Interspeech 2022
18-22 September 2022, Incheon, Korea

Bottom-up discovery of structure and variation in response tokens ('backchannels') across diverse languages

Andreas Liesenfeld [1], *Mark Dingemanse* [1]

[1]Centre for Language Studies
Radboud University

andreas.liesenfeld@ru.nl, mark.dingemanse@ru.nl

Abstract

Response tokens (also known as backchannels, continuers, or feedback) are a frequent feature of human interaction, where they serve to display understanding and streamline turn-taking. We propose a bottom-up method to study responsive behaviour across 16 languages (8 language families). We use sequential context and recurrence of turns formats to identify candidate response tokens in a language-agnostic way across diverse conversational corpora. We then use UMAP clustering directly on speech signals to represent structure and variation. We find that (i) written orthographic annotations underrepresent the attested variation, (ii) distinctions between formats can be gradient rather than discrete, (iii) most languages appear to make available a broad distinction between a minimal nasal format 'mm' and a fuller 'yeah'-like format. Charting this aspect of human interaction contributes to our understanding of interactional infrastructure across languages and can inform the design of speech technologies.

Index Terms: backchannels, feedback, linguistic typology

1. Introduction

Response tokens like 'mm' and 'yeah' are among the most frequent and yet easily overlooked aspects of conversation. Often described as 'backchannels', they occupy a central role in scaffolding interaction, streamlining turn-taking and calibrating understanding [1, 2]. The sheer frequency and functional importance of these items means we can think of them as interactional tools: linguistic devices that accomplish social actions. Progress towards interactive language technologies hinges on understanding and modelling such interactional tools [3, 4]. While some of them have long been studied in a small number of well-resourced languages [5, 6], their structure and variation across diverse languages is mostly uncharted territory. Here we combine insights from comparative linguistics, conversation analysis and computational approaches to shed new light on a key aspect of human interactional infrastructure.

Tackling structure and variation in response tokens across languages requires addressing three interlinked challenges. The first is a dearth of data: corpora of casual speech, crucial to ensure solid foundations for diversity-aware language technology, are still quite rare. We address this by using conversational corpora from diverse sources, including language documentation archives that are increasingly available but rarely used. The second is how to achieve comparability in uncharted territory. We address this by using the sequential structure of conversation to ensure we compare like with like across unrelated languages. The third is an overreliance on orthographic representations, which risks underestimating degrees of variation and flexibility in the use of response tokens. We address this by staying as closely as possible to the speech signal in its sequential context, rather than taking written forms for granted.

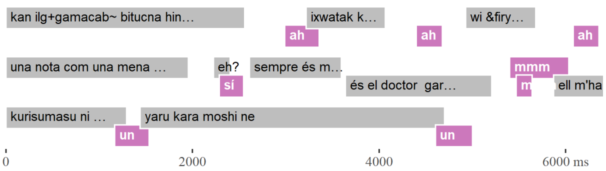


Figure 1: *Instances of response tokens in their natural environment in 3 unrelated languages: Arabic, Catalan and Japanese. We use the structural fact that such tokens tend to occur in series as a way to identify them across languages.*

2. Related work

Befitting their place at the intersection of language science and language technology, response tokens have been studied in disparate fields, from linguistics and conversation analysis to human-computer interaction and signal processing. Much work has focused on when and where response tokens occur, including efforts to identify "feedback relevance places" [7] and models to predict when participants produce response tokens in talk [8, 9, 10]. Other lines of work have used response tokens as a cue to predict the dynamics of talk and turn-taking [11] or to make inferences about mental and cognitive states [12]. A growing amount of work seeks to model feedback behaviour in human-agent interaction, including by means of response token generation [13, 14] and attentive listening systems [15, 16]. Despite considerable progress, the place of response tokens in speech technology is by no means settled: they tend to be missed by speech recognizers [17, 18, 19] and dialog managers have a hard time dealing with them [20], showing that they remain a key issue on which progress towards future generations of voice-interactive technologies and conversational user interfaces depends.

