6

A Pilot Study for a Corpus of Dutch Aphasic Speech (CoDAS)

Focusing on the orthographic transcription

ELINE WESTERHOUT AND 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. We have established the basic requirements with respect to text types, metadata, and annotation levels that CoDAS should fulfill. 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 Spoken Dutch Corpus (CGN) or CHILDES. However, they have been taken as starting point. We have investigated whether and how the procedures and protocols for the orthographic transcription and the part-of-speech tagging used for the CGN should be adapted in order to annotate and transcribe aphasic speech properly.

6.1 Introduction

The Corpus Gesproken Nederlands ('Spoken Dutch Corpus', CGN) (Oostdijk, Goedertier, van Eynde, Bovens, Martens, Moortgat and Baayen 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 speaking abilities. There is the need to develop specialized corpora that represent other types of speech. The JASMIN project has already been dedicated to extending the CGN with speech of elderly people, children and non-natives (Cucchiarini, van Hamme, van Herwijnen and Smits 2006). In our project, we performed a pilot study for the development of a corpus containing aphasic speech: 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 part-of-speech tagging.

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 taken 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. In this paper, we focus on the adaptation of the orthographic transcription protocol of the CGN.

6.2 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 impaired. The patients participating in the pilot study were all suffering from nonfluent aphasia.

6.3 Corpus Design

CoDAS can become an indispensable tool for 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 privacy issues and IPR (section 6.3.1). It is for this reason that each researcher gathers his own data and is not allowed to share it. CoDAS can 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 6.3.2). Second, the speech fragments have to be well-documented with metadata about the aphasic speakers (section 6.3.3). 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 6.3.4).

6.3.1 IPR

As already mentioned, one of the problems related to the collection of aphasic speech data is the fact that obtaining permission for recording and distributing 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. In the Netherlands, the Medical Ethics committee has to grant permission for public access to their speech.

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.

6.3.2 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 nonaphasic person and an aphasic person such as the ones carried out in the context of the Aachen Aphasia Test (AAT) could be included. Other subtests of the AAT can also be used. Conversations of the aphasic patients at home, and in aphasia centers will also be useful text types. (Westerhout and Monachesi 2006) give more information on possible text types that could be included.

6.3.3 Metadata

Metadata play an important role in enhancing the usability of the collected data, for example they 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).

6.3.4 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 would not 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. Video recordings can be helpful, because aphasic patients sometimes use gestures to explain what they mean and as a strategy to find words. As a first step, a grapheme-to-phoneme converter can be used to perform the phonetic transcription automatically. This automatically created transcription has to be corrected manually (Binnenpoorte 2006).

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 (Section 6.4.4).

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 better to have the speech and video recordings attached to the transcriptions.

6.4 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.

6.4.1 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 ŠSpontaneous Language SampleŠ of the AAT were used as speech samples for the pilot study. Table 4 shows the results on the complete AAT and on the subtest 'Spontaneous Language Sample' for the six patients.

6.4.2 Relevant corpora

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 ŠnormalŠ speech. It contains mainly speech data of young monolingual (normally developing) children interacting with their parents or siblings, but there is a small part with transcripts of children with language disorders (e.g.

	Patient						
	1	**2**	**3**	**4**	**5**	**6**	
Spontaneous speech sample	3	3	3	2	4	3	(COM)
	4	4	4	4	5	4	(ART)
	4	4	5	4	5	5	(AUT)
	3	3	4	3	4	5	(SEM)
	3	3	3	3	4	4	(PHO)
	2	2	2	1	2	2	(SYN)
Percentage Aphasia	100	100	100	100	98.4	86.5	
Percentage Broca	14.3	47.4	69.1	99.9	8.1	47.0	
Percentage Wernicke	26.2	52.6	30.8	0.1	21.4	1.2	
Percentage Anomic	59.5	0	0	0	70.5	51.8	
Aphasia type	?	?	?	Broca	Amnestic	?	

