Transcription Bottleneck of Speech Corpus Exploitation

Caren Brinckmann

While written corpora can be exploited without any linguistic annotations, speech corpora need at least a basic transcription to be of any use for linguistic research. The basic annotation of speech data usually consists of time-aligned orthographic transcriptions. To answer phonetic or phonological research questions, phonetic transcriptions are needed as well. However, manual annotation is very time-consuming and requires considerable skill and near-native competence. Therefore it can take years of speech corpus compilation and annotation before any analyses can be carried out. In this paper, approaches that address the transcription bottleneck of speech corpus exploitation are presented and discussed, including crowdsourcing the orthographic transcription, automatic phonetic alignment, and query-driven annotation. Currently, query-driven annotation and automatic phonetic alignment are being combined and applied in two speech research projects at the Institut für Deutsche Sprache (IDS), whereas crowdsourcing the orthographic transcription still awaits implementation.

1. Introduction

Generally, written corpora are more easily accessible and exploitable than speech corpora. There are many written corpora readily available for research, even very large corpora such as DeReKo[1] for German (3.4 billion words). Using the 'web as corpus', specialised corpora can be constructed with readily available tools[2]. Once compiled, written corpora can be exploited without the addition of any annotation. This annotation-free approach was applied for the extraction of higher-order collocations in the CCDB[3]. Compared to written corpora, far less speech corpora are available, and those that can be used for research are often rather small.

1 http://www.ids-mannheim.de/kl/projekte/korpora

2 E.g. Web as Corpus ToolKit: http://www.drni.de/wac-tk/

3 http://corpora.ids-mannheim.de/ccdb/

To exploit speech corpora for linguistic research, at least a basic transcription is needed. This basic transcription is usually an orthographic transcription, often without any punctuation marks. For languages without a standardised orthography, however, it is not clear which form the basic transcription should take.

For accessibility and further analyses, it is essential that the transcription is aligned with the speech signal. Most of the widely used speech annotation tools such as ELAN[4] or Praat (Boersma & Weenink 2009) do not allow a transcription without text-to-audio alignment. In several speech corpus projects (e.g. Spoken Dutch Corpus CGN[5]), 2 to 3 second inter-pause stretches of running speech are segmented and transcribed.

Most experienced speech corpus builders agree that only near-native speakers can produce a reliable orthographic transcription. This poses a problem for minority languages and dialectal speech, if near-native speakers are not available at the corpus building institution. Crowdsourcing the orthographic transcription might be a solution to this transcriber-scarcity problem and is described in detail in Section 2.

Further types of annotations can be added, such as phonetic segmentations and transcriptions for phonetic and phonological research, or prosodic labels such as ToBI[6]. When the orthographic transcription is available, all types of annotations that can be added to written corpora can be added to speech corpora as well, for example, part-of-speech, information structure, or co-references.

Manual phonetic segmentation and transcription is a very time-consuming task (ca. 1:200, that is, at least 200 hours are needed for one hour of phonetically transcribed speech). Furthermore, it requires considerable skill and training. Automatic broad phonetic alignment (discussed in Section 3) and query-driven annotation (presented in Section 4) can facilitate this task.

The overall aim of this paper is to provide an overview of these three different approaches that can be combined to overcome the transcription bottleneck of speech corpus exploitation.

4 http://www.lat-mpi.eu/tools/elan/

5 http://lands.let.kun.nl/cgn/ehome.htm

6 http://www.ling.ohio-state.edu/~tobi/

2. Crowdsourcing

The term 'crowdsourcing' was popularised by Howe (2006) and is a blend of 'crowd' and 'outsourcing'. Outsourcing is defined as subcontracting a task or process traditionally performed by an employee (such as product design or manufacturing) to a third-party company. Crowdsourcing means outsourcing a task to an undefined, generally large group of people. Even though crowdsourcing is often associated with Web 2.0, classical crowdsourcing has existed before the Internet. For example, at self-service restaurants, supermarkets, IKEA, automatic teller machines and ticket vending machines, customers perform tasks (sometimes with the help of machines) that formerly were carried out by an employee. In the Web 2.0 era, the Internet is used to publicise and manage crowdsourcing projects. The most popular crowdsourcing project is Wikipedia[7], where volunteers collaborate to write and constantly improve online lexica. Another related concept is the 'wisdom of crowds' (Surowiecki 2004), meaning that the aggregation of information in groups result in decisions that are often better than could have been made by any single member of the group.

