A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS)

Eline Westerhout, Paola Monachesi

Utrecht University, Uil-OTS
Trans 10, 3512 JK Utrecht, The Netherlands
{Eline.Westerhout, Paola.Monachesi}@let.uu.nl

Abstract

In this paper, a pilot study for the development of a corpus of Dutch Aphasic Speech (CoDAS) is presented. Given the lack of resources of this kind not only for Dutch but also for other languages, CoDAS will be able to set standards and will contribute to the future research in this area. Given the special character of the speech contained in CoDAS, we cannot simply carry over the design and annotation protocols of existing corpora, such as the Corpus Gesproken Nederlands or CHILDES. However, they have been assumed as starting point. We have investigated whether and how the procedures and protocols for the annotation (part-of-speech tagging) and transcription (orthographic and phonetic) used for the CGN should be adapted in order to annotate and transcribe aphasic speech properly. Besides, we have established the basic requirements with respect to text types, metadata, and annotation levels that CoDAS should fulfill.

1. Introduction

The Corpus Gesproken Nederlands ('Spoken Dutch Corpus', CGN) (Oostdijk et al., 2002) represents an important resource for the study of contemporary standard Dutch, as spoken by adults in the Netherlands and Flanders. However, it only contains speech from adults with intact speech abilities. There is the need to develop specialized corpora that represent other types of speech, such as aphasic speech. It is for this reason that we have carried out a pilot study for the development of such a corpus for Dutch: CoDAS, a Corpus of Dutch Aphasic Speech (Westerhout, 2006). In this study, we have established the basic requirements with respect to text types, metadata and annotation levels that this corpus should fulfill. Furthermore, we have investigated the challenges that aphasic speech poses for orthographic transcription and linguistic annotation.

Given the special character of aphasic speech, we cannot simply carry over the design and the annotation protocols of existing corpora, such as CGN or CHILDES (MacWhinney, 2000). However, they have been assumed as starting point. For the orthographic transcription, the phonetic transcription and the part-of-speech tagging, we have investigated whether and how the existing procedures and protocols written for the annotation and transcription of the CGN could be adapted in order to make them suitable for the annotation and transcription of aphasic speech.

2. Aphasia

The abilities to understand and produce spoken and written language are located in multiple areas of the brain (i.e in the left hemisphere). When one of these areas or the connection between them is damaged, the language production and comprehension becomes impaired. This language impairment is called "aphasia". In the Netherlands, about 30,000 people suffer from aphasia. In 85% of the cases, the cause of aphasia is a CVA (stroke). Other causes are traumatic brain injuries (12%) and brain tumors (3%) (Davidse and Mackenbach, 1984).

Language impairments differ depending on the location and size of the damage. As a consequence, different aphasia varieties can be distinguished. The main varieties are Broca's aphasia, Wernicke's aphasia, and global aphasia. Individuals with Broca's aphasia frequently speak in short, meaningful phrases that are produced with great effort. Broca's aphasia is thus characterized as a nonfluent aphasia. Function words such as *is*, *and*, and *the* are often omitted. Individuals with Wernicke's aphasia may speak in long sentences that have no meaning, add unnecessary words, and even create new "words". Persons suffering from global aphasia have severe communication difficulties and will be extremely limited in their ability to speak or comprehend language.

However, most aphasia patients do not neatly fit into one of the existing categories. Their speech bears characteristics of different types of aphasia. For the purpose of our investigation, it was sufficient to distinguish between fluent and nonfluent aphasia. Nonfluent aphasia is characterized by heavy syntactic disorders in which inflectional affixes and function words are often missing whereas in fluent aphasia the syntax is not the main problem, but language comprehension and language repetition are damaged. The patients participating in our pilot study were all suffering from nonfluent aphasia.

