Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1178–1192
Marseille, 20-25 June 2022

Building and curating conversational corpora for diversity-aware language science and technology

Andreas Liesenfeld, Mark Dingemanse *
Centre for Language Studies
Radboud University, Netherlands
andreas.liesenfeld@ru.nl, mark.dingemanse@ru.nl

Abstract

We present an analysis pipeline and best practice guidelines for building and curating corpora of everyday conversation in diverse languages. Surveying language documentation corpora and other resources that cover 67 languages and varieties from 28 phyla, we describe the compilation and curation process, specify minimal properties of a unified format for interactional data, and develop methods for quality control that take into account turn-taking and timing. Two case studies show the broad utility of conversational data for (i) charting human interactional infrastructure and (ii) tracing challenges and opportunities for current ASR solutions. Linguistically diverse conversational corpora can provide new insights for the language sciences and stronger empirical foundations for language technology.
Keywords: corpus creation and curation, conversation, interactional linguistics, linguistic typology, dialog systems, speech recognition

1. Introduction

Language resources that capture language use in its natural habitat of social interaction are rare despite the obvious merits of studying the very environment where we all learn language and use it everyday (Schegloff, 2006). There are multiple reasons for this. Linguists have been trained to look the other way when it comes to what is considered mere performance (Boeckx, 2010). Collecting this kind of data requires one to venture out of lab settings and other controlled environments (Enfield, 2013), and transcribing it is resource-intensive (Himmelmann, 2018). These obstacles are compounded by the fact that most NLP work focuses on a handful of well-studied languages (Joshi et al., 2020; Blasi et al., 2021). However, under the auspices of various language documentation projects, language resources have been collected in more and more communities across the world (Seifart et al., 2018), and these often include at least some conversational data. We argue such corpora harbour important insights for language science and technology.

In this paper we describe efforts to build an open and reproducible pipeline for collating and curating corpora of conversational speech. We demonstrate use of the pipeline for a growing collection of corpora covering at least 67 languages of 28 phyla (Figure 1). Around 75% of the corpora are sourced from existing language documentation projects. The remaining 25% come from other language resource platforms made available to the research community. Investigating a larger slice of the world's linguistic diversity can strengthen the empirical foundation of the language sciences and foster diversity-aware language technologies.

* Both authors contributed equally.

Figure 1: Languages currently featured in the dataset plotted by geographic location of (one of their) speech communities. Coordinates from Glottolog (Hammarström et al., 2021); full list in Appendix.

We publish a repository that collects information on content and availability of corpora of conversational interaction across many languages. Currently, the curated collections amount to over 70 open datasets, representing over 700 hours of social interaction, 1.3 million annotations and over 8 million words (Table 1, Figure 2). We also publish Python and R scripts to assess the content and quality of conversational corpora. In this paper, we detail the data analysis pipeline and formulate best practices for corpus creators to optimally prepare corpora for interoperability. We also indicate some directions for research using such corpora in the form of two case studies, one aimed at linguistic typology and the other at speech technology.

Languages	67
Phyla	28
Hours of recordings	705
Annotations (turns)	1.3 million
Tokens (estimate)	8 million

Table 1: Dataset size overview.

For sourcing conversational corpora, we have used the following three criteria. First, the resource should be *maximally naturalistic*, capturing informal, unscripted interaction between two or more participants. We assess this by looking at the dynamics of turn-taking and timing, aiming to select corpora or sub-corpora characterized by free-flowing, unscripted interaction. Second, we aim for a *maximally diverse* dataset that covers many languages and phyla beyond the usual handful of languages (mostly Indo-European) that still make up the bunk of available datasets of conversational speech. Third, in order to foster open science and reproducible research, we privilege *open* resources made available with informed consent and accessible free-of-charge to the research community. As such, the bulk of the data comes from language resource repositories, often related to language documentation projects: ELDA Shared LRs, Dobes (The Language Archive), ELAR, Talkbank, OpenSLR, and so on. In some cases, commercial platforms also host accessible data, e.g. Linguistic Data Consortium (LDC). Other sources of conversational corpora are national research projects like Spoken Dutch Corpus (CGN), Spoken British National Corpus, NINJAL in Japan and FOLK in Germany. However, some of these resources (e.g., Spoken BNC, NINJAL) do not provide turn-based timing information and were therefore not included. A complete and up to date list of all data sources can be found through osf.io/cwvbe.

2. Parsing conversational corpora

Conversational corpora come in various representation formats and levels of transcription granularity. There is no one unified representation of talk that would equally satisfy the needs of researchers working in different corners of the language sciences, be it grammar writing, conversation analysis, or phonetics (Ochs, 1979; Bolden, 2015; Couper-Kuhlen and Selting, 2017). As a result, textual representations of conversation come in a range of formats and

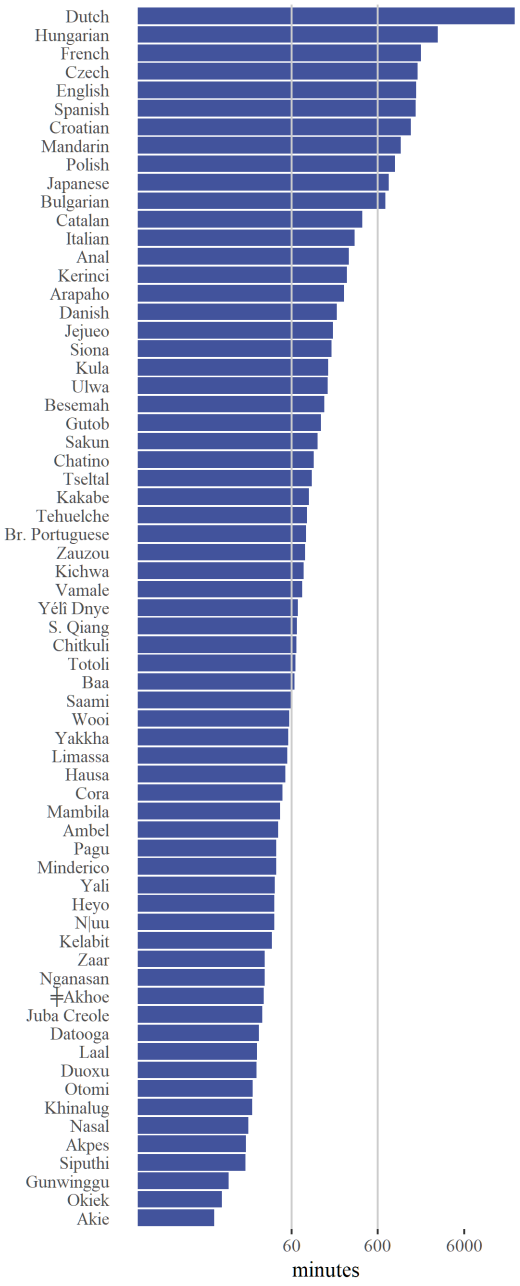

Figure 2: Language and corpus size in minutes (log scale).

with various layers of information. Many existing pipelines for working with corpora are built for textual data instead of time-aligned transcriptions of social interaction. Notable recent exceptions are the R packages `chattr` (Casillas and Scaff, 2021) and `act` (Ehmer, 2021), and a workflow for dealing with XML (Rühlemann, 2020). Less directly aimed at co-present interaction are ConvoKit (Chang et al., 2020) and the DoReCo pipeline (Paschen et al., 2020), built primarily for research into word-level time-alignment.

Despite the diversity in formats, some structural features are important for any corpus of conversation. These relate to the primary annotation level, the importance of timing and participation, the representation of supposedly marginal features, and the linking of annotations and source files. To discuss these in turn:

- **The fundamental unit of organization: turns.** People organize interactions through turns at talk, which can be characterized as communicative moves that are recognizably complete for participants in interaction (Ford and Thompson, 1996). The idealised notion of 'sentence' bears a complicated relation to empirically attested turns at talk (Sacks et al., 1974; Kempson et al., 2016). Somewhat closer to the level of turn in conversation is the notion of inter-pausal unit (IPU), which has the benefit of being automatable (Bigi, 2015), though no automated method will capture the flexibility and fluidity of human judgments about turns and the social actions they implement.

- **Represent timing and participation.** People in interaction treat the timing and duration of utterances as orderly and meaningful (Sacks et al., 1974). They minimize gaps and overlaps (Stivers et al., 2009), and are demonstrably sensitive to timing differences on the order of a few hundred milliseconds (Roberts and Francis, 2013). The accurate representation of who said what and when exactly (with at least decisecond precision) is crucial to any work on human interaction.

- **Retain relevant details.** No element of talk can be treated as discardable a priori. Conversational transcripts contain complex turns but also one-word elements such as "oh" or "um" (Buschmeier and Kopp, 2018; Williams et al., 2020), and may also capture non-verbal conduct like as breaths, laughs, sighs, or coughs (Włodarczak and Heldner, 2020; Keevallik and Ogden, 2020). If the goal is to characterize, understand, and model turns at talk, then such elements should be represented where possible and relevant.

