University of Pennsylvania Working Papers in Linguistics

Volume 28
Issue 2 *Selected Papers from NWAV 49*

Article 11

9-19-2022

(Commercial) Automatic Speech Recognition as a Tool in Sociolinguistic Research

Nina Markl
The University of Edinburgh

(Commercial) Automatic Speech Recognition as a Tool in Sociolinguistic Research

Abstract

As speech datasets used in sociolinguistic research increase in size, laborious and time-intensive manual orthographic transcription is a challenge, limiting the amount of (transcribed) data which can be analysed. In this paper, I discuss the use of (commercial) automatic speech recognition (ASR) as a tool in sociolinguistic research in the context of a case study: the Lothian Diary Project. I describe the kinds of errors produced by two commercial ASR systems for British English within the broader context of algorithmic bias in ASR, and suggest some best practices when working with ASR in sociolinguistic work.

(Commercial) Automatic Speech Recognition as a Tool in Sociolinguistic Research

Nina Markl[*]

1 Introduction

Collection, compilation, and storage of large speech datasets has, in general, become much easier in recent decades. Computational methods aiding acoustic analysis of larger datasets such as forced alignment have also become more accessible and reliable (e.g. Mackenzie and Turton 2020, Reddy and Stanford 2015), and recent developments in the application of machine learning to phonetic analysis are promising (e.g. Villarreal et al. 2020). However, most sociolinguistic analyses, whether they involve acoustic analysis or not, also require orthographic transcripts. Preparing these orthographic transcripts can pose a bottleneck: manual transcription is a very laborious process (Bird 2021). Automatic speech recognition (ASR) coupled with manual correction could potentially alleviate this work load. While ASR has recently found wide use in applications ranging from voice user interfaces in mobile devices to automatic captioning in virtual classrooms and on social media, it is not yet a widely used tool in sociolinguistic research.

In this paper, I discuss the potential advantages and problems of using ASR to facilitate transcription of sociolinguistic data, with a particular focus on commercial ASR engines provided by Amazon and Google. The specific application context is the Lothian Diary Project, an interdisciplinary research project at the University of Edinburgh, collecting audio and video diaries recorded by residents in the Scottish Lothians region documenting their experiences of the COVID-19 pandemic between May 2020 and July 2021 (Hall-Lew et al. 2022). As a member of the Lothian Diaries research team, I facilitated the transcription of the almost 200 English language recordings using ASR and manual correction.

2 Orthographic Transcription as a Task

Orthographic transcriptions are important for a wide range of sociolinguistic research methods. In (socio)phonetic research, forced alignment of (partial or complete) orthographic transcripts is now a standard method to facilitate semi-automatic segmentation and acoustic analysis (Mackenzie and Turton 2020). Even simple orthographic transcripts allow efficient search and topic and corpus analysis, while more complex transcriptions can facilitate analysis of interactions. Many sociolinguistic datasets are also of potential interest to the general public (e.g. as oral history archives) and researchers in other fields (e.g. because of the topics discussed by participants). Transcriptions make these datasets more accessible, portable and, depending on the type of recording, durable. A transcript is also often easier to reproduce in a research publication or presentation, both for practical and ethical reasons, as the voice recording, but not the transcript, may be considered "biometric data" which can be used to identify individuals (Information Commissioner's Office 2022).

Preparing "full" transcripts is a very labourious process. Importantly, it is also not a "neutral" one (Ochs 1979, Bucholtz 2000, Bird 2021, Himmelmann 2018). As has been discussed in a number of linguistic subfields, transcribers have to *interpret*, often "underdetermined" (Himmelmann 2018:35), speech signals based on their own linguistic and cultural knowledge. Even "simple" orthographic transcription, which does not account for aspects such as prosody and overlaps, involves theoretically informed decisions. For example, transcribers may establish or choose between possible conventions to represent speech and non-speech sounds (e.g. laughter, background noise), as

[*]This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences. I would like to thank Peter Bell for his help with setting up the local ASR system for the Lothian Diary Project, and Anita Klingler for her help with transcription, as well as the rest of the Lothian Diaries team for being the best. Thank you to the audience at NWAV49 for their engaged questions, and to Lauren Hall-Lew and Catherine Lai for feedback.

well as common features of spontaneous speech but not (formal) writing such as false starts and filled pauses.

