The Spoken Dutch Corpus and its Exploitation Environment

Nelleke Oostdijk
Dept. of Language & Speech
University of Nijmegen
P.O. Box 9103, 6500HD Nijmegen,
The Netherlands
n.oostdijk@let.kun.nl

Daan Broeder
Max-Planck Institute for Psycholinguistics
P.O. Box 310, 6500 XD Nijmegen,
The Netherlands
daan.broeder@mpi.nl

Abstract

The Spoken Dutch Corpus project (1998-2003) is aimed at the development of a corpus of 1,000 hours of speech. The corpus is being annotated with various types of transcriptions and annotations and will be distributed together with the speech recordings. In order for users to access the data efficiently and with relative ease, exploitation software is being developed that can handle both sound files and other types of data files. After a brief introduction in which the goals of the project are outlined, the present paper first describes the Spoken Dutch Corpus as it is presently being constructed and then goes on to describe in some detail the exploitation software. The exploitation environment makes it possible to view the data contained in the Spoken Dutch Corpus as one item in a network of other similarly structured corpora. Our outlook on such a corpus universe can be found in the 'Future Work' section.

1 Introduction

The Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) project is a five-year project that started in 1998 and aims to develop a corpus of 1,000 hours of speech originating from adult speakers of standard Dutch. The goals of the project have been set quite high. The corpus is to serve as a resource for Dutch, for use in a number of widely different fields of interest, including linguistics, language and speech technology, and education. Therefore, its design must anticipate the various research interests arising from these fields and provide for them, while the different transcriptions and annotations should be as sophisticated as possible given the present state-of-the-art. Moreover, in its construction we conform to national and international standards where available, or else follow recommendations and guidelines or adopt best practice as it has emerged from other projects. Finally, in devising protocols and procedures that will ensure the highest possible degree of accuracy and consistency we intend to contribute to setting a standard for future corpora.

All data will be orthographically transcribed, lemmatized and annotated with part-of-speech information. For part of the corpus, additional transcriptions and annotations will be available. These include an auditorily verified broad phonetic transcription, a syntactic annotation and a prosodic annotation. The corpus will be distributed together with the audio files containing the speech recordings. Within the project, exploration software is being developed that will make it possible not only to browse the data, but also to conduct potentially complex searches involving multiple annotation layers, while including a rich set of metadata. Since all transcriptions and annotations will – directly or indirectly – be aligned with the audio files, the user will be able to access the recordings from any point in the corpus.[1]

The remainder of this paper is structured as follows: in section 2 we present an account of the the construction of the Spoken Dutch Corpus which will provide the necessary background for

[1] A more detailed description of the project is given in Oostdijk (2000a).

sections 3 and 4 in which the exploitation software is described. The paper concludes with an outlook on potential future extensions.

2 Corpus Construction

2.1 Corpus Design and Data Collection

The design of the Spoken Dutch Corpus was guided by a number of considerations. First, the corpus should constitute a plausible sample of contemporary standard Dutch as spoken by speakers in the Netherlands and Flanders, that would serve the interests of rather different user groups. Second, the corpus should constitute a resource for Dutch that should hold up to international standards. With 1,000 hours of speech (approx. ten million words) the corpus will be comparable in size to, for example, the spoken component of the British National Corpus (BNC; Aston and Burnard, 1998). Third, because of the time, financial and legal constraints under which the project must operate, but also for practical reasons, it is impossible to include all possible types of speech and compromises are inevitable.[2]

In order to be able to accommodate a great many different types of user, a highly flexible design was adopted. Thus, in determining the overall structure of the corpus, the principal parameter has been the socio-situational setting in which speech occurs. As a result, the corpus comprises a number of subcorpora (ranging from spontaneous conversations to read-aloud text), each of which can be characterized in terms of its situational characteristics such as communicative goal, medium, and number of interlocutors. The design of the corpus is summarized in Table 1.[3] The specification of each of the subcorpora is given in terms of sample sizes, total number of speakers, range of topics, etc. Where this is considered to be of particular interest, speaker characteristics such as gender, age, geographical region, and socio-economic class are used as (demographic) sampling criteria; otherwise they are merely recorded as part of the metadata.

2 The project was awarded a five-year grant in the amount of 4.6 million euro. Since the corpus is distributed including the audio files, the consent of all speakers is required as well as of any other parties that hold any rights to the recorded material. This also explains why only the speech of adults is collected.

