Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6532–6541
Marseille, 11–16 May 2020

Open-source Multi-speaker Corpora of the English Accents in the British Isles

Işın Demirşahin, Oddur Kjartansson, Alexander Gutkin, Clara Rivera
Google Research
6 Pancras Square, London, N1C 4AG, United Kingdom
{isin,oddur,agutkin,rivera}@google.com

Abstract

This paper presents a dataset of transcribed high-quality audio of English sentences recorded by volunteers speaking with different accents of the British Isles. The dataset is intended for linguistic analysis as well as use for speech technologies. The recording scripts were curated specifically for accent elicitation, covering a variety of phonological phenomena and providing a high phoneme coverage. The scripts include pronunciations of global locations, major airlines and common personal names in different accents; and native speaker pronunciations of local words. Overlapping lines for all speakers were included for idiolect elicitation, which include the same or similar lines with other existing resources such as the CSTR VCTK corpus and the Speech Accent Archive to allow for easy comparison of personal and regional accents. The resulting corpora include over 31 hours of recordings from 120 volunteers who self-identify as native speakers of Southern England, Midlands, Northern England, Welsh, Scottish and Irish varieties of English.

Keywords: speech corpora, regional dialects, phonology, English, open-source

1. Introduction

English is one of the largest languages of the world, with the largest total number of native and foreign language speakers and the third largest number of native speakers (Eberhard et al., 2019). Inevitably, there is a large number of varieties of English spoken by different communities around the world. Historically major English varieties, such as American English, British English, Canadian English, Australian English, Irish English, New Zealand English have diverged from one another at different points in history, has gone through different phonological changes and came into contact with different languages (Leith, 1997; Smith, 1998).

For other varieties such as Indian English or Nigerian English, although English is an official language or the language of education, it is not necessarily the native language or the primary language of everyday communication for the majority of the speakers (Kachru, 1976; Lowenberg, 1991; Coelho, 1997). In such varieties, additional factors, such as influence from the regional languages, such as influences from Hausa, Yaruba and Igbo on Nigerian English (Olaniyi, 2014; Kperogi, 2015; Isiaka, 2019), come into play.

Usually, but not always, there is a dominant variety of English that the dialect originated from, but it is altered significantly enough that it has evolved into a new variety of English (Gut, 2008). Most of the time the speakers have one or more other native languages which interact with their English phonology and vocabulary. The major contact languages in this case are predominantly the local languages of the region. In addition, there might be influences from other remote languages or other varieties of English.

All English varieties differ in a number of ways such as phonology, morphology, syntax, and vocabulary (Kortmann et al., 2008; Trudgill and Hannah, 2017). In this work, our main focus is the accents, i.e. the phonological differences between the varieties of English, with a special focus on the accents spoken in the British Isles as the first set of recordings. Even though our first short term goal was very narrow in the great scheme of things, the recording scripts were curated in a way that they should provide high phoneme coverage and phonetic environments that will be informative for all English varieties. The recording lines consist of idiolect elicitation lines targeting a speaker's individual accent (Labov, 1972; Barlow, 2010), Global English lines targeting phonological differences in all varieties of English, and additional localized lines. A dataset of crowdsourced Nigerian English recordings following the same phonological approach and using the same Global English lines has also been released (Google, 2019a).

Here we present a freely available dataset (Google, 2019b) of high quality recordings from volunteers that self-identify as speaking with one of the following British Isles accents: Southern England, Midlands, Northern England,[1] Welsh English, Scottish English and Irish English. The accents were selected based on the number of speakers as well as the availability of volunteers. The scripts for the British Isles recordings consist of the Global English accent elicitation lines mentioned above and lines that are localized for both British Isles in general and the specific region for the accent in question in particular. This dataset is intended for the linguistic analysis of these accents, as well as to be used in speech technology applications.

This paper is organized as follows: The next section provides a brief survey of the related corpora. Section 3 describes the phonological approach followed in the building of this dataset. Section 4 provides a detailed description of the script curation. Section 5 describes the recording process and section 6 presents the details of the dataset. Finally, section 7 concludes this paper.

2. Related Corpora

English is arguably the best represented language when it comes to speech corpora. Many initiatives undertaken by the governments, academia and the industry have resulted in many databases, mostly of US English, geared towards specific applications. Besides US English, many

[1] The very broad division of England into three broad geographic regions is due to Kortmann et al. (2008). More details on our choice of speakers are provided in Section 5 .

corpora for other accents of English were collected as well, such as the Australian National Corpus (Cassidy et al., 2012), the Wellington Corpus of Spoken New Zealand English (Holmes et al., 1998), Corpus of Spoken Professional American-English (Barlow, 2000), the NIE Corpus of Spoken Singapore English (Deterding and Low, 2001), mobile database of Indian English (Agrawal et al., 2012) and many more. One of the resources which demonstrates huge phonological variability of English accents and dialects is the International Dialects of English Archive (IDEA), which is an online collection of native and non-native accents of English from around the world (Persley, 2013). The archive contains 1,500 samples from 120 countries and territories, and more than 170 hours of recordings (Meier, 1998).