3 Automatic Speech Recognition

As Himmelmann notes, “consider[ing] transcription exclusively, or even primarily, a process of mechanically converting a dynamic acoustic signal into a static graphic/visual one” would therefore be “rather naïve” (2018:35). Despite impressive advances in recent years, ASR transcriptions are not perfect, and, for most conversational speech, not on par with human transcription. This is, in part, because ASR tools, like other language technologies have only limited access to the linguistic, and perhaps more importantly, social and cultural context(s) humans can make use of when transcribing potentially ambiguous speech. They also assume that transcription *is* a straightforward mapping between a dynamic acoustic signal and a static graphic signal. In this section I will briefly introduce ASR as a task and tool and discuss some limitations, with a particular focus on language variation and algorithmic bias.

The basic task of ASR is to map an input waveform to an output character string. Conventionally, ASR systems consist of four components: a signal processing component, an acoustic model, a language model and a decoding component. The acoustic model (AM) and language model (LM) are “trained” on a (transcribed) speech corpus and text respectively. The AM is a representation of speech sounds modelling probability distributions over phones given a signal and (usually) some context,[1] while the LM is a representation of word sequences modelling probability distributions over words given a particular context. How these components are trained and implemented varies. Recent state-of-the-art “end-to-end” systems feature encoder-decoder architectures based on transformers or recurrent neural networks also used in other deep learning domains (e.g. Chan et al. 2016, Hannun et al. 2014). These systems learn to map sound sequences directly to character sequences from the speech corpus, but nevertheless usually apply an additional, larger language model to select the “most likely” output.

With the development of new algorithms and architectures as well as advances in computing, the performance of ASR systems, especially in (American) English, particular domains and speech styles, has steadily improved. Within sociolinguistic research, ASR has been incorporated in DARLA (Reddy and Stanford 2015,Coto-Solano et al. 2021), an impressive and popular tool facilitating fully automatic extraction of acoustic measurements from American English speech (using partial transcription). However, automatic accurate and full transcription of spontaneous speech is still very difficult, and not yet particularly widespread as a tool among sociolinguists (though some labs and research groups are taking it up: e.g. Wassink 2021).

As in other machine learning fields, systems are generally evaluated on well-known benchmark datasets associated with particular tasks. The fact that benchmark datasets tend to represent only a limited range of accents and speech styles (e.g. Markl 2022b, Martin 2021) calls the performance of ASR systems on “real”, conversational and “non-standard” speech into question (Szymański et al. 2020).[2] Performance of ASR systems is usually assessed using the standard metric word error rate, a simple edit-distance measure capturing the number of deletions, insertions and substitutions necessary to match the automatic transcript to a reference transcript. Recent empirical research shows that in addition to performing better on read speech (Szymański et al. 2020), automatic speech recognition systems also exhibit “predictive bias”, a systematic difference in error rates between groups (e.g. between speakers of different varieties) (Shah et al. 2020). Commercial ASR systems for US English produce significantly (and substantially) higher error rates for speakers of African American English (AAE) than Mainstream US English (Koenecke et al. 2020,Martin and Tang 2020). Koenecke et al.’s 2020 study is perhaps particularly relevant for sociolinguists, as the recordings used to test ASR systems by five big technology companies were drawn from sociolinguistic interviews drawn from the CORAAL corpus (Kendall and Farrington 2021) and Voices of California (Stanford

[1] “Phone” tends to be used in a way more similar to the linguistics use of “phoneme” rather than “phone”, i.e., in most cases it reflects an abstract category.

[2] This is a much broader crisis in machine learning (e.g. Paullada et al. 2021,Liao et al. 2021).

Linguistics n.d.). In Markl (2022a), I show that British English ASR engines sold by Google and Amazon perform significantly worse for second language speakers of English and speakers of some stigmatised British varieties. Amazon Transcribe produces very low error rates for speakers from Cambridge, and significantly higher error rates for speakers from Belfast, Newcastle, Liverpool and Bradford. This is particularly notable as, unlike Koenecke et al. (2020), I tested the systems on read speech drawn from the IViE corpus (Grabe and Nolan 2002) and Speech Accent Archive (Weinberger 2015), where we may expect better performance across the board and smaller performance differences between groups. ASR accuracy has also sometimes been shown to differ by gender, with generally better performance for female speakers, perhaps due to differences in speech styles (Markl 2022a, Koenecke et al. 2020, Adda-Decker and Lamel 2005).