3. Corpus Design

CoDAS can become an indispensable tool for linguistic research on aphasia since it will offer a considerable amount of speech data. Collecting data is a very time consuming enterprise due to the language impairment of the patients and permission issues (section 3.1.). It is for this reason that each researcher gathers his own data and is not willing to share it. CoDAS could change this state of affairs since the data included in the corpus could be made accessible to all researchers. The corpus will be relevant not only for research on language and speech processing, but also for the development of real life speech applications and for the creation of programs for diagnosing patients. Speech and language therapists could also benefit from it. Given the lack of resources of this kind not only for Dutch but also for other languages, CoDAS will be able to set standards and it will contribute to the future research in this area. Therefore, the corpus should fulfill at least the following requirements which will be discussed in more detail in the rest of this section. First, it should constitute a plausible sample of contemporary Dutch spoken by aphasic patients. Important issues are the inclusion of the different aphasia varieties and various communicational settings (section 3.2.). Second, the speech fragments have to be well-documented with metadata about the aphasic speakers (section 3.3.). Finally, the corpus should be enriched with linguistic information, such as part-of-speech tags, syntactic and prosodic annotation, as well as phonetic transcription (section 3.4.).

3.1. Obtaining permissions

As already mentioned, one of the problems related to the collection of aphasic speech data is the fact that obtaining permission for recording and using data from aphasic patients is not straightforward. Even if aphasic speakers give researchers permission to record their speech and to make it available to others, this does not automatically permit public access to their speech data. The Medical Ethics committee has to grant permission for public access to their speech.[1]

Ideally, we would like CoDAS to include authorized access to the original recordings. In case the permission for including the recordings cannot be obtained, it is important that the transcriptions are as detailed as possible. Except for privacy information, everything should be represented in the transcriptions. Whether or not the recordings are available to others will influence the choice for the transcription and annotation levels to be included (e.g. prosodic annotation is only interesting when the recordings are available).

3.2. Text types

CoDAS should encode a plausible sample of contemporary Dutch as spoken by aphasic patients, that is it should include speech representing different types of aphasia (Broca, Wernicke, global, transcortical, anomic, etc.) as well as various communication settings. Interviews between a non-aphasic person and an aphasic person such as the ones carried out in the context of the Aachen Aphasia Test (AAT) could be included.

The AAT was used to diagnose the patients involved in the pilot study. It consists of six subtests, of which five constitute text types that can be included in CoDAS. In the first subtest, the 'Spontaneous Language Sample', the patient has to answer questions on five standard topics. These samples should contain at least 300 words of the aphasic patient and three or more of the five topics should have been discussed during the conversation. The four other useful parts of the AAT involve a repetition task, a writing task, a naming task and a comprehension task. These parts are useful to researchers to gain insights on the abilities to repeat, write, name, and comprehend language.

Other possible communicational setting are conversations of the aphasic patients at home, and in aphasia centers. Depending on the type of aphasia, other types of speech can be acquired. For instance, reading aloud is only possible for a restricted group of patients.

3.3. Metadata

Metadata play an important role in enhancing the usability of the collected data, for example it can be used to define and access precisely those subsets of data that are relevant for the user. However, because of the special character of the corpus of aphasic speech, not only general information about the patients needs to be collected (e.g. age, gender, place of residence) but also some more specific features. For example: time post-onset (how long has the patient been aphasic at the time of speaking), cause of aphasia, paralysis (aphasia can be accompanied by paralysis of one or more parts of the body, most times the right part of the body is paralyzed), handedness, verbal apraxia (articulation disorder as a result of problems in planning the articulation movements), dysarthria (a speech impairment as a result of a neurological disorder), type of aphasia, and severity of aphasia (according to the AAT).

3.4. Different levels of annotation and transcription

As in other corpora, orthographic transcription is required in a Corpus of Aphasic Speech because it serves as basis for all other annotation and transcription levels.

Depending on the research questions to be answered, phonetic transcription can also be relevant. Aphasic patients often make phonetic or phonological errors and frequently encounter articulation problems. The phonetic annotation can provide users with information about these errors which wouldn't be accessible via the orthographic transcription, that makes use of standard spelling conventions. Ideally, speech and video recordings should be attached to the transcription in order to be able to listen and watch the fragments on request. A grapheme-to-phoneme converter can be used to perform the phonetic transcription automatically. The orthographic transcription forms the basis for the conversion.