3 The respective sizes of the subcorpora are given in the number of tokens they comprise.

Recordings are obtained in a variety of ways. Thus part of the data are obtained through other parties, e.g. broadcast data and interviews. In other cases, research assistants working for the project do the recording. For most recordings, however, we rely on the speakers that are recruited and asked to record conversations in the home environment or have their telephone conversations recorded. Since we wanted to include a great many different speakers and topics, we decided to include many shorter selections of the recorded material in the corpus rather than fewer lengthier ones.

2.2 Corpus Annotation

The Spoken Dutch Corpus is being compiled and annotated incrementally. Time and financial constraints prohibit that all data are annotated with the full spectrum of transcriptions and annotations. Therefore, the more advanced types of annotations (viz. broad phonetic transcriptions, syntactic and prosodic annotations) will be available for 10 per cent of the data.

Orthographic transcription

The first annotation layer is constituted by the orthographic transcription. Of all recordings a verbatim transcript is made. Following the recommendations made in den Os (1998: 170f), the transcripts to a large extent conform to the standard spelling conventions. A protocol has been developed which describes what to transcribe and how to deal with new words, dialect, mispronunciations, and so on.[4] To facilitate the transcription process, use is made of the interactive signal processing tool PRAAT.[5]

In PRAAT it is possible to listen to and visualize the speech signal and at the same time create and view an orthographic transcription. Each speaker is assigned his or her own tier. For unknown speakers an additional tier is used. While the speech of unknown speakers is transcribed, no attempt is made to distinguish between multiple unknown speakers.

During the transcription process, transcribers segment the audio files in relatively short chunks (of approximately 2 to 3 seconds each) by insert-

4 See Goedertier et al. (2000) for a more detailed description.

5 For more information on PRAAT see http://www.fon.hum.uva.nl/praat/

dialogue / multilogue	private		unscripted	direct	conversations (face-to-face)	3,000,000
					interviews	460,000
				distanced	telephone conversations	3,000,000
					business negotiations	175,000
	public	broadcast	more or less scripted		interviews and discussions	750,000
		non-broadcast	Unscripted		discuss., debates, meetings	375,000
					lectures	350,000
monologue	private		more or less scripted		descriptions of pictures	40,000
	public	broadcast	Unscripted		spontaneous commentary	250,000
			Scripted		newsreports, current affairs programmes	250,000
					news	250,000
					commentary	200,000
		non-broadcast	Scripted		lectures, speeches	275,000
					read aloud text	625,000

Table 1. Design of the Spoken Dutch Corpus

ing time markers in unfilled pauses between words. At a later stage these markers are used as anchor points for the automatic (word) alignment of the transcript and the speech file.

POS Tagging and Lemmatization

The orthographic transcription forms the point of departure for the POS tagging and lemmatization of the data. The tagging process is carried out using a POS tagger, the output of which is manually verified (Van Eynde et al., 2000). The tagger employs a tagset that was designed especially for use with the Spoken Dutch Corpus. The tagset has been inspired by the EAGLES guidelines and is compatible with the widely used authoritative Dutch reference grammar (ANS; Haeseryn et al., 1997). It consists of 316 tags and distinguishes between the major word classes, while recording additional morpho-syntactic features with each of these.

Apart from the POS tag, for each word also the associated lemma is given. In the first phase a lemmatizer is used to automatically associate with each token the appropriate lemma. The result is manually checked and corrected. At this stage the constituent parts of split verbs (e.g. *leidde* … *af*), prepositions (e.g. *van* … *uit*) and such like items are lemmatized as if they occurred independently. At a later stage, a more advanced lemmatization is undertaken in which the constituent parts are considered jointly and a lemma is associated with the combination as a whole.

Syntactic Annotation

For part of the data a syntactic annotation is provided. To this end an annotation scheme has been developed which caters for typically spoken language phenomena such as hesitations, false starts and asyndetic constructions. The syntactic analyses use two types of label: dependency labels that provide function information and node labels that give category information (Hoekstra et al., 2001). Syntactic annotation is carried out semi-automatically by means of the ANNOTATE software.[6]

Broad Phonetic Transcription

Broad phonetic transcriptions will be available for the entire corpus. However, only part of these transcriptions will be manually verified. The transcriptions make use of a set of symbols that is derived from the SAMPA set.[7] In the transcriptions a one-to-one correspondence is upheld between the tokens in the orthographic transcription and their phonetic transcriptions.

