From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology

Mark Dingemanse *
Center for Language Studies
Radboud University
mark.dingemanse@ru.nl

Andreas Liesenfeld *
Center for Language Studies
Radboud University
andreas.liesenfeld@ru.nl

Abstract

Informal social interaction is the primordial home of human language. Linguistically diverse conversational corpora are an important and largely untapped resource for computational linguistics and language technology. Through the efforts of a worldwide language documentation movement, such corpora are increasingly becoming available. We show how interactional data from 63 languages (26 families) harbours insights about turn-taking, timing, sequential structure and social action, with implications for language technology, natural language understanding, and the design of conversational interfaces. Harnessing linguistically diverse conversational corpora will provide the empirical foundations for flexible, localizable, humane language technologies of the future.

1 The natural habitat of language

The primary ecology of natural language is in real-life episodes of human interaction. This is where people learn language and where they use it to coordinate joint actions, build social relations, and exchange information (Schieffelin and Ochs, 1986; Schegloff, 2006). In contrast, when machines encounter language, it tends to be radically divorced from this habitat and reduced to large amounts of decontextualised non-interactive text (Bender and Koller, 2020; Marge et al., 2022). Natural languages are also characterized by diversity at many levels, from sound and sign systems to syntax and semantics (Nettle, 1999; Evans and Levinson, 2009). In contrast, the language samples that inform language technology tend to be limited to a handful of well-resourced languages, representing only a tiny sliver of the world's linguistic diversity (Blasi et al., 2021; Joshi et al., 2020).

The time is ripe for language technology to benefit from linguistically diverse interactional data. Insights from such data can strengthen the empirical foundations of language technology, help break down the hegemony of the resourceful few, and provide room for linguistic diversity in localized applications (Bird, 2020; Danielescu and Christian, 2018). Today there is a growing set of conversational corpora of diverse languages, thanks in large part to important primary work on language documentation and description (Seifart et al., 2018). We argue such corpora represent an important and mostly untapped resource for language technology.

Corpus size is often seen as a challenge, but data comes in levels of granularity. A well-curated corpus amounting to an hour of lively conversation may not contain enough text to train a language model. But it does provide thousands of conversational turns organized in larger sequential structures of social action, along with fine details about timing, participation and linguistic structure. Since conversational corpora are one of the few places where we can study language in a way that approaches its natural habitat, collectively, these corpora harbour important insights about human interactional infrastructure.

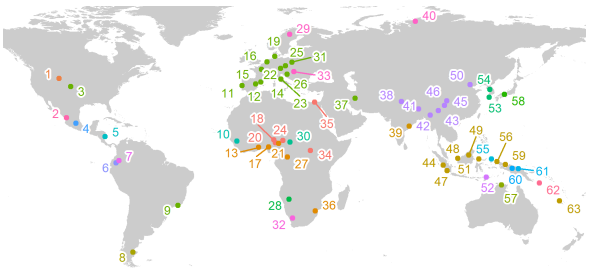

Figure 1: [1] Arapaho [2] Cora [3] English [4] Otomi [5] Ulwa [6] Kichwa [7] Siona [8] Tehuelche [9] Br. Portuguese [10] Kakabe [11] Minderico [12] Spanish [13] Siwu [14] Catalan [15] French [16] Dutch [17] Akpes [18] Hausa [19] Danish [20] Zaar [21] Baa [22] German [23] Italian [24] Sakun [25] Czech [26] Croatian [27] Limassa [28] ǂAkhoe [29] Saami [30] Laal [31] Polish [32] N|uu [33] Hungarian [34] Juba Creole [35] Arabic [36] Siputhi [37] Farsi [38] Chitkuli [39] Gutob [40] Nganasan [41] Yakkha [42] Anal [43] Zauzou [44] Kerinci [45] Duoxu [46] S. Qiang [47] Nasal [48] Sambas [49] Kelabit [50] Mandarin [51] Totoli [52] Kula [53] Jejueo [54] Korean [55] Pagu [56] Ambel [57] Gunwinggu [58] Japanese [59] Wooi [60] Yali [61] Heyo [62] Yélî Dnye [63] Vamale.

*Both authors contributed equally.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5614 - 5633

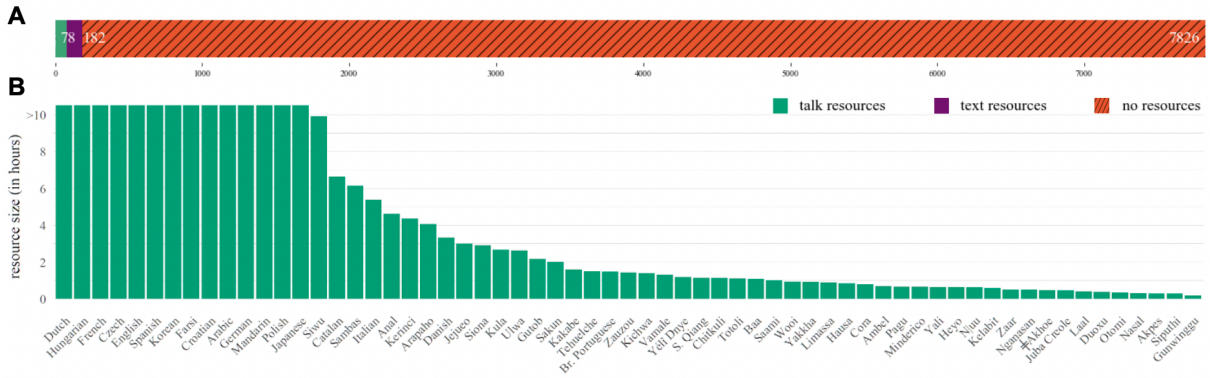

Figure 2: Language resources (corpora) and their size in relation to global language diversity. **A**: In red: total number of L1 languages in the world (estimate based on spoken and signed L1 languages in Glottolog 4.4 (Hammarström et al., 2021)); In purple: estimated languages with available textual corpora, based on number of languages used on the internet (W3techs survey; see also Pimienta et al. (2009)); In green: estimate of languages with conversational corpora made available through research-oriented language resource platforms and language documentation projects (not including scripted data like subtitles). **B**: Languages and dataset size (in hours of talk) of the subset of 63 languages used in this study. See Figure 1 for a map and Appendix B for details.

Recent work has shown the dire state of language resources in relation to linguistic diversity (Blasi et al., 2021; Joshi et al., 2020), and pointed to ways forward to increase the empirical coverage of language typology and technology (Asgari and Schütze, 2017; Bjerva and Augenstein, 2018; Deri and Knight, 2016; Duong et al., 2015; Levow et al., 2021). Work in distributional and corpus-based typology is showing how to analyse linguistic information available in text corpora (Ponti et al., 2019; Seifart et al., 2021; Levshina, 2021).

Compared to text corpora, conversational corpora are much harder to collect, annotate and transcribe, and as a result they represent a much smaller subset of data. However, we think there is reason for cautious optimism. In this paper we present a first foray into this domain. We collate conversational corpora made available for research purposes and find there is now data available for a wide range of languages, many of them not the usual suspects of NLP research. Besides well-known resources like TalkBank and the Linguistic Data Consortium, here we highlight the potential of corpora collected and archived as part of language documentation projects around the world (see Appendix B).

Our focus is specifically on corpora of informal conversations among co-present participants, transcribed and time-aligned at the level of conversational turns. Details of our curation and analysis pipeline are described in the Appendix and in Liesenfeld & Dingemanse (2022). While it is impossible to exhaustively list or estimate the size of extant conversational corpora, the quality-controlled subset we consider here represents 63 languages from 26 language families (Figure 1), and amounts to over 800 hours of talk produced by over 11.000 partipants, segmented into over 1.6 million turns (9.3 million words) (Figure 2).

In what follows, we examine aspects of this collection with scientific and technological applications in mind. In doing so, we aim to contribute towards a move from the most represented to a more representative sample of the world's languages, and to show how the study of human interaction can yield insights of relevance to linguistics, language technology and human-computer interaction.

2 From text to talk-in-interaction

Despite recent advances in speech and dialogue modelling, to date, no machine can "lead a half-decent coherent conversation with a human" (Kopp and Krämer, 2021). There are several reasons for this, including the need for complex cognitive skills like intention attribution and incremental common ground construction, but equally important is a dearth of data and domain knowledge: modern natural language processing predominantly deals with text, not talk.