Information about part-of-speech should be provided since it can shed light on questions about the word classes which are typically left out by patients. Researchers might be interested in, for example, the number of used verbs, finiteness of the verbs, used determiners, the relation between determiners and finiteness, the number of pronomina, etc.. The part-of-speech tagging can be performed automatically. For the tagging of Dutch text several taggers are available (Zavrel and Daelemans, 1999). However, existing taggers need to be adapted in order to produce a reasonable level of accuracy of aphasic speech annotation.

Syntactic annotation should also be included in a Corpus of Aphasic Speech since aphasia often influences the syntax of speech. Several parsers are available for the syntactic annotation of Dutch texts, however, also in this case they have to be adapted to be able to deal with ungrammatical sentences, uncomplete sentences and sentences with mirror constructions.

The prosody of nonfluent aphasic patients is often damaged because of the efforts the patients make in the production of speech. Just as for the phonetic transcription, it will be important to have the speech and video recordings attached to the transcriptions. Whether or not these recordings are available will influence the choice for including the prosodic annotation.

[1] This is the case in the Netherlands, the situation might be different in other countries.

4. The pilot study

A pilot study has been carried out to investigate to which extent existing annotation and transcription protocols already developed for corpora such as CGN or CHILDES could be adopted for the setup of CoDAS. To this end, speech material of aphasic patients has been collected and annotated on the basis of the existing protocols which have been revised accordingly.

4.1. Patients

Speech material of six aphasic patients has been collected. The average age of the patients was 54 and the time post onset was between three and four years. The six patients could not be assigned to one variety according to the AAT, which was conducted by a qualified Speech and Language Pathologist. However, they were all diagnosed as having a nonfluent aphasia according to this test. To determine the fluency, the sixth score on the Spontaneous Language Sample subtest indicating the syntactic structure has played a major role. The results on the subtest 'Spontaneous Language Sample' of the AAT were used as speech samples for the pilot study.

4.2. Relevant corpora for the pilot study

Two corpora have been of particular relevance for our pilot study and have been used as starting point for the definition of the transcription and annotation protocols, that is the CHILDES corpus and the CGN.

The CHILDES corpus is important because the kind of speech which has been collected within this project also deviates from 'normal' speech. It contains mainly speech data of young monolingual (normally developing) children interacting with their parents or siblings. The database has later been extended with transcripts of children with language disorders (e.g. Down syndrome, autism), bilingual children, second-language learning adults, and aphasics. However, the majority of the corpora still contain speech from English normally developing children. The CHILDES manual (MacWhinney, 2000) presents coding systems for phonology, speech acts, speech errors, morphology, and syntax. The user can create additional coding systems to serve special needs. The CHILDES guidelines have been a reference for the development of the protocols which will be used in the annotation of CoDAS.

The second corpus of interest in our pilot study is the CGN given that it is also a corpus of spoken Dutch. The CGN is a database of contemporary standard Dutch as spoken by adults in the Netherlands and Flanders. The corpus comprises approximately ten million words (about 1,000 hours of speech), two thirds of which originates from the Netherlands and one third from Flanders. It contains a large number of samples of spoken text recorded in different communicational settings. The entire corpus has been transcribed orthographically, and the transcripts have been linked to the speech files. Lemmatization and part-of-speech tagging are performed for the whole corpus. For a selection of one million words, a (verified) broad phonetic transcription has been produced, while for this part of the corpus also the alignment of the transcripts and the speech files has been verified at the word level. In addition, a selection of one million words has been annotated syntactically. Finally, for an even more restricted part of the corpus (approximately 250,000 words) a prosodic annotation is available.The extensive protocols written for the different transcription and annotation levels of the CGN were used as starting point for the pilot study.