The corpora of English dialects collected in the United Kingdom and Ireland include several very interesting projects. One of the earliest large scale collections was for the British National Corpus (BNC), a 100 million word corpus of modern British English, originally produced by a consortium of dictionary publishers and academic researchers between 1990 and 1994 (Burnard, 2002). Due to the technological limitations at the time, the proportion of written to spoken material in the BNC is about 10-to-1. The audio was recorded in many different conditions and scenarios (such as interviews, public meetings, leisure contexts). The speech portion of the corpus was recently digitized from the analogue audio cassette tapes deposited at the British Library Sound Archive and is available together with associated transcription and annotation files (Coleman et al., 2012). The Speech Accent Archive (Weinberger and Kunath, 2011) was developed for demonstrating the typology of English accents. It supports many dialects of English worldwide, but there is a limited number of samples per dialect. The IViE Corpus (Nolan and Post, 2014) was recorded to facilitate the systematic investigation within experimental phonetics of intonational variation in accents of the British Isles. The Freiburg English Dialect Corpus (Anderwald and Wagner, 2007) was developed with an added focus on non-standard morphosyntax, in addition to the attention to phonetic and phonological details. It consists of 370 texts, which total roughly 2.45 million words of text or about 300 hours of speech, excluding all interviewer utterances, collected in nine regions of the United Kingdom (such as the Hebrides and Scottish Highlands). The Scottish Corpus of Texts and Speech (Anderson et al., 2007) was developed to document language use throughout Scotland. The spoken portion of the corpus includes speech of considerable linguistic diversity from a broad linguistic continuum between Scottish Standard English on one hand and Broad Scots on the other. The Welsh English corpora are represented by the Siarad corpus of Welsh-English bilingual speech that aims to test alternative models of code switching with Welsh-English data (Deuchar et al., 2018).

The availability of the resources described above varies. Some are directly available for download with limited or no restrictions on the use, while others have stricter rules around how the resources can be used, such as non-commercial use only for the speech portion of the BNC corpus. For some of the corpora we looked at, the licenses are non-standard, which can in some cases be a hindrance for using the corpora.

While most of the corpora described so far were designed using careful methods of corpus linguistics, few of these datasets are ready for use in modern technological applications, such as text-to-speech, which places particularly high demands on the quality of audio and articulation. One of the initiatives that satisfies these requirements is the English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (VCTK) by Veaux et al. (2017), who collected a large corpus for the purpose of building statistical parametric voices (Zen et al., 2009) with an emphasis on voice morphing (Agiomyrgiannakis and Roupakia, 2016; Arik et al., 2018). While the corpus we describe in this paper is slightly smaller in size, we have undertaken a more principled phonological approach to the design of our corpus. Since both of these corpora are designed with speech applications in mind and have very similar audio quality, these corpora can be combined to form multi-speaker multi-dialect training or adaptation data for the systems such as the ones reported by Gibiansky et al. (2017) and Jia et al. (2018). We also hope that our corpus will be welcomed by the practitioners in corpus linguistics.

3. Phonological Approach

English varieties have diverged over a long span of time and across vast geographical distances (Kortmann et al., 2008; Maguire and McMahon, 2011). Over time, instances of what used to be the same phoneme can change either as a whole, or start to diverge depending on their phonetic environment. This process is called a *split*. In addition, what used to be different phonemes can end up sounding the same, resulting in *mergers*. These processes also affect the English dialects spoken across the British Isles (Maguire et al., 2010).

The spelling of the words were standardized as the language was going through major phonological changes, and while the spellings diverged a little (for example *neighbour* vs *neighbor*, *practise* vs. *practice*, *revolutionise* vs. *revolutionize* in British and American English, respectively) these changes do not reflect the phonological differences across dialects.

It is not entirely impossible to predict the pronunciation of a word from one dialect based on the other one, but it cannot be derived only from the current pronunciation or the spelling of the word without knowing the origin of the word (the historical pronunciation of the word or the language that it was borrowed from), and the phonological changes those specific dialects have gone through.

For the varieties of English where it is not the primary language of daily communication, the standardized spelling can have the opposite effect. When English is learned at school or mainly from written sources, the written form of the word may cause a divergence in pronunciation.

For example, the proposed pronunciation of *advantage* by a Nigerian English speaker is /æ d v æ n t eɪ dʒ/, as opposed to the British English pronunciation /ə d v ɑː n t ɪ dʒ/. Although the Nigerian English historically derived from British English, the first two vowels are pronounced as /æ/ like the General American English pronunciation of the word. The

Ph	01–50	51–1550	**Total**
ɪ	279	6489	6768
ə	278	5918	6196
n	229	5952	6181
t	184	5641	5825
s	159	4319	4478
l	134	3503	3637
d	133	3267	3400
r	130	3243	3373
k	122	2613	2735
z	116	2434	2550
m	69	2156	2225
e	73	2055	2128
æ	83	2036	2119
p	84	1850	1934
ð	102	1793	1895
ɒ	83	1747	1830
f	62	1569	1631
v	70	1456	1526
b	65	1281	1346
iː	60	1247	1307
w	42	1230	1272
eɪ	55	1209	1264
aɪ	42	1138	1180
j	34	1140	1174
ɔː	41	1122	1163
ʌ	26	936	962
ʊ	21	935	956
ŋ	27	858	885
əʊ	43	836	879
h	23	831	854
uː	34	816	850
g	24	692	716
ʃ	31	684	715
ɑː	20	665	685
ɜː	21	579	600
dʒ	20	563	583
tʃ	20	509	529
aʊ	22	459	481
θ	21	372	393
ʊə	20	303	323
ɔɪ	21	291	312
ʒ	24	279	303
eə	20	282	302

Table 1: Phoneme (shown in "Ph" column) frequencies for Idiolect Elicitation (EN01–EN50) and Global English (EN51–EN1551) lines.

first occurrence of the phoneme is centralized to a *schwa* /ə/ in British English. Nigerian English does not reduce vowels in the same way, so the vowel can be considered to be reconstructed to the more common pronunciation of the letter *a*. The second occurrence of the phoneme could similarly be informed by the spelling, or it could be an influence of American English, which is one of the most common accents in the media around the world.

