The Cambridge Cookie-Theft Corpus: A Corpus of Directed and Spontaneous Speech of Brain-Damaged Patients and Healthy Individuals

Caroline Williams*, Andrew Thwaites†, Paula Buttery‡, Jeroen Geertzen‡
Billi Randall*, Meredith Shafto*, Barry Devereux*, Lorraine Tyler*

*The Centre for Speech, Language and the Brain
Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK
camw3@cam.ac.uk, {billi, mshafto, barry, lktyler}@csl.psychol.cam.ac.uk

†The MRC Cognition and Brain Sciences Unit
15 Chaucer Road, Cambridge, CB2 7EF, UK
andrew.thwaites@mrc-cbu.cam.ac.uk

‡Computation, Cognition and Language Group, RCEAL
English Faculty Building, 9 West Road, Cambridge, CB3 9DP, UK
{pjb48, jg532}@cam.ac.uk

Abstract

Investigating differences in linguistic usage between individuals who have suffered brain injury (hereafter *patients*) and those who haven't can yield a number of benefits. It provides a better understanding about the precise way in which impairments affect patients' language, improves theories of how the brain processes language, and offers heuristics for diagnosing certain types of brain damage based on patients' speech. One method for investigating usage differences involves the analysis of spontaneous speech. In the work described here we construct a text corpus consisting of transcripts of individuals' speech produced during two tasks: the Boston-cookie-theft picture description task (Goodglass and Kaplan, 1983) and a spontaneous speech task, which elicits a semi-prompted monologue, and/or free speech. Interviews with patients from 19*yrs* to 89*yrs* were transcribed, as were interviews with a comparable number of healthy individuals (20*yrs* to 89*yrs*). Structural brain images are available for approximately 30% of participants. This unique data source provides a rich resource for future research in many areas of language impairment and has been constructed to facilitate analysis with natural language processing and corpus linguistics techniques.

1. Introduction

The characteristics of a population's speech can shed light on theoretical models which aim to explain how language is represented and processed in the brain. Typically, these models are based on the phonological, morphological, syntactic and discourse characteristics of language production in young healthy people and their relationship to brain function. Such models provide a baseline against which the language output of patients with brain damage can be evaluated, and can aid in the diagnosis of language impairments (Davis et al., 1998). Moreover, language changes associated both with brain damage and with neural change associated with healthy aging provide strong tests of models of language and brain. (Kemper et al., 2004). However, the development of adequate models and the ability to test them requires input data, in the form of examples of natural speech production, from a wide range of speakers, across the adult lifespan, and from brain-damaged patients with (and without) language deficits. In addition, sufficient data must be collected to allow significance testing of hypotheses based on the transcripts of the speech data. The data should also be richly annotated and easy to manipulate, so that future researchers can readily undertake further analysis of the data. The Cambridge Cookie-Theft Corpus aims to make this kind of data available to the speech and language community.

2. Background

Research at the Centre for Speech, Language and the Brain [CSLB] aims to explore the language characteristics of brain-damaged patients and possible changes in language as a function of healthy aging. Interviews with patients with specific language disorders (such as syntactic deficits (Moss et al., 1998)) and healthy participants across the adult life-span (Shafto et al., 2007)) have been recorded, providing background data on their naturalistic language use. The raw data from these recordings are highly reusable. The interviews elicit a stream of continuous speech in response to emotionally neutral, open-ended questioning. Questions addressed to patients are designed to make the participant talk about themselves and their interests. In addition, a more constrained set of speech data is obtained by asking participants to describe a picture (in this case The Cookie-theft —see below). This combination of speech samples from both naturalistic and constrained contexts can be used to investigate how language production changes due to gradual neural change (i.e. in healthy aging) and punctuate change (i.e. in aphasia).