4.3. Orthographic transcription

Transcribing spontaneous speech is a difficult task because the speech is not fluent and contains filled pauses, mispronunciations, false starts, and repetitions. In addition, it is often difficult to distinguish utterance boundaries.

Within the pilot study, the orthographic transcription protocol of the CGN has been used to transcribe aphasic speech. This protocol is based on the EAGLES guidelines developed for the transcription of spontaneous speech and has been adapted for the transcription of typical Dutch phenomena. The three criteria underlying the orthographic transcription protocol of the CGN are (Goedertier et al., 2000):

- Consistency: in order to increase consistency, standard spelling conventions are maintained. However, in a number of cases it is necessary to deviate from standard conventions to transcribe accurately what has been said. For example, when a word is not finished, only the part of the word that has been uttered should be transcribed. For indicating such problematic issues special symbols were defined.
- Accuracy: to improve the quality of the transcriptions, all orthographic transcription files were checked by a second transcriber
- Transparency: the number of transcription rules are kept down to a minimum. This makes it easier to memorize and apply them.

4.3.1. The transcription of the nonfluent aphasic speech

The orthographic transcription protocol of CGN has been adopted for transcribing the aphasic speech. Although the transparency criterion is very important, some typical problems frequently present in aphasic speech ask for additional rules. Three of such problems are discussed in more detail below, namely the interjections problem, the word finding problem and the boundaries problem.

Interjections

Nonfluent aphasic patients need much time to think and utter many interjections (most times uh and uhm). According to the CGN guidelines, all interjections have to be transcribed:

Example 4..1 (Interjections - a) *en toen **uh** ben ik **uh uh** en toen ben ik **uh** ggg ben ik **uh pff*t** toen ben ik **uh** in **uh** weet ik niet nou **uh** ga ver*a ga ver*a .*

(*and then **uh** I am **uh uh** and then I am **uh** ggg I am **uh pff*t** then I am **uh** in **uh** I don't know well **uh** go o*a go o*a .*)

Although the interjections may not seem very informative at first sight, they can give an indication of the efforts it costs to produce speech. However, leaving them out of the transcription is not a good option. The transcription of the sentence then becomes:

Example 4..2 (Interjections - b) *en toen ben ik en toen ben ik ggg ben ik toen ben ik in weet ik niet nou ga ver*a ga ver*a .*

If this option is adopted, information about the conversation is lost. Readers of the transcription get a completely wrong view of the conversation: it seems that the aphasic patient has a fluent production. The conversation also becomes more difficult to interpret because interjections can also indicate a new attempt of the aphasic speaker to convey the message in an other way.

We devised a third option to transcribe the interjections properly. First, we thought of counting the interjections and indicating in the transcription how many interjections where uttered. However, whether this would be a good way to measure speaking effort, is doubtful. A speaker can say “uh, uh, uh” a number of times in succession, but it is also possible that a speaker says “uhhhhhhhhhhh”. In this case one “uh” can last as long as five or six “uh”'s. To measure the effort, it is more relevant to know the time employed by the speaker to produce the relevant utterance. So, the best solution would be to indicate filled pauses (`<fp>`) and to link the transcriptions to the recordings, in order to include information on the timespan. The orthographic transcription then becomes:

Example 4..3 (Interjections - c) *en toen <fp> ben ik <fp> en toen ben ik <fp> ggg ben ik <fp> toen ben ik <fp> in <fp> weet ik niet nou <fp> ga ver*a ga ver*a .*

Adopting this option makes it easier to perform the orthographic transcription and little information is lost.

Word finding problems

By definition, all nonfluent aphasic patients experience word finding problems. While searching for the right word, they may produce several other related words. We believe it is relevant to mark words and phrases uttered during the word finding process since in this way we will increase the readability and make it possible to filter out these words. It will also be possible to find out which word categories typically cause word finding problems.

The patients involved in the pilot study encountered difficulties in finding numerals, geographical locations (e.g. *France*) and time indicators (e.g. *week*, *month*). In the example below, the patient searches for the numeral *twaalf* (’twelve’).

