Crowdsourcing a multi-lingual speech corpus: recording, transcription, and natural language processing

Andrew Caines[1], *Christian Bentz*[1], *Calbert Graham*[1], *Tim Polzehl*[2], *Paula Buttery*[1]

[1]The ALTA Institute, Dept. Theoretical & Applied Linguistics, University of Cambridge, U.K.
[2]Telekom Innovation Laboratories / Technische Universität Berlin, Germany

{apc38|cb696|crg29|pjb48}@cam.ac.uk; tim.polzehl@qu.tu-berlin.de

Abstract

We present a method for speech corpus collection via crowdsourcing. The target corpora feature [i] native speakers of English (CROWDISENG) and [ii] German-English bilinguals (CROWDISENGDEU) responding to business-topic questions of the type found in language learning oral tests. CROWDISENG gives us a benchmark corpus for comparison against non-native speakers undertaking similar tasks, so that we can start to address the common question in applied linguistics: 'what would a native speaker do in this situation?'. CROWDISENGDEU gives us a resource to study transfer effects from the first language, in terms of both pronunciation and lexico-grammatic selections. Recordings come from Crowdee, transcriptions and error annotations come from CrowdFlower, and we describe the use of natural language processing tools to further annotate the transcriptions with phone alignments, sentence boundaries, and morpho-syntactic labels. Upon completion, the corpora will be freely available to other researchers. We report on its current state of progress and explain how to stay informed of developments.
Index Terms: speech corpus, crowdsourcing, computational linguistics

1. Introduction

We describe two corpora collected via crowdsourcing: a native speaker corpus of English (CROWDISENG), and a corpus of German/English bilinguals (CROWDISENGDEU). Both corpora involve the speakers answering questions about selected business topics, and are designed with the following two research questions in mind:

1. With tasks and topics comparable to typical language learning oral exams, we can start to address the question, 'what would a native speaker do in this situation?' (CROWDISENG);
2. With a corpus of the same speaker undertaking the same tasks in two languages, we can investigate the effects of first language transfer in terms of phonetics, lexis and syntax (CROWDISENGDEU).

It is well-known that building speech corpora[1] is a time-consuming and expensive process: one estimate puts the cost of transcription at €1 per word, before the cost of any extra annotation [1]. Presumably the main expense in this figure is researcher time – skilled labourers with accompanying overheads. Extending the transcription work described in [2] by building a speech corpus from scratch, we present a method to collect spoken language corpora via crowdsourcing facilities, showing how we can reduce that cost considerably by distributing the work among multiple online workers.

Both corpora will be, or have been, transcribed by crowdsourcers; the preparation of both corpora is still in progress at the time of writing. We indicate how to stay informed of their development in section 5.

2. Corpus design

Both CROWDISENG and CROWDISENGDEU were designed to assess the suitability of crowdsourcing means to collect speech corpora. There is generally a shortage of spoken language corpora, whereas there is great demand for them from engineers working on automatic speech recognition (ASR), (computational) linguists intending to build natural language processing (NLP) resources trained on spoken rather than written data, and researchers across disciplines with their own research questions.

If it can be shown that crowdsourcing *works* for spoken corpus collection, then we potentially have a faster, cheaper method to access large numbers of people around the world, or subsets of these for researchers with specific interests, and a means to keep language models better up-to-date with current language trends and ongoing change – an issue common to the ubiquitous, widely-used but now aged Switchboard [3], Fisher [4] and Broadcast News [5] corpora (for instance). These high-quality, carefully-designed corpora were the outcome of huge efforts by research groups over many years. We instead propose a lightweight method (in researcher time) to collect speech corpora from 'the crowd' within months, or even weeks.

One might ask whether this lightweight method entails lower quality data. We address this issue by assessing a sample of crowdsourced soundfiles in section 4. However, ASR needs to, and does already deal with speech data captured in less-than-ideal recording environments (*e.g.* Apple's Siri, Microsoft's Cortana, Google's Voice Search). For instance, rather than laboratory conditions, data may very likely be captured by inbuilt device microphones and with much unwanted noise in the background. Thus we view this as a data type that is ecologically valid and much needed for training of resources.