As a simple illustration of the difference, compare frequency distributions of words and phrases in corpora of talk versus text in English, an Indo-European language (Figure 3). The forms most characteristic of talk are interactive interjections like *hm, uhhuh, um, yeah, okay*. Items like this

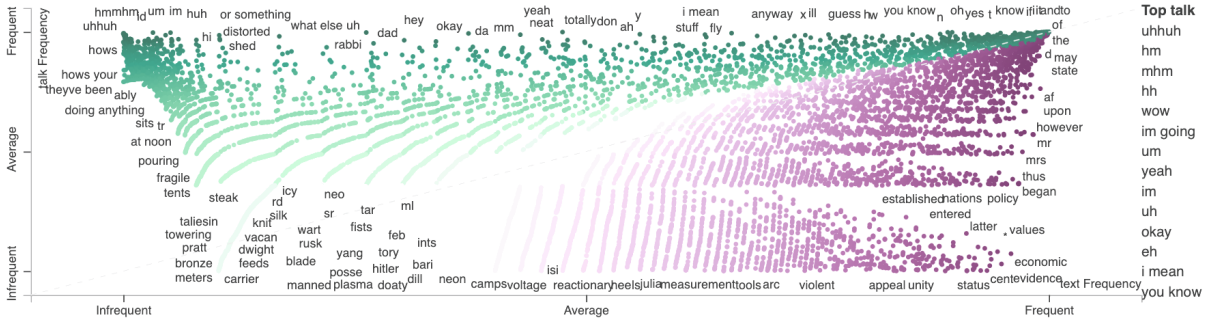

Figure 3: Words and phrases characteristic of spoken interaction (green) versus written text (purple) in English, with words most characteristic of conversational interaction in the upper left. Source data: 55k random-sampled turns and sentences from corpora of English text (Francis, 1965) and talk (Canavan and Zipperlen, 1996a; Canavan et al., 1997a), based on Scaled F score metric, plotted using *scattertext* (Kessler, 2017).

streamline conversation, calibrate mutual understanding and coordinate joint action (Clark, 1996; Bavelas et al., 2000). Yet it is precisely such items that are woefully underrepresented in the data underlying most current language models (Prevot et al., 2019). It is little surprise that conversational agents have a hard time dealing with informal conversational style (Hoegen et al., 2019) and building social bonds (Cassell, 2020), and that speech recognition easily mixes up interjections with opposite pragmatic functions (Zayats et al., 2019) if it doesn't miss them altogether (Cumbal et al., 2021).

Proposed solutions to such challenges involve imparting agents with domain-specific interactional knowledge like keyword-based scripted conversational routines and domain knowledge from Q&A databases (Bocklisch et al., 2017; Dinan et al., 2019) or with capacities for feedback generation (Oertel et al., 2016) or common ground reasoning (Kopp and Krämer, 2021). Here we propose a complementary approach: pay closer attention to how language is used in informal everyday interaction around the world. We believe this is important because just like natural language processing has long been limited to monologic texts, the science of human interaction has for the most part been based just on English and a small number of similarly well-resourced languages (Henrich et al., 2010).

If language technology is to be maximally scalable, localizable and usable, it will greatly benefit from broadening its empirical base towards more interactive data from a wider range of languages. Such data can improve our understanding of interactional infrastructure and can help us chart both language-specific routines and pragmatic universals of interaction.

3 From strings to social actions

Utterances are not just strings with probability distributions defined over them; they stand in relation to other turns, with which they form structured sequences and implement social actions. A key element of this is a socially sanctioned turn-taking system by which participants self-organize the distribution of turns over participants (Sacks et al., 1974). Foundational work on English showed that participants appear to avoid both gaps and overlaps, often achieving speaker transition in as little as 200 ms. This temporal organization is so tight that it has long puzzled psycholinguists, who observe that even planning a simple sentence in isolation may take up to 600ms, implying that language comprehension and production must run in parallel (Levinson, 2016). Indeed participants do not wait for pauses to begin their contribution, but instead start planning early, continuously weighing a range of cues to determine the likely point at which the current turn ends (de Ruiter et al., 2006).

Subsequent cross-linguistic work has confirmed this no-gap-no-overlap goal, showing that across 10 languages from 7 language families, floor transfers are usually achieved between 0 and 200ms, with language-specific means falling within 250ms on either side of the mean (Stivers et al., 2009). Currently available data allows us to replicate this in 24 languages from 12 unrelated families, more than doubling the sample size. Because our aim is to characterize the overall temporal features of quotidian interaction, we consider all turn transitions in dyadic stretches of conversation (see Appendix A.1 for a validation in question-answer sequences).

In the 24 corpora that contain at least 1000 dyadic turn transitions, we find substantially the

same finely calibrated temporal distribution of turns, suggesting that participants aim for a no-gap, no-overlap target, with the bulk of language-specific means falling within a relatively narrow bandwidth of variation (Figure 4). In the full set of 674 223 transitions, 46% of turns are produced in slight terminal overlap. This includes both fuller turns and short responsive tokens (Goodwin, 1986; Corps et al., 2022), and underlines the extent to which human interaction everywhere involves a braiding of successive and concurrent moves.

The implications for language and speech technology are considerable (Skantze, 2021; Roddy, 2021). It means that social robots that switch between listen and talk states will be behind the curve approximately half of the time: perceived as responding too slowly or switching to a listening state too late to pick up early and concurrent responses. If the aim is to facilitate fluid interaction, a first challenge is to achieve the rapid transitions that characterize human language use. This requires incremental and continuous processing (Levinson, 2016; Pitsch, 2016), representing a radical departure from classic reactive spoken dialog systems. Most work in this area is still based on English, potentially jeopardizing the generalizability of findings. Cross-linguistic conversational corpora will prove crucial to identify the most robust prosodic, lexical and interactional features that can inform continuous projections of transition relevance places (Ward et al., 2018; Roddy et al., 2018).

Even if rapid transitions may be achieved with the help of continuous, context-sensitive processing, a further layer of language-specific calibration will be necessary to account for the known range of variation (Stivers et al., 2009). Experimental work in this domain shows measurable intercultural differences in orientations to inter-turn silences (Roberts et al., 2011): across cultures, people treat gaps as meaningful beyond a threshold of a few hundred milliseconds, but the exact threshold varies with culture. Without calibration of this kind, people may easily experience conversational agents as overeager, stilted, or out of sync. Progress in this domain may hinge on endowing interactive technologies with a sense for timing and rhythm (Yu et al., 2021; Pouw et al., 2021).

An important aspect of human conversation is how rapid turn-taking enables on-the-fly calibration and coordination. This motivates a theoretical turn from singular, perfectly formulated, unambigous utterances to incremental, good enough, co-constructed understanding (Dingemanse et al., 2015; Albert and de Ruiter, 2018; van Arkel et al., 2020). Increasingly, parsers and other models of grammar and dialogue incorporate this kind of incremental perspective (Schlangen and Skantze, 2011; Vanzo et al., 2018; Buschmeier and Kopp, 2018). Promising application-oriented work in this direction exists (Ekstedt and Skantze, 2020;

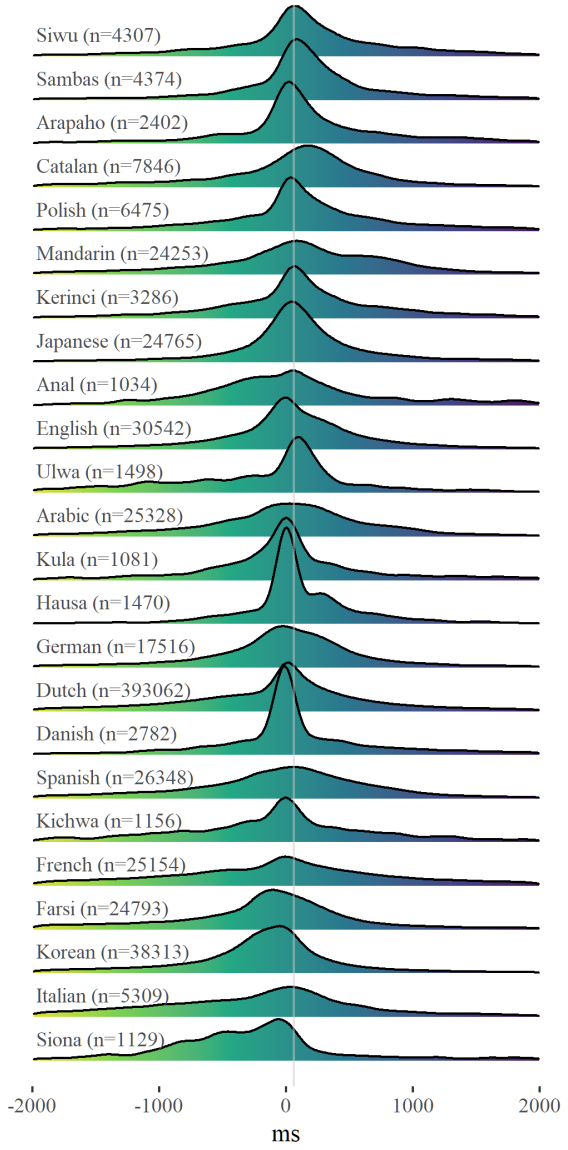

Figure 4: The timing of turn transitions in dyadic interactions in 24 languages around the world, replicating earlier findings and extending the evidence for the interplay of universals and cultural variation in turn-taking (n = number of turn transitions per corpus). Positive values represent gaps between turns; negative values represent overlaps. Across languages, the mean transition time is 59ms, and 46% of turns are produced in (slight) terminal overlap with a prior turn.

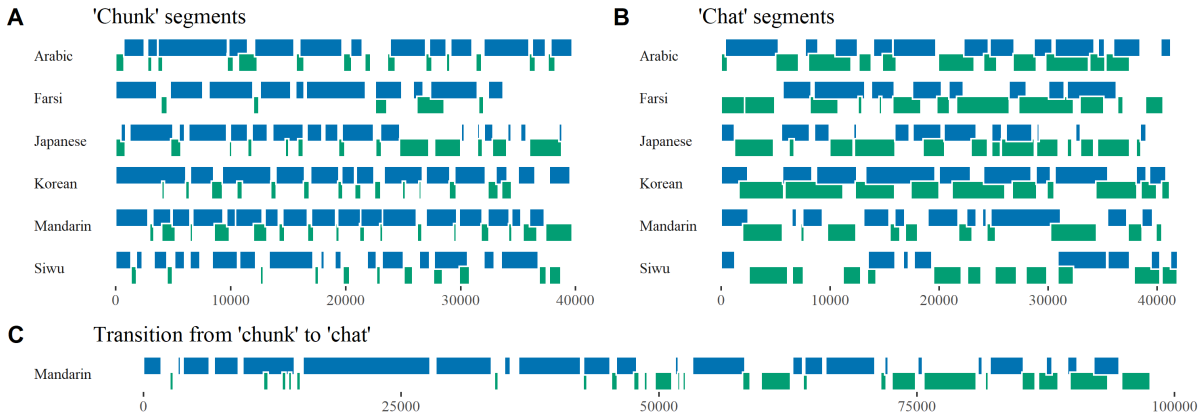

Figure 5: Two types of conversational activity in 6 unrelated languages, showing the viability of identifying broad activity types using ebbs and flows in amount of talk contributed (time in ms). **A**. Tellings ('chunks') are characterized by highly skewed relative contributions, with one participant serving as teller and the other taking on a recipient role (roles may switch, as in the Japanese example). **B**. In 'chat' segments, turns and speaking time are distributed more evenly. **C**. Shifts from one state to another are interactionally managed by participants.

