Experiences from the Spoken Dutch Corpus Project

Nelleke Oostdijk*, **Wim Goedertier**†, **Frank van Eynde**‡,
Louis Boves*, **Jean-Pierre Martens**†, **Michael Moortgat**§,
Harald Baayen¶

*Dept. of Language and Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
{N.Oostdijk, L.Boves}@let.kun.nl

†Electronics and Information Systems (ELIS)
University of Ghent, Sint-Pietersnieuwstraat 41, 9000 Belgium
{Wim.Goedertier, Martens}@elis.rug.ac.be

‡ Center for Computational Linguistics, University of Leuven
Maria-Theresiastraat 21, 3000 Leuven, Belgium
Frank.VanEynde@ccl.kuleuven.ac.be

§ University of Utrecht, OTS
Trans 10, 3512 JK Utrecht, The Netherlands
moortgat@let.uu.nl

¶ Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 XD Nijmegen, The Netherlands
baayen@mpi.nl

Abstract

This paper provides an overview of the ongoing development of a large corpus of spoken Dutch in Flanders and the Netherlands. We outline the design of this corpus and the various layers of annotation with which the speech signal is enriched. Special attention is paid to the problems we have encountered, and to the tools and protocols developed for obtaining consistent and reliable annotations. We also discuss the outcome of a recent external evaluation of our project by an international committee of experts.

1. Introduction

The Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) project aims to develop a corpus of 1,000 hours of speech originating from adult speakers of standard Dutch. 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. 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 complex searches involving multiple annotation layers, while including a rich set of meta-data. 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 (Oostdijk, 2000).

In the process of constructing the corpus, extensive use is made of tools for quality control. They include tools that support the creation of transcriptions and annotations; tools for the automatic alignment of various transcriptions and annotations; tools for checking the consistency within and across various transcription and annotation layers; and tools for validating the format of each type of transcription and annotation.

Now that the Spoken Dutch Corpus Project is approaching its fifth and final year, it seems appropriate to take stock of what has been achieved so far. In this paper we present an account of the experiences gained in the process of compiling and annotating the corpus, together with the results of the mid-term evaluation that was carried out by an international committee of experts.

2. Corpus design and data collection

The design of the Spoken Dutch Corpus, was guided by a number of considerations. First, the corpus should consti-

dialogue	private	spontaneous	face-to-face conversations	3,000,000
dialogue	private	spontaneous	interviews	460,000
dialogue	private	spontaneous	telephone conversations	3,000,000
dialogue	private	spontaneous	business negotiations	175,000
dialogue	public/broadcast	prepared	interviews and discussions	750,000
dialogue	public	spontaneous	discussions, meetings	375,000
dialogue	public	spontaneous	lessons	350,000
monologues	public/broadcast	spontaneous	spontaneous commentaries	250,000
monologues	private	prepared	route descriptions	40,000
monologues	public/broadcast	prepared	current affairs programs	250,000
monologues	public/broadcast	prepared	news	250,000
monologues	public/broadcast	prepared	opinion programs	200,000
monologues	public	prepared	lectures	275,000
monologues	public	prepared	texts read from books	625,000

Table 1: The design of the Spoken Dutch Corpus.. The text types labeled 'dialogues' can also be multilogues. The last column lists the size of each subcorpus in word tokens.

tute a plausible sample of contemporary standard Dutch as spoken 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 (approximately ten million words), the corpus will be comparable in size to, for example, the spoken component of the British National Corpus (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.

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 components, ranging from spontaneous conversations to read-aloud text, see Table 1, each of which can be characterized in terms of its situational characteristics such as communicative goal, medium, and number of interlocutors. The specification of each of the components 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 meta-data.

The meta-data are included in the text and participant headers that are available for each of the samples in the corpus. The design of the headers has been inspired by the guidelines of the Text Encoding Initiative (Sperberg-McQueen and Burnard, 1994) and the Corpus Encoding Standard (Ide, 1996). Integration in the corpus exploration software (COREX) involving a conversion to the IMDI (`http://www.mpi.nl/IMDI`) set has enabled effective access to the meta-data so that these can be used in browsing as well as in searching the corpus.