- **Keep transcriptions and source files linked.** Since annotations and transcriptions are necessarily selective and made for a particular purpose, it is important to keep source data (audio, video, and any other streams of information like kinematics or eyegaze) closely linked to textual representations (Zimmerman, 1993). This enables repeated inspection, opens up annotations and analyses to empirical scrutiny, and makes it possible to investigate aspects not captured by annotations.

Taking these properties into account, we define a minimal viable unified representation format for conversational speech in the form of a flat dataframe that features one participant turn per row and that has (at least) five columns: `begin` and `end` of the utterance, the `participant` producing the utterance, and the `utterance` content, and finally a `source` column that links the dataframe row to any corresponding media files (Figure 3). If a source corpus contains additional information such as multiple scripts, translations, or annotation layers that capture lexical, phonetic, morphological or part-of-speech information, this is stored in additional columns. This way we ensure interoperability of core columns, but also keep additional information at hand on a per-corpus basis.

	begin	end	participant	utterance	source
0	00:00:11.350	00:00:12.850	KFT	tida bisa?	AM056
1	00:00:12.370	00:00:13.610	LA	bisasuda!	AM056
2	00:00:13.420	00:00:14.170	AF	yo	AM056
3	00:00:14.155	00:00:17.605	KFT	bapa bicara bak ini	AM056
4	00:00:17.690	00:00:19.450	LA	bak ini atau ya	AM056

Figure 3: Example of the minimal viable format for conversational data (in dataframe format).

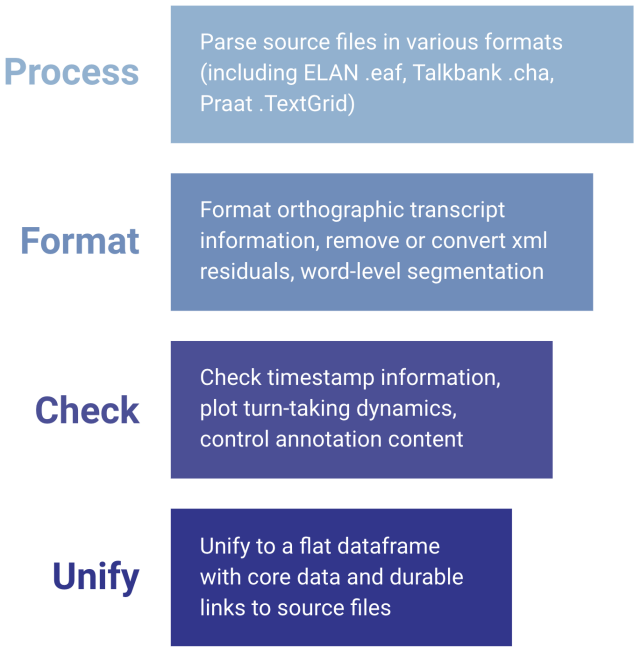

Figure 4: Overview of a four-step processing pipeline from raw transcription data to a minimal viable unified format.

The minimal viable corpus format aims to make diverse conversational corpora amenable to fundamental and applied research by enabling a basic form of sequential analysis of talk as it unfolds over time. Given the critical importance of accurate timing data and high quality transcriptions in this process, our data analysis pipeline includes a number of quality control steps that focus on assessing the accuracy and quality of available conversational corpora. The pipeline can be broken down into 4 steps (Figure 4). First we process the source transcription files by writing format-specific parsers. Then we format annotation metadata and content by converting timestamps to ms, clearing xml residuals and performing word-level segmentation. The extent of this task differs across languages: often splitting by whitespace is sufficient, but for other languages such as Chinese, we employ a parser. In a next step we check timing information quality and annotation content, using a common `[unk]` tag for missing annotations (see section 3 for details). Finally we unify the data by storing it in a flat dataframe with durable links to source files.

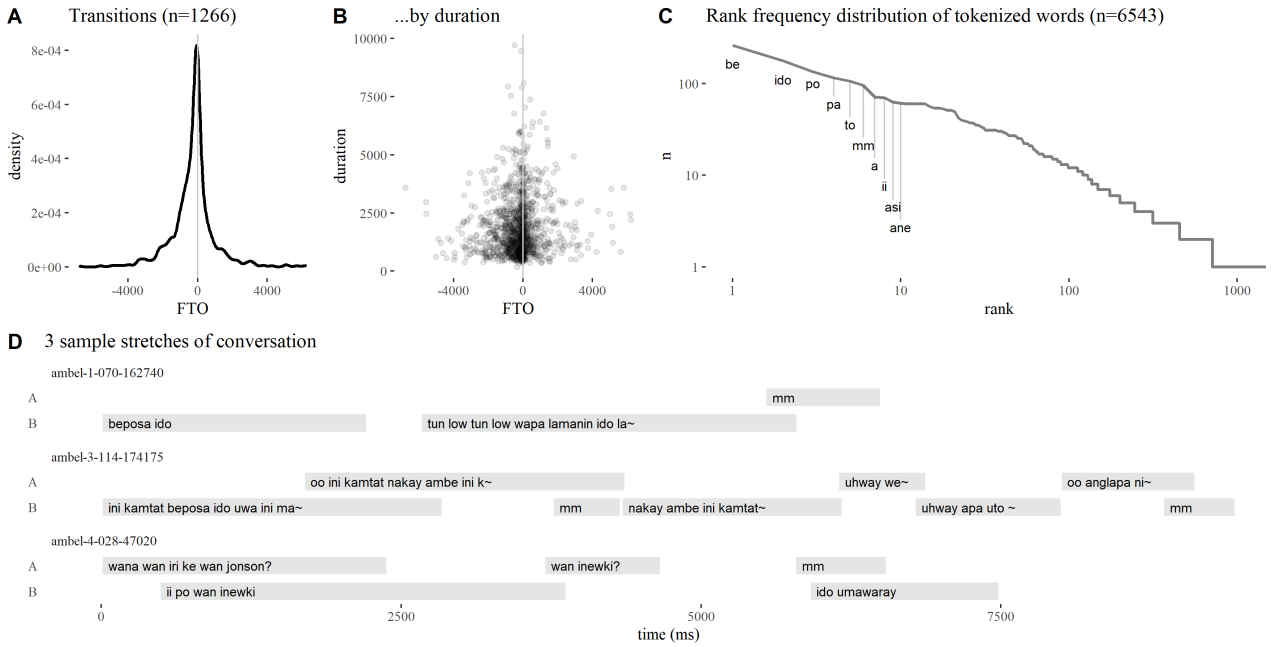

Figure 5: Example of an assessment report for conversational data, here illustrated with data from Ambel (Arnold, 2017). **A**. Distribution of the timing of dyadic turn-transitions with positive values representing gaps between turns and negative values representing overlaps. This kind of normal distribution centered around 0 ms is typical; when corpora starkly diverge from this it usually indicates non-interactive data, or segmentation methods that do not represent the actual timing of utterances. **B**. Distribution of transition time by duration, allowing the spotting of outliers and artefacts of automation (e.g. many turns of similar durations). **C**. A frequency/rank plot allows a quick sanity check of expected power law distributions and a look at the most frequent tokens in the corpus. **D**. Three randomly selected 10 second stretches of dyadic conversation give an impression of the timing and content of annotations in the corpus.

3. Curating conversational data

After parsing we perform several quality control steps to ensure that the corpus data is fit for inclusion in the curated collection. In particular, we verify annotation content, check source files, and assess timing information.

Annotation content. We verify that annotation content is transcribed using a common orthographic format. For data processing reasons, we romanize all scripts (while preserving the original). Many corpora also contain annotations tagged for bodily conduct like laughter, breathing, or coughing. We aim to preserve as much of this information as possible, converting common tags to a unified tag format, and marking other such elements using square brackets.

Source files. The baseline requirement of inclusion in the dataset is that audio data of reasonable quality exists that allows us to verify and unify transcription content using primary data. This is important to alleviate risks of grounding any subsequent analysis on transcriptions alone. We also record any other data streams (e.g. video, gaze), verify that every source reference in the data has a corresponding source file and manually inspect the overall recording quality. The resulting curated corpora can be used for automatic extraction of audio clips for further analysis, and for feature extraction of prosodic properties that are interactionally relevant like speech rate, intensity and pitch (Selting, 1996; Ward and Vega, 2012).

Timing. Given the importance of timing and participation, we aim to ensure that timing information is as complete and accurate as possible across the dataset. For the purpose of examining talk-in-interaction we define this as timestamps that accurately correspond to the beginnings and ends of conversational turns. This also includes the accurate identification of overlaps and gaps between turns. Quality control of these measures is done through a combination of manual inspection and quantitative measures, for instance by plotting the dynamics of turn-taking and timing (Figure 5A-D).