The pronunciation of the vowel in the *-age* suffix, the diphthong /eɪ/, diverges from both British and American accents. This pronunciation conforms with the common pronunciation pattern of the words that end with a silent *e*, as in the very frequent word *age* /eɪ dʒ/, in both British and American English.

Although at first glance more dependency on the spelling might make it easier to predict the pronunciation of a word, it also changes the phonetic environment for the neighboring phonemes. When combined with the natural evolution of the language and the influences from the contact languages, it causes further and less predictable divergences from other accents.

When talking about English Accents, particular phonemes are conventionally classified into *standard lexical sets* (Wells, 1982), each represented by a keyword. For example, KIT refers to a group words that have the vowel /ɪ/ in British Received Pronunciation and General American accents. Well's lexical sets target vowels. Additional lexical sets have been used in the literature to represent consonants (Hickey, 2008). These lexical sets give researchers a convenient way to talk about the phonological changes.

For example, the TRAP-BATH split refers to the process where the /æ/ changed into a /ɑː/ in certain contexts represented by the BATH words in some English accents, whereas it stayed as /æ/ in TRAP words. Broadly speaking, the Southern English accents tend to have the split, whereas the Northern accents have the same vowel for TRAP and BATH words, although the change is still in flux in some areas (Blaxter and Coates, 2019).

A FLEECE merger refers to a process in which what used to be different vowels in words such as *meat, meet, piece, peace* converged over time, putting all of them in one lexical set, FLEECE, in most varieties of English. As a result, *meet* and *meat* will be homophones in the accents that have gone through the FLEECE merger, and a minimal pair in the accents that have not. A context dependent duration variation in FLEECE words is an indicator of Scottish English (Scobbie et al., 1999), which cannot be observed in minimal pairs as they would be occurring in the same phonological context, but can be investigated through comparison of similar words such as *feet* and *feed*.

In this work, we tried to make the most use of such words. We curated a list of words that are known to typically fall into the most common lexical sets. Then we added words that constitute homophones and minimal pairs for the known phonological changes. Where such pairs could not be found, we either chose word pairs that sound as close as possible, or single words that are prototypical of a certain change.

In addition, where multiple phonological changes affect a certain sound and its environment, we added words that contain a variety of phonetic environments for the phenomena that could be indicative of certain accents. For example, about a hundred words including *suit*, *cure*, and *huge* were included in the list to investigate how the presence or lack of *yod*, i.e. /j/, affects or is affected by a variety of phonetic environments for instance in a prerhotic position (before a pronounced or dropped /r/-like sound).

Finally we added as many words as required to increase the coverage of the phonemes or diphthongs such as /ɔɪ/ or /ʒ/ that are known to occur less frequently, or usually eliminated, in most English accents. As a result we ended up with 1,200 target words that are phonologically interesting for a study of English accents to be included in the recording scripts.

4. Curation of Recording Script

Recording script design is the crucial step in developing the speech corpora. The recording materials should be well-balanced in terms of phonemic coverage, cover multiple domains in order to accommodate numerous potential application scenarios, and, crucially, pose no difficulties in natural articulation for an amateur voice talent.

The recording scripts were curated from a global English accents perspective, and with considerations for phoneme coverage, inclusion of the target words, a reasonable line length, and overlapping lines within speakers and across some similar resources.

The phoneme distributions were calculated based on a proprietary British English pronunciation lexicon, where the entries are annotated to reflect a contemporary region-neutral British English accent. Keeping in mind that the phonological differences will diverge and split some of these phoneme counts, the frequency of each phoneme, including diphthongs, were kept to a minimum of 300 occurrences. The overall phoneme count distributions are shown in the last column of Table 1.

The script lines were retrieved from a variety of sources including public domain texts such as Wikipedia, the Rainbow Passage by Fairbanks (1960), and Alice's Adventures in Wonderland (Carroll, 2011), lines that are intended to be spoken by a virtual assistant, and manually created lines to accommodate certain words. Most lines that were obtained from external resources were later edited or pruned for length and target word and phoneme coverage, while keeping the semantics and facts intact.

Idiolect Elicitation Lines Every person has their own distinctive way of using language, i.e. their idiolect, which includes their characteristic use of grammar, choice of vocabulary, pronunciation, and intonation. Here we focus on the distinctive phonological aspects of the speakers. A speaker's personal accent can be affected by a variety of factors. Some of the main influencing factors are undoubtedly the areas the speaker has lived in during their developmental years and the accents of their family and other caregivers. The accents they have been exposed at school and workplaces, through friendships and personal relations and through mass media and entertainment can also affect a speaker's accent. In addition, the gender, age, anatomy, or personal habits can affect person's accent (Peterson and Barney, 1952). In order to be able to differentiate the speaker-specific aspects of the accent of a speaker from the regional accent, we included the same 50 lines in every speaker's recording script.

The first 21 lines of the idiolect elicitation lines (line ID `EN0001`–`EN0021`) are from the Rainbow Passage. These lines are included in the VCTK corpus and IDEA as well, and therefore provide a comparison between the accents of speakers across resources.

The lines `EN0022`–`EN0025` are also shared with VCTK, as well as the Speech Accent Archive. As mentioned above, some of these lines were altered to accommodate target word and phoneme coverage. For example, the sentence "Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob." was split into two lines. The line "We also need a small plastic snake and a big toy frog for the kids." was changed to "We also should get a good book, a small plastic snake and a big toy frog for the boys." and *Wednesday* was changed to *Thursday* in "She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

Another 25 lines, `EN0026`–`EN0050`, retrieved from Wikipedia were selected for their phoneme coverage and were edited in a similar manner. The idiolect elicitation lines have minimum of 20 occurrences of each British English phoneme as shown in the second column of Table 1.