The collection and acquisition of data has appeared to be more problematic than anticipated, owing to a number of causes. Thus, while it was intended to collect parallel samples for Flanders and the Netherlands, cultural differences make it virtually impossible to collect the same kind of speech in similar situations. For example, spontaneous conversations between familiy members or friends in Flanders are not commonly conducted in standard Dutch, whereas in the Netherlands in such contexts standard Dutch is used predominantly. Moreover, the fact that for all recordings permission must be obtained before the data can be digitized, sampled, and transcribed, has appeared to be an unsettling factor which has made it impossible to obey any firm production scheme. Finally, unforeseen technical complications have delayed the collection of telephone conversations.

3. Transcriptions and annotations

3.1. Orthographic transcription

The orthographic transcription layer is the first annotation layer of the Spoken Dutch Corpus. It is of great importance that the quality of the orthographic transcriptions is high since all other annotation layers rely on it. The orthographic layer is also very essential for the users of the corpus because it constitutes the primary means for searching the corpus and accessing the speech samples. Moreover, the symbolic orthographic representation of the speech samples is much easier to handle than the audio files themselves.

The transcription rules are formalized in the Protocol for Orthographic Transcription. These rules are (as much as possible) in line with the international standards for large spoken language corpora. The final orthographic transcription aspires to be an excellent starting point for a wide range of researchers (speech and language technologists, linguists, lexicologists, phoneticians, etc.).

The orthographic transcription is a verbatim transcription. The utterances are interpreted as little as possible. There is no correction of grammatical errors, no completion of truncated words, etc. The transcription also conforms as much as possible to the spelling conventions of standard written Dutch. Common articulation phenomena (acceptable deletions, insertions and substitutions of sounds) are not indicated. In general, normalized word forms are used, although there are some spoken language phenomena that are indicated with a non-normalized form or by means of a mark-up symbol. The protocol contains a limited list of reduced word forms (like *'k*, *'ns*, *d'r*, *zo'n*, etc.) and pronunciation variants (like *goeie*, *ik snij*, etc.). It also provides a list of mark-up symbols. We use *z for standard Dutch words that are heavily regionally accented. Dialect words and constructions are indicated by *d. There are also symbols for truncated words (*a), intentional or non-intentional mispronunciations and onomatopoeia (*u), interjections (*t), foreign words (*v), and hardly-intelligible words (*x), as well as as symbols for non-linguistic speaker sounds such as coughing, laughing, etc. (ggg), unintelligible words (xxx) and unintelligible proper names (Xxx).

Some other deviations of the conventional spelling are the restriction of the punctuations to full stops (.), question marks (?) and continuations (...), the absence of upper case letters in the beginning of sentences and the use of upper case letters as a mark-up for proper names, titles and abbreviations in a more systematic way than in the standard spelling conventions.

The orthographic transcription layer makes use of several tiers: one tier for every speaker in the fragment of speech, a comment tier, and a background tier. The speaker tiers are divided into chunks. Every chunk corresponds to a particular part of an audiofile and contains the transcription of that part. Chunk boundaries are placed between words which are clearly separated acoustically. The average length of the chunks is about 2 seconds and 95% is less than 4 seconds long (99% is shorter than 6.5 seconds). These chunks play an important role during automatic word segmentation (see below). The background tier is used for describing clearly audible or meaningful background noises. The comment tier is used for commenting on the acoustic characteristics of the recording as a whole.

In general, an orthographic transcription is produced from scratch by one student or freelance worker and is subsequently verified by someone else. Following additional verification with a separate software tool that checks for illegal sequences and executes or sometimes suggests some substitutions, the orthographic transcription is ready as a starting point for the next annotation layers. Based on the feedback of groups working on other annotation layers, many of the remaining errors and inconsistencies are subsequently corrected. This feedback can lead to substitution rules that can be applied to all the orthographic transcriptions, the current ones as well as future transcriptions. Some of these substitution rules can be applied automatically, others need some human interaction.

