Crowdsourcing a multilingual speech corpus: recording, transcription and annotation of the CROWDED CORPUS

Andrew Caines[1], **Christian Bentz**[2], **Calbert Graham**[1], **Tim Polzehl**[3], **Paula Buttery**[1]
[1]The ALTA Institute, Dept. Theoretical & Applied Linguistics, University of Cambridge, U.K.
[2]Institute of Linguistics, Universität Tübingen, Germany
[3]Telekom Innovation Laboratories / Technische Universität Berlin, Germany
apc38@cam.ac.uk, chris@christianbentz.de, crg29@cam.ac.uk, tim.polzehl@qu.tu-berlin.de, pjb48@cam.ac.uk

Abstract

We announce the release of the CROWDED CORPUS: a pair of speech corpora collected via crowdsourcing, containing a native speaker corpus of English (CROWDED_ENGLISH), and a corpus of German/English bilinguals (CROWDED_BILINGUAL). Release 1 of the CROWDED CORPUS contains 1000 recordings amounting to 33,400 tokens collected from 80 speakers and is freely available to other researchers. We recruited participants via the Crowdee application for Android. Recruits were prompted to respond to business-topic questions of the type found in language learning oral tests. We then used the CrowdFlower web application to pass these recordings to crowdworkers for transcription and annotation of errors and sentence boundaries. Finally, the sentences were tagged and parsed using standard natural language processing tools. We propose that crowdsourcing is a valid and economical method for corpus collection, and discuss the advantages and disadvantages of this approach.

Keywords: crowdsourcing, speech corpus, annotation

1 Introduction

We announce the release of two corpora collected via crowdsourcing: a native speaker corpus of English (CROWDED_ENGLISH), and a corpus of German/English bilinguals (CROWDED_BILINGUAL)[1]. We refer to these datasets as the CROWDED CORPUS, where *Crowd* refers to our data source, $E = English$, and $D = Deutsch$. Both corpora involve the speakers answering questions about selected business topics, and were motivated with the following two research questions in mind:

a. With tasks and topics comparable to typical language learning oral exams, we can start to address the question, 'what would a native speaker say in a similar situation?' Hence we collected CROWDED_ENGLISH;

b. 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. Hence we collected CROWDED_BILINGUAL.

It is well-known that building speech corpora[2] 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 (Ballier and Martin, 2013). Presumably the main expense in this figure is researcher time – skilled labourers with accompanying overheads. Extending the crowdsourced transcription work described in (Cooper et al., 2014; van Dalen et al., 2015) by crowdsourcing both recordings and transcriptions, 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.

Concerns have been expressed as to the quality of crowdsourced data, which some assess as part of a trade-off for speed and economy (Snow et al., 2008; Madnani et al., 2011; Ball, 2014), with others describing methods to filter out errors (Gadiraju et al., 2015; Schmidt et al., 2015), or indeed encouraging researchers to 'embrace' error (Jamison and Gurevych, 2015; Krishna et al., 2016). On the other hand, as the old adage goes, you only get what you pay for. Crowdsourcing may be markedly cheaper than 'expert' labour, but as with any such market, higher rates of pay will generally yield better output (Sabou et al., 2014; Litman et al., 2015). We describe our quality assessments of our crowdsourced recordings and transcription: while acknowledging that there are some problems with crowdsourced data we conclude that indeed 'they can help', given the low cost and access to a more distributed population than is usually the case in academic research.

Both corpora are transcribed by crowdsourcers and annotated for a number of features: grammatical error, sentence boundaries, part-of-speech tags and grammatical relations. The corpora are freely available to other researchers under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 international licence (CC BY-NC-SA 4.0)[3]. We are continuously adding new material and will do so until the money runs out[4]. Release 1 of the CROWDED CORPUS contains 80 individual contributions, amounting to 1000 short recordings and approximately 33,400 tokens.

[1]A poster with a similar title was presented at Corpus Linguistics 2015: since then we have extended our evaluations of data quality, collected more recordings, and made the first data release public.

[2]We note that an occasional distinction is made between 'speech corpora' and 'spoken corpora' (Ballier and Martin, 2013) but use the terms interchangeably here to mean 'a collection of spoken/speech data'.

[3]http://creativecommons.org/licenses/by-nc-sa/4.0

[4]We will apply for more funding if the corpus is well received.