[6] More information on ANNOTATE can be found at http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

[7] The SAMPA set can be found at http://www.phon.ucl.ac.uk/home/sampa/dutch.htm

Prosodic Annotation

A small portion of the data is being annotated prosodically. The annotation is best characterized as perceptually-based and can serve as a starting-point for a more detailed prosodic labelling. The prosodic annotation is constituted by marking syllables as either prominent or non-prominent, and by marking the unusual lengthening of vowels and consonants not carrying prominence (Buhmann et al., 2002).

2.3 Metadata for Corpora

With the Spoken Dutch Corpus, metadata are collected about the recordings (where they originated from, with what type of equipment and under what conditions, what type of speech they exemplify: monologue or dialogue/multilogue, spontaneous speech, prepared, or more or less scripted speech, formal or informal, broadcast or non-broadcast, etc.) and about the speakers (their sex, age (group), and level of education, the geographic region from which they originate, their present domicile, occupation, etc.). With the corpus also metadata will be made available that relate to the fashion in which the data were processed. Thus, for the recordings and for each of level of transcription and annotation, the metadata describe what has been recorded, transcribed or annotated, what procedure was followed and what protocol was used, what revisions were made at what stage, and who is responsible for the data in that specific state. Moreover, the metadata also include cross-reference information with regard to data available from other projects or in other forms.

The content of the metadata has been inspired by the guidelines of the Text Encoding Initiative (TEI; Sperberg-McQueen and Burnard, Eds., 1994) and the Corpus Encoding Standard (CES; Ide, 1996) and is transformed into the IMDI format with specific extentions for the CGN. [8]

2.4 Data Formats

All audio data except for the telephone recordings have a sampling frequency of 16 kHz and a 16-bit linear resolution. The telephone recordings have a sampling frequency of 8 kHZ in the 8-bit A-law format. For the transcriptions and annotations plain text format is used in the production phase, while an XML format is used for archiving use by COREX and exchange with other tools.

3 Corpus Exploitation

Although the corpus data as such is already valuable, without possibilities of selection and access to the data itself no research can be done. Expert users that have programmer skills and have knowledge of the physical structure of the corpus usually can manipulate the data themselves by creating special scripts. More naïve users are better served with special software that has a user-friendly GUI and is helpful in exploring the corpus without first studying its structure in depth. As an example of such user-friendly systems we mention ICECUP (Nelson, 1998).

When we consider the case of multi-media corpora, the need for special exploitation and exploration software becomes even more urgent since the user will want to see the annotations and media streams in a synchronised way.

In general corpus exploitation software is created specially for a specific corpus and thus mirrors its design and idiosyncracies. Although exploration software should be able to fully manage and exploit the specific corpus it is aimed at, in these days of the Internet and huge storage capabilities it should be possible to have a corpus exploitation environment that can handle multiple corpora and allow users to compare the target corpus with other corpora. On the Internet it has become good practice to use metadata to locate resources of interest and several initiatives have surfaced that aim at encouraging the use of Internet available metadata also for language resources (OLAC; IMDI) such as the Spoken Dutch Corpus.

This use of metadata enables the localisation of resources that fulfil specific requirements and can then be downloaded for further processing. The user though, has to know in advance what he is looking for. A way to explore unknown corpora is to include human readable texts as part of the metadata. This manner of creating browsable corpus structures has been introduced by the IMDI initiative where a corpus is structured in different parallel hierarchical structures and

[8] A description of the metadata in the Spoken Dutch Corpus can be found in Oostdijk (2000b).