Skantze, 2017), though two critical challenges remain: (i) text corpora of asynchronous interaction are much less piecemeal and incremental than co-present interaction, and (ii) the interactional disruptiveness of timing discrepancies can be masked by the flexibility of human participants, who soon learn to revert to simpler forms of robot-directed talk (Suchman, 2007; Seibt, 2017).

4 Keeping unity and diversity in sight

While text corpora are sometimes treated as shapeless collections of strings, conversational data is not flat but richly structured (Goodwin, 1981; Couper-Kuhlen and Selting, 2017). Each turn at talk builds on what came before and shapes the possibility space of what comes next (Firth, 1935; Heritage, 1984). Conversation analysts call this *sequence organization* (Schegloff, 2007), and cross-linguistic work has uncovered a number of basic sequential positions along with slots for inserts and expansions (Kendrick et al., 2020). Sequences are one of the major tools for organizing social action.

Studying conversational sequences across diverse languages poses considerable challenges, because it requires access not just to form but also to social action or intent (Bender and Koller, 2020). While annotated corpora of dialog acts (Jurafsky et al., 1998) are available for a small number of well-resourced languages (Bunt et al., 2020), they invite an overly categorical view of what is in fact fluid and emergent action ascription. The open-endedness of social actions in casual conversation (Levinson, 2013) places severe constraints on the utility of slot-filling approaches (Papaioannou et al., 2018), which have their origin in narrow task-oriented interactions (Liu et al., 2021).

Here we probe conversational sequencing by starting from coarse-grained but robust structural facts about the relative distribution of turns and talk. Casual interaction often combine lively spates of equitable exchange with more lopsided moments such as tellings in which one participant secures the floor and the other assumes a recipient role (Schegloff, 1982; Goodwin, 1995). Some work on English has captured this as *chat* versus *chunk*, where a chunk is defined as 'a segment where one speaker takes the floor and is allowed to dominate the conversation for an extended period' (Eggins and Slade, 2004; Gilmartin et al., 2018). Using a measure of relative skew in contributions in a moving 10 second window, we can identify stretches corresponding to such a distinction, as well as transitions from one state to another, across unrelated languages (Figure 5A-C).

Knowing about such states and transitions between them is of great relevance to language technology and dialog systems. For instance, the relative predictability of responses differs strongly across states (Gilmartin, 2021). Our results suggest that it is possible to reliably identify at least some broad activity types across languages, opening up possibilities for investigating the linguistic resources that characterize them, and the ways in which people transition between them. The notions of 'chat' and 'chunk' should not be reified, but the distinction points to a data-driven way to get ana-

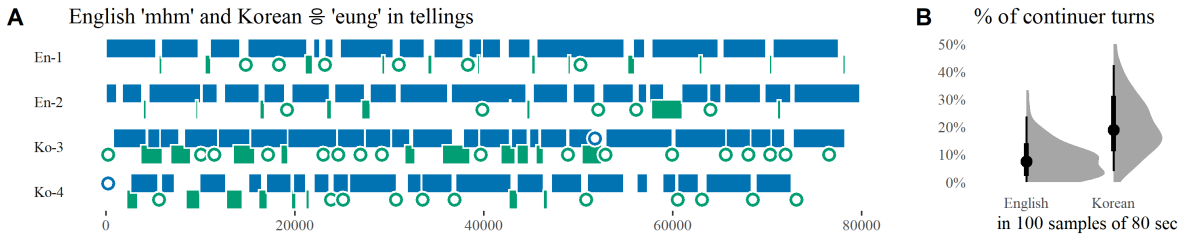

Figure 6: **A.** Continuers (marked ○) are among the most frequent recipient behaviours in tellings ('chunks') in both English and Korean, shown here in two 80 second segments each with a strong skew in contributions. **B.** However, their relative frequency is about twice as high in Korean based on 100 random samples of 80 second segments in both languages: on average, 21% of turns are continuers in Korean, against 9% of turns in English. Segments were sampled from 53 (55) distinct conversations involving 106 (95) distinct speakers in Korean (English). Measures are expressed relative to turns instead of time to control for speech rate differences.

lytical grip on structural features of activity types in conversation (Levinson, 1979).

Work on English has found that tellings can be recognized not just by their skewed division of labour, but also by the use of continuers like *mhm* (Howes and Eshghi, 2021; Schegloff, 1982) at places where turn transition would be relevant. We find that this is the case in languages in our sample too, so a simple conclusion could be that we have found a way to unearth universal aspects of tellings, or 'chunks', with possible implications for the design of, say, dialog systems sensitive to the interactional achievement of dialog states.

However, on closer look, the data also provides reason to take linguistic diversity seriously. Figure 6A zooms in on four longer stretches of conversation in English and Korean. Here, circles highlight the use of the most frequent continuer in the language, which is 'mhm' in English and 응 'eung' in Korean. What is already apparent in the four conversations shown in panel A is also borne out in a quantitative analysis of 100 random samples of 80 second stretches of English and Korean conversations: while 8% of turns are continuers in English, this is 21% in Korean (Figure 6B). This higher frequency also comes with higher susceptibility to overlap: whereas in English, 39% of continuer tokens occurs in full or partial overlap, in Korean this is 73%. The difference does not appear to be reducible to transcription conventions; for instance, in both corpora, continuers repeated in quick succession are transcribed as a distinct format (*mhm mhm*, 응응 *eung eung*) and excluded from these counts; and in both corpora, the average number of words per turn lies around 6 (Korean: 5.7; English: 6.9) and the average number of turns per 10 second window is 5.6 (Korean: 5.6; English: 5.6).

One implication of this is that continuers are apparently relevant at more points during interaction in Korean than in English (Kim, 1999), which has consequences for the design of dialog systems, incremental parsers and conversational agents. For instance, a conversational agent in Korean might have to issue more displays of recipiency and should be prepared to deal with incoming feedback at a higher pace; in the same context, an agent calibrated to English might need different conversation design. The observed variation is extreme enough to warrant a critical look at the notion of feedback relevance spaces (Howes and Eshghi, 2021): perhaps this notion needs to be relativized to cover attested cross-linguistic diversity, as has been suggested in qualitative conversation analytic research (White, 1989; Clancy et al., 1996; Young and Lee, 2004).

We have touched here only on some coarse-grained aspects of sequential structure by way of demonstrating the utility of conversational corpora representing diverse languages. Plenty of other phenomena are ripe for similar treatment.

5 Interactional tools

A key finding of linguistics going back to Estoup and Zipf (Estoup, 1917; Zipf, 1935) is that a small number of items tends to be used for a large amount of work. Power law distributions are ubiquitous in linguistic data and well-studied across a range of languages (see Piantadosi 2014 for review). Most analyses in this line of work tokenize textual data on the basis of the observation that sentences are built out of reusable elements. For such tokenised items (roughly, 'words'), we have come to expect the rank-frequency distribution to look linear on a log/log scale. Yet language does not come in stray words, but in *turns* at talk: communicative moves

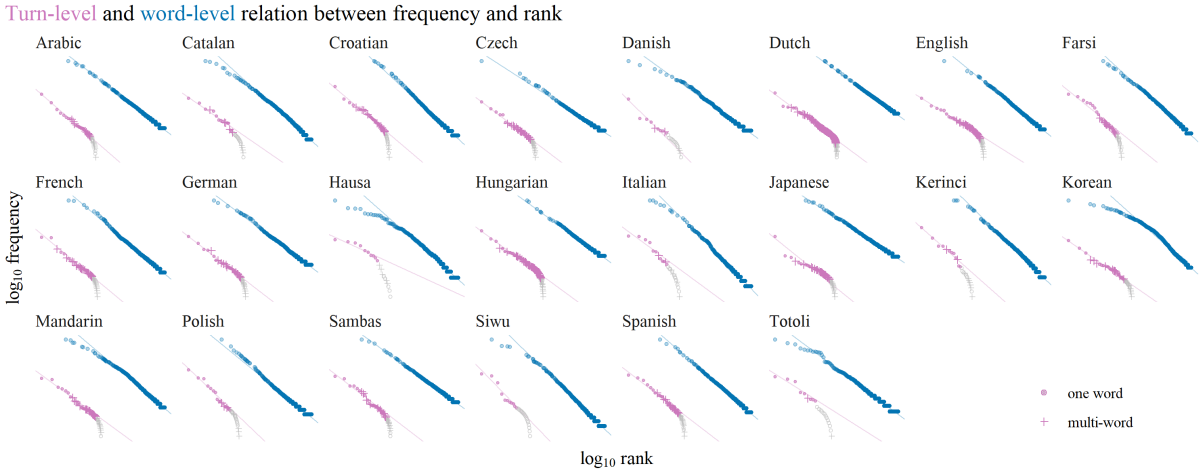

Figure 7: Frequency/rank distributions of tokenized items ('words') and recurring turn formats in conversational corpora with at least 20 such turn formats, representing 22 languages (8 phyla). Tokenized items (blue) show a linear frequency/rank relation in log/log space. Recurring turn formats (whether one-word ○ or multi-word +) appear to obey a similar frequency/rank distribution for the 20% of turns that occur >20 times (purple), tapering off towards lower frequencies and unique turns (grey). Fit fluctuates with corpus size and the parallelism of distributions is most apparent in larger corpora. Data for 21 smaller corpora (12 further phyla) in Appendix A.2.

of varying complexity and conventionality.