One reason for predictive bias is a mismatch between training data and test data. If the acoustic model and/or language model components of the system are trained on a different speech style, domain or variety than the one the system is ultimately applied to, the system will likely do poorly. For instance, AAE appears to be under-represented in the training data for acoustic models (Koenecke et al. 2020) and language models (Martin and Tang 2020) used in U.S. English commercial ASR systems.

4 A Case Study: the Lothian Diary Project

Between May 2020 and July 2021, the Lothian Diary Project collected 195 audio and video diaries (Hall-Lew et al. 2022). We invited residents of Edinburgh and the Lothians region in Scotland of all backgrounds to reflect on their experiences of the COVID-19 lockdowns. In addition to linguistic variation, we were also particularly interested in understanding and reporting on the ways people were impacted by the pandemic and the changes to day-to-day lives it brought, as well as attitudes towards government policy (which fed into a report for the Scottish Government[3]). Due to the scale and time-sensitive nature of the dataset and this interest in the attitudes and experiences expressed in the recordings, a relatively quick way of preparing full and accurate transcripts was crucial. We opted for a pipeline involving a custom, locally-run ASR system and manual correction.

4.1 Local Setup

To facilitate faster transcription of the diaries, we developed a pipeline involving ASR and manual correction. We opted for the widely used open-source ASR toolkit Kaldi (Povey et al. 2011). The acoustic model and baseline language model were provided by colleagues at the Centre for Speech Technology Research at the University of Edinburgh. As most of our recordings are in a British variety of English, we used an acoustic model trained on recordings drawn from BBC broadcasts.[4] The language model was adapted and enhanced with (manually produced) transcripts of 37 Lothian diaries, as well as text scraped from the RSS feed of the BBC news service (published between March and July 2020) and posts relating to COVID-19 on the Edinburgh subreddit.[5] These additions allowed for better inclusion of terms related to the pandemic, local issues and place names and greatly improved performance of the ASR system.

Kaldi requires recordings and auxiliary files according to specific requirements (see Chodroff 2018, for an excellent tutorial). This involves fairly complex (but automatable) creation of utterance chunks and several text files linking each utterance to a time-stamp and a speaker. Once these files are prepared, Kaldi is run locally from the command line. To facilitate manual correction, the output of the ASR system can be saved as a tab-delimited file with time-stamps and imported to ELAN for further processing. Each utterance chunk can then be reviewed and corrected if necessary. Overall,

[3] `https://lothianlockdown.org/parliamentreport/`

[4] Developed by the Centre for Speech Technology Research as part of the Multi-Genre Broadcast challenge in 2015: `http://www.mgb-challenge.org/`

[5] `https://www.reddit.com/r/Edinburgh/`: Posts containing the terms “Covid”, “Lockdown”, “Coronavirus” or “Quarantine” and all their comments: After removing duplicate posts, the corpus consists of 288 individual posts and 4833 comments. Only 2 were created before January 2020.

this pipeline is more efficient than manual chunking and transcribing, though the quality of the automatic transcript differed greatly depending on accent, recording quality and background noise.

There are clear advantages to using a custom-built system, such as control over training data and a variety of system settings. Local processing also side-steps potential ethical or data protection issues associated with transferring recordings to a third party. However, using (and in particular, training) Kaldi is a non-trivial task requiring significant coding expertise, some computing resources on a unix system (Kaldi is not supported for Windows) and, of course, data to train the acoustic model and language model. There are also more user-friendly tools building on Kaldi, such as ELPIS (Foley et al. 2018)[6] which allow relatively simple training and use of custom ASR tools, as well as alternative open-source toolkits such as DeepSpeech.[7]

4.2 Commercial Setup

While we did not rely on commercial ASR engines for the transcription in the project, the Lothian Diaries dataset is also a very interesting “test set” for commercial ASR systems. As noted above, ASR systems generally perform best on the varieties they are trained on - very often these appear to be standard varieties (Markl 2022a,Markl and McNulty 2022). They also perform much better on read speech than conversational speech and are very sensitive to differences in recording conditions and background noise. Overall, our dataset of unique self-recordings features participants from a wide range of linguistic and social backgrounds, covering a variety of topics, is particularly challenging for ASR engines (of any kind). It is also more reflective of the kind of data many sociolinguists are working with than standard benchmark datasets. I will discuss some of the different error types we discovered in exploratory analysis (see also Markl and Lai 2021).