For this Special Session on *Advanced Crowdsourcing for Speech and Beyond* we were awarded funding by Crowdee[2] and CrowdFlower[3] to carry out the corpus collection project detailed below. Crowdee is a crowdsourcing application for mobile and tablet devices using the Android operating system, and was identified as our source of crowd recordings. CrowdFlower

[1]We note that an occasional distinction is made between 'speech corpora' and 'spoken corpora' [1] but use the terms interchangeably here to mean 'a collection of spoken/speech data'.

[2]http://www.crowdee.de

[3]http://www.crowdflower.com

acts as an online platform for multiple crowdsourcing services and is used here for transcription, basic error annotation, and ratings of 'native-like-ness'.

2.1. English Corpus

Our primary motivation in proposing this project was to obtain a benchmark corpus of English native speakers undertaking tasks similar to those typically contained in learner corpora. There are many such tasks, and we decided to start with the business English domain. Hence, a majority (65%) of Crowdee funding was allocated to the recordings needed for CROWDISENG, enabling a maximum of 130 individuals to make contributions.

In the Crowdee job designed for CrowdISEng ('jobEN'), crowdworkers were required to be resident in the United Kingdom, United States or Canada, and it was a stated requirement of the task that English should be their mother tongue. They were also asked to find a quiet environment for recording, and were encouraged to attach a headset with external microphone rather than use the device's inbuilt microphone.

The general recording task was then explained, before the worker's consent was sought for the use and redistribution of their recordings for research purposes, and various metadata were collected: year-of-birth, gender, country of residence, number of years speaking English (used as the first alarm, if this total differed greatly from year-of-birth), highest level of education and degree subject if applicable, and mic type.

There were two versions of the English job (jobEN v1/v2), each of which was allocated an equal share of the money so that the same number of workers will complete each version if all goes well (*i.e.* no erroneously approved bad jobs – §3.1). And each version contains two business-related scenarios (Table 1).

	v1	v2
scen.1	starting a retail business	sports sponsorship
scen.2	hosting a business trip	starting a taxi company

Table 1: CROWDISENG*: 2 recording scenarios x 2 versions of jobEN.*

Workers were posed five questions (or 'prompts') about each scenario – for instance:

- What skills will you look for when hiring members of staff? (jobEN v1 scen.1);
- Can you suggest some appropriate gifts to give the visitors when they leave? (jobEN v1 scen.2);
- What are the benefits to companies of sponsoring sports people and sporting events? (jobEN v2 scen.1);
- Is it better to offer a 24-hour service with fewer drivers available at any one time, or a business hours service with lots of drivers on standby? (jobEN v2 scen.2).

Workers were asked to speak for approximately 15 seconds in response to each prompt. They had the facility to re-play and review their recording and were asked to do so before moving on to the next prompt. In total then, jobEN featured ten prompts and workers were expected to produce approximately 150 seconds (2 mins 30) of speech.

Workers were informed that the job would take ten minutes to complete, and were allowed up to twice this duration (*i.e.* 20 minutes) before it timed out. Payment of €2.50 was awarded to workers who provided ten recordings of sufficient duration and quality, and who apparently met the native speaker requirement (more on the quality control process in section 4 below).

2.2. German/English Corpus

The German/English task ('jobDE/EN') designed for the bilingual corpus (CROWDISENGDEU) was the same in design as the two versions of jobEN, except for the following key differences:

- workers needed to be bilingual in German/English, and their mother tongue could be either language;
- workers should be resident in Germany;
- workers were informed the job would take 15 minutes to complete (max 30 mins timeout);
- in addition to the metadata collected in jobEN, for jobDE/EN we asked for: number of years speaking German, formal instruction in English and/or German;
- the English scenarios were 1 and 2 of jobEN v1 and the German scenarios were translations of these (see Table 2);
- jobDE/EN features 20 prompts in total (10 prompts in 2 languages), and workers were therefore expected to produce approximately 300 seconds (5 mins) of speech;
- workers were paid €3.50 for completion of jobDE/EN, after quality assurance checks (§4), allowing for a maximum of 50 contributors to CROWDISENGDEU.

	EN	DE
scen.1	starting a retail business	Eröffnung eines Einzelhandels-geschäfts
scen.2	hosting a business trip	Organisieren einer Geschäftsreise

Table 2: CROWDISENGDEU*: 2 recording scenarios x 2 languages in jobDE/EN.*

3. Corpus collection

We now explain the supervised pipeline set up to collect and process the CROWDISENG and CROWDISENGDEU corpora. In broad overview, the steps are as follows:

1. collection of audio recordings via Crowdee;
2. transcription of recordings via CrowdFlower;
3. grammatical error correction via CrowdFlower;
4. forced alignment of transcriptions and soundfiles with SPPAS [6];
5. automatic tagging and parsing of transcriptions with the RASP System [7].