Doing this for over 50 corpora, we identified several recurring issues. One is incomplete transcriptions. We compute the relative annotation density in order to get a sense of the likelihood of missing annotations. Another issue is inaccurate timing information, which may arise from the use of automated transcription methods or shortcuts in manual annotation software. To diagnose such issues at scale, we generate an assessment report for every language (see Figure 5) with information on turn transition timings, turn durations, rank-frequency distributions, and sample stretches of dyadic conversation. This enables a quick assessment of the relative precision and granularity of a corpus. For instance, timing inaccuracies are often recognizable by deviations from the expected normal distribution (Stivers et al., 2009); and sample plots of conversations give an impression of interaction type and annotation content, including empty annotations. A codebase for generating such assessment reports is available in the repository.

4. Building corpora: best practices

There is still a relative dearth of conversational corpora, and most of the world's linguistic diversity remains underrepresented. The reasons for this are varied and include the fact

that conversational data is seen as hard to collect and even harder to analyse, with the language sciences preferring to focus on sanitized versions of linguistic structure and behaviour and language technology similarly focusing on text and speech data that is pristine and can be easily processed.

Broader representation is important for communities and heritage users of languages, who have long been underserved by monologic textual materials that make it hard to get a taste of what it is like to use a given language in face-to-face interaction (Amery, 2009). But it is also important for scientific purposes, as every language offers its own contribution to the tapestry of unity and diversity that characterizes our common cultural heritage (Bird, 2020).

Based on our experience building and processing conversational corpora, here we provide some recommendations for *minimally useful conversational corpora*. Good corpora can often serve multiple purposes, from language learning to linguistic research (Enfield, 2013) and from comparative investigations to supplying training data for NLP purposes. Such corpora have the following properties:

- It is collected and archived with consent of participating language users and the relevant community leaders
- It contains audio and/or video recordings of everyday face to face interaction among multiple participants
- It is time-aligned at the level of conversational turns, with turns allotted to participants
- It is transcribed in a way that provides access to the linguistic material beyond the audio/video

There are good guides for building and transcribing corpora (Allwood, 2008; Paschen et al., 2020), and increasingly, tools are available to transcribe audio and video data (Bird, 2021; Dingemanse et al., 2012) and enrich annotations (Umair et al., 2021; Zahrer et al., 2020).

5. Exploring conversational corpora

To showcase the utility of conversational corpora, we provide two case studies. The first is focused on linguistic typology and investigates the use of conversational data for comparative studies of language structure. The second is focused on language technology and compares conversational corpora to typical speech recognition data.

5.1. Continuers and repair initiators

One aspect of interactive language use that should be of considerable interest to language science and technology is the common occurrence of linguistic devices with primarily interactional functions (Allwood et al., 1990; Norrick, 2009; Liesenfeld, 2019). Consider two stand-alone turn formats that are particularly frequent yet have very different functions (Figure 6A). A *continuer* signals an understanding that the other party is producing a series of turns and an expectation that more is coming; a *repair initiator* signals a request for clarification and requires both parties to halt the conversation and interactively resolve the communicative trouble, often with a redoing of the prior turn as a result.

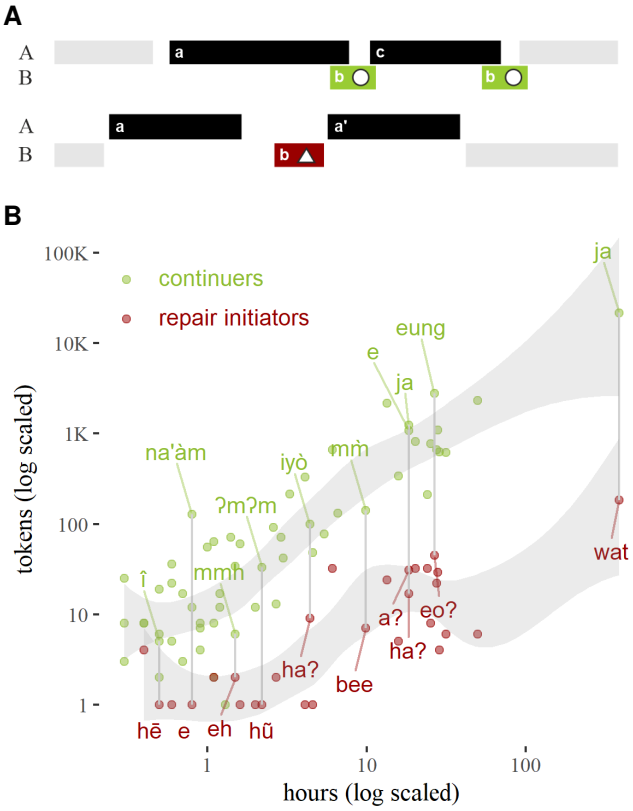

Figure 6: **A**: Typical sequential structures for continuers versus repair initiators. Continuers are recurring items found in alternation with unique turns (*a, c*). Repair initiators are recurring items found between a unique turn *a* and its near-copy *a'*. **B**: Prevalence of sequentially identified candidate continuers and repair initiators, demonstrating the potential of using sequential patterns to identify them in language-agnostic ways. Most frequent formats exemplified in 10 languages (9 phyla), from left to right: ǂAkhoe Hai||om, Hausa, Tehuelche, Gutob, Kerinci, Siwu, Mandarin, German, Korean, Dutch.

If these items occur at all in text data, they are divorced from their interactional context; indeed, they are underrepresented even in scripted conversations (Prevot et al., 2018; Prevot et al., 2019), making them hard to identify a priori by systems that have access to form alone (Bender and Koller, 2020). How then can we identify them for cross-linguistic comparison and work towards their naturalistic implementation in cross-linguistically informed language technology?

The solution is to think in terms of the sequential structure of interaction (Jefferson, 1978; Couper-Kuhlen and Selting, 2017; Dingemanse and Liesenfeld, 2022). Using only sequential and frequency information, we can define the prototypical continuer as a recurrent turn format that occurs in alternation with near-unique turns by another speaker, and the prototypical repair initiation as a recurrent turn format found between a near-unique turn and its near-copy (Figure 6A). Implementing these language-agnostic definitions as a sequential search template and using a normalised Levenshtein distance of < 0.20 to identify near-similar turns, we are able to identify candidate continuers and repair initiators across corpora (Figure 6B). Though identified in a fully language-agnostic way, the resources

identified here appear to fit available linguistic descriptions quite well. For instance, in the repair initiators we recognize the 'huh?'-like interjections and 'what?'-based question words known from prior work in pragmatic typology (Enfield et al., 2013), and among the continuers we find a similar mix of minimal 'mhm'-like interjections and affirmative answers (Dingemanse, 2021).

The sequential search method is susceptible to fluctuations in corpus size and its interpretation for specific corpora will always require careful qualitative work. However, it can be used to quickly gauge the likely forms of key interactional tools, and can provide a lower bound on the amount of data needed to identify interactional tools in conversational corpora. For instance, in this case, we may conclude that an hour of conversational data can be sufficient to identify the most important interactional tools. More generally, corpora like this can also inform careful interactional linguistic work that respecifies traditional linguistic concepts (Ozerov, 2022). Interactional data analysed in terms of sequential and distributional properties is likely to be a crucial element in the toolbox of conversational NLP, providing definitions and distinctions that are easily lost in tokenisation, part-of-speech tagging and machine translation.

5.2. Conversational vs. ASR corpora

While corpora of informal conversation are relatively rare, *speech corpora* are standard fare in language technology and particularly automatic speech recognition (ASR). Indeed readers familiar with work in this domain may wonder why we have not included such speech resources, many of which are openly available.

The most important reason for this is that few if any ASR corpora are based on conversational interaction. Instead, ASR corpora usually are derived from carefully read speech samples, whether audiobooks as in Librispeech (Panayotov et al., 2015), elicited text prompts from Wikipedia as in CommonVoice (Ardila et al., 2020), or European Parliament proceedings as in VoxPopuli (Wang et al., 2021). This makes ASR corpora very useful for recordings that share key features with the training data (i.e., audio that is monologic and comes in fairly long sentences). But the character of the training data may limit the quality and application potential of ASR in other domains such as conversational user interfaces.

Some of the differences are obvious (Figure 7). First of all, ASR corpora come in chunks that are optimally sized for current ASR training solutions, meaning they are rarely longer than 35 seconds and rarely shorter than 2 seconds (Panayotov et al., 2015). These chunks are sentences or fragments of monologic textual corpora. Conversational corpora on the other hand come in turns that are optimally sized (by communicating participants) for the delivery of social actions (Enfield, 2009). Typical ASR training chunks are much longer than turns at talk in conversation. The modal length of CommonVoice recordings for Hungarian, Dutch and Catalan is 4.6 seconds; for conversational corpora for these same three languages it is 0.5 seconds. The differences are striking enough that the overall distributions of utterances and sentences overlap only for a small part of the data (Figure 7L). This pattern is not unique to the corpora compared here, or indeed to the CommonVoice dataset: the mean length of LibriSpeech English (Panayotov et al., 2015) and of RuLS Russian sentences (Bakhturina et al., 2021) is also around 3 seconds. For comparison, in Figure 8 we provide the distributions of turn lengths in 22 conversational corpora from our dataset. This shows that conversational turns typically are at least twice as short as the typical sentences found in ASR training datasets.