2. 1 Examples of crowdsourcing

2. 1. 1 CastingWords @ Amazon Mechanical Turk

In 2005 Amazon launched a web-based service called the 'Amazon Mechanical Turk'[8]. The name is derived from Wolfgang von Kempelen's 18th-century chess-playing machine, which turned out to be hiding a human chess master inside. Anyone with a US-billing address can become a 'requester' and create so-called 'human intelligence tasks' (HITs). These tasks can usually be carried out easily by humans but are still hard to automatise. HITs are often small tasks worth less than a dollar, sometimes even only a few cents. Anyone with an Amazon account can become a 'turker', complete a task and claim the advertised reward.

The US company CastingWords offers transcription services and uses the Mechanical Turk to crowdsource all necessary tasks: "After a transcription assignment is accepted by a worker, and completed, it goes back out on Mturk.com for quality assurance, where another worker is paid a few cents to verify that it's a faithful transcript of the

7 http://www.wikipedia.org/

8 https://www.mturk.com/

audio. Then, the transcript goes back on Mturk.com a third time for editing, and even a fourth time for a quality assurance check" (Mieszkowski 2006: 2).

2. 1. 2 Distributed Proofreaders

Distributed Proofreaders[9] is a non-commercial platform that provides a Web-based method to support the conversion of public domain books from the Project Gutenberg[10] into electronic texts. Volunteers proofread pages that have been scanned and converted with optical character recognition (OCR) software. As of March 2009, more than 15,000 books have been converted into electronic texts.

2. 1. 3 Ph@ttSessionz

In order to document regional variation in speech, usually speakers from different dialectal regions who live near the recording site are recorded (cf. König 1989: 20), or the researchers visit the respective areas to carry out the recordings (e.g. 'German Today', see Section 4.3). This is a very time-consuming process. A different solution is provided by the software SpeechRecorder, which is now part of the WikiSpeech[11] system (Draxler & Jänsch 2008: 1647). SpeechRecorder was used for Ph@ttSessionz, a project carried out at the Institute of Phonetics and Speech Processing at LMU Munich. The aim of the project was to collect teenage speech from all over Germany. Participating schools were sent a recording device to be connected to a PC with broadband Internet connection. The recordings were prompted via the Web-based SpeechRecorder, which immediately uploaded the recorded speech to a central server. Thus, the speakers themselves carried out all recordings without the presence of a researcher or technician at the recording sites.

2. 1. 4 The ESP Game and Google Image Labeler

The aim of the ESP game (von Ahn & Dabbish 2004) is to collect labels for images, a task that is still hard for computers, but easy for humans. This rather dull task is turned into a game, where two persons are shown the same image and have to guess what the other person is typing. The players are given points for each image for which

9 http://pgdp.net/

10 http://www.gutenberg.org/

11 http://wikispeech.org/

they agree on a word. Players can pass five skill levels based on the number of accumulated points, which adds an additional goal-based motivation to the game. The ESP game was licensed to Google for the Google image labeler[12] in 2006. As of July 2008, 200,000 players had contributed more than 50 million labels (von Ahn & Dabbish 2008: 60).

2. 1. 5 reCAPTCHA

Von Ahn and colleagues also invented reCAPTCHA[13] (von Ahn et al. 2008). A CAPTCHA is a test to determine whether the user is a human or a computer. Usually the user is asked to decipher several severely distorted characters—a task that is still hard for computers—, and many websites use CAPTCHAs as security measures against spam. The reCAPTCHA system presents one known distorted character string together with a word from scanned texts that could not be recognised by an OCR program. This way it helps to digitise old printed material. Recently, the reCAPTCHA system has been extended to include sound files, thus helping to transcribe historic recordings.[14]

2. 2 Guidelines for successful crowdsourcing

Of course it is not sufficient to set up a webpage describing the problem at hand, hoping that some volunteers will send an e-mail offering to help. There are some guidelines to follow for successful crowdsourcing (cf. Hempel 2006):

1. *Focus*: Vaguely defined problems get vague answers. Every task should be described as clearly as possible together with a set of rules. Often it is advisable to split up a large task into several smaller ones and to provide a suitable infrastructure (Web-based platform, software, etc.).
2. *Filter*: Use the crowd and experts to extract the best answers. Even though in many social networks only 1 % of the users generate original content, another 10 % comment on it or change it, and 89 % are just passive observers, crowdsourcing often produces a wealth of material that has to be filtered. While the final filtering can

12 http://images.google.com/imagelabeler/

13 http://recaptcha.net/

14 http://blog.recaptcha.net/2008/12/new-audio-recaptcha.html

be done by the task requester, the crowd can and should be used for at least the first filtering step.