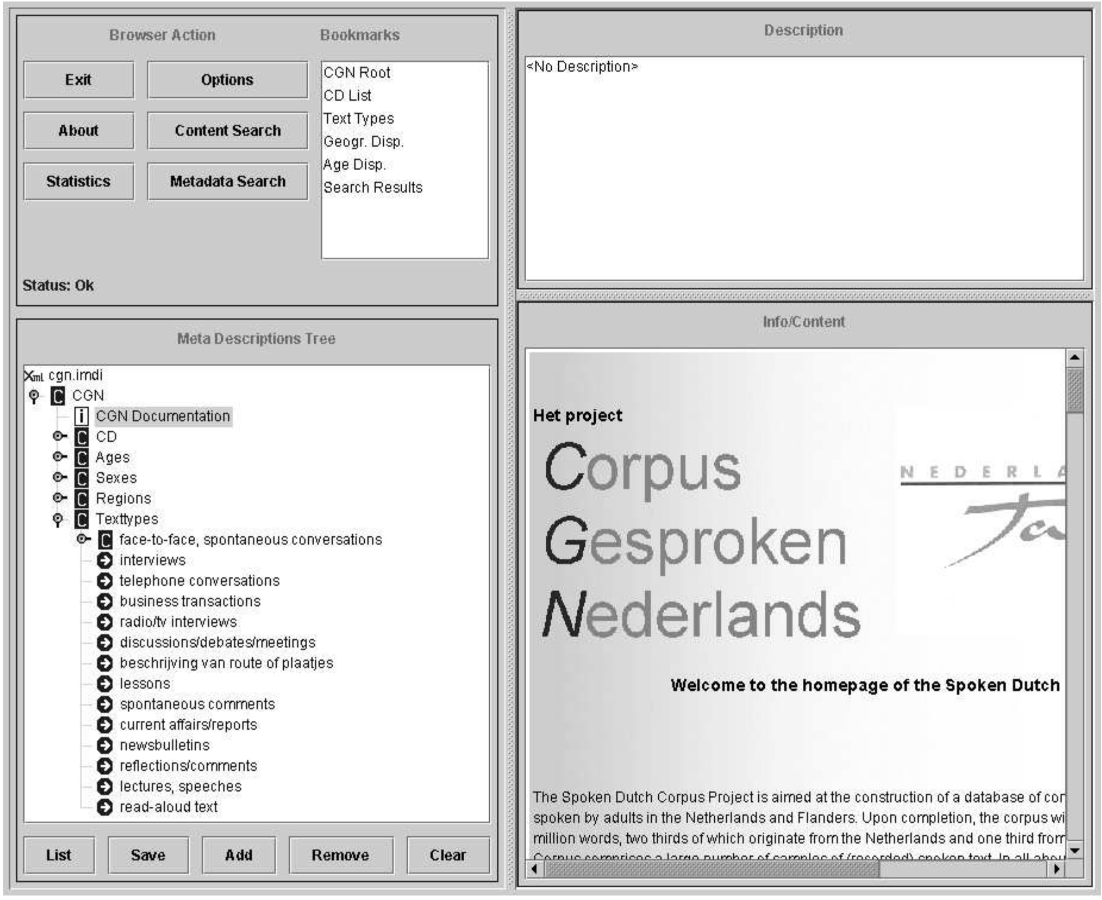

Figure 1. The Corpus Browser Window

metadata are used to create these structures and bind related resources together.

Another important extension to the use of metadata is their use to enable the starting-up of specific viewers and other tools that exploit the resources e.g. the annotation and media files.

4 The Corpus Exploitation Software

Within the Spoken Dutch Corpus project dedicated software is being developed that should enable users to exploit the corpus. The COREX (Corpus Exploitation) software presents itself to the user as a number of different windows or panels. Since the software is best described by describing its most important parts as they present themselves to the user, sections 4.1 - 4.5 focus on the browser, viewer, search and statistics panels. Some general technical information is included in section 4.6 (Technical Details).

4.1 Corpus Browser Panel

The Corpus Browser Panel (Figure 1) is the main user interface of COREX and all other panels are started from this one. It shows the different (predefined) classification hierarchies of the corpus and allows the user to browse these structures and descend them until the lowest level is reached. At this level, all the annotation files and media of a sample are bundled together with the metadata pertaining to that sample. Once this level has been reached, the user can start-up different viewers for the resources (the annotations and media files), depending on their availability and possible combinations. Also external tools (not belonging to COREX) may be made available for exploitation of the resources. This can be configured by editing a configuration file that maps the available tools to the format or a combination of formats of the resources.

The user can use the mouse to select subcorpora from the different classification hierarchies and put them in a so-called “basket”. The samples in the basket can then be made subject to further metadata search or content search.

The Corpus Browser Window also has the possibility to display normal HTML pages that can be linked in the corpus structure(s) to give the user extra information. In fact, the corpus structure hierarchy can be augmented by a hierarchy of linked HTML pages that can also contain references to IMDI subcorpora. Thus the user may jump from the IMDI metadata domain to the HTML domain and vice-versa.