Length may be the most obvious difference but it is not the main issue. The content of typical chunks differs quite a lot across corpus types. Figure 7R shows how language differs according to corpus type for Dutch (comparing conversational data and the open CommonVoice dataset). The type of language most distinctive of the CommonVoice corpus mark its origin in parliamentary recordings, with formal words like "verslag" (report), "maatregelen" (measures),

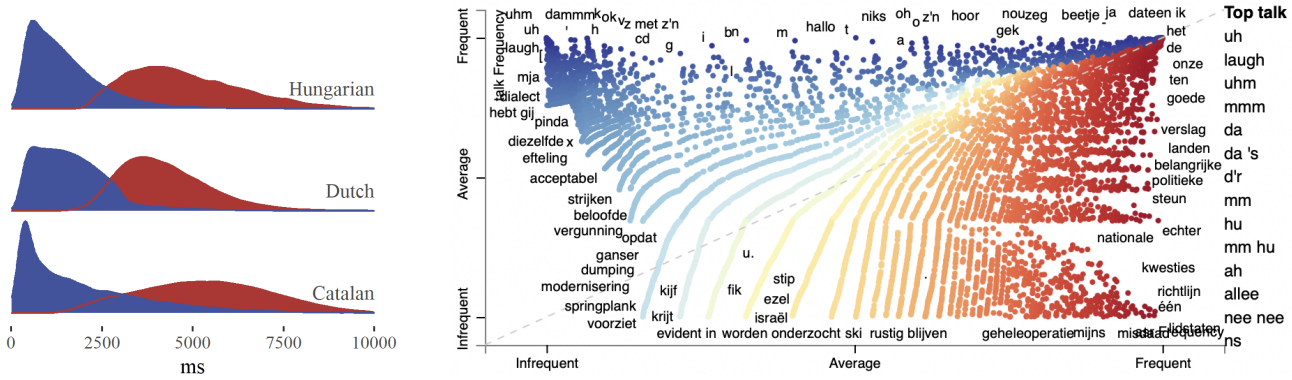

Figure 7: **L**: Distributions of durations of utterances and sentences (in ms) in corpora of informal conversation (blue) and CommonVoice ASR training sets (red) in Hungarian, Dutch, and Catalan. Modal duration and annotation content differ dramatically by data type: 496ms (6 words, 27 characters) for conversational turns and 4642ms (10 words, 58 characters) for ASR training items. **R**: Visualization of tokens that feature more prominently in conversational data (blue) and ASR training data (red) in Dutch. Source data: 80k random-sampled items from the Corpus of Spoken Dutch (Taalunie, 2014) and the Common Voice corpus for automatic speech recognition in Dutch (Ardila et al., 2020), based on Scaled F score metric, plotted using *scattertext* (Kessler, 2017)

"steun" (support) and "standpunt" (position). The type of language most typical of the conversational data marks its more informal and interactional nature, with interjections like "oh", "uh" and "uhm" (the latter two delay markers), polar response particles like "mm" (yes) and "nee" (no), and pronouns like "ge", "gij" and "m'n" (my). These are exactly the kind of interactional tools we saw in §5.1 above: little words that are frequently used and that streamline conversation.

One implication of this is that many of the short and highly frequent turn formats in conversational speech are not well represented in ASR training data, with detrimental consequences for the ability of ASR models to deal with conversational speech. And indeed there are indications that ASR performs less well for such data. One study comparing Google, Microsoft and HuggingFace ASR models for Swedish found that "for all spontaneous speech, the ASRs frequently fail to produce a transcription for short utterances" (Cumbal et al., 2021). In a study comparing gold standard human transcripts with ASR output, it is precisely the words that serve as continuers, feedback signals and other metacommunicative signals that are most frequently missed (Zayats et al., 2019).

Missing or incorrectly transcribing short utterances may not be a big problem for speech recognition models whose main function is to deal with relatively clean recordings of non-conversational speech (such as speeches, radio programs, parliamentary meetings and other highly institutionalized text types). But it does spell serious trouble for the use of speech recognition in more interactive contexts such as voice user interfaces and conversational agents. If one goal of ASR is to be able to deal with human speech input in interactive situations, then the current training data may not optimally prepare it for this job.

Recent work in two areas has claimed some territory here. Small corpus size need not be a problem: there are promising ASR results for corpora that amount to only an hour of speech, albeit non-interactive (Tyers and Meyer, 2021). Further, under the banner of 'textless' ASR, Nguyen et al. (2022) investigate how speech features can be learned from conversational English telephone data using a pipeline that relies less on textual representations. Such techniques may be extended to other types of conversational data, although this inevitably requires dealing with hurdles like noise and non-separate channels.

Indeed a serious challenge for wider reliance on conversational corpora is that the speech signal in such corpora is in many cases not as pristine as typical ASR data: it includes overlapping speech, non-speech, background noises and comes with all of the complexities generated by different recording environments and equipments. However, to the extent this is a problem, we submit that this is not so much an issue to be solved on the input side, but a matter of ecological validity. If we want ASR systems that are able to deal with the contingencies and exigencies of human interaction in its natural environment (Baumann et al., 2017; Rivière and Dupoux, 2021), then we better make sure the training data prepare them for this. We anticipate that carefully time-aligned corpora of the type we curate and describe here will play a key role in this. Very likely, selective augmentation of ASR training corpora will provide a scaleable partial solution: including truly interactional data in ASR training sets will meaningfully improve speech recognition in challenging interactional contexts.

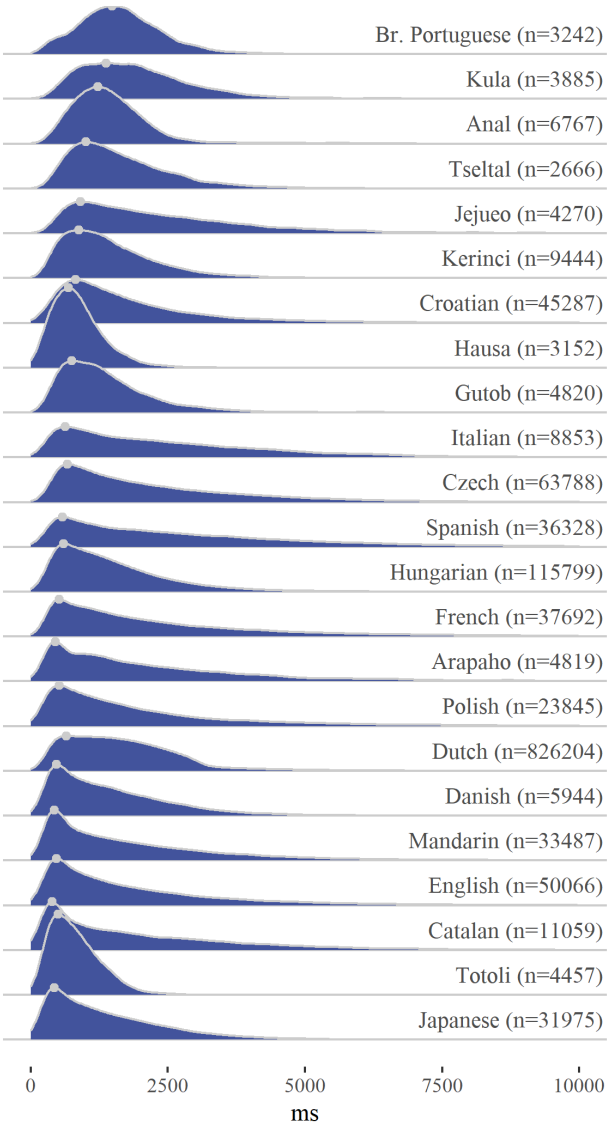

Figure 8: Distributions of turn durations (ms) in conversational corpora for 22 languages with at least 2500 turns. Across all languages (950k turns in total), the modal duration is 495ms (mean 1760ms, sd 1414ms). Recall that the modal duration of training items in most ASR corpora is an order of magnitude larger.

6. Discussion

Conversational corpora are crucial for furthering our understanding of language in its most natural habitat and for building diversity-aware language technology. An increasing number of such corpora is available as a result of decades of concerted efforts in the field of language documentation to compile and preserve linguistic primary data (Himmelmann, 2006; Seifart et al., 2018).

Yet even though interactional data is gaining prominence, it is still not standard fare in corpus linguistics and NLP research. This results in a kind of double bind: con-

versational data is hard to find and clear examples of the importance of such data for language science and technology are rare. One of our aims is to free the field from this double bind by compiling open data about available resources, specifying key properties of an interoperable format for interactional data, providing elements of an open data processing pipeline, and pointing out promising research directions. However, this work also comes with ethical considerations, and this is what we discuss first.