Global English Accents Lines The lines `EN0026`–`EN1550` are selected to reflect the phonological aspects of a wide range of Global English accents. They cover all target words that were not already present in the idiolect elicitation lines. They provide enough phoneme coverage to increase the minimum phoneme frequency to at least 300 in total (as shown in the third column of Table 1). In addition, they include the names of the most populated cities of the world, major airlines, the most popular US cities, and popular English personal names.

This set includes 760 lines from Wikipedia, 700 lines from virtual assistant dialogues, 15 sentences from Alice's Adventures in Wonderland, and 25 manually created sentences where the target word and phoneme counts could not be accommodated by the lines from other sources.

British Isles Lines Lines `EN0001`–`EN1550` are intended to be included all English accents scripts. They are supported by a minimum of 500 localized lines, as it was done for the Nigerian English lines `NG0001`–`NG0550` in the Nigerian English corpus (Google, 2019a). For British Isles accents, the localized lines were split between a common British Isles set, and further localized sets for each accent region.

Lines `BI0001`–`BI0250` are shared by all British Isles regions. They include the names of the most populated settlements in the British Isles, most used railway stations, and most used European airports, as well as some common personal names. There are 35 sentences from British Isles related Wikipedia entries, 190 virtual assistant lines, and 25 manually created sentences.

Accent Region Lines Each accent region has additional localized lines to bring up the total number of lines to 2,050. The 250 lines with IDs starting with `LN` prefix complete the Southern England set. These sentences were created by filling in templates such as "`LINE` trains won't be calling at `STATION` this weekend." They cover all Transport for London (TfL) lines and railway stations in London that were not attested in the Global English lines.

The lines starting with `NE` prefix were created in the same way to cover the Northern England train lines and most used railway stations. These 250 lines complete the Northern England and the Midlands sets.

The 50 `GC` lines cover some popular Gaelic names, and were created by filling in templates such as "You have messages from `NAME` and `NAME`." They are included in the Irish English, Scottish English, and Welsh English lines.

Another 200 localized lines were included for each of these accent regions. Line IDs that start with `IR` include common Irish names and all railway stations as well as all localities in the Republic of Ireland and Northern Ireland that are not attested in the common lines. `SC` lines cover common Scottish names not covered in other datasets. All localities and most used railway stations in Scotland are also covered in these lines. `WL` lines include common Welsh names not covered in other datasets and all localities and most used railway stations in Wales. Additional 8 sentences from Wales-related Wikipedia entries were added to bring up the Welsh English lines to 200.

5. Recording Process Details

Volunteers and Recruitment The participants were all volunteers above 21 years of age. The volunteers were recruited by two separate efforts. The first one focused on Google employees in London. The second effort was a collaboration with Cardiff University, where the participants were the students, friends and family of the collaborators. A total of 101 volunteers were recorded in London and 19 volunteers in Cardiff.

For the volunteer recruitment of Google employees, an office-wide email and call for participation posters were used. Participants were asked to fill out a short survey where they reported growing up in or near London, Essex, Manchester, Leeds, Cardiff, Edinburgh, Glasgow, with the option to add another location. The locations were then grouped into regions based on Kortmann et al. (2008). The speakers reviewed the regions before the recordings.

At Cardiff University, a more direct approach was used where the collaborators either knew the participants, or reached out to groups which might be willing to contribute to the project.

One of the oversights of this project was the omission of broad sociolinguistic profile of the participants that would have provided additional useful information for placing their particular accent on the dialectal map of the varieties of British English (Kerswill, 2003; Hughes et al., 2013). This information is not available to us at present.

Recording Equipment and Environment The recordings were all done using the same recording equipment which consists of a Rode M5 microphone, an Blue Icicle XLR-USB A/D converter as well as a fanless ASUS Zenbook laptop. This rather affordable and portable hardware setup forms the core part of our inventory for collecting high-quality speech data for low-resource languages and dialects across the world.

We recorded the audio using a web-based recording tool developed in-house. The software displays the sentences which are to be recorded and provides a simple user interface for controlling the recording. The audio is stored at 48kHz with a depth of 16-bits per sample. The recording tool gives full power to its users who can control the pace of the recording. This feature is necessary because the users of this software typically record themselves. An external observer can listen to the recordings once they have been saved.

For the recordings at the London Google offices, a sound insulated recording room was used. For the recordings in Cardiff quiet empty rooms were used, where noise was kept to a minimum.

Recording Process The recording process follows the guidelines that were designed as part of our corpus collection program over the years for crowd-sourced collections of high-quality speech corpora for text-to-speech applications, such as the recording process for Bangla described by Gutkin et al. (2016), construction of South African corpora by van Niekerk et al. (2017), and Sundanese and Javanese recordings in Indonesia described by Wibawa et al. (2018). Over the years we found these guidelines to serve us well.

At the Google offices in London, volunteers were given a chance of signing up for recording slots. All volunteers signed a data release consent form allowing their recorded utterances to be placed in public domain with no restriction on academic, commercial or private use. We made sure the distance between the mouth of the speaker and the microphone was about 30 cm by asking the volunteers to keep this distance consistent and restart the recordings if this requirement was not met. The further requirements for the position of the microphone were as follows: The microphone should point below the speaker's forehead and above their chin, and the diaphragm of microphone should be pointing directly at the mouth. The volunteers had a 30 cm ruler which they could use in order to re-align their position if needed. All volunteers were asked to record a total of 150 lines and take a break after the first 20 minutes into the process. Most volunteers were able to finish the work in about an hour.