3. *Reward*: Since many of the crowd-sourced tasks are rather dull, it is indispensible to offer incentives. Depending on the task and the requester this can be money (mturk), recognition (dp, Ph@ttSessionz), or fun (ESP game).

2. 3 Possible application: Crowdsourcing the orthographic transcription of the speech corpus 'German Today'

The IDS speech corpus project 'German Today' aims at determining the amount of regional variation in (near-)standard German spoken by young and older educated adults and to identify and locate regional features (Brinckmann et al. 2008). To this end, secondary school students and 50-to-60-year-old locals were recorded in 195 cities throughout the German speaking area of Europe (Germany, Austria, Switzerland, Liechtenstein, Luxemburg, South Tyrol, and East Belgium). More than 800 speakers read a number of short texts and a word list, named pictures, translated words and sentences from English, answered questions in a sociobiographic interview, and took part in a Map Task experiment (Anderson et al. 1991). The resulting corpus comprises over 1200 hours of read and spontaneous speech.

Currently, only the read speech and the interview data are transcribed. The Map Task data contains some rather dialectal speech and can only be transcribed reliably by near-native speakers of the dialect who are not available at the IDS. Therefore, the Map Task data has currently not been processed at all, which is rather unfortunate.

The proposed system for crowdsourcing the orthographic transcription consists of a central database and a software application controlling the transcription process (see Figure 1). The database stores the speech signals, metadata, transcripts, and information about the transcribers and the transcription process.

- *Database of speech files and metadata*: As a first step, the corpus provider fills the central database with the speech signals that are to be transcribed and the corresponding metadata. It might be necessary to pre-process the speech signals before storing them in the database. For example, 'German Today' Map Task speech signals consist of two channels that have to be transcribed separately, and therefore split up into two separate mono files. The metadata should contain basic information about the speech file, for example, recorded task, number of speakers, sex and

age of speakers, and place of recording. The latter is crucial for dialectal speech when the transcribers can opt for certain regions they are able to transcribe.

- *Task definition*: The database also contains clear and detailed task descriptions for each corpus, including conventions regarding transcription, grading and correction tasks. These conventions contain several examples and are presented to each registered transcriber of the corpus.
- *Process control*: The corpus providers have to define some parameters of the transcription process, for example, the maximum duration of presented speech signals, the number of human transcribers who have to transcribe each speech signal in parallel, grades for rating the quality of a transcription, and the awarded points for each task.
- *Database of human transcribers*: The human transcribers have to register before they can start transcribing, and for dialectal speech they are able to select certain regions by listening to examples. The database also stores which speech files each transcriber has already transcribed, graded or corrected, as well as the grades she/he received for transcriptions.
- *Transcription process*: When a transcriber is logged on, the process controlling software presents all available tasks (filtered by regions, where applicable):
  - *Initial transcription*: The speech files are presented in inter-pause stretches that do not exceed the specified maximum duration (e.g. 3 seconds). The Web-based transcription tool can be implemented in a very simple fashion using a publicly available media player[15] and an input form for the transcription. Solutions that allow more control include WebTranscribe (Draxler 2005).
  - *Grading*: Each transcription of an inter-pause stretch is graded by another human transcriber according to a predefined scale (e.g. 0=severe errors, 1=some error(s), 2=error-free)
  - *Correction*: If parallel transcriptions differ or the received grade is not 'error-free', the transcription is corrected and graded again. If all (corrected) transcriptions are identical and graded as error-free, the transcription is stored as final version.
- *Rewards*: Transcribers earn points for each task modified by the received grade. These points are posted as high-score lists on the website, for example, as All-Time Top Transcribers or Today's Top Transcribers. Virtual titles can be awarded as well as real-world incentives (such as a visit to the corpus-providing research institute).