Example 4..4 (Word finding - a) *uh toen uh toen uh ben ik uh xxx Rijndam en toen ben ik negen negen tien nee uh negen m*a uh negen nee geen negen elf tien elf twaalf twaalf weken nee maanden twaalf maanden uh uh hoe heet dat in uh Rijndam geweest .*

(*uh then uh then uh I have uh xxx Rijndam and then I have been nine nine ten no uh nine m*a uh nine no not nine eleven ten eleven twelve twelve weeks no months twelve months uh uh how do you call it in uh Rijndam .*)

In the orthographic transcription according to the CGN guidelines, it is not possible to indicate that all the uttered numbers are produced during the word finding process of the number *twaalf*. In one of the CHILDES corpora, the Holland Corpus, this is encoded by putting the words that are uttered during the word finding process between angle brackets. This makes it easy to filter out the relevant words. Below a possible way to indicate this is shown:

Example 4..5 (Word finding - b) *uh toen uh toen uh ben ik uh xxx Rijndam en toen ben ik negen*wf(twaalf) negen*wf(twaalf) tien*wf(twaalf) nee uh negen*wf(twaalf) m*a uh negen*wf(twaalf) nee geen negen elf*wf(twaalf) tien*wf(twaalf) elf*wf(twaalf) twaalf*wf(twaalf) twaalf weken nee maanden twaalf maanden uh uh hoe heet dat in uh Rijndam geweest .*

It is also possible that a word is not found at all. Words produced during the word finding process can be marked then with `*wf`, without the word to be found indicated between brackets thereafter.

Distinguishing utterances

Nonfluent aphasics patients speak in short, often ungrammatical phrases with many pauses. They generally leave out function words and word order is disturbed. It is very difficult to specify utterance boundaries since sentences are often not completed or finished after another sentence has been produced. It would help the transcriber if guidelines to detect the boundaries are given.

Although distinguishing utterances will always remain a subjective issue, it is possible to define some guidelines that could be used to decide where a new utterance starts. One possibility is to look for a topic shift. When this would be the case, it could be a clue to start a new utterance. Another option is to look for pauses. When a long pause is ’heard’, this could be a clue for starting a new utterance. However, while this might be a good clue in speech from persons without speech disabilities, this is not always the case in aphasic speech. Pauses are very common in this kind of speech, since they are also used within utterances. Even in normal speech a pause does not always mark a boundary. A third clue could be the intonation pattern (Wijckmans and Zwaga, 2005): a decreasing intonation pattern indicates an utterance boundary. However, intonation might be disturbed since aphasic patients often speak in a rather monotonous tone.

4.4. Phonetic transcription

The phonetic transcription reflects the exact pronunciation of words and sentences. Phonemic transcription, or ’broad phonetic transcription’, is the most common type of phonetic transcription which is also used in the CGN and we have assumed it for the the aphasic speech samples, as well. The grapheme-to-phoneme conversion program TreeTalk was used to generate the phonetic transcriptions. This is a memory-based word phonemization system trained on CELEX. It takes as its input the spelling of words, and produces as its output the phonemic transcription according to the rules implicit in the training data (Daelemans and van den Bosch, 2001).

4.4.1. The transcription of the nonfluent aphasic speech

The program TreeTalk was used to generate the phonetic transcriptions of the aphasic speech automatically. Therefore, as for the CGN, the phonetic transcription of the nonfluent speech is restricted to a broad phonemic level. For a small part of the data, the automatically generated transcriptions have been checked manually and we could not detect relevant problems in this respect. However, the decision for including the phonetic transcription level within a Corpus of Aphasic Speech should be based on whether the audio files can be linked to the corpus.