Since communicative turns are rarely studied as holistic units, it is an open question to what extent they may or may not show evidence of linguistic laws. Such an organization may seem prima facie unlikely: after all, we know we build complex turns out of simpler elements like words and phrases, and the unlimited expressive power generated by this compositionality is rightly celebrated as one of the hallmarks of human language (Hockett, 1960). On the other hand, as Firth (1935) noted, "Conversation is much more of a roughly prescribed ritual than most people think". Indeed a look at conversational data shows that many turns are not one-offs: at least 28% of the utterances in our sample (436 367 out of 1 532 915 across 63 languages) occur more than once, and over 21% (329 548) occur more than 20 times. Many of these recurring turn formats are interjections and other pragmatic devices that help manage the flow of interaction and calibrate understanding (Yngve, 1970; Jefferson, 1985; Allwood et al., 1990; Ward, 2006; Norrick, 2009). The ubiquity and communicative importance of these items opens up the possibility of power law-like distributions at turn level for some subset of turns.

Here we compare rank-frequency distributions of tokenized items and standalone turn formats in the subset of 22 languages with conversational corpora large enough to feature at least 20 recurring standalone turn formats (Figure 7). We find that tokenized items, as expected, reproduce some well-known structural properties of rank-frequency distributions, including their linear nature on a log-log plot and a systematic deviation from this linearity for the highest frequency (lowest rank) words. For standalone turns, distributions trail off sharply towards the lowest frequencies, reflective of the creative and compositional nature of many utterances. However, the considerable subset of recurring turn formats (Figure 5, purple) may also suggest a partial power law distribution: though the data is sparser, a log-log line fitted to the 20% of turns used at least 20 times has a comparable slope in most corpora.

The result cannot simply be reduced to the fact that standalone turns are drawn from the larger population of single words. Recurrent turn formats tend to have specialized discourse-level functions, and while many are single words like 'm-hm', 'huh?' or 'oh', one out of three are multi-word expressions like English 'but um', Japanese あそうなんだ *a soo nanda* 'oh really' or Hungarian *nem tudom* 'I dunno'. If such recurring formats obey a power law distribution, this provides novel, interactionally motivated evidence in support of the claim that the phrase rather than the word may be a privileged locus for Zipf's law of frequency (Ryland Williams et al., 2015). In this context it is worth recalling that Zipf motivated his observa-

tions in terms of tools-for-jobs (Zipf, 1949). Just as the tools of artisans are constructed and arranged in ways that support efficient use, so the tools of language are organized to optimally carry out their jobs. In this sense, we can speak of recurring turn formats as *interactional tools*.

Even if interactional tools make up a significant proportion of turns in any exchange (as we saw in §4, continuers alone may account for 10 to 20% of turns at talk), they are easily obscured by premature tokenization or erased by seemingly innocuous procedures like stopword removal. And yet it is precisely these interactional tools that may prove essential to understanding and modelling interactional infrastructure within and across languages. Getting at these tools and charting their universality and variability represents a key goal for human language technologies.

Overlooking interactional tools and the details of their deployment comes with immediate adverse consequences. A recent user study reported that a significant number of participants ran into interactional turbulence and overlap when interacting with a neural conversational agent through an English-based voice user interface (Hoegen et al., 2019). The turbulence was traced to the agent making segmentation errors and responding to every single utterance detected. This in turn made it harder for human participants to predict when the agent was done speaking, leading to cascades of overlap and confusion. The study proposed two solutions to deal with this (casting the interactional scuffles as situations to be avoided rather than as the rapid and flexible recalibrations they represent in human interaction). The first is to return the floor to a participant as soon as overlap is detected. This seems to assume that any vocalization by a participant is an attempt to take the floor (rather than, say, a minimal display of understanding-so-far). The second proposal is to "filter out stop words and interjections from the participant" on the grounds that the agent responding to these can confuse participants, "since people often do not even realize they are using stop words or are interjecting" (p. 117).

However, people do not produce interjections stochastically, but wield them as interactional tools in the service of calibrating mutual understanding and coordinating joint action (Dingemanse, 2017). A continuer like *mhm* shows understanding, while a repair initiator like *huh?* requests clarification. Indiscriminately filtering out such utterances robs conversational agents of direct access to public displays of understanding and misunderstanding. It also robs people of the very tools they use to co-construct interdependence and understanding, and therefore of a significant part of their linguistic agency. Filtering out interjections to avoid interactional turbulence is like removing all pedestrian crossings to deal with self-driving cars crashing into people. The result may be an incident-free zone, but at significant cost to human flexibility and agency (Illich, 1973).

More work is needed to explore the distributional properties of recurring turn formats, but at least we can conclude that every corpus in our dataset has a subset of recurrent turn formats with metacommunicative functions whose organization suggests a power-law distribution. Their importance in human interaction and by extension human-computer interfaces can hardly be overstated. To build flexible conversational agents (Buschmeier and Kopp, 2018) and localizable conversational interfaces (AbuShawar and Atwell, 2016), we need a solid grip both on possibly universal aspects as well as on the full range of cross-linguistic diversity.

6 Ways forward

Recent work has argued that text-based stochastic models may be running into dimishing returns (Bender and Koller, 2020), has stressed the dearth of relevant conversational data (Gilmartin, 2021), and has pointed to formidable challenges in the creation of truly interactive systems (Marge et al., 2022). Progress will come from multiple fronts, but careful and mindful data curation must be a fundamental part of it (Rogers, 2021). This requires a reconceptualization not just of what counts as "NLP work", but also of what counts as data. Here we have shown how linguistically diverse corpora of co-present conversation may contribute to such a reconceptualization. Now is the time to pivot from text to talk; for few things other than the careful study of interactive language use can bring us closer to an understanding of how language augments human cognition and supports fluid and flexible action coordination. This understanding, in turn, will be critical to make meaningful progress in any domain that involves human language technologies and interactive interfaces.

Fortunately there are good ways forward. Here we summarise three principles to foster a robust and diversity-aware science of human interaction

that can underpin engineering solutions, inform language models, and contribute to human-centered applications:

1. *Maximise ecological validity.* To understand and model human interaction, start from rich data that is as close as possible to the natural habitat of language: co-present social interaction. Audio and video corpora of informal conversation are increasingly available for many languages and provide an excellent starting point. What such corpora may lack in breadth they make up for in depth: terabytes of text cannot replace the intricacies of multimodal communication and fluid participation.

2. *Represent interactional infrastructure.* Fine-grained temporal organization, radical interdependency and emergent social action are characteristics of human interaction that cannot be reduced to stochastic properties of text. The timing, co-construction and sequential positioning of turns is as consequential to their meaning and interpretation as their form. The complex and socially distributed nature of sequence organization exceeds the powers of slot filling approaches and requires renewed attention to interactional tools: the metacommunicative resources people use to construct and calibrate mutual understanding on the fly.

3. *Design for diversity.* To escape the reign of the resourceful few, use linguistically diverse data and anticipate a combination of universal and language-specific design principles. This not only ensures broad empirical coverage and enables new discoveries; it also benefits diversity and inclusion, as it enables language technology development that serves the needs of diverse communities.

Our aim in this position paper has been to sketch how these principles, fuelled by insights from the study of dialogue, linguistic typology, conversation analysis, and a range of other fields, can provide the conceptual foundations for novel work on human language technologies and human interaction.

7 Conclusions

Cross-linguistically diverse corpora of conversation are increasingly available and can help us to better understand basic interactional patterns and build more flexible, context-sensitive language technologies. For this to work, it is important to keep both linguistic diversity and potential universals in sight (Sidnell and Enfield, 2012; Enfield et al., 2013). We cannot assume that a given piece of interactional infrastructure is universal just based on a handful of languages. Encouragingly, our results suggest that even relatively small corpora can support robust generalizations about key aspects of interactional infrastructure.

One reason this matters is *empirical grounding*. Cross-linguistic and comparative work on human interaction has barely started (Floyd, 2021; Ameka and Terkourafi, 2019). There may be more universals of interaction; equally likely is that there are more patterns of unrecognized diversity. Both types of outcomes are important for how they shed light on the structure of human interaction, and both have implications for language technology and human-computer interfaces. More fundamental work in pragmatic typology is needed — and computational approaches to low-resource languages provide a promising starting point.

But an equally important reason to consider linguistic diversity in language technology and natural language processing is one of *linguistic agency* (Di Paolo et al., 2018; Nguyen et al., 2016; Suchman, 2020). Designing interfaces that allow people to flexibly wield their preferred communicative resources lessens the hegemony of any one language and makes technology more inclusive, more humane and more convivial for a larger range of possible users (Munn, 2018; Voinea, 2018). Localizing user interface elements is only a first step; diversity in how and when basic interactional structures are deployed must ultimately be reflected in the design of conversational user interfaces.