3.1. Recordings via Crowdee

Recordings were collected from crowd workers via Crowdee per the procedure described in section 2. The authors were sent notifications of any new job submissions and, having obtained a results CSV file via the Crowdee API, ran a supervised R program [8] to quality check each worker's soundfiles. The

program, made available in a public GitHub repository[4], makes various system calls to SoX[5] and FFmpeg[6] to obtain soundfile statistics and apply maximal amplification without clipping, and to convert the files from the MP4s received from Crowdee to the MP3s required by CrowdFlower (§3.2) and WAVs offered in the public release – all of which unfortunately implies a certain loss of sound quality (§4).

If a soundfile is found to be shorter than 10 seconds, or appears to be insufficiently loud (<0.01 mean normalized amplitude), the supervisor is alerted to the fact then prompted to review and approve or reject the file. If more than half of a worker's submitted files (their 'answer') are of insufficient quality, volume or quantity, the whole answer was rejected via Crowdee API along with an explanation why, the files were not put forward for transcription on CrowdFlower, and the worker did not receive payment.

Otherwise, if all appeared to be fine, and the worker was indeed perceived as a native speaker of the relevant language (English for jobEN; German or English for jobDE/EN), an approval status was posted to the Crowdee API, the worker received payment, and the soundfiles were put forward to CrowdFlower for the next stage in the pipeline. We acknowledge that perception of 'native-like-ness' is a subjective judgement; thus we were generous in our assessment, and hence we asked for further judgements from CrowdFlower workers.

3.2. Transcription via CrowdFlower

Approved Crowdee soundfiles were uploaded to CrowdFlower, where workers were asked to complete four tasks:

1. confirm that there is spoken content in the soundfile;
2. transcribe the speech content as faithfully as possible, using full stops to divide the text 'so that it makes most sense';
3. write a corrected version of the transcribed text;
4. how likely they think it is that English/German is the speaker's mother tongue? (scale of 1 to 5).

Each 'row' (recording) was 'judged' (*i.e.* worked on) by two different workers. There were ten rows to a 'page', upon completion of which, the worker would receive 0.90 US$[7]. Since this was a survey type of job and CrowdFlower imposes no delay on payment, there was less facility for quality control and approval/rejection with these jobs. CrowdFlower has the facility to 'quiz' workers with pre-determined gold standard questions, but this approach does not suit our task (as tasks 2-4 are to some degree subjective). Instead, we restricted the job to CrowdFlower's highest quality 'level 3' workers, and set a minimum threshold of 100 seconds working time per page. We could not specify mother tongue of workers, and therefore settled for residency requirements for any English language data from Crowdee jobEN and jobDE/EN – Australia, Canada, South Africa, U.K., U.S.A. – and German 'language skills' for the remaining German recordings from jobDE/EN.

[4]http://github.com/cainesap/crowdIScorpora

[5]http://www.ffmpeg.org

[6]http://sox.sourceforge.net

[7]Despite our concerns that CrowdFlower payment was too low per page, at 90¢, given previously expressed ethical concerns as to exploitation of crowdworkers and failure to at least match minimum wage rates [9], 'pay' was in fact the most positively rated aspect of our CrowdFlower task, scoring 4.5/5 in a survey of 30 respondents so far.

Upon completion, results were collected and analysed for answers to the above 4 questions. We present preliminary results in section 5, and we indicate how we plan to evaluate the agreement between the two workers. For 1 and 4, evaluation is a straightforward calculation over two numerical values; for 2 and 3, we opt to combine transcriptions following the ASR-based method described in [2], but will also make both transcription versions available in the corpus release.

3.3. Automatic annotation

Once the CrowdFlower transcriptions have been collected, various annotation layers can be added. Firstly, error annotations may be obtained by taking the difference of the transcribed and corrected texts. This gives us 'error zones' in the transcriptions which may contain one or two corrected hypotheses (workers do not always agree that a correction is needed). These could in turn be the subject of future crowdsourced judgements as to (a) validity and (b) selecting between hypotheses.