3.2. POS tagging and lemmatization

The first layers of linguistic annotation concern the assignment of a lemma and a tag to each of the ten million tokens. The lemma is the base form of a word; for most words it is identified with the stem, i.e. the word without inflectional affixes; for verbs, however, it is identified with the infinitive. A tag consists of a part of speech and a number of morphosyntactic features which are associated with that part of speech, such as number for nouns, tense for verbs, and degree for adjectives. For the part-of-speech distinction we employ the classical classification into ten parts of speech, which is also used in the standard reference grammar for Dutch *Algemene Nederlandse Spraakkunst* (ANS; Haeseryn et al., 1997). For the addition of extra features, we follow the recommendations of EAGLES, the Expert Advisory Group on Language Engineering Standards. Adapting them to the specifics of the Dutch language, we have defined a tagset with a relatively high degree of granularity, consisting of 316 different tags. A full description of the tagset and the guidelines for lemmatization is provided in Van Eynde (2001).

For the assignment of tags to tokens we adopt the following principles. First, the units to which the tags are assigned coincide with the units of the orthographic transcription. The two words in *ter plaatse* (at-the-DATIVE place-DATIVE), for instance, are each assigned a tag, preposition and noun respectively. Conversely, a form such as *daarlangs* (there-along) is treated as a single word and is not further analysed as an adverbial pronoun followed by a postposition. Second, the assignment of the tags is governed by formal and morphological criteria, rather than by functional or semantic ones. For instance, due to its morphological characteristics and its distribution, the word *maandag* (Monday) is invariably tagged as a noun, also when it occurs in an adverbial position, as in *ik ga maandag naar Leuven* (litt. I go Monday to Leuven). Third, for all words which are potentially ambiguous with respect to the tagset, we only assign the tag which is relevant in context.

Given the size of the corpus, it is not possible to carry out the tagging and lemmatization in a purely manual way. Therefore, we undertook an evaluation of publicly available tools for automatic tagging and lemmatization, see Zavrel and Daelemans (1999). This led to the selection of MBMA (Memory Based Morphological Analyser) for lemmatization. For tagging, we adopted a system which combines the results of four individual taggers, i.e., the HMM-based TnT tagger, a memory based tagger, a maximum entropy tagger and a Brill tagger. The results of the automatic tagging and lemmatization, which takes place in Tilburg, are manually checked and — if necessary — corrected. This is done in Nijmegen for the Dutch data and in Leuven for the Flemish data. Both centers recruit student-assistants to prepare the data for the releases under supervision of local co-ordinators.

The corrected data are not only used for dissemination, but also for retraining the tagger. Table 2 shows the effect of this retraining on the performance of the individual taggers and the combi-tagger. A summary of the tagset, the guidelines for lemmatization, and the evaluation of automatic taggers and lemmatisers can be found in Van Eynde et al. (2000).

3.3. Lexicon coupling for multi-word units

For the assignment of lemmata we adopt the same word-for-word principle as for the tags. This means that the units to which the lemmata are assigned coincide with the units of orthographic transcription. These, however, do not always coincide with what are intuitively felt to be lexical units. For this reason, we add another layer of annotation

date	26/11/99	06/02/00	08/03/00	12/07/00	23/01/01	08/02/02
number of words	10802	21475	39304	95246	553226	2762712
TnT	89.1	91.6	92.7	93.9	95.3	96.2
MBT	86.5	89.4	91.2	92.0	94.3	95.6
maxent	83.6	89.4	90.1	92.6	95.2	-
Brill	83.3	86.3	87.9	89.9	-	-
Timbl combiner	94.2	94.3	94.3	95.6	96.2	96.6

Table 2: The effect of retraining on the performance of the individual taggers and the combi-tagger.

in which multi-word expressions can be treated as single lexical units. Since it is not always clear which multi-word expressions qualify as lexical units, we limit the identification to three clear-cut cases. First, discontinuous combinations of a verb and a particle, as in *hij belt je op* (litt. he calls you up), are identified with a single verb, i.c. *opbellen*. Second, names which consist of two or more words, such as *Den Haag* and *Gaston Van Den Berghe*, are treated as single units. Third, the same applies to combinations which entirely consist of foreign words, such as *chili con carne* and *ad hoc*.

