Meta-Data in the Spoken Dutch Corpus Project

Nelleke Oostdijk

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

Abstract

The Spoken Dutch Corpus that is currently under construction will constitute a 10-million-word corpus of contemporary Dutch as spoken in Flanders and the Netherlands. A collection of extremely varied data for extremely varied users, the Spoken Dutch Corpus constitutes an ideal case study for evaluating proposals for encoding standards. The paper discusses the nature of the meta-data that are deemed to be relevant for the various user groups of the Spoken Dutch Corpus. It also addresses issues such as how - from a users' point of view - these meta-data should preferably be structured. In addition, the paper evaluates the extent to which available standard proposals are adequate or need to be adapted to suit these needs.

1. Introduction

In order to exploit a language resource such as a corpus to the full, it must be documented in some form or other. While the documentation should provide any (general) background information that might be helpful, such as the context in which the resource was created, its availability and conditions of use, it should also describe in detail various aspects of its design, compilation and annotation. Preferably the protocols that were used in the creation of the resource together with a description of the procedures that were followed should be made available as well.

For some information, it suffices to present it in the form of a text document and have a pointer referring to it. This applies for instance to a description of the procedure that was followed in making the recordings or in the recruiting of volunteers. Also protocols such as the one used for orthographic transcription or the manual used in part-of-speech tagging are best presented as text documents. For other types of information, however, it is important that the information is not only available for human consultation but can effectively be used automatically in exploiting the resource by allowing the user to define which subset(s) of data specifically he wants to access. Typically it is this type of information that is optimally represented in a more formalized and standardized fashion such as the corpus and text headers proposed by various mark-up schemes.

So far mark-up schemes that have been/are being used in one form or another, such as the Guidelines provided by the Text Encoding Initiative (TEI; Sperberg-McQueen and Burnard, 1994), the Corpus Document Interchange Format (CDIF; Burnard (ed.), 1995) and the Corpus Encoding Standard (CES; Ide (ed.), 1996) are fairly well-developed for use with written language corpora, but they have far less provisions for spoken resources. As more and more spoken corpora are being compiled, the need for a standard that is geared to the needs of these corpora is becoming more pressing. The nature of the Spoken Dutch Corpus, a collection of extremely varied data for extremely varied users that is currently under construction, makes it an ideal case study for evaluating proposals for encoding standards.

After a brief description of the Spoken Dutch Corpus project[1], the paper goes on to discuss the nature of the meta-data that are deemed to be relevant for the various user groups of the Spoken Dutch Corpus. It also addresses issues such as how from a users' point of view these meta-data should preferably be structured. In addition, the paper evaluates the extent to which available standard proposals are adequate or need to be adapted to suit these needs.

2. The Spoken Dutch Corpus Project

2.1. Brief overview

In June 1998 the Spoken Dutch Corpus project was started, a five-year project aimed at the compilation and annotation of a 10-million-word corpus of contemporary standard Dutch as spoken in Flanders and the Netherlands. The entire corpus will be orthographically transcribed, lemmatized and annotated with part-of-speech information. For a selection of one million words, further, more detailed annotations are envisaged, including an auditorily verified broad phonetic transcription and a syntactic annotation. A selection of 250,000 words will receive a prosodic annotation. To enable effective access to the speech recordings, the transcriptions will be enriched with pointers into the speech files. The automatic time alignment will be manually checked for that part of the corpus for which a verified phonetic transcription is available.

The corpus is intended to serve many and diverse (research) interests. Apart from the interests held by language and speech technologists, the corpus addresses the needs of linguists from various backgrounds, while another field for which the corpus is of great importance is that of education. The fact that the corpus must serve such diverse interests is reflected in the design of the corpus: it attempts to meet the different requirements that different user groups have when it comes to the quality and the quantity of the data, the numbers and types of speakers, etc. Thus the design of the corpus takes into account the

[1] For a more elaborate description, see Oostdijk (2000).

various dimensions underlying the variation that can be observed in language use. In the overall design of the corpus the principal parameter is taken to be the socio-situational setting in which language is used. This leads us to distinguish a number of components, each of which can be characterized in terms of its situational characteristics such as communicative goal, medium, number of speakers participating, and the relationship between speaker(s) and hearer(s). For each of the components a further specification 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 (group), geographical region, and socio-economic class are used as additional (demographic) sampling criteria.