For example, the following two transcripts show how two different workers may disagree on the target hypothesis for the same soundfile, with error zones in parentheses, the 'error' marked <e>, and the 'correction' marked <c>:

(i) (Ø<e>|the<c>) most effective ways to advertise a new shop nowadays is on the internet especially on social networks like facebook because there can spread information about your new shop very cheap and (easy<e>|easily<c>).

(ii) (Ø<e>|the<c>) most effective ways to advertise a new shop nowadays is (on<e>|over<c>) the internet especially on social networks like facebook because there can spread information about your new shop very cheap and easy.

We do not attempt to decide between hypothesised corrections, where there is disagreement. Instead, we present all proposed error corrections made by the workers.

Secondly, we sought agreement on sentence boundaries, which we asked CrowdFlower workers to indicate with full stops (periods). Speech is of course not neatly punctuated, like writing usually is, so this is a purely inferential task. However, NLP tools are for the most part designed for and trained on written language, with the sentence a fundamental unit of analysis. The status of 'sentence' is more doubtful in spoken language, but for the time being the best strategy available is to adapt speech data to something like normal written form, and as part of this it is preferable to segment larger texts into smaller sentence-like units, where possible.

Again, where there is disagreement, this information is retained by virtue of both transcripts being made available. In order to decide on sentence boundaries in the single combined transcription, we will use a probabilistic language model to choose between competing segmentations of the text. For instance, consider the combined transcription below, which is annotated with two proposed sets of sentence boundaries, marked <1> and <2>:

appropriate gifts could be things the country is very well known or famous for like treats food clothes .<1> just things like this .<1>,<2> yeah .<1>,<2>

In this example there is agreement on the final 2 of the 3 hypothesised sentence boundaries. We would therefore accept

these two boundaries, and treat a decision on the first proposed boundary as an empirical matter.

Thirdly, each transcript was force-aligned with the WAV soundfiles thanks to SPPAS [6]. SPPAS alignment is based on the Julius Speech Recognition Engine and HTK-ASCII acoustic models. The output will be available both in XML and TextGrid formats, the latter for those wishing to work with Praat [10].

Finally, both workers' transcript versions for each soundfile, plus the single combined version, are tagged and parsed with the RASP System [7] with the following options:

```
$ rasp.sh -m -p'-n3 -oGITR -ph -s'-w
```

These options are explained as follows:

- `-m` allow multiple part-of-speech tags per word;
- `-p'-n3'` parse with maximum 3 possible trees;
- `-p'-oGITR'` outputting grammatical relations (G), inside-outside grammatical relation weighting (I), trees labelled with grammatical relations (T), and RMRS format (R; [11]), with CAPS for output in XML format;
- `-p'-ph'` for human-readable XML;
- `-p'-s'` use subcategorization frame probabilities [12, 13, 14];
- `-w` add character position input spans.

All transcripts will include the information provided by RASP; we provide example output for the string, "A Mercedes for instance would represent power", in our GitHub wiki[8].

As a result of these four steps, each soundfile is then rendered into written form as XML files containing maximally three transcriptions: two crowdsourced versions (where both exist), and an automatically-produced combination of the two. This set of XML files will be made available along with the WAV soundfiles in the public release of CROWDISENG and CROWDISENGDEU (§5).

4. Quality assurance

Quality assurance (QA) checks include the following:

1. By Crowdee workers (see also §3.1):
   (a) asked to use an external mic if possible, asked to find a quiet environment, asked to listen back to their recordings and re-do if of poor quality.
2. By CrowdFlower workers (see also §3.2):
   (a) asked if the soundfile has content;
   (b) asked to rate the speaker's 'native-like-ness'.
3. By the authors of this paper:
   (a) transcribe a sample of Crowdee soundfiles, treat this as 'gold' version as reference for crowdsourced transcription word-error-rates (WER);
   (b) inspect the transcription set of each CrowdFlower worker to check for whole-job failure;
   (c) manual inspection of SPPAS output for a sample of transcripts, for phone alignment error rates.

QA of (1a) was addressed via semi-automated inspection of soundfiles (§3.1); at the time of writing, 14% of Crowdee submissions were rejected for various reasons relating to recording quality, recording durations, and apparent non-suitability (*i.e.* non-nativeness) of the worker for the job.