15 For example, http://www.longtailvideo.com/players/jw-flv-player/

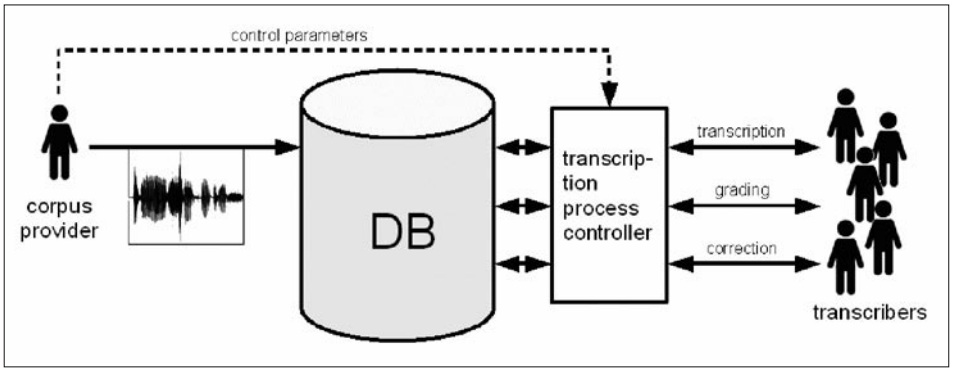

Figure 1: Simplified architecture of the proposed system for crowdsourcing transcriptions

Potential transcribers for 'German Today' could be recruited by contacting the schools where the original recordings were made. The students whose speech was recorded are usually interested in the project and have the time needed to participate as transcribers of regional speech. However, crowdsourcing the orthographic transcription is only feasible in countries with many broadband connections, which might not be the case for some lesser-used languages.

3. Automatic phonetic alignment

To facilitate (or even replace) manual phonetic segmentation and labeling, automatic phonetic alignment software has been developed. These systems usually produce time-aligned broad phonetic transcriptions (without fine phonetic detail) and need the following:

- speech signals
- orthographic transcriptions of the speech signals
- canonical/phonemic transcription of all words in the corpus: this can be provided by simple look-up in a pronunciation lexicon (e.g. CELEX: Baayen et al. 1995) or by a grapheme-to-phoneme converter. The former approach is only feasible for small corpora with a limited lexicon (e.g. read speech). Often, a grapheme-to-phoneme converter is combined with a lexicon of pronunciation exceptions.
- language-specific phoneme models (often trained HMMs), which are usually part of the alignment software.

Often, the automatic aligner can operate in two different modes: forced alignment and alignment with post-lexical phonological processes. In forced alignment, the given transcription of each word is not changed and 'forced' onto the speech signal. If a sound present in the transcription has been deleted by the speaker, the system will nonetheless label a part of the speech signal with the sound's symbol. The forced alignment mode is useful if one's analysis is based on the canonical transcription or if a detailed manual phonetic transcription is available that has to be time-aligned. The other mode tries to model post-lexical phonological processes (deletions, replacements, and insertions) and changes the given transcription accordingly. For all analyses that depend on the realised form, this is the preferred mode.

Van Bael et al. (2006, 2007) compared 10 aligners for Dutch with a manually-obtained reference transcription as 'gold-standard'. They found that a system that used canonical transcriptions and models post-lexical phonological processes with a decision tree performed best. Furthermore, the number of remaining disagreements with the reference transcription (14.6% for spontaneous speech, 8.1% for read speech) was only slightly higher than human inter-labeler disagreement scores reported in the literature.

For the project 'German Today' (see Section 4.3) and a research project on word-internal boundary effects, we carried out an informal task-based evaluation of the Munich Automatic Segmentation System MAUS[16] for German (Schiel 2004). During evaluation we found that MAUS produces some obvious errors such as extreme duration outliers, but these can easily be detected automatically and marked for manual correction. Apart from that, it depends very much on the task at hand whether an automatic segmentation with MAUS is useful or not:

- *Corpus access*: For the corpus 'German Today' MAUS proved to be especially useful for accessing specific portions of the speech signal for further manual annotation.
- *Analyses of segmental durations* can be based on the automatically aligned segment boundaries as well. However, we found that only large significant effects can be detected by relying on the automatic segmentation. Therefore, it is useful for a first gauge of the effect.
- For *analyses in the frequency domain* (e.g. formant slope), which depend on accurate segmental boundaries, the automatically set boundaries should always be corrected by hand.