TABLE 4 The scores on the AAT of the patients involved in the pilot study

Down syndrome, autism), bilingual children, second-language learning adults, and aphasics. 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 speech samples recorded in different communicational settings. The extensive protocols written for the different transcription and annotation levels of the CGN were used as starting point for the pilot study.

6.4.3   Orthographic transcription

Transcribing spontaneous speech is quite complicated, because it is not fluent and contains filled pauses, mispronunciations, false starts, and repetitions. Besides, it is often difficult to distinguish utterance boundaries. For the transcription of the aphasic speech data the protocols used for the transcription of the CGN and CHILDES have been used and were adapted to make them suitable for the transcription of aphasic speech.

CGN and CHILDES

The orthographic transcription protocol of the CGN is based on the EAGLES guidelines developed for the transcription of spontaneous speech. The protocol is based on three criteria, which were kept in mind while adapting the protocol to make it suitable for the orthographic transcription of aphasic speech. The three criteria underlying the orthographic transcription protocol of the CGN are (Goedertier, Goddijn and Martens 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.

The guidelines for the orthographic transcription of the CGN and CHILDES are both almost entirely based on the EAGLES guidelines. However, at some points complementary guidelines are required to deal with typical Dutch phenomena. Besides, for the transcription of non-speech acoustic events (such as coughing and relevant background noise) guidelines are needed. So the guidelines can be divided into three groups: 1) spelling guidelines corresponding to the EAGLES guidelines, 2) complementary spelling guidelines, 3) guidelines for dealing with non-speech material (e.g. coughing, not finished words).

Spelling guidelines corresponding to the EAGLES guidelines

Reduced word forms: For the CGN, the lexicon contains the most common reduced forms (e.g. *'k* for *ik* ('I'), *da's* for *dat is* ('that is')). The orthographic transcribers have to use the forms that are on this list when they are heard instead of the full forms. In the CHILDES project, parentheses are used to deal with this phenomenon. The sounds that are dropped are shown between brackets, so when a transcriber hears *bout* instead of *about* he transcribes *(a)bout*. For the transcription of Dutch abbreviations, the same procedure has been followed as for the English shortened forms. When a person says *es* instead of *eens* (*just*) this is transcribed as *e(en)s*.

Dialect forms: Dialect words and constructions that do not consist in standard Dutch but are of a rather dialectal nature are followed by *d according to the CGN protocol. Besides typical dialect words, such as *keuje*d* for *varken* ("pig"), there are several constructions that are typical for a specific region. An example of such a dialectal construction is the inflection of articles, pronouns, adjectives, and substantives in the South of the Netherlands (e.g. *nen*d blauwen*d auto* for *een blauwe auto* ("a blue car")). Words that belong to standard Dutch, but are pronounced dialectically, are followed by *z (e.g. *jou* ("you") pronounced as /ju/ is transcribed as *jou*z*). The CHILDES system lets annotators choose one out of four options for annotating dialect forms. The four possibilities are (1) Adding each variant to the lexicon file; (2) Adding the standard form after each variant form; (3) Creating a full phonological transcription of the whole interaction and linking this to an audio file; or (4) Ignoring dialectal variation and transcribing the standard form.

Numbers: For both the CGN and the CHILDES system, numbers have to be written out in words. When a number can be pronounced in more ways, the number should be tran-

scribed in the way it is pronounced (e.g. 1837 can be either *achttienhonderd zevenender-tig* ("eighteen hundred thirty-seven") or *achttien zevenendertig* ("eighteen thirty-seven")). Within the CGN, there also is a second option for transcribing numbers: the numbers 0 up to and including 99, the hundreds, thousands and hundred thousands can also be written as figures (e.g. 1837 can be either 1800 37 or 18 37). These numbers will then be converted automatically into the written form.