2 Corpus design

We propose that with the CROWDED CORPUS we can assess the suitability of crowdsourcing as a means to collect speech corpora. Such resources are generally in short supply, 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 various 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, and a means to keep language models better up-to-date with current language trends and ongoing change in spoken usage – an issue common to the ubiquitous, widely-used but now aged Switchboard (Godfrey et al., 1992), Fisher (Cieri et al., 2004) and Broadcast News (Garofolo et al., 1997) 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 unwanted background noise. Thus we view this as a data type that is ecologically valid and much needed for training of resources.

Our work is generally part of the Automated Language Teaching & Assessment project, funded by Cambridge English Language Assessment[5]. In addition we were awarded specific funds by Crowdee[6] and CrowdFlower[7] to carry out the corpus collection project detailed below. Crowdee is a crowdsourcing application for mobile devices using the Android operating system, and was identified as our source of crowd recordings. CrowdFlower 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 Monolingual English corpus: CROWDED_ENGLISH

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 focus initially on the business domain.

In `jobEN`, the Crowdee 'job' designed for CROWDED_ENGLISH, 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. The native speaker status of our recruits was subsequently assessed by the authors plus the crowdworkers who transcribed their recordings. The recruits 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, plus microphone type.

There were two versions of the English job (`jobEN` v1/v2), each of which was allocated an equal share of available funds. Each version contains two business-related scenarios (scen.1, scen.2; Table 1) about which the crowdworkers were asked to respond to five questions (or 'prompts'). For example –

- 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).

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

Table 1: CROWDED_ENGLISH: two recording scenarios and two versions of `jobEN`.

Workers were asked to speak for approximately 15 seconds in response to each prompt. They had the facility to replay 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).

[5] `http://alta.cambridgeenglish.org`
[6] `http://www.crowdee.de`
[7] `http://www.crowdflower.com`

2.2 Bilingual German/English corpus: CROWDED_BILINGUAL

The German/English task (jobDE/EN) designed for the bilingual corpus (CROWDED_BILINGUAL) was similar in design to jobEN, except for the following key differences:

- Workers needed to define themselves as German/English bilinguals, and their mother tongue could be either language;
- In addition to the metadata collected in jobEN, for jobDE/EN we asked for: number of years speaking German, formal instruction in English (*yes/no*), formal instruction in German (*yes/no*);
- The English part of jobDE/EN involved scen.1 and scen.2 from jobEN v1 (Table 1) and the German scenarios were translations of these (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 (section 4).

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

Table 2: CROWDED_BILINGUAL: two recording scenarios and two languages in jobDE/EN.

3 Corpus collation

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

i. Collection of audio recordings via Crowdee;

ii. Transcription of recordings via CrowdFlower;

iii. Annotation of errors via CrowdFlower;

iv. Annotation of speech-unit boundaries via CrowdFlower;

v. Automatic part-of-speech tagging and parsing of transcriptions using Stanford Core NLP (Chen and Manning, 2014).

3.1 Recordings via Crowdee

Recordings were collected from crowdworkers via Crowdee per the procedure described in section 2. The authors were notified of any new job submissions and, having obtained a metadata file for the new recordings from the Crowdee API, we ran a supervised R program (R Core Team, 2015) to quality check each worker's soundfiles. The program makes various system calls to SoX[8] and FFmpeg[9] 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 (section 3.2) and WAVs offered in the public release.

If a soundfile is found to be shorter than 10 seconds, or appears to be insufficiently loud (a mean normalized amplitude <0.01 decibels), the supervisor is alerted to the fact and 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 with an explanation why, the files were not put forward for transcription on CrowdFlower, and the worker did not receive payment.

Additionally, we verified that the speaker 'sounded like' a native speaker. Even though listeners are known to excel at the task of native speaker versus non-native speaker distinction (McCullough and Clopper, 2016), we acknowledge that perception of nativelikeness remains a somewhat subjective judgement. Thus we were cautious in our assessment, and decided to ask for further judgements from CrowdFlower workers which we could refer to if in doubt (section 3.2).

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 corpus collation process.

Release 1 of the corpus contains a total of 80 individual contributions, amounting to 1000 short recordings with a mean duration of 16.5 seconds. Table 3 shows the breakdown per corpus, and Figure 1 illustrates the demography of corpus contributors: fairly balanced for gender, mainly 20-40 years old, resident in Germany, the U.K. and U.S.A., and educated to university level.