The identification of these multi-word units will be done for the entire corpus. The Dutch data will be processed in Nijmegen and the Flemish data in Leuven. Since it is more efficient to do this in one pass, as soon as the POS tags and lemmata of all the data are available, the work on this task has so far been of a preparatory nature.

3.4. Further enrichments

For about one million words, four additional annotations, namely a broad phonetic transcription, a manually checked word segmentation, a prosodic annotation and a syntactic analysis will be provided.

3.4.1. Syntactic annotation

Syntactic annotation is carried out by the CCL-group in Leuven for the Flemish data and by the Utrecht OTS-group for the Dutch part. The annotation provides two types of information: categorial information at the level of syntactic constituency, and dependency information to capture the semantic connections between constituents. The annotation format uses datastructures expressive enough to naturally encode dependency relations, also where they are at odds with syntactic constituent structure.

Formally, the annotation structures are directed acyclic graphs (DAGs). The vertices are decorated with a syntactic category label: a POS label for the leaves, a phrasal label for the internal nodes. The edges carry dependency labels. They capture the grammatical function of the immediate constituents of a phrase, distinguishing head, complements and adjuncts.

The CGN tagset tries to strike a balance between informativity and practical usability. It uses 25 phrasal category labels and 34 dependency labels. Conciseness is obtained by giving the labels a context-sensitive interpretation. The MOD label, for example, denotes adverbial modification in verbal domains, but also adnominal adjuncts in noun phrases. Levels of granularity that are bound to lead to inter-annotator discrepancies (such as the twenty kinds of adverbial phrases distinguished in the ANS grammar) are avoided. The rich POS tagset (with 316 labels) is reduced to some 50 distinctions relevant for the dependency annotation. On the other hand, special provisions are made for the annotation of phenomena typical of spoken language. The category label DU (discourse unit) for example, allows for an articulation in terms of dependency notions such as nucleus versus satellite, tags or discourse links. An overview of the tagset can be found in Hoekstra et al. (2001), the full annotation manual is in Moortgat et al. (2001).

The annotation makes full use of the expressivity of DAGs as compared to trees. Discontinuous dependencies result in crossing branches that would be problematic in a conventional syntactic constituent structure format. Allowing items to simultaneously carry multiple dependency roles results in a simple annotation schema for phenomena that would require 'movement' or similar devices in tree-based theoretical frameworks. Finally, annotation graphs with disconnected components are useful to provide partial analyses for interrupted phrases, interpolations and the like.

The syntactic annotation procedure, which like the POS tagging is performed semi-automatically, uses the interactive annotation environment developed within the German NEGRA project (`http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html`). A simple visualisation tool for the annotation graphs is freely available from the Utrecht CGN site (`http://cgn.let.uu.nl`). In a later phase of the project, the CGN exploitation software will provide more advanced display and search facilities for the syntactic annotation.

3.4.2. Phonetic transcription

For many research aims a reliable narrow phonetic transcription of the full CGN would be a major asset. However, providing such transcriptions would require resources far beyond the budget. Moreover, not everybody in the research community is convinced that the concept of a 'reliable narrow phonetic transcription' is at all realistic. Many believe that the degree of detail that one would require from a narrow phonetic transcription strongly depends on the aims and requirements of a specific research project. For example, an investigation focusing on regional differences in the degree of diphthongisation of phonemically monophthong vowels would require another type of detail than a study into the degree of devoicing of fricatives in syllable-initial position. Many researchers believe that it would be better to have a coarse — yet reliable — transcription as part of the corpus, which can be augmented later on by adding the details that are required by a specific project.