Abbreviations and spelled words: In the CGN, spelled words or separate letters are written in capital letters (e.g. *laf* ("cowardly") becomes L A F). When letters are spelled in an alternative way, they are not written in capital letters but in the way they are pronounced, *u is assigned to each separate letter (e.g. *laf* can also be "spelled" as le*u a*u fe*u). Abbreviations do not get a special symbol and are written in the way they are used. When the component letters of a abbreviation are pronounced separately (e.g. as in *t.z.t.* ("in due time")), they are written in capital letters as one word, without white spaces between the component letters (e.g. *BTW* for *BTW* ("VAT"), *TZT* for *t.z.t.*). Acronyms are written in the way the standard spelling prescribes, but always completely in capital letters (e.g. *NASA*, *TROS*).

The CHILDES guidelines differ slightly from the CGN guidelines. For words that are spelled out each separate letter gets the symbol @l (e.g. *word* becomes *w@l o@l r@l d@l*). Acronyms are transcribed by using the component letters as a part of a linked form, the @l marking is not used for acronyms (e.g. *USA* becomes *U_S_A*). Acronyms that are not spelled out when produced are written as words (e.g. *Benelux*). Abbreviations for titles are also written out in their full form (e.g. *Mister* instead of *Mr.*).

Interjections: Both protocol for the orthographic transcription contain a list of frequently used interjections (e.g. *uh*, *hè*). In addition, the CGN protocol has the option to mark interjections that do not appear on the list with *t.

Complementary spelling guidelines - CGN

Use of capital letters: Proper names, such as cities, persons, brands and companies, start with a capital letter. When a proper name consists of more words, each word starts with a capital, even when this is not according to the standard spelling rules (e.g. *Anne Marie Van De Zande*). For titles of books, songs, films, etc., the same rules apply as for proper names.

Pronunciation: All words that are not contained in the lexicon of the CGN and also do not belong to any of the other types are marked with *u. Within these category three kinds of words can be distinguished. The first group are the onomatopoeic words (e.g. *boink*u*). The second group contains the words that are pronounced wrongly, either by accident or on purpose (e.g. *toekenbas*u* instead of *boekentas* ("book bag"), *alduns*u* instead of *aldus* ("thus")). The third group are the mispronunciations and resumptions within words (e.g. *gewee-weest*u* for *geweest* ("been"), *ver-uh-kocht*u* for *verkocht* ("sold")).

Complementary spelling guidelines - CHILDES

Phrasal combinations: In phrasal combinations of different word classes are combined. These include book titles (e.g. *Wuthering Heights*), names of places (e.g. *University of Oxford*), and lines from songs (*With a little help from my friends*). To indicate that these words form a phrasal combination the underscore character is used (*Wuthering_Heights*, *University_of_Oxford* and *With_a_little_help_from_my_friends*).

Unidentifiable material

Unintelligible speech: Words or phrases that are difficult to understand are marked with *x in the CGN. When they are completely unintelligible, the transcriber uses xxx instead of the word or phrase. This corresponds with the way CHILDES deals with unintelligible speech in which this is represented with "xxx" or "xx". The string "xxx" will be ignored when computing the mean length of utterances and other counts. The string "xx" will be counted as one word. To indicate that a transcribed word or phrase is a best guess the word or phrase is followed by "[?]" (e.g. *I want a frog [?].*, transcriber is not sure of the word *frog*).

Non-speech acoustic events: CHILDES and the CGN both provide rules for transcribing non-speech acoustic events. For the CGN clearly audible speaker sounds, such as laughter, crying, screaming or coughing, are represented by ggg (when relevant for the conversation) whereas in CHILDES this is transcribed by "0".

Phonological fragments: The CGN protocol provides the characters *a to mark phonological fragments (e.g. *ik ga mo*a nee overmorgen naar de tandarts.* ("to-m*a no the day after to-morrow I'm going to the dentist.")). When a complete word is repeated, the word is not marked (*wat wat deed je daar dan?* ("what what were you doing there?")). In CHILDES, phonological fragments are proceeded by "&" (e.g. *&t &t &k can't you go?*).

The nonfluent speech

The orthographic transcription protocol of the CGN has been used for transcribing the aphasic speech. However, although the transparency criterion is very important, some typical problems frequently present in aphasic speech ask for additional rules. These problematic phenomena - the interjections problem, the word finding problem, the produced versus intended utterance problem, the boundaries problem and the gestures problem - are discussed in more detail.