6.1. Ethical considerations

Ethical considerations start with the question of where to focus our efforts and what kind of research to pursue. The field of natural language processing is famously data-hungry, and current incentives are aligned to harvest more of the same: large amounts of mostly monologic textual data (Bender et al., 2021). In the face of this, it may seem quaint to focus on linguistically diverse, hard-won, small to medium-sized corpora of conversational interaction with relatively few opportunities for the rapid scaling up we have come to expect from web-scrapeable data. But we believe this is a worthwhile, even essential direction for language science and technology. Serious engagement with the cultural, linguistic and interactional diversity embedded in conversational corpora can be a step towards decolonizing the computational sciences (Birhane and Guest, 2021; Bird, 2020). A true understanding of the workings of language in interaction requires deep engagement with this kind of data (Dingemanse and Liesenfeld, 2022). For anyone seeking to model open-domain conversation, conversational corpora offer a richer and more challenging model than threaded forum posts. For engineers working on conversational agents, actual records of conversation are the best place to study the foundations of interactional infrastructure. For conversation designers, there is no better place to appreciate the sheer flexibility and open-endedness of human interaction than records of people talking.

Ethical considerations extend to the data sourcing and curation process (Rogers, 2021). Language resources can have helpful uses but also harmful ones. Language documentation corpora come with their own possibilities and risks for such *dual use* (Hovy and Spruit, 2016; Levow et al., 2021). For datasets archived with community consent and made openly available, there is always the possibility of secondary uses of datasets not foreseen by compilers and communities (Seyfeddinipur et al., 2019; Rogers et al., 2021). This may include applications such as the possibility of using conversational data to inform pragmatic typology or to improve speech recognition technology, as we have shown above. While neither of these use cases relies on personally identifiable information, more questionable uses might also possible. For instance, audio and video data necessarily includes personally identifiable information (people's voices and likenesses), and the content of quotidian conversations may occassionally feature information that may be subject to privacy considerations.

For this reason and others, data sourcing and distribution has to be regulated. In the case of language documentation corpora, an important part of this responsibility is carried by language resource archives, which typically provide data use agreements, and by corpus compilers, who typically record the sources, goals and circumstances of data collection as well as point out limitations and restrictions on use (Seyfeddinipur et al., 2019). The field of language documentation has long emphasized the importance of vigilance about ethical data collection, informed consent and reproducible research (Dwyer, 2006; Bowern, 2007; Berez-Kroeker et al., 2018; Good, 2018). The use of data sheets for NLP data sets (Gebru et al., 2021) represents an important convergent development, and may provide inspiration to language documentation archives and corpus compilers.

Two complicating factors here are worth noting. First, it is the very nature of secondary uses that they cannot be fully foreseen at the point of data collection. This makes it all the more important that researchers are rooted in the communities they work with, and that they clearly communicate the possible implications of having research data archived in internet-accessible repositories (Levow et al., 2021). Second, ethical notions cannot be assumed to be culturally neutral; for instance, Ameka & Terkourafi point out that Western ethical frameworks privilege autonomy and privacy, whereas "in some communities, research participants are happy, indeed expect, to be fully identified" (Ameka and Terkourafi, 2019). As they note, ideally, questions about data archiving including anonymization practices should be informed by *local* ethical standards. Throughout, a guiding principle should be that data is archived, and its availability and conditions on reuse set, in accordance with the wishes of participants and communities (Nathan, 2013).

6.2. Conclusions

We have documented here a first effort at sourcing a maximally diverse set of openly available conversational corpora. We publish an up to date survey of available corpora that provides a headstart for people looking to work with more diverse data sets. And we share specifications and code for an analysis pipeline to enable others to use similar methods for building and curating conversational corpora. In order to further the goal of resource creation, we have also formulated some simple criteria for creating minimally viable conversational corpora.

Having generalisable methods and data representations for dealing with interactional data serves the interest of many communities working with language resources. It contributes towards alleviating problems of resource inequality in two ways: by making visible the relative diversity of corpora already available, and by showing how such data can be productively used. Conversational data enables new research questions in linguistic typology and brings into view new applications for language technology. In time, the increasing availability of interactional data in interoperable formats will provide the foundations for novel work at the intersection of language resources and human language technologies.

7. Acknowledgements

Funding for the work reported here comes from Dutch Research Council grant NWO 016.vidi.185.205 to MD. We thank Ada Lopez for work on the processing pipeline and Matheus Azevedo for help in collating corpus metadata.

8. Bibliographical References

Allwood, J., Nivre, J., and Ahlsén, E. (1990). Speech Management—on the Non-written Life of Speech. *Nordic Journal of Linguistics*, 13(01):3–48.

Allwood, J. (2008). Multimodal Corpora. In *Corpus Linguistics. An International Handbook*, pages 207–225. Mouton de Gruyter, Berlin.

Ameka, F. K. and Terkourafi, M. (2019). What if…? Imagining non-Western perspectives on pragmatic theory and practice. *Journal of Pragmatics*, 145:72–82, May.

Amery, R. (2009). Phoenix or Relic? Documentation of Languages with Revitalization in Mind. *Language Documentation & Conservation*, 3(2):138–148.

Bakhturina, E., Lavrukhin, V., and Ginsburg, B. (2021). A Toolbox for Construction and Analysis of Speech Datasets. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, December.

Baumann, T., Kennington, C., Hough, J., and Schlangen, D. (2017). Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. In *Dialogues with social robots*, pages 421–432. Springer.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July. Association for Computational Linguistics.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (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*, pages 610–623, Virtual Event Canada, March. ACM.

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., and Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1):1–18.

Bigi, B. (2015). SPPAS - Multi-Lingual Approaches to the Automatic Annotation of Speech. *The Phonetician. Journal of the International Society of Phonetic Sciences*, 111–112(ISSN:0741-6164):54–69.

Bird, S. (2020). Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Bird, S. (2021). Sparse Transcription. *Computational Linguistics*, 46(4):713–744, February.

Birhane, A. and Guest, O. (2021). Towards Decolonising Computational Sciences. *Kvinder, Køn & Forskning*, (2):60–73.

Blasi, D., Anastasopoulos, A., and Neubig, G. (2021). Systematic Inequalities in Language Technology Performance across the World's Languages. *arXiv:2110.06733 [cs]*, October.

Boeckx, C. (2010). *Language in Cognition: Uncovering Mental Structures and the Rules behind Them*. Wiley-Blackwell, Malden, MA.

Boersma, P. and Weenink, D. (2013). Praat: Doing phonetics by computer.

Bolden, G. B. (2015). Transcribing as Research: "Manual" Transcription and Conversation Analysis. *Research on Language and Social Interaction*, 48(3):276–280, July.

Bowern, C. (2007). *Linguistic Fieldwork: A Practical Guide*. Palgrave MacMillan, New York.

Buschmeier, H. and Kopp, S. (2018). Communicative listener feedback in human–agent interaction: Artificial speakers need to be attentive and adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*.

Casillas, M. and Scaff, C. (2021). Analyzing contingent interactions in R with chattr. In *Proceedings of CogSci 2021*, pages 2540–2546. Cognitive Science Society.

Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J., and Danescu-Niculescu-Mizil, C. (2020). ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting, July. Association for Computational Linguistics.

Couper-Kuhlen, E. and Selting, M. (2017). *Interactional Linguistics: An Introduction to Language in Social Interaction*. Cambridge University Press, Cambridge.

Cumbal, R., Moell, B., Lopes, J., and Engwall, O. (2021). "You don't understand me!": Comparing ASR results for L1 and L2 speakers of Swedish. *Proc. Interspeech 2021*, pages 4463–4467.

de Vos, C., Casillas, M., Uittenbogert, T., Crasborn, O., and Levinson, S. C. (2021). Predicting conversational turns: Signers' and nonsigners' sensitivity to language-specific and globally accessible cues. *Language*, 98(1):35–62.

Dingemanse, M. and Liesenfeld, A. (2022). From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Dingemanse, M., Hammond, J., Stehouwer, H., Somasundaram, A., and Drude, S. (2012). A high speed transcription interface for annotating primary linguistic data. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 7–12, Avignon, France, March.

Dingemanse, M. (2021). Interjections. *The Oxford Handbook of Word Classes*.

Dwyer, A. M. (2006). Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, et al., editors, *Essentials of Language Documentation*, page 31.

Ehmer, O. (2021). Act: Aligned Corpus Toolkit.