3. Participants

The aetiology of patients includes stroke, brain tumours, infarction, haemorrhage, aneurysm, ischaemia, haematoma and medical excisions. The damage is mainly left lateralised focusing on the frontal and temporal cortices; these

are thought to be critical to language (Binder et al., 1997). The age range of the patients is 70 years, with the youngest (at the time of recording) 19*yrs* and the oldest 89*yrs*. Table 1 shows the number of interviews transcribed for the corpus. In total there are recordings of 107 cookie thefts from 99 different patients and 129 spontaneous speech recordings from 89 different patients. 78 patients completed both tasks at least once, of whom 41 also have structural brain scans using MRI. The scans provide important additional information to the brief diagnosis provided with each transcription. Patients were selected from a variety of sources: from the neuroscience panel at the CBU (these will normally have detailed medical notes from a clinician), self-referrals, community/self-help groups, or from country-wide memory clinics.

The healthy individuals were volunteers both at the CSBL and the CBU, most of whom were part of a wider panel recruited for other behavioural and neuroimaging studies. There are currently 222 healthy cookie theft recordings and 82 spontaneous speeches, from 244 subjects. T1, T2, DTI scans have been obtained for 82 of the healthy individuals.

4. The recordings

The interview recordings include two tasks, as described above. In the first task, the subject is asked questions designed to elicit spontaneous speech, either in the form of a semi-prompted monologue (where the participant answers general non-intrusive questions about their lives and hobbies), and/or free speech (where an initial question is asked and no secondary prompting is required). This task produces a wide range of speech styles, including genuine dialogue, prompted speech, and connected narrative. In the cookie-theft task, the participant is asked to "describe what's going on in" or "tell me about" a picture depicting a complex household scene, which includes the notable feature of a child stealing cookies off a high shelf. The cookie theft picture was selected because it is widely used in the study of aphasia (Giles et al., 1996), being included in a popular aphasia diagnostic protocol (Goodglass and Kaplan, 1983). Whereas the free speech task allows participants to use whatever strategies they have at their disposal to hide any deficits, and thus to show how fluently they can talk, the cookie-theft tasks constrains them to particular lexical items (*cookie, stool, boy*) and grammatical constructions (present tense forms), thus highlighting deficits. Similarly, the free speech task obtains speech in a variety of styles, which is useful for analysis of naturalistic language use, whereas the cookie-theft task provides the controlled context which is important in terms of reducing confounds in the analysis.

It should be noted that, unlike most spoken corpora, substantial overlap is relatively rare, since the interviewers were focussing eliciting speech from the participants. Overlap of backchannels is common, but extensive sections of overlap are infrequent. The vast majority of recordings also contain no more than two people, and the maximum number is four (where there were two interviewers, and a patient's family member was present).

The length of the patient spontaneous speech samples ranges from 28 seconds to 14 minutes with most being between 1–5 minutes long. They are therefore substantially shorter than the recordings from healthy participants, which are typically around 10 minutes duration. Impressionistically, they also tend to contain less linguistic content, due to higher incidence of pausing and false starts. Due to resource constraints, only two minutes of each spontaneous speech file have been transcribed, starting from the midpoint of the file, in order to maximise the number of participants whose speech was transcribed. The two groups of participants also differ in terms of the cookie-theft files, with healthy participants producing fairly homogeneous recordings of between 45*s* and 2 minutes, whereas patient recordings range between 13 seconds and 10 minutes. The 10 minute recordings are from patients with Herpes Simplex Encephalitis who were unable to stay on task. No more than three minutes of these recordings was transcribed.

The interviews were conducted at the CSLB and the MRC Cognition and Brain Sciences Unit [CBU], except for those patients who wished to be interviewed at their homes (sometimes with family members unavoidably present). All healthy individuals were interviewed at the CBU or the CSLB. Insofar as was practical, these recordings were carried out in an isolated environment such as a sound-attenuated interview room. The recordings are stored as mp3s and wav files.