I considered the off-the-shelf British English ASR systems from Google (Google Cloud Speech-to-Text) and Amazon (Amazon Transcribe). These systems can be accessed through APIs in various programming languages or user-friendly interfaces. Using the ASR system does not require pre-processing of the data or programming skills. Like most commercially available ASR engines, they rely on cloud-computing. This can be advantageous as they do not require much local computing resources. As a result audio files need to be uploaded to the cloud storage of the respective provider (usually only for the duration of the processing).

4.2.1 Errors

As noted above, ASR systems are usually evaluated using word error rate. This standard metric fails to account for the context in which errors occur: some errors distort the intended message more than others (e.g., substituting *can’t* for *can*). In qualitative, “context-sensitive” exploratory error analysis of the commercial ASR systems, we identified several error types.

Some errors appear to be caused by phonetic differences between the training and test data. Figure 1 shows a comparison between three ASR models applied to recordings by two Scottish English speakers (a man from Edinburgh and a woman from Glasgow). Google only offers one model for “British English”, Amazon offers “British English” and “Scottish English”. While the models produce different errors, reduced phonetic forms appear particularly challenging for them. As can be seen in Figure 1a, all three models tend to delete tokens they identify as filled pauses. This affects actual filled pauses (here transcribed as *er* and *erm*) and reduced forms of word (e.g., *I* in *but I live alone*). In Figure 1a, we can also see how this interacts with phonetic variation: one of the filled pauses is identified as such and deleted by Google’s model and Amazon’s Scottish English model. Amazon’s British English model, however, transcribes this this token, which is produced as /eː/ as *a*, presumably because this realisation is much more similar to, for example, Southern British English pronunciations of *a* than it is to those of *uh*. The reduced centralised vowel in *I* in *I live with a cat but I live alone* appears to be misclassified as a filled pause by all three models (and subsequently deleted). *Cat*, which is produced with final glottal stop, is also mistranscribed by all three models

[6] Developed for language documentation: `https://elpis.readthedocs.io/en/latest/`

[7] `https://deepspeech.readthedocs.io/en/r0.9/`

as *car* or *cap*. Word-medial glottal replacement in *isolating* appears less challenging for two of the systems, perhaps because there are fewer phonetically similar alternatives. Though notably, Amazon's British English system also fails to transcribe this correctly. This seems particularly surprising as glottal replacement is prevalent in most British varieties, including those in South of England which this system appears to be mostly trained on.

Substitutions are often morphologically related forms. Sometimes these are also phonetically similar and could be the result of reduction (e.g., *hardest* > *hard* in Figure 1b) and other times they are less similar (e.g., *found* > *find* in Figure 2b). These errors are particularly disruptive where they significantly change the meaning. For example, *We couldn't even go out* was transcribed as *We could even go out* by Google's ASR tool (see Figure 2b). These errors may be in part driven by the language model, which is used to decode the recognised sound sequences into utterances. When decoding a sound sequence using a language model, all words (types) present in the language model get assigned a probability. This probability is conditional on the sound sequence, (usually) some linguistic context (e.g., the words preceding and following) and the baseline probability of the word in the corpus. This means that, depending on the decoding algorithm, words which are very infrequent in the corpus used to train the language model[8] may be less likely to be accurately decoded than high-frequency words (which may, conversely, be assigned too high a probability). In the context of the Lothian Diaries, we can see this is in frequent errors around terms like "lockdown", "Covid", "social distancing" and other words and names related to current affairs. We also see the influence of the language model in utterances with repetitions. For example in Figure 2c, the sentence *I miss my family, I miss my friends* is simplified to *I miss my family and my friends* by the Google ASR model. Both models delete a false start in that same utterance (*it's – I shouldn't be here* > *I shouldn't be here*). Another potential influence of the language model is the decoding of contractions as full forms (e.g. *I'm* > *I am* in Figure 2b). The deletion of filled pauses discussed above may also be exacerbated by language models trained on text which do not contain them.