	ENGLISH	BILINGUAL
contributors	33	47
recordings	296	704
recording duration (μ)	21.6	14.3
recording duration (σ)	6.0	1.8
recording duration (max.)	30.0	15.0
recording duration (min.)	10.3	5.5

Table 3: Recordings via Crowdee in the CROWDED corpora.

[8] http://www.ffmpeg.org
[9] http://sox.sourceforge.net

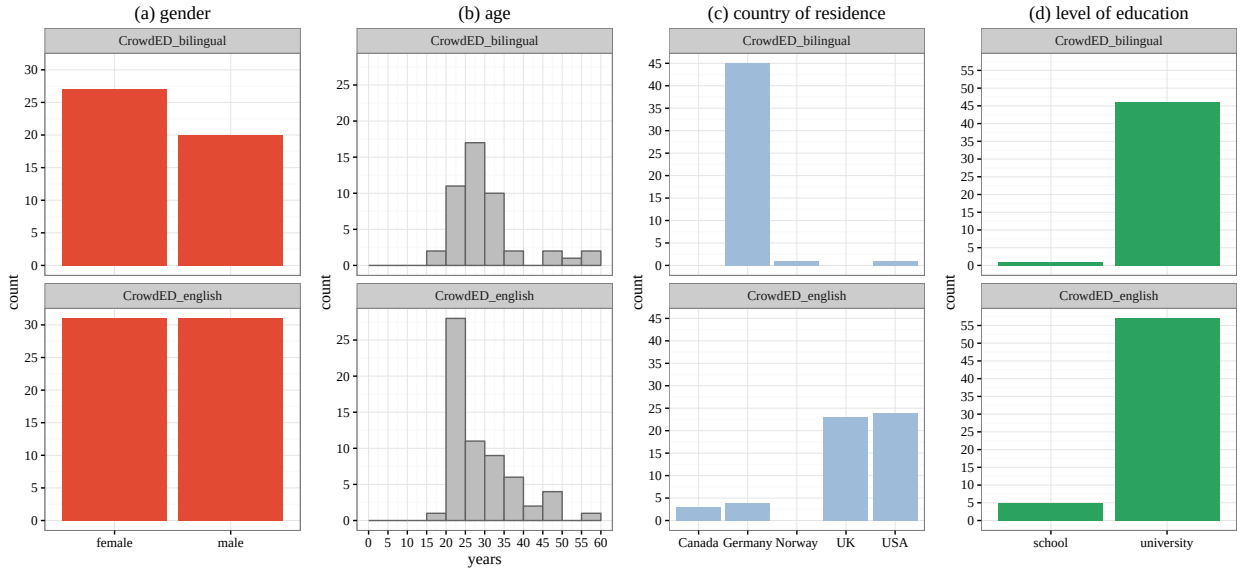

Figure 1: Demographics of contributors to the CROWDED corpora.

3.2 Transcription via CrowdFlower

Approved Crowdee soundfiles were uploaded to CrowdFlower, where workers were asked to complete four tasks, with the instructions either in English or German accordingly (Figure 2):

T1. Confirm that there is spoken content in the soundfile (*true/false*);

T2. Transcribe the speech content as faithfully as possible, using full stops to divide the text 'so that it makes most sense' (*free text*);

T3. Write a corrected version of the transcribed text (*free text*);

T4. How likely they think it is that English/German is the speaker's mother tongue (*scale 1 to 5*).

Each 'row' (recording) was 'judged' (worked on) by two different workers. There were ten rows to a page, upon completion of which, the worker would receive 0.90 USD[10]. CrowdFlower offers the facility to 'quiz' workers with pre-determined gold standard questions, but this approach does not suit our task (as T2,T3,T4 are to some degree subjective). Thus, as CrowdFlower imposes no delay on payment, there was less facility for quality control and approval/rejection with these jobs. We return to this issue in section 4.

To achieve some kind of prior quality control, we restricted the job to CrowdFlower's most 'trusted' level 3 workers, and set a minimum threshold of 300 seconds working time per page. We could not specify the workers' mother tongue, 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' plus residency in Germany for the remaining German recordings from `jobDE/EN`.

We inspected every worker's set of ten transcripts, visually distinguishing between apparently substantial output and evidently substandard efforts: for example, one worker simply supplied single words or small phrases as a transcript of each recording (an unlikely set of transcripts, given the mean recording durations seen in Table 3). Transcript sets such as these were removed from the corpus, and the recordings in question were resubmitted to CrowdFlower for transcription and annotation at additional expense. With these bad transcripts removed, we found that workers had spent an average of 33 (German) and 37 (English) minutes to write their ten transcriptions.