2.2. Role of the meta-data

With a language resource such as the Spoken Dutch Corpus, the meta-data play a very important role in the enhancing the usability of the data. The meta-data constitute the means to define and access precisely those subsets of the data contained in the corpus that are considered to be relevant for a specific task.[2] Thus speech technologists wishing to use the corpus are likely to have outspoken wishes with regard to the quality of the recordings and/or the conditions under which these were made, the speakers involved, the time they speak, etc. They want to be able to select from the corpus only those samples that meet their requirements, and disregard any other samples. In a similar fashion, a sociolinguist will put high value on being able to select groups of speakers from particular backgrounds and compare their language use in specific settings. This is only possible if the meta-data make it possible to give an exact specification of the required selection.

2.3. Nature of the meta-data

With the Spoken Dutch Corpus, meta-data are collected about the recordings (where they originated from, how they were made, with what equipment and under what conditions, what type of speech they exemplify – monologue/dialogue/multilogue, spontaneous speech/ prepared/more or less scripted, formal/informal, broad-cast/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 meta-data will be made available that relate to the fashion in which the data were processed. Thus, for the recordings and for each level of transcription and annotation, the meta-data 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 meta-data also include cross-reference information with regard to data available from other projects or in other forms. For example, in another Flemish-Dutch project other types of spoken data are being collected. To the extent that the same speakers occur in these collections it is indicated that additional data are available. Where – as in the case of read aloud text – a printed text was read out, the reference to this text might be worth including in the meta-data as well.

3. Meta-data from a users' point of view

The information that accompanies the corpus and that we want the user to be able to access is vast and of a rather varied nature. Users of the corpus should be able to access information relating to the contents of the corpus, the texts of which it is composed, the speakers involved, etc. Apart from the information that relates to individual texts and specific speakers, there is also information that relates to the corpus as a whole. This not only includes general information about the creation of the corpus, its copyright holders, conditions of use, etc. but also information about aspects of the corpus' design and the sampling criteria that were used.

Information relating to the overall design of the corpus and the sampling strategy used is essential to a user since it gives insight into the composition of the corpus and the nature of the subsets that are worth defining for exploration purposes. Preferably users should be able to access the corpus through exploration software that immediately ties in with the description of the composition of the corpus. A good example in this respect is the ICECUP software used with the International Corpus of English (ICE; Greenbaum (ed.), 1996). In ICECUP the so-called 'corpus map' provides the user with a menu that presents an overview of the corpus in terms of its components.[3] The user can navigate through the corpus and specify the selection he wants to explore.

As observed above, the information provided with the corpus is of a varied nature. Below different types of information are distinguished and discussed in the light of how users may put them to use.

Descriptive vs classifying information

With the wealth of information that we collect and want the users of the corpus to have access to and to be able address in the specification of a query, a distinction should be made between descriptive information on the one hand, and classifying information on the other hand. Descriptive information has a tendency to become unwieldy, especially when the descriptions are not constrained in any way. Classifying information by its very nature is easier to control. Information that has been used as a sampling criterion typically is of this type. Thus, in sampling, a speaker's age group membership will be used as a criterion rather than his specific age. For users this

[2] An additional advantage is that the meta-data can also be used in the presentation of, for example, frequency counts, so that not only total results can be given but also (sub)totals for various subsets of the data.

[3] For a description of the ICECUP software, see Aarts *et al*. (1998) and also http://www.ucl.ac.uk/english-usage/

also has the advantage that they can immediately use the information to contrast the speech of speakers from different age groups. This is not to say that in exploring the corpus (predefined) classifying information is to be preferred to the exclusion of descriptive information. To have a maximum of flexibility, the user should be able to have access to both descriptive and classifying information,[4] so that whenever he is happy with the predefined classes (if only to see whether the hypotheses underlying the corpus sampling hold or not) he can use these, while in those situations where he would prefer some other classification he can define his own.

Hierarchical structuring of information
In relation also to the previous point, it can be observed that with certain types of information it is desirable to have some sort of structuring. This is true for instance when for the Spoken Dutch Corpus we consider the information relating to the place or region from which a speaker originates. For sampling purposes only a limited number of regions[5] have been distinguished. Additionally, for each speaker it has been recorded as part of the meta-data what his place of origin is. The users interests are well-served by distinguishing sub-regions as an intermediate level, as of the other levels one level (the one used in sampling) is too crude and the other too fine-grained to be practical for exploration purposes.