Current results for (2a) indicate that 2% of soundfiles submitted to CrowdFlower had no content, indicating that our inspection of Crowdee soundfiles is not bulletproof. As for (2b), where the Crowdee worker's mother tongue is expected to be English, CrowdFlower workers have so far indicated that 98% of soundfiles are 'very likely' to have been recorded by a native speaker of the task language (English); where the mother tongue is expected to be German but the task is still in English, only 18% of CrowdFlower judgements were 'very likely' or 'quite likely' in this regard.

Initial sampling of 2.5% of English and 1.25% of German transcriptions received from CrowdFlower results in mean WERs of 30% and 42% respectively (3a). The English WER is comparable to the WER reported in [2], and further work remains to be done to investigate whether errors are systematic and/or determined by recording factors such as mother tongue of speaker, soundfile quality, *etc*. With (3b), we found that our original 'German language skill' setting was insufficiently strict, with a worker rejection rate of 68% due to incomplete or missing transcriptions; we have now added 'location:Germany' to the conditions, and hope for stricter control and lower WER as a result. In comparison, the rejection rate for missing transcription was just 5% for the English CrowdFlower task.

5. Corpus readiness

As mentioned, the collection of corpus data is a work-in-progress. At the time of writing, in March 2015, we have received 90% (if the available funds are used optimally) of CROWDISENGDEU recordings and 12% of CROWDISENG recordings. The small number of submissions to jobEN thus far are a result of new status outside of Germany; we are working on spreading the word about our jobs and Crowdee across multiple outlets.

All soundfiles received so far have been transcribed via CrowdFlower and at time of writing, we are preparing XML and TextGrid files for dissemination alongside the WAV recordings. Their full availability will be announced at a dedicated Speech and Language Data Repository URL[9], where you can currently find sample files and the latest descriptive statistics about the corpora.

6. Summary of contributions

- Two new speech corpora, to be made freely available to other researchers: CROWDISENG, an English native speaker corpus of crowdsourced answers to business-topic questions; CROWDISENGDEU, a German/English corpus containing bilingual speakers answering the same questions;
- Resources for corpus preparation and quality assurance, made available via a GitHub repository;
- Demonstration of the use of crowdsourcing to collect language resources.

7. Acknowledgements

This paper reports on research supported by Crowdee, CrowdFlower, and Cambridge English, University of Cambridge.

[8] http://github.com/cainesap/crowdIScorpora/wiki/RASP-Output

[9] http://sldr.org/ortolang-000913, http://sldr.org/ortolang-000914

8. References

[1] N. Ballier and P. Martin, "Developing corpus interoperability for phonetic investigation of learner corpora," in *Automatic treatment and analysis of learner corpus data*, A. Díaz-Negrillo, N. Ballier, and P. Thompson, Eds. Amsterdam: John Benjamins, 2013.

[2] R. van Dalen, K. Knill, P. Tsiakoulis, and M. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers, 2015.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proceedings of Acoustics, Speech, and Signal Processing (ICASSP-92)*. IEEE, 1992.

[4] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a resource for the next generations of speech-to-text," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, 2004.

[5] J. Garofolo, J. Fiscus, and W. Fisher, "Design and preparation of the 1996 Hub-4 Broadcast News benchmark test corpora," in *Proceedings of the DARPA Speech Recognition Workshop*, 1997.

[6] B. Bigi, "SPPAS: a tool for the phonetic segmentation of speech," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, 2012.

[7] T. Briscoe, J. Carroll, and R. Watson, "The second release of the RASP System," in *Proceedings of the COLING/ACL 2006 Interactive Presentations Session*. Association for Computational Linguistics, 2006.

[8] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org

[9] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through Crowdsourcing: towards best practice guidelines," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, 2014.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 5.4.08, retrieved 24 March 2015," 2015, uRL http://www.praat.org.

[11] A. Copestake, "Semantic composition with (Robust) Minimal Recursion Semantics," in *Proceedings of the ACL Workshop on Deep Linguistic Processing*. Association for Computational Linguistics, 2007.

[12] T. Briscoe and J. Carroll, "Automatic extraction of subcategorization from corpora," in *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, 1997.

[13] T. Briscoe, "Dictionary and System Subcategorisation Code Mappings," 2000, unpublished manuscript, University of Cambridge Computer Laboratory.

[14] P. Buttery and A. Caines, "Reclassifying subcategorization frames for experimental analysis and stimulus generation," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, 2012.