The first few lines recorded were listened to in order to make sure the recordings were good and noise-free, and spot checks were made during the recordings to make sure these were consistent. Volunteers could listen to the recordings immediately after they were recorded, but were asked to limit that and rather re-record the audio if issues such as ambient noise, coughing, breaks or disfluency in articulation were noticed.

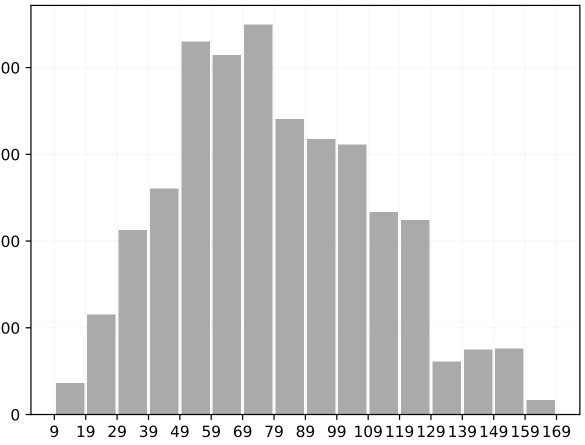


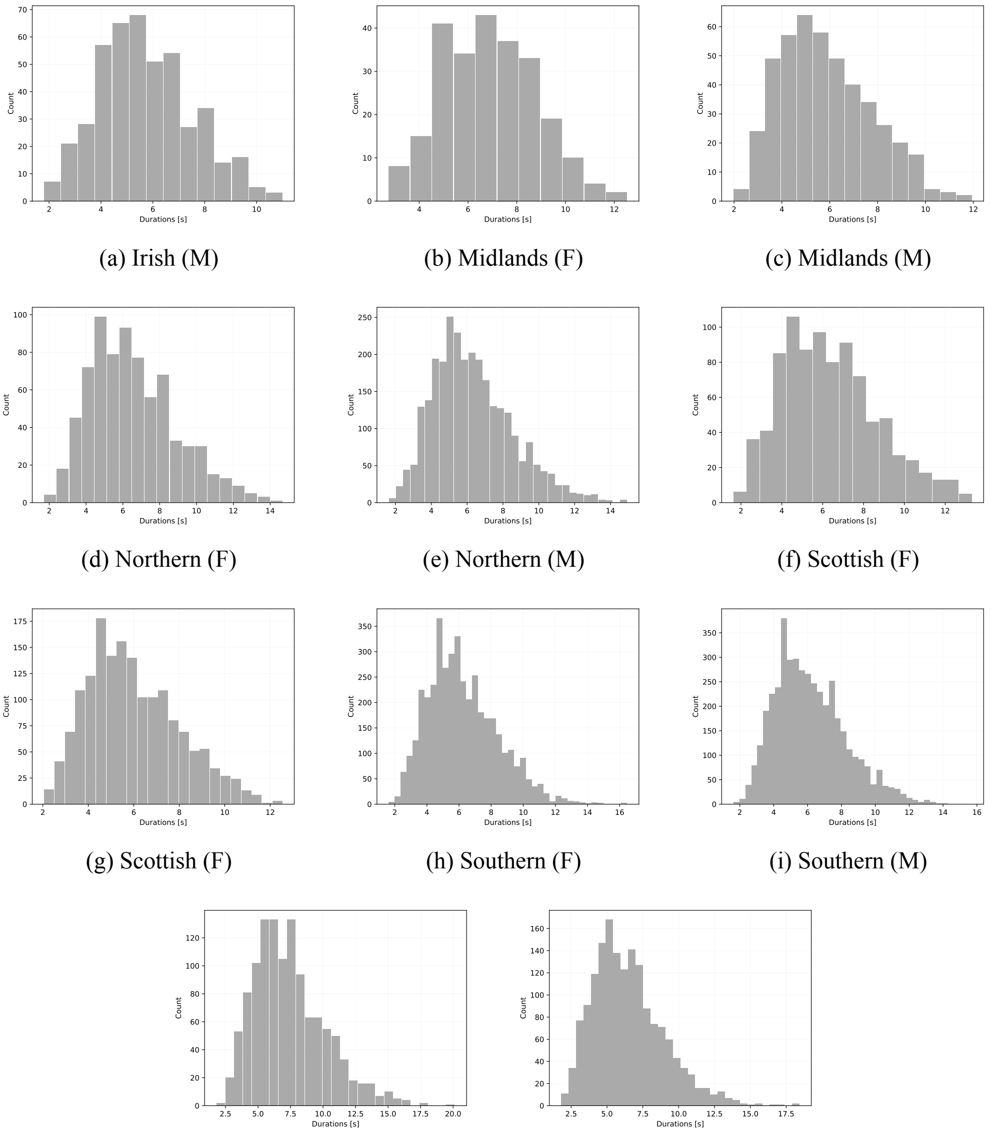
Figure 1: Histogram of the line lengths (measured in number of characters) for all the lines recorded. The line length is on the x-axis and the count on the y-axis.

6. Corpora Details

The corpora contains a total of 17,877 recordings of six dialects with the associated transcriptions. A total of 120 volunteers were recorded, 49 female and 71 male volunteers. Even though transcriptions mostly contain sequences of natural language words, because they have not been text normalized they also contain non-standard word (NSW) token expressions (Sproat et al., 2001), such as numbers. Therefore, here and below we refer to the constituent elements of transcriptions as "tokens" or "word tokens" rather than words. For instance, a numeric token may correspond to one or more natural language words, e.g., "25" → "twenty five". The combined recording script contains a total of 244,558 tokens out of which the 7,646 tokens are unique. The total duration of the corpora is over 31 hours of recorded audio. A full overview of the corpus is shown in Table 2, where the breakdown of number of participants, number of recorded utterances, duration of the audio recordings, average utterance duration (in seconds) and (word) token counts are shown for each gender of each recorded dialect. Overall, the combined corpora contains 17,877 transcriptions (consisting of 244,558 word tokens) corresponding to over 31 hours of audio recordings.