A combination of budgetary and scientific considerations has resulted in the decision to restrict the phonetic transcription to a broad phonemic level. The starting point for the transcriptions is a phonemic representation of the orthographic transcription that is generated fully automatically. Work is under way to develop automatic transcription procedures that maximise the 'quality' of the automatic transcription. Automatic phonemic transcriptions will be provided for the full CGN. For approximately one million words the automatic phonemic transcriptions will be checked and corrected by students trained for this task. The work of the students is supervised by trained phoneticians. The procedure for this manual checking is defined in a detailed protocol (Gillis, 2001). The set of symbols that can be used in the transcriptions is derived from the SAMPA set (`www.phon.ucl.ac.uk/home/sampa/dutch.htm`). This set does not contain diacritics, so that the transcription is truly limited to the broad phonemic level.

The design of the internal data structures of the CGN are completely based on the concept of words as units delimited by blank spaces. This principle was carried over to the level of phonemic transcription. However, it is well known that cross-word assimilations and degeminations abound in continuous speech. To retain the one-to-one correspondence between the orthographic words and the phonemic transcriptions, a special notation had to be developed for cross-word degemination. For example, the word sequence *op pad* (on the way) in Dutch is likely to be pronounced as /OpAt/. To restore the link with the orthographic level the notation in the CGN is /Op_pAt/. The /..p_p../ notation stands for a single phoneme /p/, of which it is impossible to say whether it is the word final phoneme of the first, or the word initial phoneme of the second word. Phoneme insertion at word boundaries is handled in the same way: Underscores are used to link the inserted phone to both its left and right neighbour word.

It has taken extensive and lengthy discussions to reach agreement on a protocol that is at the same time sufficiently detailed as well as practical. However, the resulting protocol has now been in use for over a year, and our experience is very positive. Transcribers encounter few problems, and if problems do occur, supervisors find it easy to arbitrate.

Evidently, one would want to have a precise estimate of the 'quality' of the manual phonemic transcriptions. However, as yet there are no generally agreed procedures for a formal evaluation of the quality of phonetic transcriptions (Cucchiarini et al., 2001). It remains to be seen whether such procedures become available before the end of the CGN project.

3.4.3. Word segmentation

Effective and efficient access to specific utterances in the corpus, and especially to specific words or types of words requires some kind of low level 'segmentation' of the recordings. In other words, time markers must be added to the recordings at a sufficiently fine-grained level. One might think that the chunk level provides sufficient detail, but there are reasons to believe that time markers at the level of individual words are needed for a wide range of research goals. Therefore, it was decided to add time markers at all word boundaries. These markers are generated automatically for the whole corpus. For the part of the corpus that comes with manual phonemic transcription the time markers are checked — and if necessary corrected — by hand. The details of the automatic and manual provisioning of time markers is discussed in detail in another paper in this proceedings (Martens et al., 2002). Therefore, it suffices to present general information in this paper.

The coupling of manual phonemic transcription and manual verification of time alignment is easy to explain: The automatic aligner takes the phonemic transcription, and finds the best alignment between the speech signal and the sequence of phonemes. This alignment is expected to become more precise as the phonemic transcription is a better representation of the actual speech signal. Moreover, the task of manual verification of word boundaries becomes quite fuzzy if one cannot rely on a verified phonemic transcription. If only an automatic transcription were available, correcting alignments would amount to correcting transcriptions. It follows immediately that the word boundaries that are provided fully automatically cannot be more precise than the automatic phonemic transcription. All discrepancies between these transcriptions and the speech signal at or near word boundaries will be reflected in the word segmentation.

At the start of the CGN project it was investigated whether the quality of this automatic alignment could be improved if the results of two different alignment systems were combined. To that end a fusion system was built that obtains its input from a conventional Hidden Markov Model system and a Speech Segment Model based system. It appeared that the output of the fusion system was virtually identical to one of its inputs, in this case the input of the HMM based system. Therefore, it was decided to use just the HMM aligner to generate the automatic word segmentation.