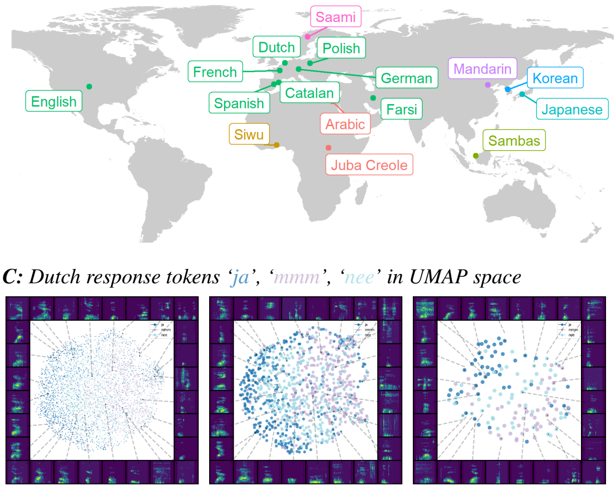
Observational work on forms and functions of response tokens in human interaction is an important empirical foundation of any speech technology intended for human use. Prior work includes in-depth studies of the response token system in English, Japanese and a handful of other languages [8, 5, 21] as well as comparative work on prosodic and multimodal aspects [22, 23, 24]. Besides a small number of single-language descriptive studies [25, 26, 27], the bulk of empirical, experimental and computational work has focused on a handful of well-resourced Indo-European and East Asian languages. We cannot assume that findings based on this small sample of languages apply across the board. Getting a handle on the true extent of diversity will be a critical stepping stone towards speech technologies that can cater to human needs around the globe.

10.21437/Interspeech.2022-11288

A: Languages, corpus size, and sequentially identified response tokens

Language (Glottocode)	Size (hrs)	Response tokens
Dutch (dutc1256)	387.6	ja, nee, mmm
French (stan1290)	31.4	ouais, hm
English (nort3314)	28	yeah, mhm, uhhuh
Spanish (stan1288)	27.6	sí, mmm, vale
Korean (kore1280)	26.6	eung, eo, ye
Farsi (west2369)	25.3	AhAn, mhm, KHob
Arabic (egyp1253)	20.3	ah, M, mhm
Mandarin (mand1415)	18.6	e, en, ai
German (stan1295)	18.6	ja, mhm, aha
Polish (poli1260)	15.8	mhm, tak, aha
Japanese (nucl1643)	13.4	un, e, un un
Siwu (siwu1238)	9.9	mm, ɛ̃ɛ̃
Catalan (stan1289)	6.7	sí, mmm, vale
Sambas (kend1254)	6.1	eeq, aoq, oh
Pite Saami (pite1240)	1	ja, mmm, nå
Juba Creole (suda1237)	0.5	mːːː, aj

B: Location of largest speech community



C: Dutch response tokens 'ja', 'mmm', 'nee' in UMAP space

(a) *total n = 20140* (b) *n=1000* (c) *n=150*

Figure 2: ***A**: Overview of included languages with dataset size in hours and top 3 sequentially identified response tokens as transcribed in the corpus. **B**: Location of largest speech community. **C**: Assessing the impact of sparse data on UMAP projections using three samples of Dutch response tokens. A look at the full dataset (a) and random-sampled subsets of decreasing size (b, c) suggests isomorphism across scales and interpretability of clustering solutions as small as 150 tokens.*

3. Data and methods

For the quantitative and inductive analysis that we envision, we need relatively large and maximally diverse language resources with time-aligned transcriptions. Rather than working with non-interactive data sources or collecting new data, here we explore the potential of language resources collected by the global language documentation movement [28, 4].

We curate corpora of unscripted conversation made available in language documentation archives. Corpora are assessed for factors that directly impact the feasibility of signal processing at scale: corpus size, transcription density, timestamp accuracy, and noise levels. For details on the data curation process see [29] and the repository at osf.io/7t9pn. The current dataset consists of corpora for 16 languages (8 phyla) (Fig. 2).