Dialect	Gender	Participants	Lines	Audio Duration		Word Tokens	
				Total (h:m:s)	Average (s)	Total	Unique
Irish	M	3	450	0:42:56	5.72	6,042	1,888
Midlands	F	2	246	0:28:12	6.88	3,468	1,395
	M	3	450	0:43:55	5.86	6,310	1,978
Northern	F	5	750	1:22:11	6.58	10,215	2,707
	M	14	2,097	3:37:42	6.23	28,594	5,438
Scottish	F	6	894	1:35:05	6.38	12,187	3,069
	M	11	1,649	2:44:42	5.99	22,194	4,539
Southern	F	28	4,161	7:11:17	6.22	57,508	6,781
	M	29	4,331	7:24:49	6.16	59,697	6,804
Welsh	F	8	1,199	2:28:12	7.42	16,139	3,425
	M	11	1,650	2:58:13	6.48	22,204	4,355
Total	–	120	**17,877**	31:17:19	–	244,558	7,646

Table 2: Overview of the datasets.



(a) Irish (M)

(b) Midlands (F)

(c) Midlands (M)

(d) Northern (F)

(e) Northern (M)

(f) Scottish (F)

(g) Scottish (F)

(h) Southern (F)

(i) Southern (M)

(j) Welsh (F)

(k) Welsh (M)

Figure 2: Histograms of the utterance durations by language and gender (x-axis shows duration, y-axis the frequency).

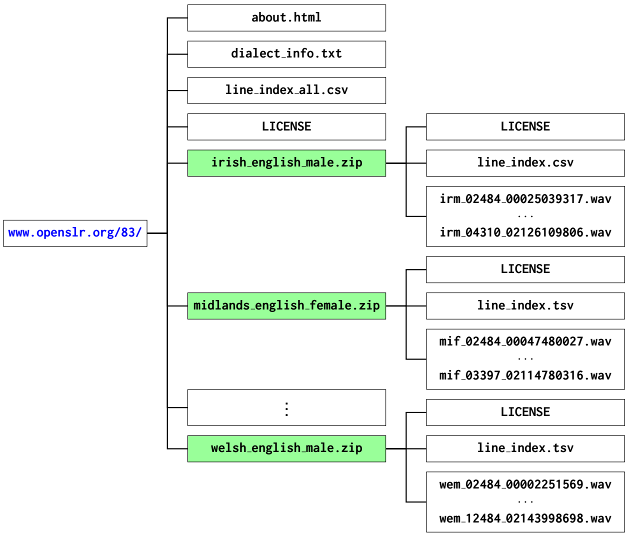


Figure 3: Layout of the corpus.

Sentence ID	Utterance ID
EN1223	wef_12484_01482829612
EN1223	wef_02484_01570891971
EN1223	wem_12484_01128399768
EN1223	wem_02484_01873501110
EN1223	nom_01523_01471201148
EN1223	sof_02121_01343324547
EN1223	som_02121_00413603831

Table 3: Entries in the master index file associated with the EN1223 sentence key corresponding to the sentence “The sun provides energy”.

The lengths of utterance transcriptions range between 9 and 169 characters, with the overwhelming majority of transcriptions being between 39 and 119 characters. A histogram of line lengths computed for all the recorded datasets is shown in Figure 1. The utterances are on average 6.3 seconds long, with the longest utterance being 20.1 seconds long and the shortest utterance being 1.62 seconds long. Figure 2 gives a full overview of the utterance durations (in seconds) displayed as histograms for each gender of each of the recorded dialects. Comparing Figure 1 and Figure 2, the durations are in line with what is to be expected. The line lengths (in characters or tokens) and durations (in seconds) of the files can be used to estimate the average rate of speech for each of the dialects, which is a useful metric for estimating the speech tempo (Kubina et al., 2008). Globally across all datasets, the average rate of speech ranges from 10.5 to 13.6 characters per second or from 110.3 to 142.8 tokens per minute, respectively.

Corpus Contents A schematic depiction of corpora structure is shown in Figure 3. The file line_index_all.csv is a comma-separated text file that represents the master index for all the available dialect datasets and contains three columns. The first column contains the line identifier in the original curated script (described in Section 4) that can be used to retrieve all the renditions of the same line across all the dialects. For example, the identifier EN1223 serves as a key to all possible pronunciations of the sentence “The sun provides energy“ in all the datasets. The second column contains the unique utterance identifier which consists of a three-letter prefix followed by a five digit user identification number and a unique 11 digit line identifier. For example, the original sentence EN1223 when spoken by the Welsh male 02484 is stored with an utterance identifier wem_02484_01873501110 in our database. The third column contains the transcription of the audio files, which have the same name as the utterance identifiers. In other words, there is a one-to-many mapping between the first column that identifies the transcription and the second column that points to the actual speaker utterance, as shown in Table 3. Please note that each sentence is not necessarily available for *all* the genders, speakers and dialects.

ISLRN	OpenSLR ID	URL Link
204-161-521-586-9	SLR83	openslr.org/83/

Table 4: Corpus identifiers and the URL.

The recordings for each gender of each dialect are stored separately each in their own archive file created using zip compressor. For the six dialects this amounts to 11 archive files, as the Irish female dataset has not been recorded. Each archive contains gender and dialect-specific index file line_index.csv the format of which is identical to the structure of the master index file. Each archive also stores the corresponding audio files which are released as 48 kHz single-channel (mono) in 16-bit linar PCM RIFF (.wav) format.