The protocol for manual verification of the automatic alignment is presently only available in Dutch (Binnenpoorte, 2002). The task of the verifiers — who are trained on the job before they start working — is to check whether the automatically generated word boundaries are reasonable. The major criterion is the ability to recognise the words if only the signal between the boundaries is played back. As expected, it appeared that this criterion cannot be strictly enforced. Especially short and often reduced function words may be difficult to recognise. In many cases this failure can already be predicted from the transcription, e.g., when a word only appears as a clitic sound. The verifiers know that the phonemic transcriptions from which the word boundaries are derived have been approved by experts. Therefore, they are only allowed to report obvious errors that have escaped attention. In all other cases their task is to find the best possible alignment between the signal and the phonemes in the transcriptions that appear in word initial/final position.

3.4.4. Prosodic annotation

Understanding the prosodic mechanisms dominating the spoken communication between humans is of great importance for the further development of human-machine dialog systems. In order to make progress in this area, one needs large prosodically labeled corpora. As there are no such corpora for Dutch as yet, we decided to perform a prosodic annotation of one quarter of the one million words that were selected for further enrichment.

After having consulted potential users of the prosodic annotation, it was decided to strive for a perceptually based annotation as, e.g., in Portele and Heuft (1995) and Grover et al. (1998). This annotation can serve as a starting point for further detailed prosodic labelings, e.g., a ToBI labeling as used by Wightman and Rose (1999).

The aims of the proposed annotation are : (i) to mark syllables carrying a clear prominence, (ii) to locate important between-word and within-word interruptions of the normal speech stream (henceforth called 'breaks'), and (iii) to mark unusual lengthening of individual vowels and consonants not carrying prominence. To simplify things even further, a syllable is either marked as prominent or not (no different degrees of prominence), and a break can either be weak or strong. Clearly, the proposed annotation scheme constitutes a compromise between what is desirable information for a large number of users, and what can actually be labeled with a sufficiently high degree of consistency at a limited cost.

During manual annotation, the transcribers are looking at a computer screen with a display of the signal together with an orthographic transcript. So as to preserve as much as possible the perceptual nature of the annotation, and to reduce any bias towards putting breaks at syntactic boundaries, all punctuation is removed from the orthography. On the other hand, as the manually checked word segmentation also delimits clear pauses between words, it was possible to split up the speech in phrase-like units in an automatic way on the basis of this information. The displayed orthography was therefore synchronized with the signal at the level of the phrases. The automatic phrasing is designed in such a way that it produces units that are no longer than 10 seconds, and that are separated by long pauses (typically longer than 300 ms) which always correspond to strong breaks.

Since the prosodic annotation is to be performed by non-expert transcribers (students) working at four different sites, under the direction of four different supervisors, it is very important to install mechanisms for enforcing a maximum degree of consistency between students and sites. Two important actions were taken in this respect.

1. Since prominence and break strengths are basically ordinal variables which are to be labeled on a 2 and 3-point scale respectively, it is important to reach a common understanding of these labels. Therefore, we developed a written protocol providing examples and describing the general rules and procedures to follow during the annotation.

2. Since the textual examples in the protocol are mainly suggestive, they are supplemented with real examples of speech fragments and their prosodic annotation. These real examples are supplied in the form of a learning corpus for which the supervisors created a concensus annotation.

Two learning corpora (one for Dutch and one for Flemish), each containing 15 minutes of speech were designed to test and refine the protocol. The main findings (Buhmann et al., 2002) are that correlations between annotations of a single transcriber and a consensus of three others range from 0.65 to 0.80 for prominence, and from 0.90 to 0.95 for break strengths, and that it takes on average 40 minutes to annotate one minute of speech. Based on the experiences gathered during the pilot study, a plan for the production of the annotations has been worked out in which two independent annotations made at different sites of each file will be provided.

4. Quality control and consistency

To maintain consistency between the annotation levels and to obtain optimal quality control, we have developed procedures for validation and bug-reporting such as:

- Transcriptions and annotations of one transcriber/annotator are checked by another transcriber/annotator.

- In so far as one type of annotation builds on an other type (as POS tagging on orthographic transcription, but also — for part of the material — syntactic annotation on POS tagging), this automatically involves a verification of the output of a previous annotation; Upon the detection of what is perceived to be an error, a bug report is filed with the team responsible for the annotation.