Seeking agreement between workers was our next challenge. For T1 and T4, evaluation is a straightforward calculation over two numerical values – $sum(x)$ for T1 and $mean(x)$ for T4, where x is a vector of numbers, we seek a value of 2 or more for T1, and report the rounded mean for T4. For T2 and T3 we make both transcription versions available in the corpus release, and may opt in future development to combine transcriptions following the ASR-based method described in (van Dalen et al., 2015).

3.3 Error annotation

As well as a faithful transcription of the recording, we also asked CrowdFlower workers to write a 'corrected' version, opting to leave this concept underspecified so as not to predispose the workers to seek out particular features, or to overly constrain what they would correct. Viewing error as another subjective notion, we wanted to see what the workers would come up with, and it remains an open question whether this is the optimal approach to crowdsourced error

[10]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 (Sabou et al., 2014), 'pay' was in fact the most positively rated aspect of our CrowdFlower task in the post-job participant survey (the other dimensions being, 'instructions clear', 'test questions fair', 'ease of job').

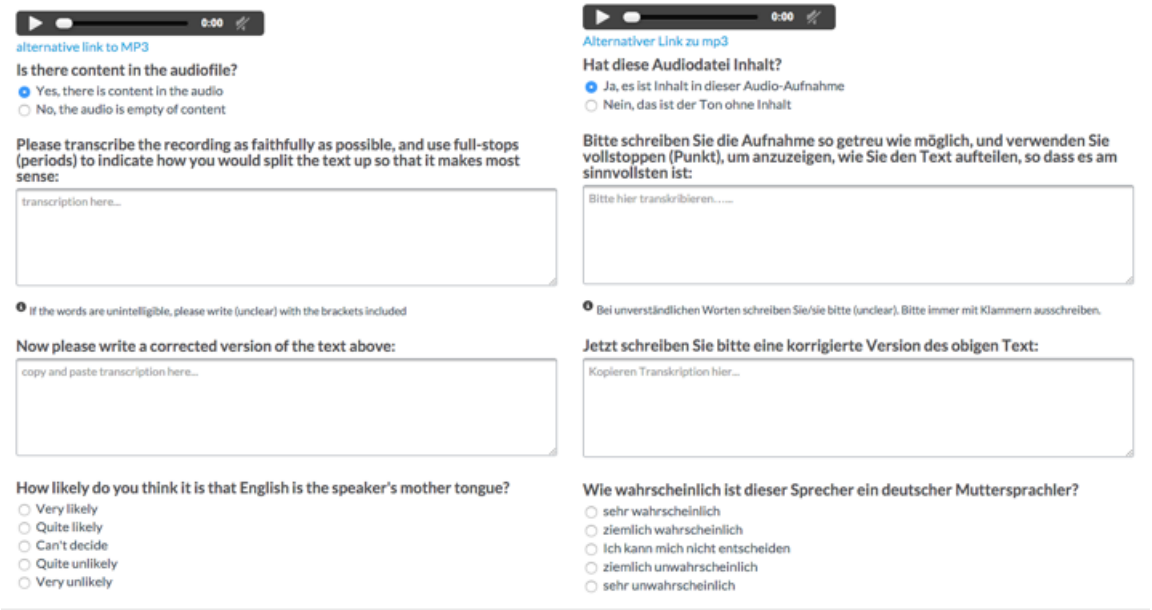

Figure 2: CrowdFlower transcription jobs for English (left) and German (right) recordings.

annotation.

Since at least two crowdworkers worked on each recording, we have multiple corrected versions of the transcript. We do not attempt to decide between hypothesised corrections, where there is disagreement. Instead, we present all proposed error corrections made by the workers. For example, texts (1) and (2) show how two different workers may disagree on the target hypothesis for the same recording, with 'error zones' in parentheses, the 'error' marked <e>, and the 'correction' marked <c>:

(1) (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>).

(2) (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 note in this example that the workers have opted to focus on lexical corrections only, opting to ignore the possible syntactic error that is, *there can spread information*. We propose that a full evaluation of the error annotations in the CROWDED CORPUS would be a fruitful future project, involving further crowdsourced judgements as to correction validity and to select between hypotheses.