The Corpus Browser Window is also the place where the Spoken Dutch Corpus can be seen in the context of other corpora. This is done by presenting the Spoken Dutch Corpus as just one subcorpus of a bigger domain. All corpora that use IMDI metadata and for which at least one classification structure has been defined can be incorporated in this domain.

4.2 COREX Viewer

The most important tool that the user can start-up from the Corpus Browser is the COREX viewer. This is a panel that can display all available annotations in different combinations. It can also show the annotation synchronised with the playing audio signal. A running display of the waveform is also possible. Figure 2 shows the COREX viewer with two annotation tiers: the orthographic transcription and the POS tagging.

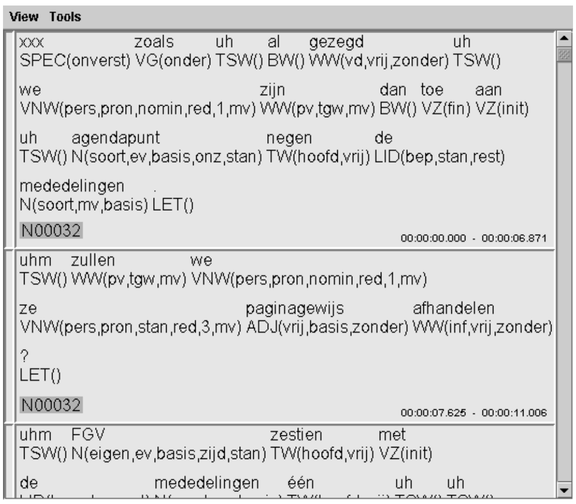

Figure 2. The COREX Annotation Viewer

4.3 Metadata Search Panel

The Metadata Search Panel allows the user to search for samples that conform to specific requirements e.g. “Give me all samples with female speakers under 25 years of age that have had their education in the city of Utrecht”. The panel offers the user guidance into what metadata fields are available. The user can immediately inspect the results by selecting them whereupon a new Corpus Browser Panel appears with the found corpus sample. Results from a metadata query can also be saved as a “result corpus” that is stored for future use. Result corpora can be further processed, for instance by content search.

4.4 Content Search Panel

Content search is the process of searching for items or combinations of items on the different annotation tiers. The Content Search Panel offers the user a helpful environment for specifying such queries. It has knowledge about the different annotation tiers available and in the case of for instance the POS tier it knows what tags belong to the tag set and can be used in a query. Initially the results are only shown in their orthographical context but they can be inspected fully by starting a COREX Annotation Viewer by clicking on the hit. The user may also save the results in a file for future reference.

4.5 Statistics Panel

The Statistics Panel allows a user to specify multiple content search queries to be executed on different subcorpora. The results are displayed in the form of a frequency table. An export function permits saving the result table in text format for further processing by for instance a statistical software package.

4.6 Technical Details

Here we include some technical details of COREX that may be of interest:

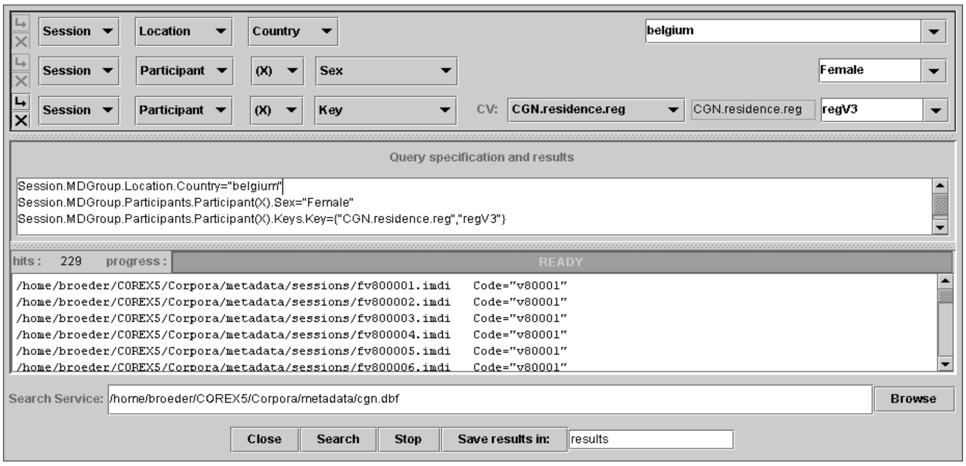

Figure 3. The Metadata Search Panel

- All software is written in Java and has been tested on Windows 98/NT/2000/XT, Solaris, Linux, MacOSX
- At installation the COREX programme with all annotation data is copied on the users computer leaving the computers CD drive available for CD's with the audio data.
- No special database software is required. Perl scripts searching through files in an optimised format aided in some cases by index files perform content search and metadata search.
- COREX allows the incorporation of external tools to work on individual or combinations of resources. Configuration of such extensions can be performed at the user level. An example is the Praat programme for signal analysis purposes.
- All annotation files and most of the metadata are in compressed format. The Corpus Browser Panel and the COREX Annotation Viewer decompress the data on the fly when needed. Thus all the text data and the COREX software itself fit on a single CD.