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 6.4.1 (Interjections - 1a). ***uh uh*** *de bed helemaal* ***uh uh*** *vliegen* ***uh*** *nou zes*

*stoelen en **uh** en een gordijntje d'r omheen **uh** .*

(**uh uh** *the bed all **uh uh** fly **uh** well six chairs and **uh** and a curtain around it **uh** .*)

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

Example 6.4.2 (Interjections - 1b). *de bed helemaal vliegen nou zes stoelen en en een gordijntje d'r omheen .*

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 another 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 were 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"Šs. 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 (this is also done in the CGN). The orthographic transcription then becomes:

Example 6.4.3 (Interjections - 1c). *<fp> de bed helemaal <fp> vliegen <fp> nou zes stoelen en <fp> en een gordijntje d'r om heen <fp> .*

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, and time indicators. In the example below, the patient searches for the country *Frankrijk* ('France').

```
<i> ok en waar was je precies ? </i>
```

```
<p> Valdorand . </p>
<i> en waar ligt dat ? </i>
<p> uh in Zwitserland niet maar uh uh Valdorand uh Duitsland
    Oostenrijk Zwitserland uh Oostenrijk Oostenrijk Zwitserland .
    </p>
<i> nee hij is even weg ? </i>
<p> nee uh Duitsland uh Zwitser*a he Valdorand uh . </p>
<i> in het buitenland . </i>
<p> ja uh Oostenrijk niet Zwitserland niet Spanje niet . </p>
<i> je hebt ze allemaal voor je de landen maar . </i>
<i> hij is even weg ? </i>
<i> nou misschien dat je d'r zo op komt . </i>
<p> ja . </p>
```

In the orthographic transcription according to the CGN guidelines, it is not possible to indicate that all countries (*Zwitserland* ('Switzerland'), *Oostenrijk* ('Austria'), *Duitsland* ('Germany') and *Spanje* ('Spain')) are produced during the word finding process of the country *Frankrijk*. 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 possible to filter out only the relevant words. Another way of indicating that a word was produced during the word finding process is to mark it with *wf followed by the intended word. The orthographic transcription of the relevant part of the example would then be:

```
<p> uh in Zwitserland*wf(Frankrijk) niet maar uh uh Valdorand uh
    Duitsland*wf(Frankrijk) Oostenrijk*wf(Frankrijk)
    Zwitserland*wf(Frankrijk) uh Oostenrijk*wf(Frankrijk)
    Oostenrijk*wf(Frankrijk) Zwitserland*wf(Frankrijk) . </p>
<i> nee hij is even weg ? </i>
<p> nee uh Duitsland*wf(Frankrijk) uh Zwitser*a he Valdorand uh .
    </p>
<i> in het buitenland . </i>
<p> ja uh Oostenrijk*wf(Frankrijk) niet Zwitserland*wf(Frankrijk)
    niet Spanje*wf(Frankrijk) niet . </p>
```

The DTD of CGN XML can be extended to make it possible to mark these words with a special markedness category, e.g. "word_finding". The word to be found can also obtain an attribute, e.g. "wordtobefound". When we would transcribe word finding difficulties in this way, the word "Zwitserland" in the example would be transcribed as:

```
<w id="fn..." marked="word_finding" wordtobefound="Frankrijk">
Zwitserland</w>
```

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.

Produced utterance vs. intended utterance

Produced words are sometimes (slightly) different from the intended words, it is clear what the speaker wants to say, but the realization of the word is not completely correct (e.g. *legepodie* instead of *logopedie* (speech therapy)). Such errors are marked with `*u`, the marking used in the orthographic transcription of the CGN to indicate that a word is a mispronunciation (either by accident or on purpose) or an onomatopoeic word. Although the errors of the aphasic speakers are not exactly the same as the mispronunciations produced by speakers with intact speech abilities, this is the category that comes most close. It would be better if such errors could be marked in a special way, for example with `*i`.