In the rush for better language technology we should avoid being driven into the arms of only the best-resourced languages and the easiest-to-get data. We need language models that are representative of the actual ways in which people use language, and conversational interfaces that give people the feeling they do not have to leave their own linguistic identities at the door. Comparative and computational work on conversational corpora from a wide range of languages is crucial to develop a strong foundational understanding of universals and diversity in interactional infrastructure, and to ensure we can build the humane and diversity-aware language technologies of the future.

Acknowledgements

We thank Calle Börstell, Riccardo Fusaroli, Wim Pouw, Marlou Rasenberg and Marieke Woensdregt for helpful comments. Funding for the work reported here comes from Dutch Research Council grant NWO 016.vidi.185.205 to MD.

References

Bayan AbuShawar and Eric Atwell. 2016. Usefulness, localizability, humanness, and language-benefit: Additional evaluation criteria for natural language dialogue systems. *International Journal of Speech Technology*, 19(2):373–383.

Saul Albert and J. P. de Ruiter. 2018. Improving Human Interaction Research through Ecological Grounding. *Collabra: Psychology*, 4(1).

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

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

Laura Arnold. 2017. The documentation of Ambel, an Austronesian language of Eastern Indonesia. http://hdl.handle.net/2196/00-0000-0000-000C-E849-2.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124.

Russell Barlow. 2017. Documentation of Ulwa, an endangered language of Papua New Guinea. https://hdl.handle.net/1839/b1796725-1a49-48ee-93ea-75e5b440c7bc.

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender and Alexander Koller. 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. Association for Computational Linguistics.

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

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

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic Inequalities in Language Technology Performance across the World's Languages. *arXiv:2110.06733 [cs]*.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv:1712.05181 [cs.CL]*.

Maria Brykina, Valentin Gusev, Sándor Szeverényi, and Beáta Wagner-Nagy. 2018. Nganasan Spoken Language Corpus (NSLC). http://hdl.handle.net/11022/0000-0007-C6F2-8.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO Standard for Dialogue Act Annotation, Second Edition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

Hendrik Buschmeier and Stefan Kopp. 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*.

Alexandra Canavan, David Graff, and George Zipperlen. 1997a. CALLHOME American English Speech.

Alexandra Canavan, David Graff, and George Zipperlen. 1997b. CALLHOME German Speech.

Alexandra Canavan and George Zipperlen. 1996a. CALLFRIEND American English-Non-Southern Dialect.

Alexandra Canavan and George Zipperlen. 1996b. CALLFRIEND Korean.

Alexandra Canavan and George Zipperlen. 1996c. CALLHOME Mandarin Chinese Speech.

Alexandra Canavan and George Zipperlen. 1996d. CALLHOME Spanish Speech.

Alexandra Canavan, George Zipperlen, and David Graff. 1997c. CALLHOME Egyptian Arabic Speech.

Alexandra Canavan, George Zipperlen, and David Graff. 2014. CALLFRIEND Farsi Second Edition Speech.

Bernard Caron. 2016. Hausa collection in LLACAN. http://www.language-archives.org/language/hau.

Bernard Caron, Marvellous S. Davan, and Justin M. B. Ali. 2014. Zaar collection in LLACAN. http://www.language-archives.org/language/say.

Vera Alexandra Carvalho Ferreira, Sabine Wurm, Peter Bouda, Annette Endruschat, Regina Rosa Hämmerle, Ioana Fugaru, Ramón Rodriguez, Henrique Lobo Ferreira, and Katharina Knuffmann. 2011. Minderico, An Endangered Language in Portugal. https://hdl.handle.net/1839/f47b19bd-ac9c-434c-b559-c6ea00485f3c.

Justine Cassell. 2020. The ties that bind: Social interaction in conversational agents. *Reseaux*, 220221(2):21–45.

Katia Chirkova and Zhengkang Han. 2017. Duoxu: Documentation of a Critically Endangered Language of South-West China. https://cocoon.huma-num.fr/exist/crdo/meta2/crdo-COLLECTION_CHK_DUOXU.

Patricia M. Clancy, Sandra A. Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26(3):355–387.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Ruth E. Corps, Birgit Knudsen, and Antje S. Meyer. 2022. Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. *Cognition*, 223:105037.

Elizabeth Couper-Kuhlen and Margret Selting. 2017. *Interactional Linguistics: An Introduction to Language in Social Interaction*. Cambridge University Press, Cambridge.

Andrew Cowell. 2010. A Conversational Database of the Arapaho Language in Video Format. http://hdl.handle.net/2196/3bba11be-a5e2-47dd-bfe5-42f2ee9e0bf4.

Ronald Cumbal, Birger Moell, José Lopes, and Olov Engwall. 2021. "You don't understand me!": Comparing ASR results for L1 and L2 speakers of Swedish. In *Proceeding of Interspeech 2021*, pages 4463–4467.

Luiz Antônio da Silva. 1996. Projeto da Norma Urbana Linguística Culta. https://fale.ufal.br/projeto/nurcdigital/.

Andreea Danielescu and Gwen Christian. 2018. A bot is not a polyglot: Designing personalities for multilingual conversational agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–9. Association for Computing Machinery.

J. P. de Ruiter, Holger Mitterer, and N. J. Enfield. 2006. Projecting the end of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Language*, 82(3):515–535.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408.

Ezequiel A. Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. 2018. *Linguistic Bodies: The Continuity between Life and Language*. MIT Press, Cambridge, MA.

Thomas Diaz. 2018. Documentation of Heyo [auk], a Torricelli language of Papua New Guinea. http://hdl.handle.net/2196/18d3c47e-db5b-492c-853f-1a000aa19606.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mark Dingemanse. 2017. Brain-to-brain interfaces and the role of language in distributing agency. In N. J. Enfield and Paul Kockelman, editors, *Distributed Agency*, pages 59–66. Oxford University Press, Oxford.

Mark Dingemanse and Ordime Kanairoh. 2012. Collection Siwu. https://hdl.handle.net/1839/c410de17-81eb-4477-ae0d-d43ff1aea085.

Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. Universal Principles in the Repair of Communication Problems. *PLOS ONE*, 10(9):e0136100.

Javier Domingo. 2019. Tehuelche Language Collection. http://hdl.handle.net/2196/00-0000-0000-0011-F549-B.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. A neural network model for low-resource universal dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 339–348.

Suzanne Eggins and Diana Slade. 2004. *Analysing Casual Conversation*. Equinox Publishing Ltd.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: A Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online. Association for Computational Linguistics.

N. J. Enfield, Mark Dingemanse, Julija Baranova, Joe Blythe, Penelope Brown, Tyko Dirksmeyer, Paul Drew, Simeon Floyd, Sonja Gipper, Rósa Gísladóttir, Gertie Hoymann, Kobin H. Kendrick, Stephen C. Levinson, Lilla Magyari, Elizabeth Manrique, Giovanni Rossi, Lila San Roque, and Francisco Torreira. 2013. Huh? What? – A first survey in twenty-one languages. In Makoto Hayashi, Geoffrey Raymond, and Jack Sidnell, editors, *Conversational Repair and Human Understanding*, pages 343–380. Cambridge University Press, Cambridge.

Mirjam Ernestus, Lucie Kočková-Amortová, and Petr Pollak. 2014. The Nijmegen corpus of casual Czech. In *LREC 2014: 9th International Conference on Language Resources and Evaluation*, pages 365–370.

Jean-Baptiste Estoup. 1917. Les mots usuels. *Journal de la société statistique de Paris*, 58:137–140.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32:429–492.

Rekinan Fadlul, Timothy Mckinnon, David Gil, and Bradley Taylor. 2016. Kerinci (Sungai Penuh) Database. https://hdl.handle.net/1839/00-0000-0000-0022-654E-D.

J. R. Firth. 1935. The Technique of Semantics. *Transactions of the Philological Society*, 34(1):36–73.

Simeon Floyd. 2021. Conversation and Culture. *Annual Review of Anthropology*, 50(1):219–240.

W. Nelson Francis. 1965. A standard corpus of edited present-day American English. *College English*, 26(4):267–273.

Juan María Garrido, David Escudero, Lourdes Aguilar, Valentín Cardeñoso, Emma Rodero, Carme De-La-Mota, César González, Carlos Vivaracho, Sílvia Rustullet, Olatz Larrea, et al. 2013. Glissando: A corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language resources and evaluation*, 47(4):945–971.

Emer Gilmartin. 2021. What's Chat and Where to Find It. In Erik Marchi, Sabato Marco Siniscalchi, Sandro Cumani, Valerio Mario Salerno, and Haizhou Li, editors, *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, volume 714, pages 261–265. Springer Singapore, Singapore.

Emer Gilmartin, Benjamin R. Cowan, Carl Vogel, and Nick Campbell. 2018. Explorations in multiparty casual social talk and its relevance for social human machine dialogue. *Journal on Multimodal User Interfaces*, 12(4):297–308.

Charles Goodwin. 1981. *Conversational Organization: Interaction Between Speakers and Hearers*. Language, Thought, and Culture. Academic Press, New York.

Charles Goodwin. 1986. Between and within: Alternative Sequential Treatments of Continuers and Assessments. *Human Studies*, 9(2/3):205–217.

Charles Goodwin. 1995. The negotiation of coherence within conversation. In Morton Ann Gernsbacher and Talmy Givón, editors, *Coherence in Spontaneous Text*, pages 117–137. John Benjamins, Amsterdam / Philadelphia.