16 MAUS can be downloaded from http://www.phonetik.uni-muenchen.de/forschung/Verbmobil/VM14.7eng.html

While phonetic aligners are available for all major languages, this is not the case for less-resourced languages. One solution is to use an existing alignment system and train your own language-specific phoneme-models (e.g. with the Hidden Markov Model Toolkit HTK[17]). The catch is that at least one hour of phonetically segmented and labeled speech data is usually needed as training material. Another solution might be to find an alignment system for a language that is phonetically similar to the target language. Its pre-built phoneme-models could be used, adding a mapping between the phonemes of the target language and the language modeled by the existing aligner.

4. Query-driven Annotation

The traditional corpus annotation process consists of three tasks. First, an annotation schema is developed, then the actual annotation is performed and finally, when the whole corpus is annotated, the corpus can be queried and analysed (see Figure 2). One major problem of this sequential approach is that it is too time-consuming. Large corpora require many years of annotation work before the corpus can be exploited and any results can be published. Furthermore, due to coder drift during the long annotation process (see Gut & Bayerl 2004: 567), the reliability of the annotations can be rather limited. Finally, the corpus queries are restricted to those phenomena which have been annotated beforehand. Some queries might be impossible due to the structure of annotations, and it might be necessary to re-annotate the whole corpus.

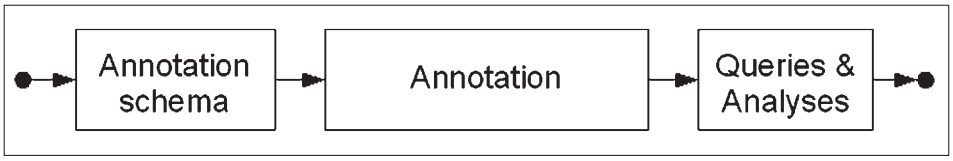

Figure 2: Traditional sequential process of corpus annotation

To overcome these problems, Voormann and Gut (2008) suggest an approach they call 'agile corpus creation', where the traditional corpus annotation process is replaced by a cyclic and iterative corpus annotation process. As shown in Figure 3, each annotation cycle starts with the formulation of a query. For example, the duration of schwa

17 http://htk.eng.cam.ac.uk/

in word-final /C@n/[18] sequences shall be compared between German words ending in a suffix and those not ending in a suffix. Then the annotation schema necessary for this query is specified and the annotation is carried out. After a successful analysis, the next query is formulated and the cycle starts afresh. Sometimes it becomes clear during the annotation or the analysis that the annotation scheme has to be modified, and jumping back within the cycle becomes necessary.

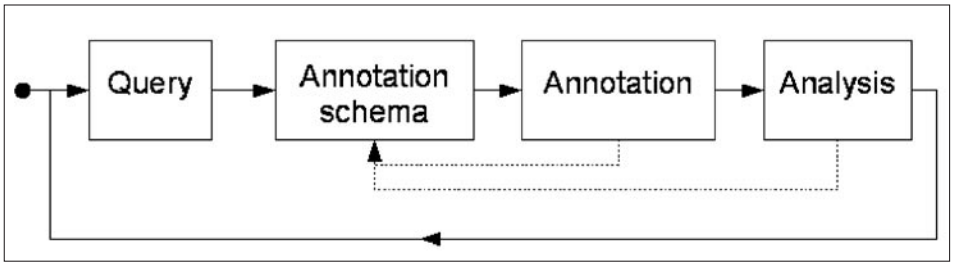

Figure 3: Query-driven corpus annotation process

In this query-driven approach, successive cycles improve the annotation schema and limit it to the elements necessary for the queries. Suitability and consistency of each addition to the annotation schema are immediately put to the test, so that only small amounts of data have to be re-annotated or discarded. Since the annotators focus on one particular annotation task within each cycle, coder drift is seldom a problem. Most importantly, results can be published shortly after the completion of the first annotation cycles. While the query-driven approach can be applied to a corpus of any size, for large corpora compiled for speech research there is almost no alternative.