5. Orthographic Transcription

5.1. Producing a machine-parseable transcription

Given our research aims, the transcriptions needed to be easily machine-parseable, but it is also useful to retain easy access to the original recordings, rather than relying on the transcripts. This is especially important for a corpus including patient speech, since even more so with normal speech, there is often more than one possible interpretation of what has been said. To this end, the data were transcribed using Praat (Boersma and Weenink, 2005) and the output automatically converted to XML (see Figure 1 for an example). The use of Praat makes it easy to navigate the recordings using the transcriptions, and provides a raft of temporal information which can be used to calculating rate of speech and pauses automatically. Automatically converting Praat's output to XML makes the transcriptions more accessible to parsers, since they are accompanied by a DTD. In an ideal world it would have been possible to carry out a phonological transcription as well as an orthographic one, but this was not possible given the resources available. The use of Praat, however, means that it would be easy to add a phonological and prosodic transcription at a later date.

The design of an appropriate XML schema presented an interesting challenge, since the ideal input for many parsers is written text, complete with punctuation, and without repetitions, hesitations, false-starts, and rephrasings. The information which in writing is conveyed with punctuation, variant spelling, and phrases such as "he whispered", is conveyed in speech through pauses, pronunciation, varying speed rate, changes in voice quality and particular pitch contours, in short, through full use of the gamut of segmental and prosodic realisation options. It is possible to 'clean up' speech so that it looks like writing, but doing so removes the point of analysing speech in the first place.

Age range	Brain-injured patients		Healthy individuals	
	Cookie-Theft	Spontaneous speech	Cookie-Theft	Spontaneous speech
0-19	0	0	33	8
20-29	5	3	50	28
30-39	7	6	12	6
40-49	14	3	0	3
50-59	18	7	8	6
60-69	22	10	48	21
70-79	18	10	61	9
80-89	3	1	10	1
90-09	0	0	0	0

Table 1: Number of transcriptions per age-range

This is particularly important when working with participants who have language disorders, as it is often difficult to tell from any given sample whether the problem is one of articulation and phonology, or of lexical retrieval and or syntax. Similarly, one can impose punctuation upon speech, but fundamentally speech is not structured in the same way as written text. One can impose clauses and sentences onto it, but that does not change the fact that speech is organised into prosodic and discourse units which do not map onto the written concept of clause and sentence, as described so well in MacWhinney (2007). As Edwards (1993) discusses, the way we represent speech substantially affects how we interpret and analyse it, so it is important to avoid imposing structure upon it which is not there. Given that it was not feasible to produce a phonetic, phonological, or prosodic transcription of the corpus, the schema therefore had to negotiate the partially conflicting goals of producing something which could be parsed automatically, yet which also adequately represented the speech on the recordings. The CompLex project has the advantage that the transcription was designed and carried out by one person (the first author), which made it easier to ensure consistency of coding, but in order for the corpus to be extended and analysed further, it was essential that the system be relatively easy to learn and apply.

5.2. Comparable corpora and existing guidelines

Although COBUILD (Payne, 1995) and the British National Corpus (Crowdy, 1995) are both substantial collections of spoken language, they were not designed for disordered speech. Perhaps the most obviously comparable corpus is the CHILDES corpus (MacWhinney, 2007), as child language can be just as fragmented and distorted as that of patients with severe language deficits. Several corpora of aphasic speech also exist, including the Dutch Corpus of Aphasic Speech (Westerhout and Monachesi, 2006) and PerLA (Paúls, 2004), which provide useful overviews of the issues involved in transcribing and parsing aphasic speech.

In terms of commonly agreed guidelines, the Text Encoding Initiative (the TEI Consortium, 2007,) provides a set of recommendations for the digital representation of texts, be they spoken or written, and specifically contains extensive guidance on the representation of corpora. The current version of TEI is XML-based. The BNC XML edition is now compatible with these recommendations, as is the corpus of British Academic Spoken English (Nesi and Thompson, 2006). The EAGLES guidelines (Llisterri, 1996) also provide instructive discussion of what constitutes a corpus and of the different levels of transcription possible.

In the transcription of this corpus we broadly follow the TEI approach for compatibility with other corpora and interoperability with other parsers. Our XML does not conform to their schemas, however, as we only implement a subset of their elements, and aspects of our format vary. As an example, `desc` is treated as an attribute on elements, rather than an element to be nested, and temporal information is included through `start` and `end` attributes on all structural units. It is generally true, however, that we follow the TEI terminology and definitions. To avoid imposing written structure on the transcriptions, we follow the PerLA and BASE corpora and transcribe without punctuation, dividing up the text only into utterances and 'segments' (see below). Analysis of the structure of the speech is treated as a separate task to the transcription.