4.5. Part-of-Speech Tagging

For the part-of-speech tagging, the approach of the CGN was adopted. The results of the automatically performed tagging of the aphasic speech where compared to the results obtained within the CGN project. Because of the size of the CGN, part-of-speech tagging was automated as much as possible. The TiMBL (Tilburg Memory-Based Learner) combitagger was used (Daelemans et al., 2004). This tagger systematically compares the results of four separate working taggers in order to obtain a result that is more accurate then the results the individual taggers can give. The result of the automatic tagging and lemmatization has been verified and corrected manually. The performance of the combitagger on the CGN after retraining was 96.6 % (Oostdijk et al., 2002).

4.5.1. The annotation of the nonfluent aphasic speech

Aphasic nonfluent speech differs from spontaneous speech by persons with intact speech abilities. To investigate how automatic part-of-speech taggers actually perform on nonfluent speech, a subset of the automatically tagged data has been checked manually. The used tagger is one of the four taggers that was incorporated into the combitagger that has been used for the annotation of the CGN, namely the Memory-Based Tagger (MBT) (Daelemans et al., 1996).

One third of the data has been verified manually. For each word, it is indicated whether it is spoken by the aphasic patient or by the interviewer. All tagged words are classified as correct, wrong, interjection or punctuation mark. The interjections and punctuation marks have been separated from normal words because for the aphasic patients 36.6 % of the words consists of interjections and punctuation marks, whereas for the interviewer this is only 19.7 %. The interjections - as far as they are recognized by the tagger - and punctuation marks are always tagged correct. In the comparison of the utterances of the two groups (patients and interviewer), they are left out in order to prevent that the results are influenced by the large number of interjections used. The percentage of words that are assigned a wrong tag is 21.3 % (183/860) for the patients whereas this percentage for the interviewer is only 15.8 % (90/570). This difference is significant, $X^2(1, N = 1430) = 6.688$, $p \leq 0.05$, so the tagger performs better on the utterances of the interviewer.

	Correctness		
Subject	Not correct	Correct	Total
Interviewer	90 (15.8 %)	480 (84.2 %)	570
Patients	183 (21.3 %)	677 (78.7 %)	860
Total	273 (19.1 %)	1157 (80.9 %)	1430

Table 1: Tagger correctness for interviewer and patients

Further evaluation of the data showed that the errors can be divided roughly in five categories (Table 2). The main error categories differed for the two kinds of speech. Within these categories, subcategories can be distinguished. The most occurring problem in the interviewer's speech was tagging the pronoun *je* ('you', 44.4 % within "Same POS-tag"). The problem with *je* was that the tagger often tagged it as an indefinite pronoun instead of a personal pronoun. Within the same category the tagging of capital letters was most problematic for the speech of the aphasic patients (71.9 % of the cases).

The three main reasons for assigning a wrong tag in the aphasic speech were:

- Words marked with a * in the orthographic transcription (29.5 %)
- Unknown interjections, most times *uhm* or *ok* (11.5 %)
- Capitals, e.g. N, A, D (14.2 %)

For the speech produced by the interviewer the main problems were:

- Unknown interjections, most times *uhm* or *ok* (34.4 %)
- Tagging the pronoun *je* as an indefinite pronoun instead of a personal pronoun (13.3 %)

All other errors did not occur frequently and involved, among others, words with diacritic marks (e.g. *één* ('one')) and ambiguous words (e.g. *vier*, which means either "four" or "celebrate").

4.5.2. Improving the performance of the Memory-Based Tagger

There are several ways to improve the performance of MBT on the speech of both the aphasic patients and the interviewer. The performance of MBT heavily depends on the quality of the training corpus. Therefore, the best way to improve the over-all performance accuracy, is to base the tagger on a manually tagged training corpus of the target speech, in this case on speech produced by aphasic patients. This will probably result in a lower error rate, mainly in the common error categories, such as the tagging of capitals. The problem of unknown interjections can be solved by adding them to the vocabulary of interjections. Words marked with an * should be excluded from the tagging process and get no tag at all. Dealing with abbreviated words, such as *da's* (that is) and *'t* (it), should be improved. The abbreviations consisting of two words (e.g. *da's*) should be separated during the tokenization process and tagged as two words. Abbreviations of one word (e.g.