To further speed up and facilitate the annotation process, the query-driven approach can be combined with an automatic phonetic alignment. First, the whole corpus is automatically segmented and labeled. These automatically set labels are then used to access those parts of the corpus that are to be annotated manually. Figure 4 shows a small excerpt from a corpus compiled for the study of word-internal boundary effects (Raffelsiefen & Brinckmann 2007: 1442). The third tier (shaded in grey) of the Praat TextGrid contains the orthographic transcription of the complete sentences that were read by the subjects. This orthographic transcription served as input for the phonetic aligner MAUS. Its output was converted to the Praat TextGrid format and is shown

18 All phonetic symbols in this paper are given in SAMPA: http://www.phon.ucl.ac.uk/home/sampa/

in the first tier (word boundaries) and second tier (phoneme labels and boundaries). A customised Praat script allows the human annotator to comfortably access those parts of the corpus which match a certain query. In the displayed example, the annotator wants to correct the automatic phonetic annotation of all word-initial consonant-[E6]-consonant sequences. The Praat script repeatedly jumps to the respective parts of the corpus and copies the matching phoneme sequences to the fourth tier where the annotator can easily correct the location of the segment boundaries.

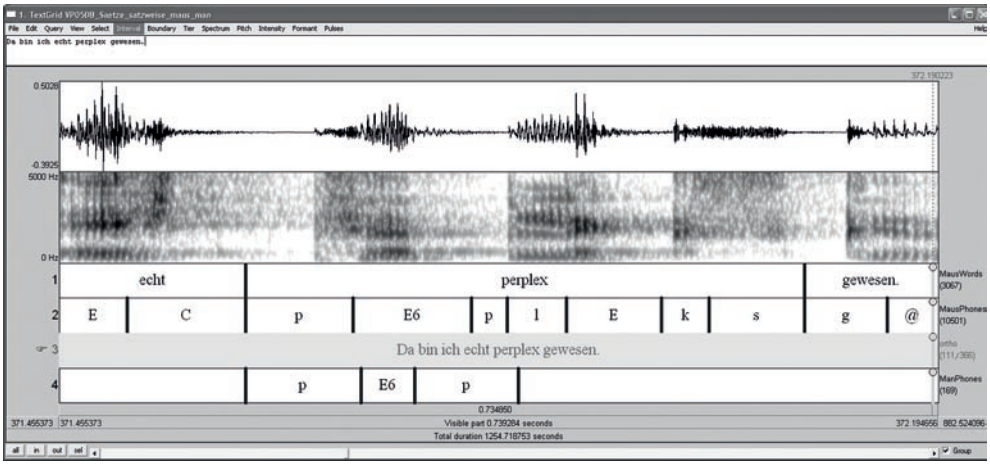

Figure 4: Combining query-driven annotation and automatic phonetic alignment

5. Conclusion

Compared to written corpora, there are far fewer speech corpora available for linguistic research. Speech corpora are not only more difficult to compile, their exploitation is further hindered by the fact that at least a basic orthographic transcription has to be added to the speech signal. Manual transcription and further linguistic annotation tasks are very time-consuming and require considerable skill. Three approaches addressing this 'transcription bottleneck' have been presented in this paper: crowdsourcing the orthographic transcription, automatic phonetic alignment, and query-driven annotation.

Producing a reliable orthographic transcription is almost impossible for non-native speakers. For dialectal speech or minority languages, transcribers with near-native competence might not be available at the corpus producing institution. Crowdsourcing the orthographic transcription might prove a viable solution as it has been successfully applied for commercial applications. To my knowledge, however, for speech research corpora this is a concept still awaiting implementation.

Once the orthographic transcription is available, the corpus can be treated as a written corpus, adding annotations on different linguistic levels. Phonetic and phonological research requires annotations based on the speech signal, such as phonemic/phonetic segmentation and labeling, which is even more time-consuming than orthographic transcription. Here, an automatic phonetic alignment software is useful. For some tasks, the automatically set segmental boundaries can be used.; for others they have to be corrected manually, which is nonetheless faster than segmenting and labeling the signal from scratch.

For large corpora, the traditional sequential corpus annotation process requires many years of annotation work before any analyses can be carried out. Using the query-driven cyclic approach, results can be published shortly after the completion of the first annotation cycles. At the Institut für Deutsche Sprache, we combine the query-driven annotation process with automatic phonetic alignment in two speech-corpus-based research projects.