3.4 Speech unit boundary annotation

We also sought annotation of sentence-like boundaries, which we asked CrowdFlower workers to indicate with full stops (periods). Speech is of course not neatly punctuated as writing usually is. Furthermore, NLP tools are for the most part trained on and designed to work best with written language, in which the sentence is a fundamental unit of analysis. Until speech-specific NLP tools are produced, the best strategy available is to adapt speech data to something like normal written form (Caines and Buttery, 2014; Moore et al., 2015), and as part of this it is preferable to segment larger texts into smaller sentence-like units, where possible. However, the status of the *sentence* concept is more doubtful in spoken language, and thus following guidelines set out by the Linguistic Data Consortium we refer to the *SU* ('speech unit') as it carries less implication of *grammaticality* than 'sentence' does (Strassel, 2003).

As with error annotation, where there is disagreement this information is retained by virtue of both transcripts being made available, thereby acknowledging that SU delimitation is a highly subjective task. For instance, consider the combined transcription below, which has been annotated with different hypothesised SU boundaries by two different crowdworkers, 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 two of the three hypothesised sentence boundaries. We would therefore accept these two boundaries, and treat a decision on the first proposed boundary (after *clothes*) as an empirical matter. This is another potential task for future work in which one might use, for example, a probabilistic language model to choose between competing SU hypotheses.

3.5 Part-of-speech tagging and parsing

Finally we processed the faithful transcript (as opposed to the corrected one) of every soundfile with Stanford Core

NLP (Manning et al., 2014). The XML output for both German and English transcripts gives part-of-speech tags, a constituency parse in XML format, and an additional dependency parse in Universal Dependencies format (Marneffe et al., 2014).

4 Corpus quality

Corpus quality assurance (QA) checks included the following:

1. By Crowdee workers:
   i. Asked to use an external microphone if possible;
   ii. Asked to find a quiet environment;
   iii. Asked to listen back to their recordings and re-do if of poor quality.
2. By CrowdFlower workers:
   i. Asked if the soundfile has content;
   ii. Asked to rate the speaker's nativelikeness.
3. By the authors of this paper:
   i. Obtain soundfile statistics using SoX and the inspection of any recordings deemed to be too short (<10sec.) or quiet (<0.01dB);
   ii. Inspect the transcription set of each CrowdFlower worker to check for whole-job failure;
   iii. Transcribe a sample of Crowdee soundfiles, treat these as the gold standard reference transcripts for calculation of word-error-rates (WER).

4.1 Crowdee recordings

At the time of writing, 21% 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. In most cases we were able to reject the recordings before 'auto approval' kicked in three days after submission. In those few cases for which we responded too late, this unfortunately has to be assigned as an extra cost in the price of collecting the corpus.

We actually put 1724 Crowdee recordings forward for transcription in CrowdFlower, but for reasons explained in the next section we do not have two transcriptions for all of them, and so only 1000 of these feature in the first release of the CROWDED CORPUS. Figure 3 shows the nativelikeness ratings given by CrowdFlower workers having heard each recording. In CROWDED_BILINGUAL in which the speakers could be either German or English native speakers so long as they were bilingual, we see that the majority were perceived to be German native speakers, which is probably indicative of Crowdee's origins as a Berlin-based application with a local workforce. CROWDED_ENGLISH on the other hand is confirmed as a predominantly English native speaker corpus.

4.2 CrowdFlower transcriptions

We had a greater problem with the quality of CrowdFlower data: lack of the facility to withhold payment until we could review workers' output meant that we paid for many transcripts that were too brief, nonsensical, or completely missing[11]. We realise that part of the appeal of the crowdsourcing method is that it is a relatively unsupervised, non-labour-intensive process for the corpus collators, and we do not wish to undermine that. However, even a cursory visual check on a set of transcripts immediately distinguishes the very bad data from the normal, good quality data. Unfortunately, even such a speedy task could not be completed before payment had already been made to the CrowdFlower workers. Alternatively at source, an automatic check using regular expressions would enable filtration of bad data at the point of worker submission, pre-payment; however, no such facility was available in CrowdFlower at the time of writing, even though it would be greatly beneficial to quality control.

Transcription quality from CrowdFlower workers has been somewhat problematic: we have rejected six-in-ten of the 3300 transcriptions received so far on the basis of their poor quality. Thus we have two good quality transcriptions for 1000 of our Crowdee recordings, and this is the dataset we release first. For the remaining 724 recordings, the relevant soundfile has been re-uploaded to CrowdFlower for another attempt at crowdsourced transcription. Good quality transcription and recording sets will be made available in a future release of the CROWDED CORPUS.