Enfield, N. J., Dingemanse, M., Baranova, J., Blythe, J., Brown, P., Dirksmeyer, T., Drew, P., Floyd, S., Gipper, S., Gísladóttir, R., Hoymann, G., Kendrick, K. H., Levinson, S. C., Magyari, L., Manrique, E., Rossi, G.,

San Roque, L., and Torreira, F. (2013). Huh? What? – A first survey in twenty-one languages. In Makoto Hayashi, et al., editors, *Conversational Repair and Human Understanding*, pages 343–380. Cambridge University Press, Cambridge.

Enfield, N. J. (2009). *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances*. Cambridge University Press, Cambridge.

Enfield, N. J. (2013). Doing fieldwork on the body, language, and communication. In Cornelia Müller, et al., editors, *Handbook Body – Language – Communication*, pages 974–981. Mouton De Gruyter, Berlin.

Ford, C. E. and Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resourcers for the management of turns. In Elinor Ochs, et al., editors, *Interaction and Grammar*, number 13 in Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, November.

Good, J. (2018). Ethics in language documentation and revitalisation. In Kenneth Regh et al., editors, *Oxford Handbook of Endangered Languages*, pages 419–440. Oxford University Press, Oxford.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). Glottolog/glottolog: Glottolog database 4.4, May.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for. In Jost Gippert, et al., editors, *Essentials of Language Documentation*, pages 1–30.

Himmelmann, N. P. (2018). Meeting the Transcription Challenge. In Bradley McDonnell, et al., editors, *Reflections on Language Documentation 20 Years after Himmelmann 1998*, pages 33–40. University of Hawai'i Press, Honolulu, December.

Hovy, D. and Spruit, S. L. (2016). The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August. Association for Computational Linguistics.

Jefferson, G. (1978). Sequential aspects of storytelling in conversation. In J. Schenkein, editor, *Studies in the Organization of Conversational Interaction*, pages 219–248. Academic Press, New York.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Keevallik, L. and Ogden, R. (2020). Sounds on the Margins of Language at the Heart of Interaction. *Research on Language and Social Interaction*, 53(1):1–18, January.

Kempson, R., Cann, R., Gregoromichelaki, E., and Chatzikyriakidis, S. (2016). Language as Mechanisms for Interaction. *Theoretical Linguistics*, 42(3-4):203–276.

Kessler, J. (2017). Scattertext: A Browser-Based Tool for Visualizing how Corpora Differ. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, pages 85–90.

Levow, G.-A., Ahn, E. P., and Bender, E. M. (2021). Developing a Shared Task for Speech Processing on Endangered Languages. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1(2).

Liesenfeld, A., Parti, G., and Huang, C.-R. (2021). Scikit-talk: A toolkit for processing real-world conversational speech data. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 252–256.

Liesenfeld, A. (2019). Cantonese turn-initial minimal particles: Annotation of discourse-interactional functions in dialog corpora. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 471–479. Waseda Institute for the Study of Language and Information.

Nathan, D. (2013). Access and Accessibility at ELAR, a Social Networking Archive for Endangered Languages Documentation. In Mark Turin, et al., editors, *Oral Literature in the Digital Age: Archiving Orality and Connecting with Communities*, pages 21–40. Open Book Publishers.

Nguyen, T. A., Kharitonov, E., Copet, J., Adi, Y., Hsu, W.-N., Elkahky, A., Tomasello, P., Algayres, R., Sagot, B., Mohamed, A., and Dupoux, E. (2022). Generative Spoken Dialogue Language Modeling. *arXiv:2203.16502 [cs, eess]*, March.

Norrick, N. R. (2009). Interjections as pragmatic markers. *Journal of Pragmatics*, 41(5):866–891, May.

Ochs, E. (1979). Transcription as theory. In Elinor Ochs et al., editors, *Developmental Pragmatics*, page 43–72. Academic Press, New York.

Ozerov, P. (2022). This research topic of yours — Is it a research topic at all? Using comparative interactional data for a fine-grained reanalysis of traditional concepts. In Geoffrey Haig, et al., editors, *Doing Corpus-Based Typology with Spoken Language Data: State of the Art*, pages 233–280. University of Hawai'i Press, Honolulu.

Paschen, L., Delafontaine, F., Draxler, C., Fuchs, S., Stave, M., and Seifart, F. (2020). Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2657–2666, Marseille, France, May. European Language Resources Association.

Prevot, L., Stali, M., and Tseng, S.-C. (2018). Grouping conversational markers across languages by exploiting large comparable corpora and unsupervised segmentation. In *11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan, May.

Prevot, L., Magistry, P., and Lison, P. (2019). Should we use movie subtitles to study linguistic patterns of conver-

sational speech? A study based on French, English and Taiwan Mandarin. In *Third International Symposium on Linguitic Patters of Spontaneous Speech*, Proceedings of Third International Symposium on Linguitic Patters of Spontaneous Speech, Taipei, Taiwan, November.

Rivière, M. and Dupoux, E. (2021). Towards Unsupervised Learning of Speech Features in the Wild. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 156–163, January.

Roberts, F. and Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–EL477.

Rogers, A., Baldwin, T., and Leins, K. (2021). 'Just What do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Rogers, A. (2021). Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2194, Online, August. Association for Computational Linguistics.

Rühlemann, C. (2020). *Visual Linguistics with R: A Practical Introduction to Quantitative Interactional Linguistics*. John Benjamins Publishing Company, Amsterdam, July.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735, December.

Schegloff, E. A. (2006). Interaction: The Infrastructure for Social Institutions, the Natural Ecological Niche for Language, and the Arena in which Culture is Enacted. In Nick J. Enfield et al., editors, *Roots of human sociality: Culture, cognition, and human interaction*, pages 70–96. Berg, Oxford.

Schmidt, T. and Wörner, K. (2014). EXMARaLDA. In *Handbook on Coprus Phonology*, pages 402–419. Oxford University Press, Oxford.

Seifart, F., Evans, N., Hammarström, H., and Levinson, S. C. (2018). Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Selting, M. (1996). Prosody as an activity-type distinctive cue in conversation: The case of so-called 'astonished' questions in repair initiation. In Elizabeth Couper-Kuhlen et al., editors, *Prosody in Conversation: Interactional Studies*, pages 231–270. Cambridge University Press, Cambridge / New York.

Seyfeddinipur, M., Ameka, F., Bolton, L., Blumtritt, J., Carpenter, B., Cruz, H., Drude, S., Epps, P., Ferreira, V., Galucio, A., Hellwig, B., Hinte, O., Holton, G., Jung, D., Buddeberg, I. K., Krifka, M., Kung, S., Monroig, M., Neba, A. N., Nordhoff, S., Pakendorf, B., von Prince, K., Rau, F., Rice, K., Riessler, M., Brenig, V., Thieberger, N., Trilsbeek, P., van der Voort, H., and Woodbury, T. (2019). Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation*, 13:545–536.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, June.

Tyers, F. M. and Meyer, J. (2021). What shall we do with an hour of data? speech recognition for the un- and under-served languages of common voice. *arXiv preprint arXiv:2105.04674*.

Umair, M., Mertens, J., Albert, S., and de Ruiter, J. P. (2021). GailBot: An automatic transcription system for Conversation Analysis. Technical report, OSF.

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. (2021). VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *arXiv:2101.00390 [cs, eess]*, July.

Ward, N. G. and Vega, A. (2012). A Bottom-Up Exploration of the Dimensions of Dialog State in Spoken Interaction. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 198–206, Seoul, South Korea, July. Association for Computational Linguistics.

Williams, N., Stenzel, K., and Fox, B. (2020). Parsing particles in Wa'ikhana. *Revista Linguítica*, 16(Esp.):356–382, November.

Włodarczak, M. and Heldner, M. (2020). Breathing in Conversation. *Frontiers in Psychology*, 11:575566, October.

Zahrer, A., Zgank, A., and Schuppler, B. (2020). Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2893–2900, Marseille, France, May. European Language Resources Association.

Zayats, V., Tran, T., Wright, R., Mansfield, C., and Ostendorf, M. (2019). Disfluencies and Human Speech Transcription Errors. In *Proceedings of Interspeech 2019*, pages 3088–3092. ISCA, September.

Zimmerman, D. H. (1993). Acknowledgment Tokens and Speakership Incipiency Revisited. *Research on Language & Social Interaction*, 26(2):179–194, April.

9. Language Resource References[1]

Amith, J. D., Alcántara, A. D., Osollo, H. S., Castañeda, C. S., and Salazar, E. G. (2009). Audio corpus of Sierra Nororiental and Sierra Norte de Puebla Nahuat(l) with accompanying time-code transcriptions in ELAN. *OpenSLR*, (link).

[1]Language resources are cited here according to the latest LREC citation style. We note that this style makes it hard to cite web resources in the way recommended by archives. To make available both the archive name (e.g., Endangered Language Archive) as well as a durable URL, we have resorted to a workaround: we encode the URL as an href attribute in the number field. Full and correct .bib metadata is available in our repository.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222.

Arnold, L. (2017). The documentation of Ambel, an Austronesian language of Eastern Indonesia. *Endangered Languages Archive*, (link).