Distribution and Licensing The corpus is released under Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license (Creative Commons, 2019) and is made available for download from Open Speech and Language Resources repository (OpenSLR) (Povey, 2019) as shown in Table 4 along with the International Standard Language Resource Number (ISLRNs) (Mapelli et al., 2016) and the OpenSLR Speech and Language Resource (SLR) identifier. The ISLRN is a 13-digit number that uniquely identifies the corpus and serves as official identification schema endorsed by several organizations, such as ELRA (European Language Resources Association) and LDC (Linguistic Data Consortium).

7. Conclusion

In this paper, we presented an open-source, multi-speaker speech corpora for Southern England, Midlands, Northern England, Welsh English, Scottish English and Irish English accents.

The corpora consists of volunteers reading a recording script that has been curated specifically for English accent elicitation. An overlapping set of 50 lines read by each speaker provides cross-resource comparison, as well as making it possible to separate personal accents from regional accents. Another set of Global English lines cover 1,200 target words that are aimed at revealing the phonological profile of the recorded accents, and provide pronunciations of highly populated world locations, major airlines, and popular names in a variety of English accents. Finally a localized set of 500 sentences provide the pronunciation of British Isles locations, railway lines and stations, and popular names in the local accents.

The corpora are intended for speech technologies as well as linguistic studies and are released with an open-source license with no limitations on academic or commercial use.

8. Acknowledgments

The authors would like to thank Dawn Knight, Anna Jones and Alex Thomas from Cardiff University for their assistance in collecting the Welsh English data presented in this paper. The authors also thank Richard Sproat for his comments on the earlier drafts of this paper. Finally, the authors thank the anonymous reviewers for many helpful suggestions.

9. Bibliographical References

Agiomyrgiannakis, Y. and Roupakia, Z. (2016). Voice morphing that improves TTS quality using an optimal dynamic frequency warping-and-weighting transform. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5650–5654, Shanghai, China, March. IEEE.

Agrawal, S. S., Sinha, S., Singh, P., and Olsen, J. Ø. (2012). Development of Text and Speech database for Hindi and Indian English specific to Mobile Communication environment. In *Proc. of LREC*, pages 3415–3421, Istanbul, Turkey.

Anderson, J., Beavan, D., and Kay, C. (2007). SCOTS: Scottish Corpus of Texts and Speech. In J. Beal, et al., editors, *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*, pages 17–34. Springer.

Anderwald, L. and Wagner, S. (2007). FRED – The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data. In J. Beal, et al., editors, *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*, pages 35–53. Springer.

Arik, S., Chen, J., Peng, K., Ping, W., and Zhou, Y. (2018). Neural Voice Cloning with a Few Samples. In *Advances in Neural Information Processing Systems*, pages 10019–10029.

Barlow, M. (2000). *Corpus of Spoken, Professional American-English*. Rice University. Corpus available from: `http://www.linguistics.ucsb.edu/research/santa-barbara-corpus`.

Barlow, M. (2010). Individual usage: a corpus-based study of idiolects. In *Proc. 34th International LAUD Symposium*, Landau, Germany, March.

Blaxter, T. and Coates, R. (2019). The Trap–Bath Split in Bristol English. *English Language & Linguistics*, pages 1–38.

Burnard, L. (2002). Where did we go wrong? A retrospective look at the British National Corpus. In *Teaching and Learning by Doing Corpus Analysis*, pages 51–70. Brill Rodopi.

Carroll, L. (2011). *Alice's Adventures in Wonderland*. Broadview Press.

Cassidy, S., Haugh, M., Peters, P., Fallu, M., et al. (2012). The Australian National Corpus: National Infrastructure for Language Resources. In *Proc. of 8th International Conference on Language Resources and Evaluation (LREC)*, pages 3295–3299, Istanbul, Turkey. Corpus available from: `http://www.ausnc.org.au`.

Coelho, G. M. (1997). Anglo-Indian English: A nativized variety of Indian English. *Language in Society*, 26(4):561–589.

Coleman, J., Baghai-Ravary, L., Pybus, J., and Grau, S. (2012). Audio BNC: the audio edition of the Spoken British National Corpus. *Phonetics Laboratory, University of Oxford*. Available from: `http://www.phon.ox.ac.uk/AudioBNC`.

Creative Commons. (2019). Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). `http://creativecommons.org/licenses/by-sa/4.0/deed.en`.

Deterding, D. and Low, E. L. (2001). The NIE corpus of spoken Singapore English (NIECSSE). *SAAL Quarterly*, 56(1):2–5. Corpus available from: `http://www.linguistics.ucsb.edu/research/santa-barbara-corpus`.

Deuchar, M., Webb-Davies, P., and Donnelly, K. (2018). *Building and using the Siarad corpus: bilingual conversations in Welsh and English*, volume 81 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). *Ethnologue: Languages of the world*. SIL International, Dallas, Texas, 22nd edition.

Fairbanks, G. (1960). *Voice and Articulation Drillbook*. Harper & Row, 2nd edition. `https://www.dialectsarchive.com/the-rainbow-passage`.

Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y. (2017). Deep Voice 2: Multi-Speaker Neural Text-to-Speech. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2962–2970.

Gut, U. B. (2008). Nigerian English: Phonology. In Rajend Mesthrie, editor, *Varieties of English: Africa, South and Southeast Asia*, pages 35–54. Mouton de Gruyter, Berlin.

Gutkin, A., Ha, L., Jansche, M., Kjartansson, O., Pipatsrisawat, K., and Sproat, R. (2016). Building Statistical Parametric Multi-Speaker Synthesis for Bangladeshi Bangla. In *Proc. of 5th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, volume 81, pages 194–200, Yogyakarta, Indonesia, May. Elsevier.