5.3. Meta-data

The CSLB has extensive background information on all participants, but for the purposes of this corpus, only the following items are recorded in the transcription: the patient's unique id, their diagnosis (i.e. stroke, aphasia, agrammatism, etc.), aetiology (i.e., haemorrhage, infarction, aneurysm, excisions, etc.), area of damage, date of birth, gender and recording date. Not yet publically available are T1, T2 and DTI scans of a large proportion of the patients and healthy individuals. These structural MR scans were either carried out at the CBU or at the Wolfson Brain Imaging Centre.

5.4. Structural units

Time-stamping in Praat was applied liberally, with timestamps being inserted wherever it would facilitate transcription, or to delimit any stretch of speech which might be analytically interesting (e.g. a repetition, a mispronunciation etc). These short stretches of speech are therefore the smallest unit in the transcription: the sub-segment. They are not theoretically meaningful and simply reflect the temporal divisions in the transcription editor. The next largest unit is the 'segment', a stand-alone chunk of text, defined either by pauses, or by the clear rising/falling completion of

```
<file file_id="AB_CBU123_CT_12345678" length="70.3">
 <subject subj_id="AB_CBU123">
  <type>Agrammatic frontal</type>
  <aetiology>Aneursym/ischaemia</aetiology>
  <brain_damage>Left anteromedial temporal pole, LIFG, orbitofrontal, MTG/STG, parietal</brain_damage>
  <dob>1960-01</dob>
 </subject>

 <participants>
  <person role="investigator1" initials="AB" sex="m" />
  <person role="subject" initials="CD" sex="m" />
 </participants>

 <task type="cookie theft" topic="cookie theft" recording_date="2003-03-15">
  <comments></comments>
  <u who="subject" start="0.0" end="52.5">
   <seg start="0.0" end="0.9">
    <subseg start="0.0" end="0.9">erm </subseg>
   </seg>
   <seg start="1.6" end="2.2">
    <subseg start="1.6" end="2.2">mum </subseg>
   </seg>
   <seg start="2.7" end="3.8">
    <subseg start="2.7" end="3.8">washing up </subseg>
   </seg>
   <seg start="5.7" end="6.6">
    <subseg start="5.7" end="6.6">erm </subseg>
   </seg>
   <seg start="6.8" end="8.7">
    <subseg start="6.8" end="8.7">the sink is </subseg>
   </seg>
   <seg start="14.3" end="14.6">
    <subseg start="14.3" end="14.6"><tr>s</tr> </subseg>
   </seg>
   <seg start="19.3" end="20.1">
    <subseg start="19.3" end="20.1"><tr target="flooding">blʌdm</tr> </subseg>
   </seg>
  </u>
 </task>
</file>
```

Figure 1: Cookie-theft XML transcription for a brain-damaged patient (abridged, with biographical details changed)

an intonational phrase. Although not corresponding tightly with any particular theoretical definition, they were found to correlate quite highly with syntactic boundaries and were therefore considered useful for the parser, while also giving an impressionistic sense of the flow of speech. The largest unit in the transcriptions is the utterance, defined as '*a stretch of speech usually preceded and followed by silence or by a change of speaker*' as per the TEI guidelines. It should be noted, however, from the point of view of discourse analysis, this is actually closer to the definition of the conversational turn than the utterance, because it is not related to topics or themes (Crookes (1990)). The 'who' attribute from TEI is also adopted, and automatically completed from the tier names in the transcription editor. In total, the corpus contains 1331 utterances, 15248 segments, and 18840 sub-segments.

