
- Informatie over het gebruik van de *E-ANS*...

Het woord

Klanken - letters - woorden
Het werkwoord (verbum)
Het substantief (zelfstandig naamwoord)
Het lidwoord (artikel)
Het voornaamwoord (pronomen)
⇨ Het adjectief (bijvoeglijk naamwoord)
Het telwoord (numerale)
Het bijwoord (adverbium)
Het voorzetsel (de prepositie)
Het voegwoord (de conjunctie)
Het tussenwerpsel (de interjectie)

De constituent

De constituent: algemeen
De naamwoordelijke constituent
De adjectivische constituent
De bijwoordelijke constituent
De voorzetselconstituent
De werkwoordelijke (verbale) constituent

Figure 3: Screenshot of the opening page of the demo version of the ANS

6. Concluding Remarks

In this paper, we have considered the advantages a digitised descriptive grammar has over its traditional print version. To sum up, the advantages of digitising a metalanguage resource like the ANS for the different kinds of users are:

- The users are offered better ways of navigation through the text.
- The accessibility for the users is improved.

Not only human users benefit from this new development. The electronic ANS also offers new means of access to the metalinguistic information for machine users. The ANS can be used as a bench mark and data mining source for computational grammars, and its case lists can be extracted for use in NLP tools. Moreover, the ANS can be an authoritative reference source for language consultants or grammar checkers.

The editorial staff is offered better possibilities for efficient content management and control over the intake of edited material (validation).

7. Acknowledgements

We would like to thank the members of the steering committee of this project for their comments and suggestions on the work presented in this paper: Gosse Bouma, Walter Daelemans, Carel Jansen, Gerard Kempen, Luuk Van Waes.

8. References

De Vriend, F. (1999). Naar een elektronische ANS. Vooronderzoek naar ergonomische aspecten en implementatie. Undergraduate thesis, General Linguistics. University of Nijmegen (online version: www.kun.nl/e-ans/content/publicaties/scriptie).

Haeseryn et al. (1997). Algemene Nederlandse Spraakkunst (ANS) (2nd edition). Groningen/Deurne: Martinus Nijhoff uitgevers/Wolters Plantyn.

Horton, W. (1994). Designing and Writing Online Documentation. Hypermedia for Self-Supporting Products (second edition). New York: John Wiley & Sons, Inc.