3.1. Sequential search method

Working with diverse corpora raises the issue of how to identify tokens of interest. Some transcriptions are more orthographic and regularized, others are more oriented towards phonetics, and few are devised with the study of response tokens in mind. For instance, annotations such as English 'yeah' often conflate variations that may or may not have interactional import [30, 31, 32]. Instead of taking any one representation format at face value, we use a language-agnostic *sequential search method* that allows us to inductively identify candidate tokens.

A key aspect of this method is that we do not search corpora for forms that sound like (or are translated as) 'yeah' or 'hmm'. Instead we define structural facts about how turns follow one another to pinpoint responsive behaviours in language-agnostic and form-agnostic ways (Fig. 1). In particular, we look for items that: (i) feature in the top decile of frequency counts by turn format per corpus; and (ii) occur at least once in a series of at least two produced by the same speaker. These search criteria reflect two basic observations about response tokens: their high frequency in naturally occurring talk [4], and the fact that they often occur in series of consecutive response tokens [33]. In short, we use patterns in how turn formats recur and follow one another to identify items of interest. Given this abstract structural characterization, we can achieve comparability across corpora [34].

3.2. Exploratory clustering

We conduct a bottom-up, exploratory analysis of structure and variation using UMAP dimensionality reduction [35], a method that is conceptually similar to PCA, MDS, and t-SNE. Like prior clustering methods, it builds a topological representation of the data in higher dimensional space and then reduces it to a two dimensional projection while preserving as much of the graph structure as possible. We use UMAP because it better represents similarity structure across datasets [36].

We start by using the sequential search method to identify the subset of human annotations that fit the profile of response tokens. We then use the annotation timestamps to generate spectrograms from the source audio. After normalizing and log-rescaling the spectrograms we apply UMAP clustering (for details, see osf.io/7t9pn).

Since some datasets are relatively small, a first question is to what extent lower numbers of response tokens impact the interpretability of clustering results. We use random-sampled successively smaller subsets of our largest corpus, Dutch, to assess the interpretability of clustering solutions for 20000, 1000 and 150 tokens (Figure 2C).

Since even 150 tokens makes for interpretable projections, and since our goal is to maximize the diversity of our sample, we select languages that have on the order of 10^2 tokens. There are 16 languages where the sequential search method yields sufficient numbers of tokens (minimum 86, maximum 20140). To facilitate quality control and visual exploration, we plot spectrograms alongside the clustering projections.