5.5. Representing the nature of speech

Dictionary spellings, abbreviations and contractions were used, in accordance with EAGLES guidelines and the BNC lists where possible. Contractions are used to represent the full spectrum of possible reductions; full-forms are only used if the auxiliary really is completely realised. In addition, filled pauses are kept and lexically transcribed, using a control list amended from Crowdy (1994). Westerhout and Monachesi (2006) suggest transcribing them as <fp/ >, but it was felt that keeping their lexical forms gives more of a sense of the original recording, and would also be more useful for those researching discourse. Numbers were transcribed in text rather than numerals, so as to preserve information as how the number was said, e.g. twenty-ten versus two thousand and ten.

Repetitions present a challenge for parsers since they generate ungrammatical strings. In a corpus of disordered speech, however, a simple string-matching filter would falsely identify cases where the speaker was making one string serve multiple discourse purposes. In order to identify genuine repetitions which can be ignored by the parser, exact repetitions are therefore marked with <rep>. The first use of a word/string is left as is, while subsequent iterations are wrapped in <rep> tags with a no attribute to record which repetition it is (not including the original). <rep> can be used for any type of repetition, including phonological, semantic, and syntactic repetition. If the repetition is not an exact repeat then <rep> is not used. Nested repetitions can sometimes be problematic due to the strict XML schema, but these are handled in a systematic way by flagging lexical repetitions at the expense of phrasal ones.

Speech errors also present difficulties for the parser as they also produce ungrammatical strings, but the precise cause of the error is often a matter for theoretical debate. Indeed, even the identification of errors is a theoretical issue - does a string which breaks the rules of grammar but goes unnoticed by both speaker and listener count as an error? Given the tendency of the human brain to mentally correct speech errors, and given the other attentional demands of the transcription task, how reliably would we spot these cases? In this corpus we therefore compromise by flagging as errors

those instances where the speaker appears to identify an error in their speech, for example when they abort a word and try again. Where the error is clearly semantic e.g *And the girl boy is on the stool* then the error is flagged as being semantic in nature by giving it type `sem`. Likewise, when it is clearly phonological, e.g. a meaningless phonological sequence is uttered which is very similar to the target sequence, then it is flagged as phonological in nature by giving it type `phon`. Phonological errors receive a phonological transcription, as described below. In the vast majority of cases, the nature of the error is debatable and is therefore not categorised. Syntactic errors are treated differently, since rephrasings and restructurings are so endemic in speech. Any string which is abandoned before generating a complete syntactic unit is therefore marked as incomplete using '. . .'. This does not imply any kind of pause (unlike the normal written convention for ellipses).

Inevitably, it is not always possible to identify with any certainty what is being said. Speech fragments for which it is not completely clear what the speaker said are therefore wrapped in `<unclear>` and given a reason, ie. `distorted phonology` or `background noise`. Where the reason is that the phonology is distorted, then a phonological transcription is given. If the value is ambiguous (e.g. "*taps*" as plural or "*tap is*") then the `reason` attribute in the `unclear` tag is 'ambiguous'. In cases where what was uttered could not be determined at all, the tag `<gap>` is used, with the `reason` attribute set to 'inaudible ' or 'unintelligible'.

As noted above, overlap is not particularly frequent in this corpus, but it is of course important, especially for those studying discourse. Stretches of speech which overlapped were each given their own sub-segment, which enables overlap to be automatically identified during later processing. As Praat is a 'partitur' editor, overlap is immediately visually obvious when working in the transcription editor itself.

5.6. Suprasegmental features

Despite the importance of prosody in understanding speech, resources were simply not available for any kind of prosodic annotation beyond the loose correspondence between segments and intonational phrases noted above. The use of Praat for transcription, however, means that later researchers can very easily use this corpus to carry out prosodic research. The detailed time-stamping, does, however, allow for some analysis of rate (given that words are transcribed in dictionary form, even where not all the dictionary syllables are realised), and also allows researchers to adopt whatever definition of pause seems appropriate (short versus long pauses, for example, may differ in patients to control subjects).