We randomly sampled 2.5% of the English and German transcriptions received from CrowdFlower, transcribed them and obtained mean word error rates of 30% and 42% respectively. In future work we could return to investigate whether transcription errors are systematic and/or determined by recording factors such as mother tongue of speaker, soundfile quality, *etc*.

Despite these transcription error rates, we maintain that crowdsourcing is an effective method for corpus collation for two reasons: first, it may be that there are ways to better control data quality at the transcription stage – whether via CrowdFlower settings, or by using a different service; secondly, it should be remembered that transcription is a highly subjective activity for which 'gold standard' is a misleading term. Instead, we can think of individual transcripts as hypotheses, each of which may capture different *truths* (and errors) from the target recording. We therefore have data akin to *crowd truth*: the idea that "measuring annotations on the same objects of interpretation across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations" (Aroyo and Welty, 2015). In future we intend to investigate whether merging multiple transcriptions to a single version improves transcription quality, as the work by van Dalen et al. (2015) suggests it should.

[11] *e.g.* One worker's ten transcripts were, `fdzvt`, `gfbh`, `srevt`, `jxfb fgh`, `dsfbzt`, `zv fyxdf`, `dsaf as`, `fdghx`, `gtsyv`, `sfdtvg`. This was not the only such example.

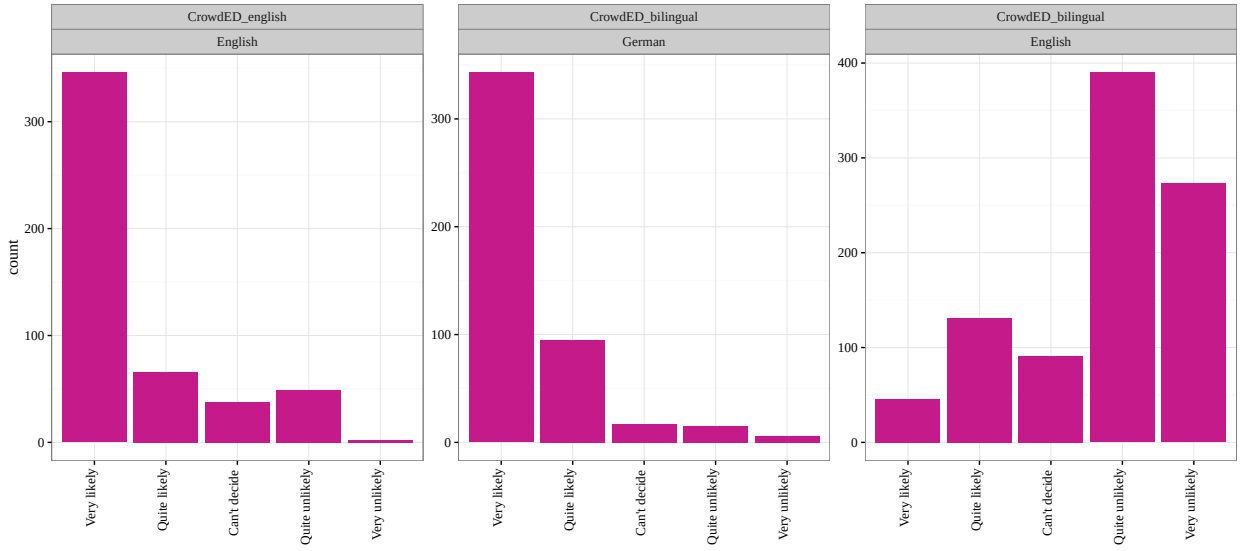

Figure 3: 'Nativelikeness' of Crowdee speakers as judged by CrowdFlower workers.

5 Summary

We have presented the CROWDED CORPUS, two new crowdsourced speech corpora in which recruits respond to business-topic questions (where constraint of topic may prove to be a useful design feature (Caines and Buttery, in press)). CROWDED_ENGLISH features native speakers of English; CROWDED_BILINGUAL features German/English bilinguals. Release 1 of the CROWDED CORPUS contains 1000 recordings from 80 speakers, which range from 5 to 30 seconds in duration (Table 4). Each recording has been transcribed at least twice, amounting to 33,400 tokens (summing over means of the transcript set for each recording).

	ENGLISH	BILINGUAL
recruited speakers	33	47
approved recordings	296	704
approved transcripts	592	1408
token count	14,570	18,848

Table 4: CROWDED CORPUS release 1.