5 Incorporating ASR Tools in Sociolinguistics

5.1 Transcription in the Age of Big Speech Data

As the storage and collection of large amounts of speech data has become easier and cheaper, several subfields of linguistics, such as acoustic phonetics, psycholinguistics and sociolinguistics have embraced the idea of working with larger corpora of speech (Liberman 2019). There are some clear advantages to using larger datasets in the quantitative analysis of language variation and change. Rare phenomena are more robustly represented and variation and change across time and space may be more easily observable. In addition to established methods of data compilation (Benjamin 2021) in variationist sociolinguistics, like sociolinguistic interviews conducted by fieldworkers with individual participants within a speech community, recent years have also seen the rise of new methods. While self-recordings created for either public audiences or just the researchers have been used in variationist, and in particular sociophonetic, research for several years (Schøning and Møller 2009, Hall-Lew and Boyd 2017, Leemann 2016, Leemann et al. 2018, Clark et al. 2016), facilitating remote recordings with speakers became for many research groups the only way to gather speech data during the COVID-19 pandemic. Some projects developed specifically in response to the pandemic (Sneller 2022), including MI Diaries (Sneller et al. 2022) and the Lothian Diary Project (Hall-Lew et al. 2022). These new ways of compiling data to analyse variation appear to be here to stay. Developing efficient data processing and data analysis pipelines is essential to make good use of these incredibly rich datasets, especially in interdisciplinary contexts where we (or perhaps our collaborators in different fields) might be just as interested in *what* is said as we are in *how* it is said.

[8] In the case of large language models like GPT-3, these corpora include petabytes worth of text crawled from the internet, e.g. CommonCrawl, a corpus which also contains a "significant amount of undesirable content, including hate speech and sexually explicit content" (Luccioni and Viviano 2021).

I've	missed it	it's	been quite a while but	erm	myself I found	er	isolating	a bit difficult
of		Mr has	been quite a while but	***	myself I found	**	isolating	a bit difficult
I've		Mr has	been quite a while but	a	myself I found	a	a silly and	a bit difficult
I've	missed it	has	been quite a while but	a	myself I found	**	isolating	a bit difficult

Er	I live alone	I live	with a	cat	but	I	live	alone	no other adults	living	in the	house
**	I live alone	* love	with a	car	but	*	live	alone	no other adults	****** ** going out		
**	I live alone	* ****	with a	cap	but	*	let	alone	lure that I don't	see you	in the	nose
**	I live alone	* love	with a	cap	but	*	live	alone	no other adults	coming	in the	house

(a) Utterance produced by a male speaker from Edinburgh (born 1975).

The	hardest	part of	lockdown	for me was definitely	missing out	on	seeing	loved ones
The	highest	part of	what time	for me was definitely	missing out	on	seeing	loved ones
The	hard	part of	what done	for me was definitely	messing it	on	seen	loved ones
The	hard	part of	lock down	for me was definitely	missing out	on	seen	loved ones

(b) Utterance produced by a female speaker from Glasgow (born 1999).

Figure 1: Comparison of reference transcript (top) with automatic transcripts. For Scottish speakers, there are three models to compare: Google (British English) with errors highlighted in blue, Amazon (British English) with errors highlighted in orange and Amazon (Scottish English) with errors highlighted in yellow. Deletions are marked with asterisk (*).

5.2 Algorithmic Bias, Practical Considerations, and Trade-offs

As (socio)linguists, we should be particularly aware of the ways language technologies can fail. As noted above, algorithmic bias is a real and urgent problem in speech and language technologies (see also Blodgett et al. 2020, Koenecke et al. 2020, Bender et al. 2021, Markl 2022a). In practice, this means that speech and language technologies tend to be designed for prestigious language varieties and perform worse or not at all for marginalised groups. Another important shortcoming of ASR technologies today is their limited capacity to deal with conversational speech, noisy backgrounds, code-switching and multiple and overlapping speakers. In other words, many of the speech styles that we as sociolinguists are *most* interested in are also the *most challenging* for current ASR technology.

Nevertheless, our experience with the Lothian Diary Project suggests that incorporating automatic speech recognition can significantly reduce the time involved in transcription. As I've outlined, there may be practical reasons to opt for custom or off-the-shelf (commercial or open-source) systems. Important considerations here could be sharing permissions of the recording data, the varieties and speech styles used by the speakers, local computing resources and programming skills, cost of various options, among others. For example, for recordings of read speech in a "standard" variety of English, ASR transcripts might be very good. If the recordings contain multiple speakers with a significant amount of overlap, identifying and separating utterances by different speakers ("speaker diarisation") may be very challenging in and of itself. Of course, some of the contexts that are most difficult for an ASR system are also most difficult for human transcribers.