Barlow, R. (2017). Documentation of Ulwa, an endangered language of Papua New Guinea. *Endangered Languages Archive*, (link).

Brykina, M., Gusev, V., Szeverényi, S., and Wagner-Nagy, B. (2018). Nganasan Spoken Language Corpus (NSLC). *University of Hamburg*, (link).

Canavan, A. and Zipperlen, G. (1996a). CALLFRIEND American English-Non-Southern Dialect. page 1508304 KB.

Canavan, A. and Zipperlen, G. (1996b). CALLFRIEND Korean.

Canavan, A. and Zipperlen, G. (1996c). CALLHOME Mandarin Chinese Speech. page 1080128 KB.

Canavan, A. and Zipperlen, G. (1996d). CALLHOME Spanish Speech.

Canavan, A., Graff, D., and Zipperlen, G. (1997a). CALLHOME German Speech.

Canavan, A., Zipperlen, G., and Graff, D. (1997b). CALLHOME Egyptian Arabic Speech. page 1807744 KB.

Canavan, A., Zipperlen, G., and Graff, D. (2014). CALLFRIEND Farsi Second Edition Speech. page 1914713 KB, January.

Caron, B., Davan, M. S., and Ali, J. M. B. (2014). Zaar collection in LLACAN. *COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO)*, (link).

Caron, B. (2016). Hausa collection in LLACAN. (link).

Carvalho Ferreira, V. A., Wurm, S., Bouda, P., Endruschat, A., Hämmerle, R. R., Fugaru, I., Rodriguez, R., Ferreira, H. L., and Knuffmann, K. (2011). Minderico, An Endangered Language in Portugal. *DOBES*, (link).

Cowell, A. (2010). A Conversational Database of the Arapaho Language in Video Format. *Endangered Languages Archive*, (link).

da Silva, L. A. (1996). Projeto da Norma Urbana Linguística Culta. *Linha D'Água*, (link).

Diaz, T. (2018). Documentation of Heyo [auk], a Torricelli language of Papua New Guinea. *DOBES*, (link).

Domingo, J. (2019). Tehuelche Language Collection. *Endangered Languages Archive*, (link).

Ernestus, M., Kočková-Amortová, L., and Pollak, P. (2014). The Nijmegen corpus of casual Czech. In *LREC 2014: 9th International Conference on Language Resources and Evaluation*, pages 365–370.

Fadlul, R., Mckinnon, T., Gil, D., and Taylor, B. (2016). Kerinci (Sungai Penuh) Database. *DOBES*, (link).

Garrido, J. M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., De-La-Mota, C., González, C., Vivaracho, C., Rustullet, S., Larrea, O., et al. (2013). Glissando: A corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language resources and evaluation*, 47(4):945–971.

Gil, D. (2015). A documentation of Besemah, Malayic Languages of Sumatra. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and Universitas Bung Hatta, Padang. *DOBES*, (link).

Griscom, R. (2018). Documentation of Isimjeeg Datooga. *Endangered Languages Archive*, (link).

Grzech, K. (2020). Upper Napo Kichwa: Documentation of language and culture. *Endangered Languages Archive*, (link).

Güldemann, T. and Witzlack-Makarevich, A. (2014). Text documentation of N—uu. *Endangered Languages Archive*, (link).

Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2020). MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release.

Hemmings, C. (2017). Documentation of the Kelabit Language, Sarawak, Malaysia. *Endangered Languages Archive*, (link).

Hernandez-Green, N. (2009). Documentation of San Jerónimo Acazulco Otomi, Ocoyoacac, Mexico. *Endangered Languages Archive*, (link).

Hisyam, M., Dwi Purwoko, U., and Peranginangin, D. (2013). A project of LIPI (the Indonesian Institute of Sciences) on Documenting and Revitalizing Endangered Languages and Cultures in Eastern Indonesia. *DOBES*, (link).

Hou, L. and Mesh, K. (2018). Documenting Chatino Sign Language. *DOBES*, (link).

Hunyadi, L., Váradi, T., Kovács, G., Szekrényes, I., Kiss, H., and Takács, K. (2018). Human-human, human-machine communication: On the HuComTech multimodal corpus. In *Selected Papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, pages 56–65. Linköping University Electronic Press, Linköpings universitet.

Kim, S.-U. (2018). A multi-modal documentation of Jejuan conversations. *Endangered Languages Archive*, (link).

Kopf, M., Schulder, M., and Hanke, T. (2021). Overview of Datasets for the Sign Languages of Europe. https://www.fdr.uni-hamburg.de/record/9561, July.

Kuvač Kraljević, J. and Hržica, G. (2016). Croatian adult spoken language corpus (HrAL). *FLUMINENSIA: Časopis za filološka istraživanja*, 28(2):87–102.

Lau, J. (2019). Documenting Àbèsàbèsì. *Endangered Languages Archive*, (link).

Legère, K., König, C., Heine, B., and Micheli, I. (2019). Collection Akie. *DOBES*, (link).

Leto, C., Alamudi, W. S., Himmelmann, N. P., and Riesberg, S. (2010). Collection Totoli. *DOBES*, (link).

Levinson, S. C., Casillas, M., Armstrong, W., and Torreira, F. (2019). Collection Yélî Dnye. *DOBES*, (link).

Li, Y. (2017). Documentation of Zauzou, an endangered language in China. *Endangered Languages Archive*, (link).

Lionnet, F., Hoinathy, R., and Loncke, S. (2020). Laal language documentation project. *DOBES*, (link).

Manfredi, S. (2016). Juba Creole collection in LLACAN. *OLAC resources*, (link).

Martine, B. (2012). Documentation of Ecuadorian Siona. *Endangered Languages Archive*, (link).

Martinez, P. A. (2020). Documentary Corpus of Chhitkul-Rakchham, an endangered Tibeto-Burman language of Northern India. *Endangered Languages Archive*, (link).

McDonnell, B. (2017). Documentation of Nasal: An overlooked Malayo-Polynesian isolate of southwest Sumatra. *DOBES*, (link).

Meurant, L. (2015). Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB).

Möller Nwadigo, M. (2016). A documentation project of Baa, a language of Nigeria. *Endangered Languages Archive*, (link).

Nakamura, T. and Granadillo, T. (2005). CABank Japanese CallFriend Corpus. *Talkbank*, (link).

Ogunsola, B. (2018). Documentation of Len-Mambila. *Endangered Languages Archive*, (link).

Ozerov, P. (2018). A community-driven documentation of natural discourse in Anal, an endangered Tibeto-Burman language. *Endangered Languages Archive*, (link).

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Parker, W. H. (2020). Documentation of Cora in San Juan Corapan. *Endangered Languages Archive*, (link).

Pezik, P. and Dróżdż, Ł. (2011). PELCRA Polish spoken corpus. *ELDA*, (link).

Polian, G. (2010). Tseltal Documentation Project of Gilles Polian. *Endangered Languages Archive*, (link).

Riesberg, S., Himmelmann, N. P., Walianggen, K., and Arilaha, A. (2015). Yali Summits Collection in collection "CELD Papua". *DOBES*, (link).

Rind-Pawlowski, M., Alxas, R., Babayev, N., Khvisiashvili, T., Bahddinov, A., hmdov, N., Aliyev, A., Ibrahimov, V., hmdov, A., Babayev, A., and Ağayev, H. (2016). Khinalug, A documentation project in Azerbaijan. *DOBES*, (link).

Rohleder, J. (2018). Documentation and description of Vamale, an endangered language of New Caledonia. *Endangered Languages Archive*, (link).

Schackow, D. (2014). Documentation and grammatical description of Yakkha, Nepal. *Endangered Languages Archive*, (link).

Shah, S. (2019). A multimedia corpus of siPhuthi. *Endangered Languages Archive*, (link).

Si, A. (2014). Audio and video recordings of Kune, a Bininj Gunwok dialect spoken in Buluhkaduru Outstation near Maningrida, Northern Territory. *Paradisec*, (link).

Sims, N. (2018). Documentation of Yonghe Qiang language and culture. *DOBES*, (link).

Taalunie. (2014). Corpus Gesproken Nederlands - CGN (Version 2.0.3). *Vlaamse en Nederlandse regering en NWO*, (link).

Tadmor, U. (2007). Languages of Western Borneo Documentation Project. *DOBES*, (link).

Tano, A. (2013). Documentation and description of a sign language in Côte d'Ivoire. *DOBES*, (link).

Thomas, M. (2014). Sakun (Sukur) Language Documentation. *Endangered Languages Archive*, (link).

Tisheva, Y., Dzhonova, M., and Hauge, K. R. (2013). The Corpus of Spoken Bulgarian. (link).

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3):201–212.

Unterladstetter, V., Loch, A., Morigerowsky, F., and Sawaki, Y. (2013). Collection Wooi. *DOBES*, (link).

Voß, J. (2018). Documentation and grammar of Gutob (Munda). *Endangered Languages Archive*, (link).