Figure 4. The Content Search Panel

5 Future Work

In December 2002 the latest version of COREX was published together with release 6 of the Spoken Dutch Corpus. The final COREX release is planned for the final release (release 7) of the Corpus at the end of 2003. Additions to COREX that are under consideration include: Searching in syntax annotations, incorporation of the Spoken Dutch Corpus lexicon, and viewing prosodic annotations.

We will also pursue making the Dutch Spoken Corpus available within larger corpora frameworks using IMDI technology with browsable hierarchies as sketched in section 3. This could be realised in the INTERA project [INTERA].

Also under consideration are the way annotations and audio data are distributed. Tests have been performed that show that it is very well possible to store the data centrally and access them over a broadband Internet connection (DSL/ ADSL).

Acknowledgement

This publication was supported by the Netherlands Organisation for Scientific Research (NWO) under grant number 014-17-510.

References

Guy Aston and Lou Burnard. 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Jeska Buhmann, Johanneke Caspers, Vincent van Heuven, Heleen Hoekstra, Jean Pierre Martens, and Marc Swerts. 2002. Annotation of Prominent Words, Prosodic Boundaries and Segmental Lengthening by Non-Expert Transcribers in the Spoken Dutch Corpus. In *Proceedings LREC 2002*. Vol.III: 779-785. Paris: ELDA.

Catia Cucchiarini, Diana Binnenpoorte, and Simo Goddijn. 2001. Phonetic Transcriptions in the Spoken Dutch Corpus. How to Combine Efficiency and Good Transcription Quality. In *Proceedings Eurospeech 2001*. 1679-1682. Aalborg, Denmark.

Wim Goedertier, Simo Goddijn, and Jean Pierre Martens. 2000. Orthographic Transcription of the Spoken Dutch Corpus. In *Proceedings LREC 2002*. Vol. II: 909-914. Paris: ELDA.

Walter Haeseryn, Kirsten Romijn, Guido Geerts, Jacobus de Rooij, and Maarten van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Groningen: Martinus Nijhoff.

Heleen Hoekstra, Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. 2001. Syntactic Annotation for the Spoken Dutch Corpus Project (CGN). In Walter Daelemans, K. Sima'an, Jorn Veenstra, and Jakub Zavrel (eds.) *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*. 73-87. Amsterdam: Rodopi.

Nancy Ide. 1996. *Corpus Encoding Standard. Document CES 1*. Version 1.4. http://www.cs.vassar.edu/CES/

Gerarld Nelson. 1998. *The International Corpus of English. The British Component. ICE-GB. Getting Started*. London: Survey of English Usage, University College London.

Nelleke Oostdijk. 2000a. The Spoken Dutch Corpus. Overview and first Evaluation. In *Proceedings LREC 2002*. Vol. II: 887-893. Paris: ELDA.

Nelleke Oostdijk. 2000b. Meta-Data in the Spoken Dutch Corpus Project. In *LREC 2000 Workshop Proceedings. Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources*. 21-25. Paris: ELDA.

Els den Os. 1998. SL Corpus Representation. In Dafydd Gibbon, Roger Moore and Richard Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems. Vol. IV. Spoken Language Systems and Corpus Design*. 146-174. Berlin – New York: Mouton de Gruyter.

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

Frank Van Eynde, Jakub Zavrel, and Walter Daelemans. 2000. Part of Speech Tagging for the Spoken Dutch Corpus. In *Proceedings LREC 2002*. Vol. II: 1427-1433. Paris: ELDA.

OLAC: The Open Language Archive Community. http://www.language-archives.org

IMDI: ISLE Metadata Initiative. http://www.mpi.nl/ISLE

INTERA: Integrated European language data Resource Area. http://www.mpi.nl/INTERA (future website).