Some para- and extra-linguistic information is included, in order to help refine rate analysis, and for the purpose of discourse analysis. The `<shift desc="speech_type">` tag encodes the point at which normal speech has moved to obviously modulated speech such as laughing, reported speech or read. `<shift desc="normal" />` signifies the return to normal speech, and is assumed to be the default value if no `<shift>`s are present. The very few cases of non-English speech are accommodated by a variant on this: a `<shift>` with the special attribute `lang` which states the language. Coughs, grunts, groans, etc, are recorded using the `<vocal>` element. Gestures are recorded using the `<kinesic>` element, but these are of course rare, since these are audio recordings only. Sometimes earlier transcriptions do exist, however, and these occasionally note gestures.

Background noises are only recorded if the participants give any indication of hearing them. Thus, a truck going by would not be recorded unless one speaker referred to it, either directly, or by repeating what they had just said louder. Background noises are recorded using the `<incident>` element.

5.7. Segmental information

It is important to retain at least some phonological information, because there are some speakers for whom articulation issues represent a large portion of the impairment, and also because, as described above, the intended meaning is not always clear, and a phonological transcription enables researchers to look back to what was actually said, rather than taking the 'best guess' as the definitive value (where, of course, the original recordings are not available).

Phonological transcriptions (`<tr target='orthographic string'>`International Phonetic Alphabet in unicode`</tr>`) were inserted in the following cases:

a) where the target is unknown but a transcription can be produced.

b) where the phonology is non-standard and appears to be a property of the impairment, not part of a dialect (e.g. "cookie gar"). This is often paired with a `<unclear reason='distorted phonology'>` tag, and accompanied by the `target` attribute (see below).

c) where the phonology is non-standard and it is not clear whether this is due to impairment or dialect.

d) for incomplete words or isolated phonemes. These are surrounded by `<trunc>` `</trunc>` and add the target attribute if it is clear what was intended. These truncated words do not trigger repetition tags.

Because IPA transcriptions are not useful for an automatic parser, wherever possible the `target` attribute was used to insert the word which the transcriber thought was intended. This information is placed in an attribute to remind researchers that it is often an educated guess, and therefore subject to doubt. Ultimately, the presence of a phonological transcription is a guide to the researcher to revert to the original recordings.

The transcription tries to be as faithful to the subject's speech as possible, even though on some occasions this means making an assumption about what was intended. Specifically, when a speaker with phonological difficulties but apparently no semantic difficulties aims for one word and produces another (e.g. *'off of the stool'* is articulated as *'off of the tool'*), then this will be transcribed

as '<tr target = 'stool'>tul</tr>', in an attempt to show that this is highly likely to be an articulation difficulty rather than a semantic problem. Of course this production could be the result of an error in lexical retrieval rather than an error due to articulation difficulties or faulty phonological representation, but if the speaker otherwise seems to show no problems retrieving semantically appropriate words, this representation is less misleading than putting in a semantically inappropriate word. This also applies to grammatical words: if a patient appears to have mostly intact syntax/morphology but very distorted articulation/phonology, and produces '*he's*' with the vowel of '*his*' then it is transcribed as '*he's*' with <tr> tags rather than '*his*', as the latter would imply a grammatical error which is in all likelihood not present. Some patients have such severe articulation difficulties that no attempt is meant to transcribe every distorted word. Their difficulties are flagged in the meta-data as requiring manual analysis.

5.8. Anonymisation

For privacy reasons, identifying names such as personal names and names of home towns/counties were replaced with a <gap> *number* < /gap>, and the 'reason' attribute was set to 'place' or 'name' as appropriate (the 'sex' attribute is also used if applicable, to facilitate future research on gender agreement for pronouns). The number refers to the referent rather than the form, thus '*Cathy or Catherine as she was then and I went to the cinema*' would be '*1 or 1 as she was then and I went to the cinema*'. Articulation issues with proper names are therefore not flagged, but semantic issues can be, as a proper noun with a different referent would have a different number.

6. Future work

There are two current shortcomings in the corpus, both concern the issue of data sparsity. The first is the current gap in ages for healthy individuals with the cookie-theft task between the ages of 25*yrs* and 63*yrs*, for which there are only 33 recordings. The second is a shortfall in the number of instances within each aetiology (for instance, only two patients have semantic dementia, as this, fortunately, is a very rare condition) and damage type. This is due to each of the patients having very different stages and instances of damage. In future, additions to the corpus will focus on these areas.