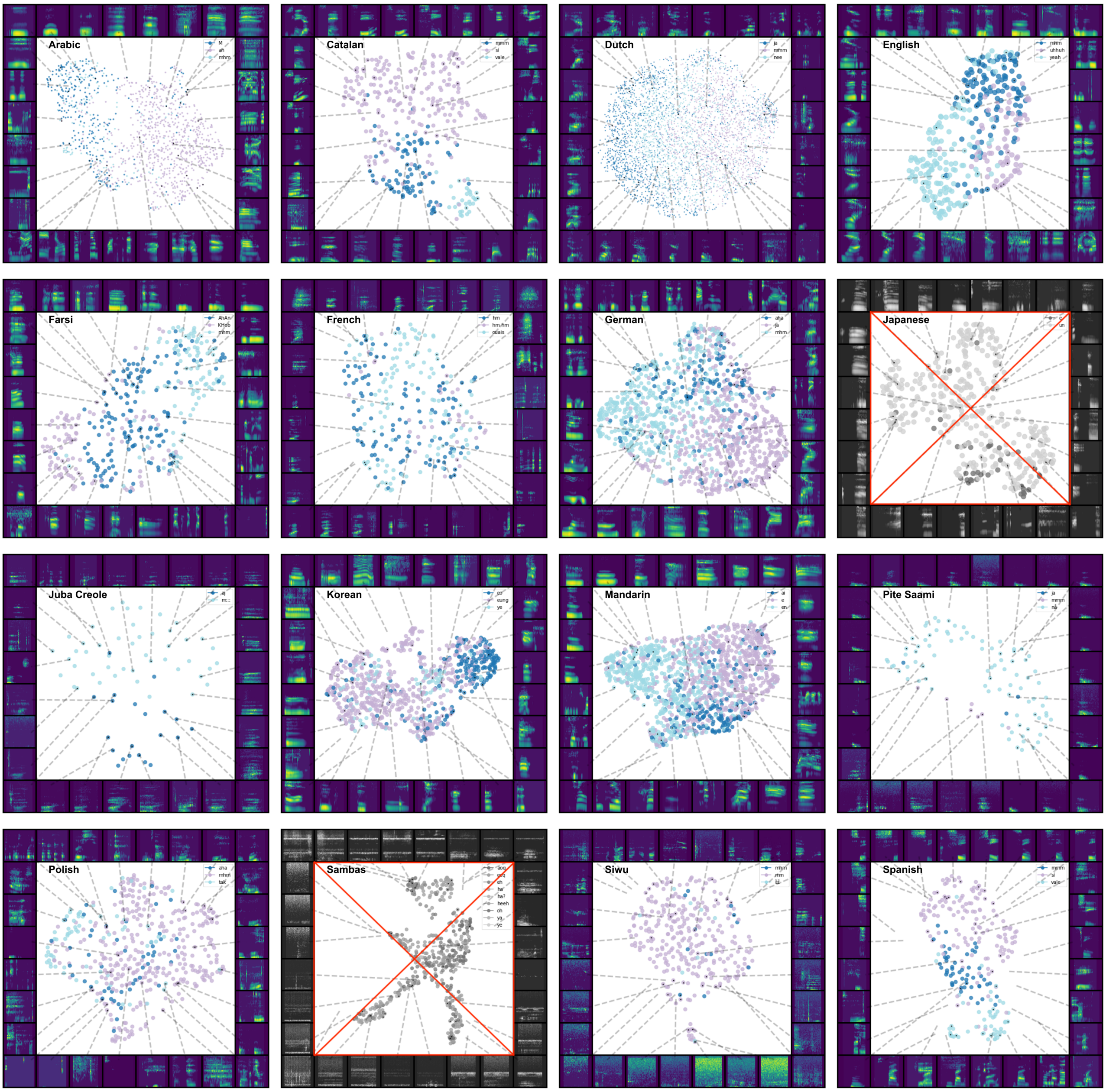


Figure 3: *Exploratory clustering projections of top 3 response tokens in 16 languages. Center: tokens in UMAP space ordered by decreasing corpus frequency. Frame: Sample spectrograms from key areas of the projection to allow visual inspection. Total number of tokens (estimated number of speakers): Arabic n=1434 (8), Catalan n=768 (24), Dutch = 20140 (1266), English n=1210 (32), Farsi n=324 (6), French n=164 (8), German n=797 (14), Juba Creole n=87 (4), Korean n=865 (9), Mandarin Chinese n=1092 (8), Pite Saami n=86 (7), Polish n=430 (42), Siwu n=234 (18), Spanish n=870 (24). Japanese and Sambas, greyed out, demonstrate two challenges of human-annotated speech data: timestamp inaccuracy and high noise levels (both visible in the spectrograms).*

4. Results

Applying dimensionality reduction techniques to response tokens in unfolding conversations allows a closer look at structure and variation at the signal level. Here we draw attention to findings in four areas.

Formats. Despite the variety of response token formats (Fig. 2A), one basic distinction that appears to be available in all languages is that between a minimal monosyllabic nasal format and one or more fuller forms that feature vowels and consonants other than nasals. The distinction shows up in most of the clustering projections, with the most distinct formats pulled apart across the space in each language.

Gradience. At the same time, some formats appear to bear more gradient relations to one another. For instance, in languages with multiple nasal or nasalized formats, distributions often overlap at least partly (see Arabic, English, Farsi, German, Siwu). Given good enough audio quality, it is conceivable that clustering gives us a handle on the gradience in form —corresponding for instance to different degrees of mouth opening—that speakers can exploit interactionally.