References

Anderson, A. / Bader, M. / Bard, E. / Boyle, E. / Doherty, G. / Garrod, S. / Isard, S. / Kowtko, J. / McAllister, J. / Miller, J. / Sotillo, C. / Thompson, H. S. / Weinert, R. (1991). "The HCRC Map Task Corpus", *Language and Speech,* 34 (4), 351–366.

Baayen, R. H. / Piepenbrock, R. / Gulikers, L. (1995). *The CELEX Lexical Database* (Release 2). CD-ROM: Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA [Distributor].

Boersma, P. / Weenink, D. (2009). *Praat: doing phonetics by computer* (Version 5.1.02) [Computer program]. Retrieved March 9, 2009 from http://www.praat.org/

Brinckmann, C. / Kleiner, S. / Knöbl, R. / Berend, N. (2008). "German Today: an areally extensive corpus of spoken Standard German" in *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Morocco. Retrieved March 9, 2009 from http://www.lrec-conf.org/proceedings/lrec2008/pdf/806_paper.pdf

Draxler, C. (2005). "WebTranscribe—an extensible web-based speech annotation framework" in *Proceedings of the 8th International Conference on Text, Speech and Dialogue* (TSD 2005), Karlovy Vary, Czech Republic, 61–68.

Draxler, C. / Jänsch, K. (2008). "WikiSpeech—A Content Management System for Speech Databases" in *Proceedings of Interspeech 2008*, Brisbane, Australia, 1646–1649.

Gut, U. / Bayerl, P. S. (2004). "Measuring the Reliability of Manual Annotations of Speech Corpora" in *Proceedings of Speech Prosody 2004*, Nara, Japan, 565–568. Retrieved March 9, 2009 from http://www.isca-speech.org/archive/sp2004/sp04_565.pdf

Hempel, J. (2006). "Crowdsourcing: Milk the masses for inspiration", *BusinessWeek*, September 25, 2006. Retrieved March 9, 2009 from http://www.businessweek.com/magazine/content/06_39/b4002422.htm

Howe, J. (2006). "The Rise of Crowdsourcing", *Wired,* 14.06. Retrieved March 9, 2009 from http://www.wired.com/wired/archive/14.06/crowds.html

Keibel, H. / Belica, C. (2007). "CCDB: a corpus-linguistic research and development workbench" in *Proceedings of Corpus Linguistics 2007*, Birmingham, United Kingdom. Retrieved March 9, 2009 from http://www.corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf

König, W. (1989). *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland. Band 1: Text.* Ismaning: Hueber.

Mieszkowski , K. (2006). "I make $1.45 a week and I love it". Retrieved March 9, 2009 from http://www.salon.com/tech/feature/2006/07/24/turks/

Raffelsiefen, R. / Brinckmann, C. (2007). "Evaluating phonological status: significance of paradigm uniformity vs. prosodic grouping effects" in *Proceedings of the 16th International Congress of Phonetic Sciences* (ICPhS XVI), Saarbrücken, Germany, 1441–1444. Retrieved March 9, 2009 from http://www.icphs2007.de/conference/Papers/1684/1684.pdf

Schiel, F. (2004). "MAUS Goes Iterative" in *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, Portugal, 1015–1018.

Surowiecki, J. (2004). *The Wisdom of Crowds*. Boston: Little, Brown.

Van Bael, C. / Boves, L. / van den Heuvel, H. / Strik, H. (2006). "Automatic phonetic transcription of large speech corpora" in *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), Genoa, Italy, 4–11.

Van Bael, C. / Boves, L. / van den Heuvel, H. / Strik, H. (2007). "Automatic phonetic transcription of large speech corpora", *Computer Speech and Language*, 21 (4), 652–668.

von Ahn, L. / Dabbish, L. (2004). "Labeling Images with a Computer Game" in *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI 2004), Vienna, Austria, 319–326.

von Ahn, L. / Dabbish, L. (2008). "General Techniques for Designing Games with a Purpose", *Communications of the ACM*, 51 (8), 58–67.

von Ahn, L. / Maurer, B. / McMillen, C. / Abraham, D. / Blum, M. (2008). "reCAPTCHA: Human-Based Character Recognition via Web Security Measures", *Science* 321, 1465–1469.

Voormann, H. / Gut, U. (2008). "Agile corpus creation", *Corpus Linguistics and Linguistic Theory* 4 (2), 235–251.