Distinguishing utterances

Nonfluent aphasic 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 can 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. Topics often contain more then one utterance, so it is still possible to miss boundaries in this way. 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 for some aphasic patients, sometimes they speak in a rather monotonous tone.

Gestures

For the encoding of gestures, the Holland corpus gives a possible solution. The non-speech acoustic events that influence the conversation are clearly encoded in the Holland transcripts. In the Holland corpus, fragments of non-speech acoustic events are coded by `[% non-speech acoustic event]`, e.g. `[% laugh]` for encoding laughing. Not only actions as laughing are transcribed, but also all other non-speech events, such as taking something. Unfortunately, the Holland corpus has not been converted to XML. In XML, a solution could be to use a seperate tier parallel to the speech to annotate the gestures. For the annotation of gestures, it would be of much help if it would be allowed to make video recordings of the conversations.

6.4.4 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) combi-tagger was used (Daelemans, Zavrel, van der Sloot and van den Bosch 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).

Nonfluent 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 and van den Bosch 1996)

MBT uses a memory-based learning approach to tagging. In this approach, a set of example cases is kept in memory. Each example case consists of a focus word with preceding and following context (two positions to the left and two positions to the right) and the category for that word in that specific context. New sentences are tagged by mapping each word to the most similar example case. For the construction of a POS-tagger for a specific corpus, an annotated corpus is needed. From this annotated corpus three data structures are extracted: a lexicon, a case base for known words, and a case base for unknown words. During tagging of new text, each word is looked up in the lexicon. When a word is found, it is disambiguated using the context to decide what the most similar case is. When a word is not contained in the lexicon, the tag for that word is based on its form, its context, and the most similar cases in the lexicon. The output is a best guess of the category for the word in its current context.

After tagging with MBT, 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.

Subject	Correctness		Total
	Not correct	Correct	
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 5 Tagger correctness for interviewer and patients

Further evaluation of the data showed that the errors can be divided roughly in five categories. 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 error category "Same POS-tag"). The problem with *je* was that the tagger often tagged it as an indefinite pronoun instead of a personal pronoun. For the speech of the aphasic speakers the most problematic in the within error category was the tagging of capital letters (71.9% of all errors).

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").

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. *'t*) should be learned from the training corpus. However, for tagging the CGN these abbreviations were not problematic, so maybe our bad results on this point are due to using only one of the taggers of the combitagger. Finally, the tagger should be able to deal with words with diacritic marks.

6.5   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.

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.

Besides orthographic transcription and part-of-speech tagging, we also investigated in the pilot study whether the phonetic transcription procedure of the CGN could be adopted. For a small part of the data, the automatically generated transcriptions have been checked globally. At first sight there seemed to be few problems. A more detailed investigation of the results is needed to draw strong conclusions (Westerhout 2006). The investigation of the problems that aphasic speech constitute for syntactic and prosodic annotation is left for future research.

References

Binnenpoorte, D.(2006), *Phonetic Transcriptions of Large Speech Corpora*, PhD thesis, Radboud University.

Cucchiarini, C., van Hamme, H., van Herwijnen, H. and Smits, F.(2006), Jasmin-cgn: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality, *Proc. 5th International*

Conference on Language Resources and Evaluation, Genoa, Italy, pp. 135–138.

Daelemans, W. and van den Bosch, A.(1996), Language-independent Data-oriented Grapheme-to-phoneme Conversion, *in* J. Van Santen, R. Sproat, J. Olive and J. Hirschberg (eds), *Progress in Speech Synthesis*, Springer Verlag, pp. 77–90.

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

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

Goedertier, W., Goddijn, S. and Martens, J.(2000), Orthographic transcription of the Spoken Dutch Corpus, *in* M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (eds), *Proceedings of LREC 2000*, Vol. II, pp. 909–914.

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

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

Westerhout, E.(2006), *A corpus of dutch aphasic speech: Sketching the design and performing a pilot study*, Master's thesis, Department of Linguistics, Utrecht University, Utrecht, The Netherlands.

Westerhout, E. and Monachesi, P.(2006), A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS), *Proc. 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 1648–1653.

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

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