Subject	Category					
	Same POS-tag	Other POS-tag	Unknown interjection	*-word	Other problems	Total
Interviewer	27 (30.0 %)	22 (24.4 %)	31 (34.4 %)	1 (1.1 %)	9 (10.0 %)	90
Patients	32 (17.5 %)	49 (26.8 %)	21 (11.5 %)	54 (29.5 %)	27 (14.8 %)	183
Total	59 (21.6 %)	71 (26.0 %)	52 (19.0 %)	55 (20.1 %)	36 (13.1 %)	273

Table 2: Problematic tagging categories

't) should be learned from the training corpus. Finally, the tagger should be able to deal with words with diacritic marks.

5. Conclusions

The pilot study we have carried out is a preliminary investigation for the setup of a Corpus of Dutch Aphasic Speech. Corpus design issues have been examined and we have especially focused on whether existing annotation and transcription protocols such as those developed within the CGN project or CHILDES could be employed within CoDAS.

We can conclude that the orthographic transcription protocol of the CGN is not completely suited for aphasic speech and special attention has been dedicated to features that are typical of this kind of speech such as interjections, word finding difficulties and the problem of distinguishing utterances. On the other hand, the phonetic transcription program TreeTalk seems to perform quite well on the available data, even though our study is not conclusive since manual investigation has been carried out only on a small set of the data.

The performance of MBT, one of the four automatic part-of-speech taggers used for the tagging of the CGN, on the tagging of the orthographic transcriptions of the Dutch aphasic speech, was worse than the performance of the combitagger on CGN annotation. Some main error categories can be distinguished. Training MBT on a corpus of manually tagged aphasic speech will probably result in a better performance of the tagger. Especially the type of errors contained in the main error categories will cause less problems if the tagger is trained on aphasic speech. The investigation of the problems that aphasic speech constitute for syntactic and prosodic annotation is left for future research.

6. References

W. Daelemans and A. van den Bosch. 2001. TreeTalk: Memory-Based Word Phonemisation. In R. Damper, editor, *Data-Driven Techniques in Speech Synthesis*, pages 149–172. Kluwer Academic Publishers.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. 1996. MBT: A Memory-Based Part of Speech Tagger-Generator. In E. Ejerhed and I. Dagan, editors, *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 14–27.

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2004. TiMBL: Tilburg Memory-Based Learner. Technical report, Induction of Linguistic Knowledge (ILK), Tilburg University.

W Davidse and J.P. Mackenbach. 1984. Aphasia in the Netherlands; Extent of the Problem. *Tijdschrift voor Gerontologie en Geriatrie*, 15(3):99–104.

W. Goedertier, S. Goddijn, and J.P. Martens. 2000. Orthographic transcription of the Spoken Dutch Corpus. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of LREC-2000*, volume II, pages 909–914.

B. MacWhinney. 2000. *Transcription Format and Programs*, volume 1 of *The CHILDES project: tools for analyzing talk*. Lawrence Erlbaum.

N. Oostdijk, W. Goedertier, F. van Eynde, L. Bovens, J.P. Martens, M. Moortgat, and H. Baayen. 2002. Experiences from the Spoken Dutch Corpus Project. In M. Gonzalez Rodriguez and C. Paz Saurez Araujo, editors, *Proceedings of LREC-2002*, pages 340–347.

E.N. Westerhout. 2006. *A Corpus of Dutch Aphasic Speech: Sketching the Design and Performing a Pilot Study*. Ph.d. diss., Department of Linguistics, Utrecht University, Utrecht, The Netherlands.

E. Wijckmans and M. Zwaga. 2005. ASTA: Analyse voor Spontane Taal bij Afasie. Standaard volgens de VKL.

J. Zavrel and W. Daelemans. 1999. Evaluatie van Part-of-Speech taggers voor het Corpus Gesproken Nederlands. Rapport CGN: werkgroep corpusannotatie, Tilburg University.