Vydrina, A. (2013). Description and documentation of the Kakabe language. *Endangered Languages Archive*, (link).

Wagner, J. and Maegaard, B. (2017). SamtaleBank, Danish spoken language component of the DK/CLARIN project. *Talkbank*, (link).

Wilbur, J. (2009). Pite Saami: Documenting the language and culture. *DOBES*, (link).

Williams, N. (2017). Documenting Language and Interaction in Kula. *Endangered Languages Archive*, (link).

Winkhart, B. (2016). A documentation of the remnant Baka-Gundi language Limassa. *Endangered Languages Archive*, (link).

Yim, F. B. S., Xiao, M. W., and Yiu, A. L. W. (2014). Preliminary Documentation of Macau Sign Language. *DOBES*, (link).

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., and Alikhani, M. (2021). Including Signed Languages in Natural Language Processing. *arXiv:2105.05222 [cs]*, May.

Appendices

9.1. Data access

All included datasets are available to the research community, but most do not allow direct redistribution. Further, conversational data such as this is subject to important ethical considerations (see §6.1). We provide a detailed overview of name, location, as well as authorship and usage rights information for each dataset at osf.io/cwvbe. We plan to maintain this repository of information about open datasets of conversational speech as a resource for anyone interested in compiling a dataset of similar nature or in reproducing the dataset described in this paper. Many of the datasets are distributed by language documentation archives that require the creation of a dedicated user account as well as the signing of a data use agreement. For this reason we are not able to provide direct access to the full dataset directly.

9.2. Tools for processing conversational data

Language resource platforms provide files in a range of formats, all of which need parsing to access transcription content. Language documentation corpora often come as ELAN .eaf files, an XML-based format that links utterance transcription content to timing information (with varying levels of precision). Tier structures in ELAN are subject to considerable customization by corpus creators, so they require manual inspection prior to parsing to identify the desired annotation levels. The web-based "ELAN inventory" tool provided by the ARC CoEDL is useful for providing a quick look at the tier structure of ELAN files.

Other common formats of conversational speech corpora are the Talkbank CHAT .cha format, Praat .TextGrid (Boersma and Weenink, 2013), and Exmeralda .exb (Schmidt and Wörner, 2014). This whole range of formats is usually accompanied by parsing tools for various operating systems and scripting languages. We use the Python-based `scikit-talk` which works with parsers for most formats we encountered while collecting the corpora that make up the current dataset (Liesenfeld et al., 2021).

9.3. List of languages with openly available conversational corpora

The table below presents the languages and corpora surveyed in this paper, with glottocodes and families according to Glottolog (Hammarström et al., 2021) and with citations according to source archives. Full details, including corpus statistics, sample annotations and links, are in the study repository at osf.io/cwvbe.

Language (glottocode)	Family	Citation
Akie (mosi1247)	Nilotic	(Legère et al., 2019)
Akpes (akpe1248)	Atlantic-Congo	(Lau, 2019)
Ambel (waig1244)	Austronesian	(Arnold, 2017)
Anal (anal1239)	Sino-Tibetan	(Ozerov, 2018)
Arabic (egyp1253)	Afro-Asiatic	(Canavan et al., 1997b)
Arapaho (arap1274)	Algic	(Cowell, 2010)
Baa (kwaa1262)	Atlantic-Congo	(Möller Nwadigo, 2016)
Besemah (musi1241)	Austronesian	(Gil, 2015)
Br.Portuguese (braz1246)	Indo-European	(da Silva, 1996)
Bulgarian (bulg1262)	Indo-European	(Tisheva et al., 2013)
Catalan (stan1289)	Indo-European	(Garrido et al., 2013)
Chitkuli (chit1279)	Sino-Tibetan	(Martinez, 2020)
Cora (sant1424)	Uto-Aztecan	(Parker, 2020)
Croatian (croa1245)	Indo-European	(Kuvač Kraljević and Hržica, 2016)
Czech (czec1258)	Indo-European	(Ernestus et al., 2014)
Danish (dani1285)	Indo-European	(Wagner and Maegaard, 2017)
Datooga (isim1234)	Nilotic	(Griscom, 2018)
Dutch (dutc1256)	Indo-European	(Taalunie, 2014)
English (nort3314)	Indo-European	(Canavan and Zipperlen, 1996a)
Farsi (west2369)	Indo-European	(Canavan et al., 2014)
French (stan1290)	Indo-European	(Torreira et al., 2010)
German (stan1295)	Indo-European	(Canavan et al., 1997a)
Gunwinguu (gunw1252)	Gunwinyguan	(Si, 2014)
Gutob (bodo1267)	Austroasiatic	(Voß, 2018)
Hausa (haus1257)	Afro-Asiatic	(Caron, 2016)
Heyo (heyo1240)	Nuclear Torricelli	(Diaz, 2018)
Hungarian (hung1274)	Uralic	(Hunyadi et al., 2018)
Japanese (nucl1643)	Japonic	(Nakamura and Granadillo, 2005)
Jejueo (jeju1234)	Koreanic	(Kim, 2018)
Juba Creole (suda1237)	Afro-Asiatic	(Manfredi, 2016)

Kakabe (kaka1265)	Mande	(Vydrina, 2013)
Kelabit (kela1258)	Austronesian	(Hemmings, 2017)
Kerinci (keri1250)	Austronesian	(Fadlul et al., 2016)
Khinalug (khin1240)	Nakh-Daghestanian	(Rind-Pawlowski et al., 2016)
Kichwa (tena1240)	Quechuan	(Grzech, 2020)
Korean (kore1280)	Koreanic	(Canavan and Zipperlen, 1996b)
Kula (kula1280)	Timor-Alor-Pantar	(Williams, 2017)
Laal (laal1242)	Laal	(Lionnet et al., 2020)
Limassa (lima1246)	Atlantic-Congo	(Winkhart, 2016)
Mambila (came1252)	Atlantic-Congo	(Ogunsola, 2018)
Mandarin (mand1415)	Sino-Tibetan	(Canavan and Zipperlen, 1996c)
Minderico (mind1263)	Indo-European	(Carvalho Ferreira et al., 2011)
N—uu (nuuu1241)	Tuu	(Güldemann and Witzlack-Makarevich, 2014)
Nahuatl (cent2132)	Uto-Aztecan	(Amith et al., 2009)
Nasal (nasa1239)	Austronesian	(McDonnell, 2017)
Nganasan (ngan1291)	Uralic	(Brykina et al., 2018)
Otomi (esta1236)	Otomanguean	(Hernandez-Green, 2009)
Pagu (pagu1249)	North Halmahera	(Hisyam et al., 2013)
Polish (poli1260)	Indo-European	(Pezik and Dróżdż, 2011)
S.Qiang (sout2728)	Sino-Tibetan	(Sims, 2018)
Saami (pite1240)	Uralic	(Wilbur, 2009)
Sakun (suku1272)	Afro-Asiatic	(Thomas, 2014)
Sambas (kend1254)	Austronesian	(Tadmor, 2007)
Siona (sion1247)	Tucanoan	(Martine, 2012)
Siputhi (swat1243)	Atlantic-Congo	(Shah, 2019)
Spanish (stan1288)	Indo-European	(Canavan and Zipperlen, 1996d)
Tehuelche (tehu1242)	Chonan	(Domingo, 2019)
Totoli (toto1304)	Austronesian	(Leto et al., 2010)
Tseltal (tzel1254)	Mayan	(Polian, 2010)
Ulwa (ulwa1239)	Misumalpan	(Barlow, 2017)
Vamale (vama1243)	Austronesian	(Rohleder, 2018)
Wooi (woii1237)	Austronesian	(Unterladstetter et al., 2013)
Yakkha (yakk1236)	Sino-Tibetan	(Schackow, 2014)
Yali (pass1247)	Nuclear Trans New Guinea	(Riesberg et al., 2015)
Yélî Dnye (yele1255)	Yele	(Levinson et al., 2019)
Zaar (saya1246)	Afro-Asiatic	(Caron et al., 2014)
Zauzou (zauz1238)	Sino-Tibetan	(Li, 2017)

A note on sign languages. The table above includes only spoken languages. While we have sampled as broadly as possible, sign language corpora of conversation are still quite rare (Kopf et al., 2021; Yin et al., 2021), and the handful that are openly available are primarily organized in terms of sign-level rather than turn-level annotations. This holds for Chatino Sign Language (Hou and Mesh, 2018), Côte d'Ivoire Sign Language (Tano, 2013), French Belgian Sign Language (Meurant, 2015), German Sign Language (Hanke et al., 2020), and Macau Sign Language (Yim et al., 2014). Although there is ample evidence that sign language conversations are also turn-organized (de Vos et al., 2021), the sign-level annotations mean that additional processing steps would be required to render such corpora interoperable with the minimal data format specifications proposed here. We think this is best done in consultation with corpus compilers and language experts.