Karolina Grzech. 2020. Upper Napo Kichwa: Documentation of language and culture. http://hdl.handle.net/2196/00-0000-0000-000C-F5FB-A.

Tom Güldemann and Alena Witzlack-Makarevich. 2014. Text documentation of N|uu. http://hdl.handle.net/2196/00-0000-0000-0002-F81F-F.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. glottolog/glottolog: Glottolog database 4.4. Version Number: v4.4 Type: dataset.

Charlotte Hemmings. 2017. Documentation of the Kelabit Language, Sarawak, Malaysia. http://hdl.handle.net/2196/00-0000-0000-000F-B667-4.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The Weirdest People in the World? *Behavioral and Brain Sciences*, 33(2-3):61–83.

John Heritage. 1984. Conversation Analysis. In John Heritage, editor, *Garfinkel and Ethnomethodology*, pages 233–292. Polity Press, Cambridge; New York.

Nestor Hernandez-Green. 2009. Documentation of San Jerónimo Acazulco Otomi, Ocoyoacac, Mexico. http://hdl.handle.net/2196/00-0000-0000-0002-AA48-9.

M. Hisyam, Usman Dwi Purwoko, and Dalan Peranginangin. 2013. A project of LIPI (the Indonesian Institute of Sciences) on Documenting and Revitalizing Endangered Languages and Cultures in Eastern Indonesia. https://hdl.handle.net/1839/00-0000-0000-0022-6530-D.

Charles F. Hockett. 1960. The Origin of Speech. *Scientific American*, 203(3):89–96.

Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An End-to-End Conversational Style Matching Agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, pages 111–118, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy and Shannon L. Spruit. 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. Association for Computational Linguistics.

Christine Howes and Arash Eshghi. 2021. Feedback Relevance Spaces: Interactional Constraints on Processing Contexts in Dynamic Syntax. *Journal of Logic, Language and Information*, 30(2):331–362.

Laszlo Hunyadi, Tamás Váradi, Gy Kovács, István Szekrényes, Hermina Kiss, and Karolina Takács. 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.

Ivan Illich. 1973. *Tools for Conviviality*. Harper & Row, New York.

Gail Jefferson. 1985. Notes on a systematic Deployment of the Acknowledgement tokens 'Yeah' and 'Mmhm'. *Papers in Linguistics*, 17(2):197–216.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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.

Dan Jurafsky, Elizabeth Shriberg, Barbara Fox, and Traci Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL-COLING Workshop on Discourse Relations and Discourse Markers*, pages 114–120, Montreal.

Kobin H. Kendrick, Penelope Brown, Mark Dingemanse, Simeon Floyd, Sonja Gipper, Kaoru Hayano, Elliott Hoey, Gertie Hoymann, Elizabeth Manrique, Giovanni Rossi, and Stephen C. Levinson. 2020. Sequence organization: A universal infrastructure for social action. *Journal of Pragmatics*, 168:119–138.

Jason Kessler. 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.

Kyu-hyun Kim. 1999. Phrasal Unit Boundaries and Organization of Turns and Sequences in Korean Conversation. *Human Studies*, 22(2-4):425–446.

Soung-U. Kim. 2018. A multi-modal documentation of Jejuan conversations. http://hdl.handle.net/2196/00-0000-0000-000E-D15C-1.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2021. Overview of Datasets for the Sign Languages of Europe. https://www.fdr.uni-hamburg.de/record/9561.

Stefan Kopp and Nicole Krämer. 2021. Revisiting Human-Agent Communication: The Importance of Joint Co-construction and Understanding Mental States. *Frontiers in Psychology*, 12.

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

Jonas Lau. 2019. Documenting Àbèsàbèsì. http://hdl.handle.net/2196/a6371e17-d083-4e29-bc4a-ecb9303f9197.

Claudia Leto, Winarno Salim Alamudi, Nikolaus P. Himmelmann, and Sonja Riesberg. 2010. Collection Totoli. https://hdl.handle.net/1839/00-0000-0000-0009-2A75-5.

Stephen C. Levinson. 1979. Activity types and language. *Linguistics*, 17(5-6):365–399.

Stephen C. Levinson. 2013. Action formation and ascription. In Jack Sidnell and Tanya Stivers, editors, *Handbook of Conversation Analysis*, pages 103–130. Blackwell Publishers, Malden, MA.

Stephen C. Levinson. 2016. Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, 20(1):6–14.

Stephen C. Levinson, Marisa Casillas, W.E. Armstrong, and Francisco Torreira. 2019. Collection Yélî Dnye. https://hdl.handle.net/1839/97720d9d-927e-41cb-bd65-8e6accc42d2c.

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

Natalia Levshina. 2021. Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*.

Yu Li. 2017. Documentation of Zauzou, an endangered language in China. http://hdl.handle.net/2196/00-0000-0000-0010-8822-7.

Andreas Liesenfeld and Mark Dingemanse. 2022. Building and curating conversational corpora for diversity-aware language science and technology. *arXiv:2203.03399 [cs]*.

Florian Lionnet, Remadji Hoinathy, and Sandrine Loncke. 2020. Laal language documentation project. https://hdl.handle.net/1839/93472197-4462-489c-8cee-0d9a3587f3e5.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking Natural Language Understanding Services for Building Conversational Agents. In Erik Marchi, Sabato Marco Siniscalchi, Sandro Cumani, Valerio Mario Salerno, and Haizhou Li, editors, *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, volume 714, pages 165–183. Springer Singapore, Singapore.

Stefano Manfredi. 2016. Juba Creole collection in LLACAN. http://www.language-archives.org/language/pga.

Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David Traum, and Zhou Yu. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, 71:101255.

Bruil Martine. 2012. Documentation of Ecuadorian Siona. https://hdl.handle.net/1839/b1796725-1a49-48ee-93ea-75e5b440c7bc.

Philippe Antoine Martinez. 2020. Documentary Corpus of Chhitkul-Rakchham, an endangered Tibeto-Burman language of Northern India. http://hdl.handle.net/2196/23cecd70-2879-412f-bde1-456b0ea0ef1a.

Bradley McDonnell. 2017. Documentation of Nasal: An overlooked Malayo-Polynesian isolate of southwest Sumatra. http://hdl.handle.net/2196/00-0000-0000-0010-798B-E.

Daniela Mereu and Alessandro Vietti. 2021. Dialogic ItAlian: The creation of a corpus of Italian spontaneous speech. *Speech Communication*, 130:1–14.

Mirjam Möller Nwadigo. 2016. A documentation project of Baa, a language of Nigeria. http://hdl.handle.net/2196/e050a2cd-f61d-435e-824e-93d24877bbaa.

Luke Munn. 2018. Alexa and the intersectional interface. *Angles. New Perspectives on the Anglophone World*, (77).

Tomoe Nakamura and Tania Granadillo. 2005. CABank Japanese CallFriend Corpus. https://ca.talkbank.org/access/CallFriend/jpn.html.

Daniel Nettle. 1999. *Linguistic Diversity*. Oxford University Press, Oxford.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.

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

Catharine Oertel, Joakim Gustafson, and Alan W. Black. 2016. On data driven parametric backchannel synthesis for expressing attentiveness in conversational agents. In *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, MA3HMI '16, pages 43–47, New York, NY, USA. Association for Computing Machinery.

Pavel Ozerov. 2018. A community-driven documentation of natural discourse in Anal, an endangered Tibeto-Burman language. http://hdl.handle.net/2196/af2415d6-dc75-4330-ba5d-7b8122e50982.

Ioannis Papaioannou, Christian Dondrup, and Oliver Lemon. 2018. Human-Robot Interaction Requires More Than Slot Filling - Multi-Threaded Dialogue for Collaborative Tasks and Social Conversation. In *FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction*, pages 61–64. ISCA.

William H Parker. 2020. Documentation of Cora in San Juan Corapan. http://hdl.handle.net/2196/0829a3a6-92c4-4346-8e37-04845cdd1f7f.

Piotr Pęzik and Łukasz Dróżdż. 2011. PELCRA Polish spoken corpus. http://pelcra.pl/res/spoken/plec.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, pages 1–19.

Daniel Pimienta, Daniel Prado, and Álvaro Blanco. 2009. Twelve years of measuring linguistic diversity in the internet: balance and perspectives. *Report published by United Nations Educational, Scientific and Cultural Organization*.

Karola Pitsch. 2016. Limits and opportunities for mathematizing communicational conduct for social robotics in the real world? Toward enabling a robot to make use of the human's competences. *AI & SOCIETY*, 31(4):587–593.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Wim Pouw, Shannon Proksch, Linda Drijvers, Marco Gamba, Judith Holler, Christopher Kello, Rebecca S. Schaefer, and Geraint A. Wiggins. 2021. Multilevel rhythms in multimodal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1835):20200334.

Laurent Prevot, Pierre Magistry, and Pierre Lison. 2019. Should we use movie subtitles to study linguistic patterns of conversational 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.

Sonja Riesberg, Nikolaus P. Himmelmann, Kristian Walianggen, and Apriani Arilaha. 2015. Yali Summits Collection in collection "CELD Papua". https://hdl.handle.net/1839/e941f246-4842-4ff0-9535-9dbb25f9f805.

Felicia Roberts, Piera Margutti, and Shoji Takano. 2011. Judgments Concerning the Valence of Inter-Turn Silence Across Speakers of American English, Italian, and Japanese. *Discourse Processes*, 48(5):331–354.