Accurately transcribing spontaneous speech with filled pauses, overlaps, hesitations, and false

Hi my name is Rosa and today we'll be seeing how my lockdown

** ** how many murderers and today will be seeing how my lockdown

** ** how many mysteries on today we will be seeing how my lock down

(a) Utterance produced by female speaker born in 2010 in Italy who is growing up in Edinburgh.

we couldn't even go out more than once a day I mean I I'm very active I cycle and run everywhere

we could even go out more than once a day I mean I am very active I cycle *** run everywhere

we couldn't even go out more than once a day I mean I I'm very active I cycle *** run everywhere

and I found it really hard to only be able to go out once a day

and I find it really hard to only be able to go out **** * today

andi I found it really hard to only be able to go out once a day

(b) Utterance produced by a female speaker from London (born 1974).

I miss my family I miss my friends erm and i kind of feel like it's I shouldn't be here

I miss my family * and my friends *** and I kind of feel like **** I shouldn't be here

I missed my family and missed my friends *** and I kind of do you like **** I shouldn't be here

(c) Utterance produced by a female speaker from Lithuania (born 1994).

Figure 2: Comparison of reference transcript (black) with automatic transcript. For non-Scottish speakers there are two models to compare: Google (British English) with errors highlighted in blue and Amazon (British English) with errors highlighted in orange. Deletions are marked with asterisk (*).

starts is a very challenging task even for trained human transcribers. Correcting incomplete or erroneous ASR transcripts of spontaneous speech can be similarly challenging. We hand-corrected transcripts by checking each utterance in ELAN, listening to each audio chunk, reading the corresponding transcript and making changes as necessary. While many errors are easy to spot and correct, others such as missing filled pauses and reduced forms can be trickier. These errors are also unlikely to be detected when reading through the transcript without the audio, since the resulting sentences may be perfectly grammatical and meaningful. In addition to errors humans might make too, there are also errors no human transcriber would make (e.g., Figure 2a). ASR systems appear to be sensitive to issues in recording quality, noise and, as discussed above, language model biases.

There are trade-offs involved in whether and how to integrate ASR tools in transcription workflows. Especially with very diverse speech datasets like the Lothian Diaries (different accents, topics, recording conditions etc.) the same ASR tool may also perform very differently for particular speakers and recordings. Fully manual transcripts may contain fewer transcription errors (of some types: e.g. substitutions of nouns or verbs) than hand-corrected automatic transcripts. However, they also take an order of magnitude longer to prepare. At a bare minimum we have found automatic segmentation of recordings into chunks (using voice activity detection) very useful in both manual and automatic transcription.

6 Conclusion

Embedding automatic speech recognition systems in (socio)linguistic research workflows could reduce the time involved in preparing orthographic transcriptions. Based on experiences with the Lothian Diary Project, I'd recommend working where possible with experts in speech technology to establish what the most appropriate tool is depending on available resources, skills, and the particular dataset. As linguists we are particularly well-placed to use these technologies responsibly with a clear understanding of their limitations, and to contribute to improving them.

References

Adda-Decker, Martine, and Lori Lamel. 2005. Do speech recognizers prefer female speakers? *9th European Conference on Speech Communication and Technology* 2205–2208.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery.

Benjamin, Garfield. 2021. What we do with data: A performative critique of data 'collection'. *Internet Policy Review* 10.

Bird, Steven. 2021. Sparse Transcription. *Computational Linguistics* 46:713–744.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Bucholtz, Mary. 2000. The politics of transcription. *Journal of Pragmatics* 32:1439–1465.

Chan, William, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4960–4964.

Chodroff, Eleanor. 2018. Kaldi tutorial. `https://eleanorchodroff.com/tutorial/kaldi/index.html`.

Clark, Lynn, Helen MacGougan, Jennifer Hay, and Liam Walsh. 2016. "kia ora. this is my earthquake story". multiple applications of a sociolinguistic corpus 3:13–20.