Beyond orthography. Transcribers of conversational speech face the impossible choice between capturing types (and functions) versus tokens (and forms). This trade-off becomes most apparent when transcribing minimal utterances like response tokens. For instance, the distribution of German ‘aha’ tokens largely overlaps with that of ‘mhm’ in the clustering projection, a fact that becomes more intelligible once we realise that the phonetic realization of the former is closer to [ɐ̃hɐ̃]. An additional layer of complexity is posed by orthographic idiosyncrasies and different writing systems. For instance, flamboyant-looking French ‘ouais’ is phonetically [wɛ] (not so far from [mm], as suggested by the clustering projection); and both Mandarin 嗯, romanized as ‘en’ and Korean 응, romanized as ‘eung’, often are phonetic [m]. Clustering projections make visible these similarities in ways that orthographies do not. (Below we note why a retreat to full-on IPA would not solve this.)

Quality control. Our results also make visible some of the challenges of working with diverse corpora, including field recordings. For instance, we include the plots of Japanese and Sambas to show that this visualization enables a quick diagnosis of problems. For Japanese, spectrograms reveal truncated segments, meaning that original timestamps do not correspond to utterance boundaries and clustering solution should not be taken at face value. For Sambas, we’re essentially classifying types of noise, as can be seen from the spectrograms (a small set of similarly noisy tokens is usefully pulled apart from the main distribution in Siwu). Visualizing spectrograms along with clusters enables the rapid visual identification of possible problems and can therefore double as a quality control method.

4.1. Limitations

The methods pioneered here are preliminary and come with a number of limitations.

Noise. Field recordings are anything but pristine studio recordings. The levels of background sound present in these datasets impact clustering (see e.g. Pite Saami, Siwu spectrograms). Additional noise reduction steps or filtering based on noise levels may improve results here.

Timing. Another issue is inaccurate timestamps which can cause unreliable clustering results. Large-scale timing issues can be spotted easily (as in Japanese). Forced alignment may be a helpful to check and improve timing accuracy at scale [37].

Overlap. Against our better judgement, we removed response tokens that occur in overlap because this would adversely impact clustering results. Since response tokens often occur in overlap, this means we exclude 25 to 45% of response tokens per corpus. Advances in speech separation [38] may provide a solution, though most of these methods remain untested with everyday conversational data.

Untranscribed tokens. Spot checks revealed that occassionally, minimal response tokens are not transcribed at all, either because they are ’drowned’ in overlap or so minimal that they stayed under the radar of the transcribers. We see a role here for specialist ASR methods like spoken term detection [39].

5. Discussion

Our aim in this paper has been to enrich the data-driven study of response tokens within and across languages by contributing methodological and conceptual tools.

Methodologically, we propose that sequential rather than token-based search methods are crucial to ensure comparability. Given structurally comparable sets of tokens, the next methodological challenge is to find new ways to characterize their structure and variation. For this, we find that new dimensionality reduction techniques can help to visualize rich conversational data for quality control and for analytical purposes.

Conceptually, a key challenge is how to reconcile the discreteness of orthographic representations with the continuity of actual speech signals. We submit that a fruitful way to deal with this is to enable analysts to fluidly navigate between type-level abstraction and token-level precision. Surface transcriptions are informative of conventionalized linguistic resources; at the same time, their actual realizations are much more variable and gradient.

It is worthwhile to think about other solutions to the problem of characterizing structure and variation in response tokens (and more generally, interactional tools). Wouldn’t fine-grained phonetic transcription address some of the problems of orthographic idiosyncracies? While phonemizers can make a useful contribution here [40, 41] , we submit that the crucial question is not so much how to find the single perfect representation (which does not exist) but how to navigate between types and tokens, between generalization and precision, between social action formats and the sound shapes that are their vehicles.

6. Conclusion

Response tokens are a microcosm of human action coordination. We rely on them every minute to streamline social interaction. The work they do becomes visible when studying conversation across languages. It also becomes visible, in photo negative, when we see interactive interfaces struggle with them, ignore them, or erase them. This means that getting at response tokens is critical for progress in human speech technology.