Matthew Roddy. 2021. *Neural Turn-Taking Models for Spoken Dialogue Systems*. Ph.D. thesis, Trinity College Dublin, Dublin.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs. *arXiv:1806.11461 [cs]*.

Anna Rogers. 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. Association for Computational Linguistics.

Jean Rohleder. 2018. Documentation and description of Vamale, an endangered language of New Caledonia. https://hdl.handle.net/1839/b1796725-1a49-48ee-93ea-75e5b440c7bc.

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Scientific Reports*, 5.

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

Diana Schackow. 2014. Documentation and grammatical description of Yakkha, Nepal. http://hdl.handle.net/2196/00-0000-0000-0002-D744-B.

Emanuel A. Schegloff. 1982. Discourse as Interactional Achievement: Some Uses of 'uh huh' and Other Things That Come Between Sentences. In Deborah Tannen, editor, *Analyzing Discourse: Text and Talk*, pages 71–93. Georgetown University Press, Washington DC.

Emanuel A. Schegloff. 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 and Stephen C. Levinson, editors, *Roots of Human Sociality: Culture, Cognition, and Human Interaction*, pages 70–96. Berg, Oxford.

Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge University Press, Cambridge.

Bambi B. Schieffelin and Elinor Ochs, editors. 1986. *Language Socialization Across Cultures*. Cambridge University Press, Cambridge.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse*, 2(1):83–111.

Johanna Seibt. 2017. Towards an Ontology of Simulated Social Interaction: Varieties of the "As If" for Robots and Humans. In Raul Hakli and Johanna Seibt, editors, *Sociality and Normativity for Robots*, pages 11–39. Springer International Publishing, Cham.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Frank Seifart, Jan Strunk, Swintha Danielsen, Iren Hartmann, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich, Nikolaus P. Himmelmann, and Balthasar Bickel. 2021. The extent and degree of utterance-final word lengthening in spontaneous speech from 10 languages. *Linguistics Vanguard*, 7(1).

Sheena Shah. 2019. A multimedia corpus of si-Phuthi. http://hdl.handle.net/2196/00-0000-0000-0010-D126-A.

Aung Si. 2014. Audio and video recordings of Kune, a Bininj Gunwok dialect spoken in Buluhkaduru Outstation near Maningrida, Northern Territory. https://doi.org/10.4225/72/56E97A3F99539.

Jack Sidnell and N. J. Enfield. 2012. Language Diversity and Social Action. *Current Anthropology*, 53(3):302–333.

Nathaniel Sims. 2018. Documentation of Yonghe Qiang language and culture. http://hdl.handle.net/2196/91c9326c-6e76-4d8c-aafe-e3f1e7ceccab.

Gabriel Skantze. 2017. Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany. Association for Computational Linguistics.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

Tanya Stivers, N. J. Enfield, Penelope Brown, C. Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, J. P. de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Lucy A. Suchman. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*, 2nd ed edition. Cambridge University Press, Cambridge.

Lucy A. Suchman. 2020. *Agencies in Technology Design: Feminist Reconfigurations*, pages 361–375. Routledge, London.

Taalunie. 2014. Corpus Gesproken Nederlands - CGN (Version 2.0.3). http://hdl.handle.net/10032/tm-a2-d9.

Uri Tadmor. 2007. Languages of Western Borneo Documentation Project. https://hdl.handle.net/1839/00-0000-0000-0022-5D7C-E.

Michael Thomas. 2014. Sakun (Sukur) Language Documentation. http://hdl.handle.net/2196/ea080fc9-a392-4d41-a91c-0c801bbde646.

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

Mohammad Umair, Julia Mertens, Saul Albert, and J. P. de Ruiter. 2021. GailBot: An automatic transcription system for Conversation Analysis. Technical report, OSF.

Volker Unterladstetter, Alexander Loch, Freya Morigerowsky, and Yusuf Sawaki. 2013. Collection Wooi. https://hdl.handle.net/1839/eb0ab65a-e985-42d1-a9ee-fccdba47a526.

Jacqueline van Arkel, Marieke Woensdregt, Mark Dingemanse, and Mark Blokpoel. 2020. A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 177–194. Association for Computational Linguistics.

Andrea Vanzo, Jose L Part, Yanchao Yu, Daniele Nardi, and Oliver Lemon. 2018. Incrementally Learning Semantic Attributes through Dialogue Interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, pages 865–873, Stockholm, Sweden.

Cristina Voinea. 2018. Designing for conviviality. *Technology in Society*, 52:70–78.

Judith Voß. 2018. Documentation and grammar of Gutob (Munda). http://hdl.handle.net/2196/f027a3a2-d38f-4428-88ec-33b46d346cb3.

Alexandra Vydrina. 2013. Description and documentation of the Kakabe language. http://hdl.handle.net/2196/3015b4c3-1ffc-4cc5-8309-f05f9d4ce8b2.

Johannes Wagner and Bente Maegaard. 2017. SamtaleBank, Danish spoken language component of the DK/CLARIN project. https://samtalebank.talkbank.org/.

Nigel Ward. 2006. Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14:129–182.

Nigel G. Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. 2018. Turn-Taking Predictions across Languages and Genres Using an LSTM Recurrent Neural Network. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 831–837.

Sheida White. 1989. Backchannels Across Cultures: A Study of Americans and Japanese. *Language in Society*, 18(01):59–76.

Thomas Widlok, Christian Rapold, and Gertie Hoymann. 2007. Collection "+Akhoe Haillom". https://hdl.handle.net/1839/b1796725-1a49-48ee-93ea-75e5b440c7bc.

Joshua Wilbur. 2009. Pite Saami: Documenting the language and culture. http://hdl.handle.net/2196/00-0000-0000-0003-1170-E.

Nicholas Williams. 2017. Documenting Language and Interaction in Kula. http://hdl.handle.net/2196/020426e9-bffc-42da-9f8c-b67c1160a0f9.

Benedikt Winkhart. 2016. A documentation of the remnant Baka-Gundi language Limassa. http://hdl.handle.net/2196/70607394-af7c-4fe2-af53-fd40fd6cac50.

Victor Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting, Chicago Linguistic Society*, pages 567–578.

Richard F. Young and Jina Lee. 2004. Identifying units in interaction: Reactive tokens in Korean and English conversations. *Journal of Sociolinguistics*, 8(3):380–407.

Christine Guo Yu, Alan F. Blackwell, and Ian Cross. 2021. Perception of rhythmic agency for conversational labeling. *Human–Computer Interaction*, pages 1–24.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and Human Speech Transcription Errors. In *Proceedings of Interspeech 2019*, pages 3088–3092. ISCA.

George K. Zipf. 1935. *The psycho-biology of language*. Houghton Mifflin, Boston.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort; an Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA.

A Appendix

A.1 Turn-taking: validating analyses

In order to make visible the structure and timing of turn-taking in quotidian interaction, the turn-taking analysis in §3 takes all directly adjacent conversational turns in dyadic interactions in which a speaker change occurs, without regard for type of turn or social action. Stivers et al. (2009) limited their comparison to polar question-answer sequences, achieving a form of "natural control" to ensure comparability. In our dataset, the timing of such sequences does not radically differ from the overal timing distribution, for a subset of 10 languages for which we are able to automatically identify probable QA-sequences (Figure A1). Given the broad-scale comparability of the overall timing distributions (in grey) and the more controlled subset of at least 250 question-answer sequences per language (in black), we conclude that QA sequences can act as a useful proxy for timing in general (supporting Stivers et al. 2009), but also that QA-sequences are not necessary for a relatively robust impression of overall timing.

Analyses of turn-taking sometimes note that some turns are more likely to appear in overlap than others. In particular, continuers (also known as acknowledgement tokens (Jefferson, 1985) or backchannels (Yngve, 1970)) do not represent a claim to the conversational floor that is as strong as some other turns (Schegloff, 1982; Levinson, 2016). It is possible that cross-cultural variation in the frequency of sequentially defined continuers (as we report in §4) might underlie some part of the observed differences in turn-taking distributions. However, we feel that to dichotomize our resolutely cross-linguistic dataset into "continuers" versus "other turns" would be premature and would risk overinterpreting data that requires careful qualitative consideration.