7. Acknowledgments

This work is part of the Computational Natural Language Processing and the Neuro-Cognition of Language (COMPLEX) project, supported by EPSRC (grant EP/F030061/1) and by a Medical Research Council UK grant to LKT (U.1055.04.002.00001.01 and grant G0500842).

8. References

Jeffrey R. Binder, Julie A. Frost, Thomas A. Hammeke, Robert W. Cox, Stephen M. Rao, and Thomas Prieto. 1997. Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience*, 17(1):353362, January.

P. Boersma and D. Weenink, 2005. *Praat: doing phonetics by computer (Version 4.3.01) [Computer program] Retrieved from http://www.praat.org/.*

Graham Crookes. 1990. The utterance, and other basic units for second language discourse analysis. *Applied Linguistics*, 11:193–199.

Steve Crowdy. 1994. Spoken corpus transcription. *Literary and Linguistic Computing*, 9(1):25–28.

Steve Crowdy. 1995. The BNC spoken corpus. In *Spoken English on Computer: Transcription, Mark-Up, and Application*. Longman.

Barbara L. Davis, Kathy J. Jakielski, and Thomas P. Marquardt. 1998. Developmental apraxia of speech: Determiners of differential diagnosis. *Clinical Linguistics & Phonetics*, 12:p25–45.

Jane A. Edwards. 1993. Principles and contrasting systems of discourse transcription. In Jane A. Edwards and Martin D. Lampert, editors, *Talking Data: Transcription and Coding in discourse research*, chapter 1, pages 3–31. Lawrence Erlbaum.

Elaine Giles, Karalyn Patterson, and John R. Hodges. 1996. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimers type: missing information. *Aphasiology*, 10(4):395–408.

Harold Goodglass and Edith Kaplan. 1983. *Boston Diagnostic Aphasia Examination (BDAE)*. Lea and Febiger. Distributed by Psychological Assessment Resources, Odessa, FL.

S. Kemper, R. Herman, and C. Lian. 2004. Age differences in sentence production. *Journals of Gerontology: Psychological Sciences*, 58B:P220–P224.

J. Llisterri, 1996. *EAGLES Preliminary recommendations on Spoken Texts*.

Brian MacWhinney, 2007. *The* CHILDES *Project. Tools for Analyzing Talk. Electronic Edition. Part 1: The* CHAT *Transcription Format*. Carnegie Mellon University.

Helen E. Moss, Lorraine K. Tyler, Mark Durrant-Peatfield, and Elaine M. Bunn. 1998. Two eyes of a see-through: Impaired and intact semantic knowledge in a case of selective deficit for living things. *Neurocase: The Neural Basis of Cognition*, 4:291– 310.

Hilary Nesi and Paul Thompson, 2006. *The British Academic Spoken English Corpus Manual*.

B. Gallardo Paúls. 2004. La transcripcin del lenguaje afsico. In B. Gallardo and M. Veyrat, editors, *Estudios de lingstica clnica: Lingstica y patologa.*, pages 83–114. Valncia: Universitat de Valncia - Asociacin Valenciana de Lenguaje, Comunicacin y Cultura.

Jonathan Payne. 1995. The COBUILD spoken corpus: transcription conventions. In *Spoken English on Computer: Transcription, Mark-Up, and Application*. Longman.

Meredith Shafto, D. M. Burke, E. Stamatakis, P. Tam, and Lorraine Tyler. 2007. On the tip-of-the-tongue: Neural correlates of increased word-finding failures in normal aging. *Journal of Cognitive Neuroscience*, 19:2060–2070.

the TEI Consortium, 2007. TEI P*5: guidelines for elec-*

tronic text encoding and interchange. Edited by Lou Burnard and Syd Bauman.

E. Westerhout and Paola Monachesi. 2006. A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS): Focusing on the orthographic transcription. In *Proceedings of Computational Linguistics in the Netherlands 2005*, University of Amsterdam. Amsterdam.