Factual vs derived information
Most information will be factual information. From time to time, however, users interests are better served by derived information. An example is for instance the information about a speaker's place of residence. For certain users it suffices to give the factual information, naming the place or the code identifying it. For other users, however, it may be far more interesting to know whether the speaker lives in a city, town or village. In such a case it must be considered to include the derived information about the population of the speaker's place of residence.

Information concerning permanent characteristics as opposed to temporary characteristics
Especially when it comes to speaker information, a distinction must be made between permanent characteristics on the one hand, and temporary ones on the other.[6] Especially where, as in the Spoken Dutch Corpus, speakers can occur in more than one sample, collected over a period of time, it is useful to make this distinction. Permanent characteristics include a speaker's sex and date/year of birth. [7] The information given for a speaker will be identical for all samples in which the speaker occurs. Temporary characteristics, on the other hand, are subject to change. An example is the voice quality: a speaker may sound hoarse in one sample, while in another sample the voice is clear.

4. Proposed mark-up schemes

In recent years there has been a growing interest in large-scale language corpora from the language engineering community. Against this background the need for a set of standards for encoding corpora has become the more urgent. While general guidelines such as those proposed by the Text Encoding Initiative (TEI) have been available since the early 1990s, the development of an encoding standard geared specifically to the needs of language corpora, for use in language engineering research, however, is still on-going.

The TEI was established with the purpose of developing "a common encoding scheme for complex textual structures in order to reduce the diversity of existing encoding practices, simplify processing by machine, and encourage the sharing of electronic texts" (Sperberg-McQueen and Burnard, 1994: v). In the course of the project the scope was broadened "to meet the varied encoding requirements of any discipline or application" (ibid.) The TEI (revised) guidelines that emerged from this project were published in 1994. They specify the encoding conventions for a number of key text types and features. Upon publication of these guidelines it was envisaged that work would continue in order to extend the scheme "to cover additional text types and features, as well as to continue to refine its encoding recommendations on the basis of extensive experience with their actual application and use", while it was also anticipated that users of the TEI Guidelines would "in some instances adapt and extend them as necessary to suit particular needs" (ibid.).

The TEI guidelines have formed the basis for the CDIF mark-up scheme used in the case of the British National Corpus (Aston and Burnard, 1998) and also for the Corpus Encoding Standard (CES). Both the CDIF and CES are TEI compliant. While the elements and attributes in CDIF are a 'clean' subset of those proposed by TEI, the CES uses not simply a selection of the TEI set but extends it where necessary "to meet the specific needs of corpus-based work in language engineering" (Ide (ed.), 1996).

For the encoding of meta-data, CDIF and CES make use of the corpus and text headers along the lines proposed in the TEI Guidelines. The general orientation of the TEI and CES proposals towards written language corpora is noticeable from the fact that elements such as <sourceDesc>[8] are considered obligatory, whereas they

[4] It should be observed that in order to retain a maximum amount of flexibility it is necessary to have access to the information in its 'pure' form, i.e. it should be avoided that different pieces of information are combined in one attribute.

[5] In all, eight regions are distinguished, four for Flanders and four for the Netherlands.

[6] The distinction between permanent and temporary characteristics that is introduced here is not to be confused with what den Os (1998: 111) refers to as 'stable' vs 'transient' speaker characteristics.

[7] When we look at the way speaker information is handled in speech databases what we see is that information relating to more permanent characteristics is stored in a speaker database (see also Draxler, 1998: 138).

[8] The <sourceDesc> element in intended to supply a bibliographic description of the copy text(s) from which an electronic text was derived or generated. CDIF which also has to

have no application for use with spoken language corpora, except perhaps in the case of samples where a text was read out. Adaptation of the TEI header in CES includes the following changes (cf. Ide (ed.), 1996):

- elements have been added for more precision in the specifications
- attributes have been added to existing elements
- attribute values have been constrained to allow only a given set of values

The CES proposal for the encoding of spoken language corpora is still under construction.

In next sections the CDIF and CES proposals are discussed with an eye to the information that we want to encode in the case of the Spoken Dutch Corpus.

5. Encoding meta-data for the Spoken Dutch Corpus

When we consider the proposals for the encoding of meta-data we find that on the whole they provide an adequate framework. The various types of information can be accommodated and no major revisions appear to be necessary. The distinct uses of the corpus and text headers, where the one is used for providing information relating to the corpus as a whole and the other providing information that relates specifically to a given text, contributes to the transparency of the way in which the information is encoded. Both header types are discussed below.