The endeavour to chart structure and variation of social action formats and their linguistic implementations across diverse languages has just begun. We hope that the methods pioneered here may serve as a stepping stone towards untangling human interactional infrastructure. The results will further our scientific understanding of human interaction, and can help inform a next generation of human speech technologies, putting them within reach of people around the globe.

7. Acknowledgements

Funding for the work reported here comes from Dutch Research Council grant NWO 016.vidi.185.205 to MD.

8. References

[1] J. B. Bavelas, L. Coates, and T. Johnson, “Listeners as co-narrators,” *Journal of Personality and Social Psychology*, vol. 79, no. 6, pp. 941–952, 2000.

[2] V. Yngve, “On getting a word in edgewise,” in *Papers from the Sixth Regional Meeting, Chicago Linguistic Society*, 1970, pp. 567–578.

[3] L. F. D’Haro, Z. Callejas, and S. Nakamura, Eds., *Conversational Dialogue Systems for the Next Decade*, ser. Lectures Notes in Electrical Engineering. Springer, 2021.

[4] M. Dingemanse and A. Liesenfeld, “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 (Volume 1: Long Papers)*. Dublin: Association for Computational Linguistics, 2022, pp. 5614–5633.

[5] N. Ward, "Non-lexical conversational sounds in American English," *Pragmatics & Cognition*, vol. 14, pp. 129–182, 2006.

[6] R. Gardner, "The Conversation Object Mm: A Weak and Variable Acknowledging Token," *Research on Language & Social Interaction*, vol. 30, no. 2, pp. 131–156, Apr. 1997.

[7] C. Howes and A. Eshghi, "Feedback Relevance Spaces: Interactional Constraints on Processing Contexts in Dynamic Syntax," *Journal of Logic, Language and Information*, vol. 30, no. 2, pp. 331–362, Jun. 2021.

[8] N. Ward and W. Tsukahara, "Prosodic features which cue backchannel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, Jul. 2000.

[9] G. Kjellmer, "Where do we backchannel?: On the use of mm, mhm, uh huh and such like," *International Journal of Corpus Linguistics*, vol. 14, no. 1, pp. 81–112, Jan. 2009.

[10] T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. G. Ward, and T. Kawahara, "Analysis and Prediction of Morphological Patterns of Backchannels for Attentive Listening Agents," in *Proceedings of the 7th International Workshop on Spoken Dialogue Systems*, 2016.

[11] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, "Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 991–995.

[12] H. Buschmeier and S. Kopp, "Adapting Language Production to Listener Feedback Behaviour," in *Proceedings of Workshop on Feedback Behaviors in Dialog*, Stevenson, WA, 2012, pp. 7–10.

[13] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, "Speech Driven Backchannel Generation Using Deep Q-Network for Enhancing Engagement in Human-Robot Interaction," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 4445–4449.

[14] R. Poppe, K. P. Truong, and D. Heylen, "Perceptual evaluation of backchannel strategies for artificial listeners," *Autonomous Agents and Multi-Agent Systems*, vol. 27, no. 2, pp. 235–253, Sep. 2013.

[15] D. Lala, P. Milhorat, K. Inoue, M. Ishida, K. Takanashi, and T. Kawahara, "Attentive listening system with backchanneling, response generation and flexible turn-taking," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Saarbrücken, Germany: Association for Computational Linguistics, Aug. 2017, pp. 127–136.

[16] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N. Ward, "Prediction and Generation of Backchannel Form for Attentive Listening Systems," in *Interspeech 2016*. ISCA, Sep. 2016, pp. 2890–2894.

[17] V. Zayats, T. Tran, R. Wright, C. Mansfield, and M. Ostendorf, "Disfluencies and Human Speech Transcription Errors," in *Proceedings of Interspeech 2019*. ISCA, Sep. 2019, pp. 3088–3092.