All Crowdee recordings and corresponding CrowdFlower transcripts are made freely available to other researchers[12]. By using crowdsourcing services for recording, transcription and annotation, we have demonstrated a fast and efficient method to collate speech corpora. Having to date spent €500 on Crowdee recordings, $500 on CrowdFlower transcriptions, and an estimated £2000 in researcher time, at today's currency exchange rates[13] this amounts to €3539, or €0.11 per token in the corpus, well below the €1/token rate referred to by Ballier and Martin (2013). With this there are some issues of data quality, but these are not insurmountable, and may in fact be part and parcel of a move toward 'crowd truth' (Aroyo and Welty, 2015). We view the CROWDED CORPUS as useful for investigations of first language transfer, learner-native speaker comparisons, and other as yet un-anticipated purposes.

[12] http://apc38.user.srcf.net/resources/#crowded

[13] Source: http://www.xe.com, accessed 2016-03-09.

6 Acknowledgements

We gratefully acknowledge receipt of funding from Crowdee, CrowdFlower, and Cambridge English, University of Cambridge. We are grateful for the assistance of André Beyer of Technische Universität Berlin, Wil Stevens of CrowdFlower, and Dr Rogier van Dalen of Cambridge University Engineering Department. We also thank the three anonymous LREC reviewers for their helpful comments and questions.

7 Bibliographical References

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ball, P. (2014). Strength in numbers. *Nature*, 506(7489):422–423.

Ballier, N. and Martin, P. (2013). Developing corpus interoperability for phonetic investigation of learner corpora. In Ana Díaz-Negrillo, et al., editors, *Automatic treatment and analysis of learner corpus data*. John Benjamins, Amsterdam.

Caines, A. and Buttery, P. (2014). The effect of disfluencies and learner errors on the parsing of spoken learner language. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*.

Caines, A. and Buttery, P. (in press). The effect of examination topic on opportunity of use in learner corpora. In Vaclav Brezina et al., editors, *Learner Corpus Research: New perspectives and applications*. Bloomsbury, London.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Cieri, C., Miller, D., and Walker, K. (2004). 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.

Cooper, S., Jones, D., and Prys, D. (2014). Developing further speech recognition resources for Welsh. In *Proceedings of the First Celtic Language Technology Workshop*.

Gadiraju, U., Siehndel, P., Fetahu, B., and Kawase, R. (2015). Breaking bad: Understanding behavior of crowd workers in categorization microtasks. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15.

Garofolo, J., Fiscus, J., and Fisher, W. (1997). Design and preparation of the 1996 Hub-4 Broadcast News benchmark test corpora. In *Proceedings of the DARPA Speech Recognition Workshop*.

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

Jamison, E. and Gurevych, I. (2015). Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Krishna, R., Hata, K., Chen, S., Kravitz, J., Shamma, D. A., Fei-Fei, L., and Bernstein, M. S. (2016). Embracing error to enable rapid crowdsourcing. *ArXiv e-prints*.

Litman, L., Robinson, J., and Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, 47(2):519–528.

Madnani, N., Chodorow, M., Tetreault, J., and Rozovskaya, A. (2011). They can help: Using crowdsourcing to improve the evaluation of grammatical error detection Systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Marneffe, M.-C. D., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford Dependencies: a Cross-Linguistic Typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

McCullough, E. A. and Clopper, C. G. (2016). Perceptual subcategories within non-native English. *Journal of Phonetics*, 55:19–37.

Moore, R., Caines, A., Graham, C., and Buttery, P. (2015). Incremental dependency parsing and disfluency detection in spoken learner English. In *Proceedings of the 18th International Conference on Text, Speech and Dialogue (TSD)*. Berlin: Springer-Verlag.

R Core Team, (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Sabou, M., Bontcheva, K., Derczynski, L., and Scharl, A. (2014). Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Schmidt, M., Müller, M., Wagner, M., Stüker, S., Waibel, A., Hofmann, H., and Werner, S. (2015). Evaluation of crowdsourced user input data for spoken dialog systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast – But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October. Association for Computational Linguistics.

Strassel, S., (2003). *Simple metadata annotation specification*. Version 5.0.

van Dalen, R., Knill, K., Tsiakoulis, P., and Gales, M. (2015). Improving multiple-crowd-sourced transcriptions using a speech recogniser. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.