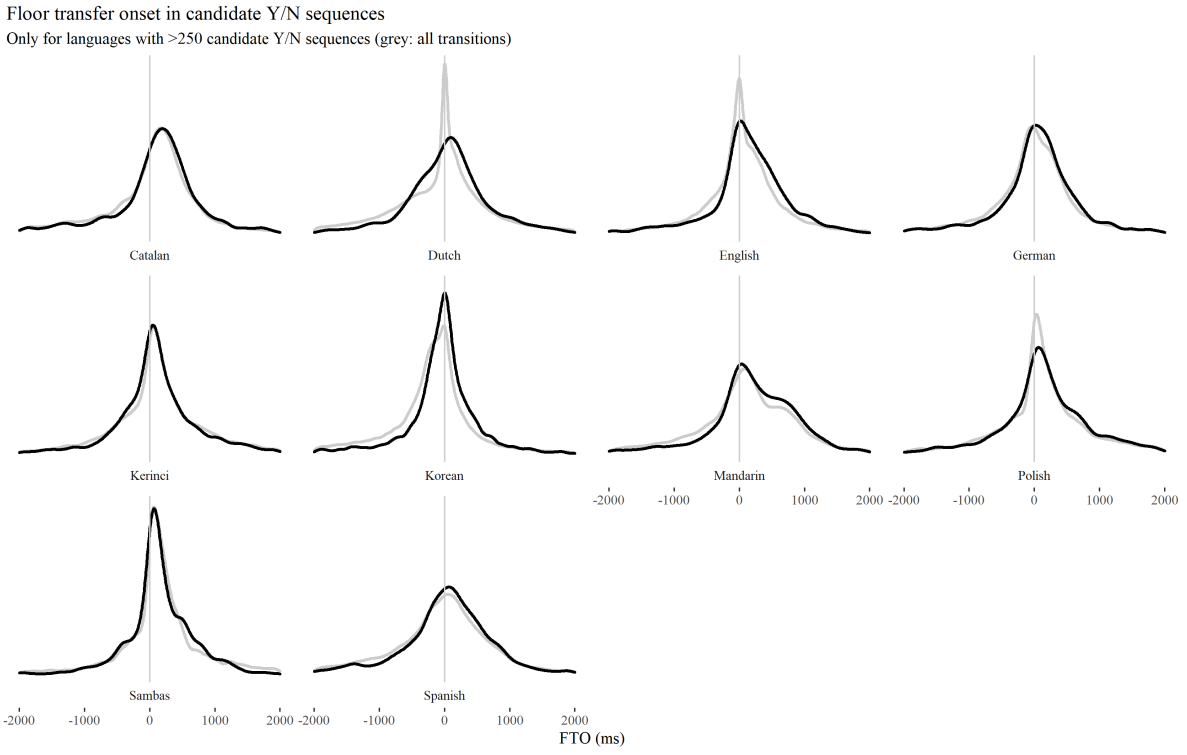

Figure A1: Timing of turn-taking in candidate polar question-answer sequences, for 10 languages with >250 such sequences. Probable sequences where identified by selecting all pairs of directly adjacent turns by two distinct participants where the first was annotated as ending in a question mark and the second was a non-unique turn format. Only languages in which questions are transcribed with a final "?" and which had at least 250 such sequences are included here.

A.2 Frequency and rank in smaller corpora

Turn-level and word-level relations between rank and frequency are sensitive to fluctuations in corpus size, so in the body of the paper we provide data for the 22 corpora that feature at least 20 turn formats that occur at least 20 times. Here we provide the same for a further 21 languages in which there are at least 9 but less than 20 recurrent turn formats (Figure A2). As in §5, turn-level fit lines are computed on the 20% of turns that occur >20 times (purple); however, fit strongly fluctuates with corpus size and should not be taken at face value.

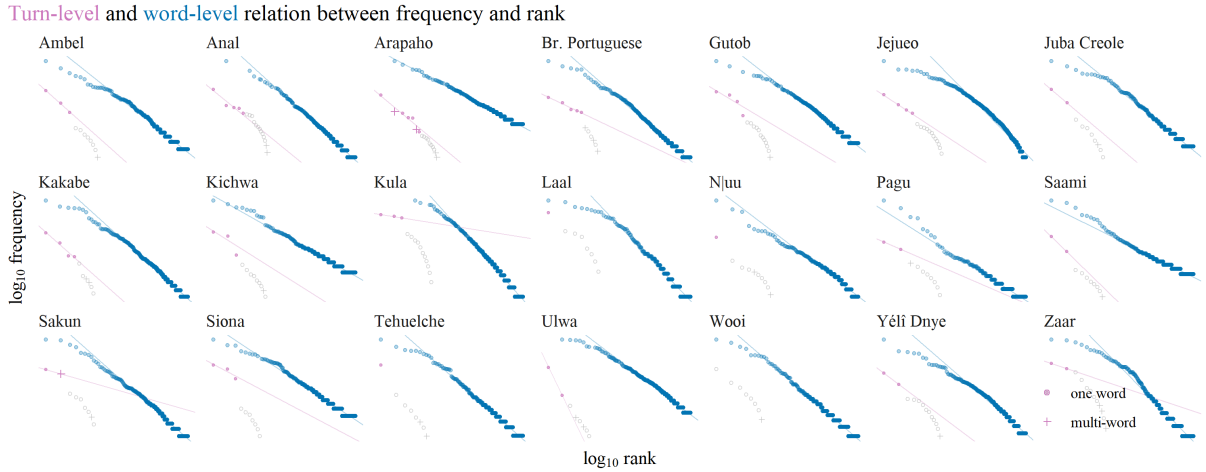

Figure A2: Frequency/rank distributions of tokenized items ('words') and recurring turn formats in conversational corpora with at least 9 such turn formats, representing 21 languages (18 phyla). Tokenized items (blue) show a linear frequency/rank relation in log/log space. Recurring turn formats (whether one-word ○ or multi-word +) taper off towards lower frequencies and unique turns (grey) but more frequent items may obey a similar frequency/rank distribution (purple).

B Dataset

B.1 Curation, reproducibility and ethical considerations

The data described and analysed here consists of a range of maximally diverse language resources collected (as primary data) over several decades with the contributions and consent of communities around the world. Because our aim is to highlight the potential of available language resources and to maximise the reproducibility of our research, we focus on existing data made available for research purposes in language archives like the Endangered Language Archive (ELAR), The Language Archive (TLA) and Paradisec, as well as through centralized services like CLARIN and the Linguistic Data Consortium (LDC).

While most of the sourced corpora come from publicly-funded and openly accessible language documentation or language resource platforms, some corpora are currently unpublished or paywalled. Due to the diverse set of licensing and publication agreements with the providers of the sourced corpora, we are unable to provide direct, unrestricted access to the dataset. To enhance transparency and reproducibility of the findings presented in the study, we provide an extensive datasheet with details on motivation, makeup, and processing steps, as well as full information on the larger set of corpora we have considered. This includes metadata on compilers, durable links to archival copies, and a quantitative overview of key properties of turn duration, timing and translations. The study repository is here: https://osf.io/zd34r/

The work reported here also comes with ethical considerations. There are at least three points at which such considerations are important when it comes to the kind of data considered here: data collection; data usage; and implications of any technologies developed on the basis of such data. With regard to the first point, we only work with corpora collected with the help and involvement of communities who have given

their informed consent for this data to be recorded, annotated, translated and archived. The other two points are discussed in detail in Levow et al. (2021), who develop a shared task reliant on corpora much like the ones considered here (albeit not focusing on co-present conversation). With regard to data usage, the corpora considered here are only those for which contributors have granted access openly or to all registered users, usually for research purposes. We cannot redistribute the dataset directly, but have strived to document the process of curation in sufficient detail to enable others to register and access the data (Liesenfeld and Dingemanse, 2022). With regard to technological applications, as with any technology, there is potential for helpful as well as harmful uses (Hovy and Spruit, 2016) and we side with Levow et al. (2021) in stressing the need for computational linguists to work closely with language communities in maximising helpful uses and minimising harmful ones. As noted above, our supplementary materials also include details in the form of a data statement (Bender and Friedman, 2018).

B.2 Inclusions and exclusions

Not all of the available corpora are represented in all analyses presented in the paper or appendices, because corpora differ in size, precision of annotation, and level of transcription. For instance, some corpora use segmentation methods that do not precisely link annotations to the corresponding communicative turn. Others may split annotations in ways that are not clearly documented and that do not seem to correspond to the turn-level annotation format that is most common across corpora. While such corpora may lend themselves to various corpus linguistic analyses, incommensurable methods of segmentation means that it would take considerable additional work to use this data in qualitative and quantitative analyses of timing, turn-taking and talk-in-interaction.

The online supplementary materials provide several illustrative examples along with a detailed accounting of reasons for exclusions. We are optimistic that in the future, language technology can be harnessed to improve time-alignment and temporal precision of existing conversational corpora (Bird, 2021; Umair et al., 2021). We also hope that annotation procedures can be designed with an eye to staying faithful to the temporal and sequential structure of the primary data.

While we have collected information on sign language corpora, the set considered here does not include any of them, as corpora of casual conversation in sign languages are exceedingly rare (Kopf et al., 2021), and annotation conventions tend to focus on the level of signs rather than utterances, sequences and timing. Incorporating sign language corpora requires careful work with sign language linguists and deaf communities to arrive at common and commensurable annotation standards that afford the cross-modal comparison of interactional structure.

B.3 List of languages and corpora

The table below presents the 63 languages included in the curated dataset, 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.

Language (glottocode)	Family	Citation
ǂAkhoe (haio1238)	Khoe-Kwadi	(Widlok et al., 2007)
Akpes (akpe1248)	Atlantic-Congo	(Lau, 2019)
Ambel (waig1244)	Austronesian	(Arnold, 2017)
Anal (anal1239)	Sino-Tibetan	(Ozerov, 2018)
Arabic (egyp1253)	Afro-Asiatic	(Canavan et al., 1997c)
Arapaho (arap1274)	Algic	(Cowell, 2010)
Baa (kwaa1262)	Atlantic-Congo	(Möller Nwadigo, 2016)
Br. Portuguese (braz1246)	Indo-European	(da Silva, 1996)
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)
Duoxu (ersu1241)	Sino-Tibetan	(Chirkova and Han, 2017)
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., 1997b)
Gunwinggu (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)
Italian (ital1282)	Indo-European	(Mereu and Vietti, 2021)
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)
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)
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)
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	(Pęzik 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)
Siwu (siwu1238)	Atlantic-Congo	(Dingemanse and Kanairoh, 2012)
Spanish (stan1288)	Indo-European	(Canavan and Zipperlen, 1996d)
Tehuelche (tehu1242)	Chonan	(Domingo, 2019)
Totoli (toto1304)	Austronesian	(Leto et al., 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)