- All words (tokens) and lemmas in the orthographic transcriptions are validated against the lexicon, as are all combinations of token-tag pairs.

- Quality checks are also made on the basis of the information in the frequency lists that are produced regularly: Low frequency items typically help to pinpoint potential errors, while alternative entries for one and the same item help to identify inconsistencies.

Tools that we have found useful for quality control and consistency are:

- A script to automatically convert a printed text version to a version that conforms to a large extent to the protocol and can be used in the transcription process,

- A customized version of a spelling checker (which helps to conform to some of the conventions adopted in the protocol for orthographic transcription),

- A script to automatically expand numbers represented as digits to their full written forms,

- an XML parser for validating the format of the datafiles,

- a POS tagger for automatically tagging the corpus and the MBMA lemmatizer,

- a tag selection program that is used for the manual verification of the tagger output,

- the Annotate software for syntactic annotation,
- a grapheme-to-phoneme converter for automatically generating a phonetic transcription,
- an automatic word segmentation tool for generating an initial word segmentation,
- and an automatic phrasing tool for generating the phrases to be annotated prosodically.

5. External evaluation of the CGN project

During the summer of 2001, a mid-term external evaluation of the CGN project was performed. The evaluation consisted of a technical evaluation of the first three intermediate releases of the corpus (the data made available in April 2001), and a scientific evaluation of the project as a whole.

5.1. Technical evaluation

The technical evaluation was performed by BAS (Bavarian Archive for Speech signals) under the direction of Christoph Draxler. In a formal validation part, BAS checked the correctness of the file names and formats, the completeness of the data, the consistency between data and metadata, and the quality of the documentation. In a content evaluation part, it checked the validity of the signals and their transcriptions (orthography, POS tags and lemmas).

For the content validation, the aim was to perform a large scale evaluation (on 3 hours of speech) that would reflect the way potential users of the CGN would assess the transcriptions. Therefore, BAS was not asked to create independent transcriptions, but rather to check the CGN transcriptions against the signal and the transcription protocols. The validation was performed by native speakers of Dutch and Flemish, and was carried out on 84 samples that were randomly selected from the 14 main components of the CGN.

The formal validation showed that the bulk of the data was formally correct. Some minor errors were discovered, however, which have already been corrected in the fourth release. The content validation clearly demonstrated that the manual annotations meet international standards. To quote from the BAS report : ”Compared to SpeechDat and comparable speech data collection efforts, the CGN corpus shows good to very good results”.

In its evaluation report, BAS formulates a number of recommendations. Some were suggestions for further increasing the usability of the corpus (e.g., provide more information on recording conditions, include format conversion tools), others for maximally enabling the addition of new enrichments to the data (e.g., make format checkers available, provide tools for subcorpus extractions, etc.), and still others for optimization of the corpus distribution (e.g., use DVD’s for the speech files, make annotations available on the Web).

5.2. Scientific evaluation

The sponsors of the CGN project also wanted a scientific evaluation of the CGN project by a panel of international experts. The chairman, Reinier Salverda (University College London), and his team consisting of Steven Bird (LDC), Jan Hajic (University Prague) and Harald Höge (Siemens, Munich) read the BAS evaluation report, and had access to all the CGN documentation that was available in English. In addition, they had a full-day discussion with the CGN Board and the CGN Steering Committee, as well as with members of the CGN User Group.

Based on all this input, the panel was requested to draw up an evaluation report. This report (Salverda et al., 2001) is publically available on the CGN website (http://www.elis.rug.ac.be/cgn). It provides answers to questions regarding the design of the corpus, the choices that were made in defining the annotation protocols, the technical evaluation, etc. It also contains recommendations for future developments (e.g., develop ideas and plans for research projects that will use the CGN, continue the validation of new annotations as soon as they become available).

In summary, the mid-term evaluation of the CGN project first of all confirmed that our product (the CGN corpus) in its present state (releases 1 to 3) meets international standards. But furthermore, the evaluation also produced a list of valuable recommendations which will definitely raise the quality of the final product.