5.1. The corpus header

Information common to all samples or specific for the corpus as a whole is included in the corpus header. With the exception of the <sourceDesc> elements, the elements proposed are unproblematic. Most information found in the corpus header is descriptive in nature, providing the user with general background information. The information is not generally used for exploration purposes. An exception is the information encoded in the classification declaration <clas(s)Decl> element which contains the descriptive taxonomy used to classify texts within the corpus.

At this point, until a standardized set of text categories –under construction by the EAGLES Corpus Working Group on Text Typology – becomes available, the need to explicitly provide a descriptive taxonomy in the header remains (cf. Ide (ed.), 1996). What can be observed then is that the text categories distinguished for the BNC are of a completely different nature than the text categories for the Spoken Dutch Corpus. While the BNC uses categories such as informative and imaginative with different subject areas, including natural & pure science, applied science, social science, and world affairs, the descriptive taxonomy employed by the Spoken Dutch Corpus is based upon a number of dimensions, including the following:

1. number of speakers actively participating: two or more vs one, thus distinguishing dialogues/ multilogues from monologues
2. situation: private vs public
3. broadcast vs non-broadcast
4. communicative goal, e.g. informing, entertaining, negotiating
5. direct (face-to-face) vs distanced
6. relationship between speaker(s) and hearer(s)
7. degree of preparedness: spontaneously spoken vs scripted

5.1.1. The text header

It is especially with the text headers that we find that the CDIF and CES mark-up schemes must be adapted somewhat to meet the needs of the Spoken Dutch Corpus. This applies more in particular to the information that is characteristic of spoken language data, more in particular about the speakers and the recordings. While CES does not as yet include a recording statement at all, CDIF's <recStmt> includes only the following attributes (cf. Burnard, 1995: 55):

- type; characterizing the recording in terms of the equipment used to make the recording
- date; specifying the date of the recording
- time; specifying the time of day the recording was made
- dur; specifying the duration of the recording (in seconds)

Adaptation of the proposed standards as we see it involves mainly an extension of the number of attributes and values, so that more detailed information can be encoded relating to how and under what conditions the recording was made (the recording device used, number and type of microphones, number of channels, whether the recording was made indoors or out in the open, in a noisy environment or not, etc.). In a similar fashion we intend to introduce additional attributes to include information about voice quality, speech rate, etc.

Finally, on a somewhat different note, we intend to include in the header (where applicable) both factual and derived information as well as descriptive and classifying information, and to structure information. This should enable the user to exploit the information provided to the full.

6. Conclusion

While the importance of meta-data is abundantly clear, what is less obvious is what information exactly (and with what level of detail) is relevant to users and how it can be made optimally accessible. When we go by the meta-data that we want to encode for the Spoken Dutch Corpus, we can conclude that the encoding standards proposed so far provide an adequate basis but need to be adapted further, more in particular for use with spoken language corpora.

---

cater for spoken data uses a corresponding <srcDesc> element which has two sub-elements, the <recStmt> and the <bibStmt>, neither of which is required in the corpus header "since each text is derived from a different source" (Burnard, 1995: 68).

7. Acknowledgement

This publication was supported by the Netherlands Organization for Scientific Research (NWO) under grant number 014-17-510. Thanks are due to Hans van Halteren for his suggestions and comments on an earlier version of this paper.

8. References

Aarts, B., G. Nelson, and S.A. Wallis, 1998. Using Fuzzy Tree Fragments to explore English grammar. In *English Today* 14: 52-56.

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

Burnard, L., Ed., 1995. *Users Reference Guide for the British National Corpus*. Oxford: Oxford University Computing Services.

Draxler, C. SL Corpus Collection. In D. Gibbon, R. Moore, and R. Winski, Eds., 1998: 119-145.

Gibbon, D., R. Moore, and R. Winski, Eds., 1998. *Handbook of Standards and Resources for Spoken Language Systems*. Vol. 1 *Spoken Language System and Corpus Design*. Berlin New York: Mouton de Gruyter.

Greenbaum, S. Ed., 1996. *Comparing English Worldwide. The International Corpus of English*. Oxford: Clarendon Press.

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

Oostdijk, N., 2000. The Spoken Dutch Corpus. Overview and first evaluation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation - LREC 2000* 31 May-2 June 2000, Athens (Greece).

den Os, E., 1998. SL Corpus Design. In D. Gibbon, R. Moore, and R. Winski, Eds., 1998: 79-118.

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