Hickey, R. (2008). Irish English: phonology. In Bernd Kortmann et al., editors, *Varieties of English: the British Isles*. Mouton de Gruyter, Berlin.

Holmes, J., Vine, B., and Johnson, G. (1998). *Guide to the Wellington corpus of spoken New Zealand English*. School of Linguistics and Applied Language Studies, Victoria University of Wellington, New Zealand.

Hughes, A., Trudgill, P., and Watt, D. (2013). *English Accents and Dialects: An Introduction to Social and Re-*

gional Varieties of English in the British Isles. Routledge, 5th edition.

Isiaka, A. L. (2019). A phono-ethnic story of Nigerian English: As told by high vowels. *Ampersand*, 6. Article 100049.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Moreno, I. L., Wu, Y., et al. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. In *Advances in Neural Information Processing Systems*, pages 4480–4490.

Kachru, B. B. (1976). Indian English: A Sociolinguistic Profile of a Transplanted Language. *ERIC: Institute of Education Sciences*, pages 1–52. Available from: https://files.eric.ed.gov/fulltext/ED132854.pdf.

Kerswill, P. (2003). Dialect levelling and geographical diffusion in British English. In D. Britain et al., editors, *Social dialectology: in honour of Peter Trudgill*, pages 223–243. John Benjamins Amsterdam, the Netherlands.

Kortmann, B., Upton, C., Schneider, E. W., Burridge, K., and Mesthrie, R. (2008). *Varieties of English*, volume 1. Mouton de Gruyter Berlin.

Kperogi, F. A. (2015). *Glocal English: The Changing Face and Forms of Nigerian English in a Global World*, volume 96 of *Berkley Insights in Linguistics and Semiotics*. Peter Lang.

Kubina, R. M., Amato, J., Schwilk, C. L., and Therrien, W. J. (2008). Comparing Performance Standards on the Retention of Words Read Correctly Per Minute. *Journal of Behavioral Education*, 17(4):328.

Labov, W. (1972). Some Principles of Linguistic Methodology. *Language in society*, 1(1):97–120.

Leith, D. (1997). *A social history of English*. Routledge.

Lowenberg, P. H. (1991). Variation in Malaysian English: The pragmatics of language in contact. *English around the world: Sociolinguistic perspectives*, pages 364–375.

Maguire, W. and McMahon, A. (2011). *Analysing variation in English*. Cambridge University Press.

Maguire, W., McMahon, A., Heggarty, P., and Dediu, D. (2010). The past, present, and future of English dialects: Quantifying convergence, divergence, and dynamic equilibrium. *Language Variation and Change*, 22(1):69–104.

Mapelli, V., Popescu, V., Liu, L., and Choukri, K. (2016). Language Resource Citation: the ISLRN Dissemination and Further Developments. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1610–1613, Portorož, Slovenia, May. ELRA.

Meier, P. (1998). IDEA The International Dialects of English Archive. Corpus available from: http://www.dialectsarchive.com/.

Nolan, F. and Post, B. (2014). The IViE Corpus. In Jacques Durand, et al., editors, *The Oxford Handbook of Corpus Phonology*, pages 475–486. Oxford University Press.

Olaniyi, O. K. (2014). The taxonomy of Nigerian varieties of spoken English. *International Journal of English and Literature*, 5(9):232–240.

Persley, N. H. (2013). An innovative IDEA: a review of the International Dialects of English Archive. *English Today*, 29(3):63–64.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184.

Povey, D. (2019). Open SLR. http://www.openslr.org/resources.php. Accessed: 2019-03-30.

Scobbie, J. M., Hewlett, N., and Turk, A. E. (1999). Standard English in Edinburgh and Glasgow: the Scottish vowel length rule revealed. *Urban voices: Variation and change in British accents*.

Smith, J. (1998). *An Historical Study of English: Function, form and change*. Routledge, London and New York.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., and Richards, C. (2001). Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Trudgill, P. and Hannah, J. (2017). *International English: A guide to varieties of English around the world*. Routledge.

van Niekerk, D., van Heerden, C., Davel, M., Kleynhans, N., Kjartansson, O., Jansche, M., and Ha, L. (2017). Rapid development of TTS corpora for four South African languages. In *Proc. Interspeech 2017*, pages 2178–2182, Stockholm, Sweden, August.

Veaux, C., Yamagishi, J., MacDonald, K., et al. (2017). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. Corpus available from: http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html.

Weinberger, S. H. and Kunath, S. A. (2011). The Speech Accent Archive: towards a typology of English accents. In *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*, pages 265–281. Brill Rodopi. Corpus available from: http://accent.gmu.edu/.

Wells, J. C. (1982). *Accents of English*, volume 1. Cambridge University Press.

Wibawa, J. A. E., Sarin, S., Li, C. F., Pipatsrisawat, K., Sodimana, K., Kjartansson, O., Gutkin, A., Jansche, M., and Ha, L. (2018). Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1610–1614, 7-12 May 2018, Miyazaki, Japan.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical Parametric Speech Synthesis. *Speech Communication*, 51(11):1039–1064.

10. Language Resource References

Google. (2019a). *Crowdsourced high-quality Nigerian English speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), http://www.openslr.org/70, Google crowd-sourced speech and language resources, 1.0, ISLRN 861-285-836-675-3.

Google. (2019b). *Crowdsourced high-quality UK and Ireland English dialect speech data set by Google*.

Google, distributed by Open Speech and Language Resources (OpenSLR), http://www.openslr.org/83, Google crowd-sourced speech and language resources, 1.0, ISLRN 204-161-521-586-9.