6. Acknowledgements

This publication was supported by the ‘Spoken Dutch Corpus’ (CGN-project) project which is funded by the Netherlands Organisation for Scientific Research (NWO) and the Flemish Government. We are indebted to Antal van den Bosch for the data of Table 2.

7. References

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

D. Binnenpoorte. 2002. Protocol voor manuele verificatie van automatisch gegenereerde woordsegmentaties. Internal publication CGN. lands.let.kun.nl/cgn/publicat.htm.

J. Buhmann, J. Caspers, V. van Heuven, H. Hoekstra, J.P. Martens, and M. Swerts. 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken Dutch corpus. In *Proceedings of LREC 2002*, Las Palmas.

C. Cucchiarini, D.M. Binnenpoorte, and S.M.A. Goddijn. 2001. Phonetic transcriptions in the spoken Dutch corpus: How to combine efficiency and good transcription quality. In *Proceedings Eurospeech 2001*, pages 1679–1682, Aalborg, Denmark.

S. Gillis. 2001. Protocol voor brede fonetische transcriptie. Internal publication CGN. lands.let.kun.nl/cgn/publicat.htm.

C. Grover, J. Facrell, H. Vereecken, J.P. Martens, and B. Van Coile. 1998. Designing prosodic databases for automatic modeling in 6 languages. In *Proceedings of the 3rd ESCA/COCOSDA workshop on Speech Synthesis*, pages 93–98, Jenolan Caves.

W. Haeseryn, K. Romijn, J. de Rooij, and M. C. Van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Wolters-Noordhoff, Groningen.

H. Hoekstra, M. Moortgat, I. Schuurman, and T. van der Wouden. 2001. Syntactic annotation for the spoken Dutch corpus project (CGN). In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, editors, *Proceedings of the 3rd ESCA/COCOSDA workshop on Speech Synthesis*, pages 73–87, Amsterdam. Rodopi.

N. Ide. 1996. Corpus encoding standard. http://www.cs.vassar.edu/CES/.

J.P. Martens, D. Binnenpoorte, K. Demuynck, R. Van Parys, T. Laureys, W. Goedertier, and J. Duchateau. 2002. Word segmentation in the spoken Dutch corpus. In *Proceedings of LREC 2002*, Las Palmas.

M. Moortgat, I. Schuurman, and T. van der Wouden. 2001. Syntactische annotatie. CGN rapport. Technical report, OTS, Universiteit Utrecht en Centrum voor Computerlinguistiek, K.U.Leuven. http://lands.let.kun.nl/cgn/publicat.htm.

N. Oostdijk. 2000. The spoken Dutch corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume 2, pages 887–893, Paris. ELRA.

T. Portele and B. Heuft. 1995. Two kinds of stress perceptions. In *Proceedings of the ICPhS*, pages 126–129, Stockholm.

R. Salverda, S. Bird, J. Hajic, and H. Hoge. 2001. Mid term evaluation of the spoken Dutch corpus project. Internal publication CGN. `lands/let/kun.nl/cgn/publicat.htm`.

C. M. Sperberg-McQueen and L. Burnard. 1994. Text encoding initiative (TEI) guidelines for electronic text encoding and interchange. ACH-ACL-ALLC.

F. Van Eynde, J. Zavrel, and W. Daelemans. 2000. Part of speech tagging and lemmatisation for the spoken Dutch corpus. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume 3, pages 1427–1433, Paris. ELRA.

F. Van Eynde. 2001. Part of speech tagging en lemmatisering. Technical report, Centrum voor Computerlinguistiek, K.U.Leuven. http://lands.let.kun.nl/cgn/publicat.htm.

C. Wightman and R. Rose. 1999. Evaluation of an efficient prosody labeling system for spontaneous speech utterances. In *Proceedings of the IEEE Automatic Speech recognition and Understanding Workshop (ASRU) (Keystone)*.

J. Zavrel and W. Daelemans. 1999. Evaluatie van part-of-speech taggers voor het corpus gesproken nederlands. Technical report, CGN rapport. K.U.Brabant, Tilburg.