[18] R. Cumbal, B. Moell, J. Lopes, and O. Engwall, ""You don't understand me!": Comparing ASR results for L1 and L2 speakers of Swedish," in *Proceeding of Interspeech 2021*, 2021, pp. 4463–4467.

[19] C. Mansfield, S. Ng, G.-A. Levow, R. A. Wright, and M. Ostendorf, "Revisiting Parity of Human vs. Machine Conversational Speech Transcription," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 1997–2001.

[20] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski, "An End-to-End Conversational Style Matching Agent," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, ser. IVA '19. New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 111–118.

[21] L. Prévot, B. Bigi, and R. Bertrand, "A quantitative view of feedback lexical markers in conversational French," in *Proceedings of the SIGDIAL 2013 Conference*. Metz, France: Association for Computational Linguistics, Aug. 2013, pp. 87–91.

[22] P. M. Clancy, S. A. Thompson, R. Suzuki, and H. Tao, "The conversational use of reactive tokens in English, Japanese, and Mandarin," *Journal of Pragmatics*, vol. 26, no. 3, pp. 355–387, Sep. 1996.

[23] J. Trouvain and K. P. Truong, "Acoustic, Morphological, and Functional Aspects of "yeah/ja" in Dutch, English and German," in *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*, El Paso, TX, 2012.

[24] G.-A. Levow and S. Duncan, "Contrasting cues to verbal and non-verbal backchannels in multi-lingual dyadic rapport," in *Interspeech 2012*. ISCA, Sep. 2012, pp. 835–838.

[25] M. Zellers, "An overview of forms, functions, and configurations of backchannels in Ruruuli/Lunyala," *Journal of Pragmatics*, vol. 175, pp. 38–52, Apr. 2021.

[26] A. Liesenfeld, "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*. Waseda Institute for the Study of Language and Information, 2019, pp. 471–479.

[27] N. Williams, K. Stenzel, and B. Fox, "Parsing particles in Wa'ikhana," *Revista Linguistica*, vol. 16, no. Esp., pp. 356–382, Nov. 2020.

[28] F. Seifart, N. Evans, H. Hammarström, and S. C. Levinson, "Language documentation twenty-five years on," *Language*, vol. 94, no. 4, pp. e324–e345, 2018.

[29] A. Liesenfeld and M. Dingemanse, "Building and curating conversational corpora for diversity-aware language science and technology," in *LREC 2022*, Jun. 2022.

[30] G. Jefferson, "Caveat Speaker: Preliminary Notes on Recipient Topic-Shift Implicature," *Research on Language & Social Interaction*, vol. 26, no. 1, pp. 1–30, Jan. 1993.

[31] K. Drummond and R. Hopper, "Some Uses of Yeah," *Research on Language & Social Interaction*, vol. 26, no. 2, pp. 203–212, Apr. 1993.

[32] D. H. Zimmerman, "Acknowledgment Tokens and Speakership Incipiency Revisited," *Research on Language & Social Interaction*, vol. 26, no. 2, pp. 179–194, Apr. 1993.

[33] K. Drummond and R. Hopper, "Acknowledgment tokens in series," *Communication Reports*, vol. 6, no. 1, pp. 47–53, Jan. 1993.

[34] D. H. Zimmerman, "Horizontal and Vertical Comparative Research in Language and Social Interaction," *Research on Language & Social Interaction*, vol. 32, no. 1-2, pp. 195–203, 1999.

[35] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv:1802.03426 [cs, stat]*, Sep. 2020.

[36] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLOS Computational Biology*, vol. 16, no. 10, p. e1008228, Oct. 2020.

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

[38] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[39] N. San, M. Bartelds, M. Browne, L. Clifford, F. Gibson, J. Mansfield, D. Nash, J. Simpson, M. Turpin, M. Vollmer *et al.*, "Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages," *arXiv preprint arXiv:2103.14583*, 2021.

[40] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for Many Languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[41] M. Bernard and H. Titeux, "Phonemizer: Text to Phones Transcription for Multiple Languages in Python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, Dec. 2021.