Coto-Solano, Rolando, James N. Stanford, and Sravana K. Reddy. 2021. Advances in Completely Automated Vowel Analysis for Sociophonetics: Using End-to-End Speech Recognition Systems With DARLA. *Frontiers in Artificial Intelligence* 4.

Foley, Ben, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*.

Grabe, Esther, and Francis Nolan. 2002. The ivie corpus: English intonation in the british isles.

Hall-Lew, Lauren, and Zac Boyd. 2017. Phonetic variation and self-recorded data 23:86–95.

Hall-Lew, Lauren, Claire Cowie, Catherine Lai, Nina Markl, Stephen Joseph McNulty, Shan-Jan Sarah Liu, Clare Llewellyn, Beatrice Alex, Zuzana Elliott, and Anita Klingler. 2022. The lothian diary project: sociolinguistic methods during the covid-19 lockdown. *Linguistics Vanguard* 8:321–330.

Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. https://arxiv.org/abs/1412.5567.

Himmelmann, Nikolaus P. 2018. Meeting the transcription challenge. In *Reflections on Language Documentation 20 Years after Himmelmann 1998. Language Documentation & Conservation Special Publication no. 15.*, ed. Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, 33–40. Honolulu: University of Hawaii Press.

Information Commissioner's Office. 2022. What is special category data? `https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-is-special-category-data/`.

Kendall, Tyler, and Charlie Farrington. 2021. The corpus of regional african american language.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117:7684–7689.

Leemann, Adrian. 2016. Analyzing geospatial variation in articulation rate using crowdsourced speech data. *Journal of Linguistic Geography* 4:76–96.

Leemann, Adrian, Marie-José Kolly, and David Britain. 2018. The english dialects app: The creation of a crowdsourced dialect corpus. *Ampersand* 5:1–17.

Liao, Thomas, Rohan Taori, Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* .

Liberman, Mark Y. 2019. Corpus Phonetics. *Annual Review of Linguistics* 5:91–107.

Luccioni, Alexandra, and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 182–189. Online: Association for Computational Linguistics.

Mackenzie, Laurel, and Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6:1–14.

Markl, Nina. 2022a. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Markl, Nina. 2022b. Mind the data gap(s): Investigating power in speech and language datasets. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 1–12. Dublin, Ireland: Association for Computational Linguistics.

Markl, Nina, and Catherine Lai. 2021. Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, 34–40. Online: Association for Computational Linguistics.

Markl, Nina, and Stephen Joseph McNulty. 2022. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

Martin, Joshua L. 2021. Spoken corpora data, automatic speech recognition, and bias against african american language: The case of habitual 'be'. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, FAccT '21, 284. New York, NY, USA: Association for Computing Machinery. Number of pages: 1 Place: Virtual Event, Canada.

Martin, Joshua L., and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Proc. Interspeech 2020*, 626–630.

Ochs, Elinor. 1979. Transcription as Theory. In *Developmental Pragmatics*, 43–72.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2:100336.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Reddy, Sravana, and James N. Stanford. 2015. Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1:15–28.

Schøning, Signe, and Janus Spindler Møller. 2009. Self-recordings as a social activity. *Nordic Journal of Linguistics* 32:245–269.

Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5248–5264. Online: Association for Computational Linguistics.

Sneller, Betsy. 2022. Covid-era sociolinguistics: introduction to the special issue. *Linguistics Vanguard* 8:303–306.

Sneller, Betsy, Suzanne Evans Wagner, and Yongqing Ye. 2022. MI Diaries: ethical and practical challenges. *Linguistics Vanguard* 8:307–319.

Stanford Linguistics. n.d. Voices of California. `http://web.stanford.edu/dept/linguistics/VoCal/`.

Szymański, Piotr, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3290–3295. Online: Association for Computational Linguistics.

Villarreal, Dan, Lynn Clark, Jennifer Hay, and Kevin Watson. 2020. From categories to gradience: Auto-coding sociophonetic variation with random forests. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11.

Wassink, Alicia Beckford. 2021. Uneven success: Racial bias in automatic speech recognition. `https://www.youtube.com/watch?v=CFKTxUmLByo`. Martin Luther King, Jr. Colloquium, University of Michigan.

Weinberger, Steven. 2015. The speech accent archive. online.

Institute for Language, Cognition and Computation
The University of Edinburgh
10 Crichton Street, EH8 9AB
Edinburgh, Scotland
nina.markl@ed.ac.uk