Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

Evaluation of Automatic Speech Recognition for Conversational Speech in Dutch, English, and German: What Goes Missing?

Alianda Lopez, Andreas Liesenfeld, Mark Dingemanse
Centre for Language Studies
Radboud University, Nijmegen, The Netherlands
{ada.lopez,andreas.liesenfeld,mark.dingemanse}@ru.nl

Abstract

As voice user interfaces and conversational agents grow in importance, automatic speech recognition (ASR) encounters increasingly free-form and informal input data. Conversational speech is at once the most challenging and the most ecologically relevant type of data for speech recognition in this context. Here we evaluate the performance of several ASR engines on conversational speech in three languages, focusing on the fate of backchannels and other interactionally relevant elements of talk. We propose forms of error analysis based on ngram salience scoring that can complement default measures like word error rates (WER) and are more informative of ASR's ability to live up to the task of accurately representing real-world interaction.

1 Introduction

Conversational agents and voice-driven virtual assistants are becoming more and more integrated into our daily lives. However, users are still dissatisfied with their conversational abilities, describing them as frustrating, stilted, and unnatural (Clark et al., 2019; Moore, 2017; Kopp and Krämer, 2021). One likely reason is that most automatic speech recognition (ASR) systems are trained on carefully read monological speech (Panayotov et al., 2015; Ardila et al., 2020) rather than on free-flowing informal conversational interaction.

One of the key ways conversational speech differs from read speech is the nature of its production: planned and produced in real-time by people together. Conversation bears the traces of its dialogical origins in the form of elements like backchannels (Yngve, 1970; Fujimoto, 2007), disfluencies (Ginzburg et al., 2014; Hough and Schlangen, 2017), and other forms of speech management (Allwood et al., 1990), collateral signals (Clark, 1996) and non-lexical conversational sounds (Ward, 2006). The variety of terms in this area highlights the disparate strands of research concerned with such phenomena, and also encodes an implicit evaluation of these elements as somehow missable, marginal, or straying from the norm. Quite some work has focused on "disfluency detection", often with the goal of 'cleaning up' transcripts for use in downstream natural language understanding pipelines or for public consumption (Hough and Schlangen, 2017; Shalyminov et al., 2018; Zayats et al., 2019). However, a recent upsurge in research shows the importance of these elements as metacommunicative tools for streamlining conversation (Buschmeier and Kopp, 2018; Kosmala and Morgenstern, 2018; Dingemanse and Liesenfeld, 2022), and this is where their relevance for some ASR applications lies. For instance, interjections like *mhmm* and *uh-huh* in English serve as a cue for the speaker to continue talking, while others like *huh?* instead indicate a need for repetition or clarification — quite an important distinction to get for voice user interfaces. Likewise, items like *uh* and *um* are easily seen as irregularities to be cleaned up, but they can also do interactional work, such as signalling upcoming complexities or interactionally delicate moments (Clark and Fox Tree, 2002; Kosmala, 2020). While there are use cases for ignoring them, there are also contexts where natural language processing pipelines can benefit from keeping them available in some form (Dinkar, 2022).

The most common methods for benchmarking ASR systems are hardly relevant to conversations. The popular metric of word error rate (WER) compares ASR output against reference transcripts in terms of insertions, deletions, and substitutions. While useful, it has its limitations (Aksënova et al., 2021; Errattahi et al., 2018). For one, it gives more weight to insertions than deletions. It also does not take into account that there are different types of words, even when work on ASR transcription errors in English showed that errors are more likely to

occur for conversational interjections (Zayats et al., 2019). Indeed, some applications of WER exclude interjections because they are not well-represented in the training data in the first place (Papadopoulos Korfiatis et al., 2022). Because WER is computed at utterance level, it fails when whole utterances go missing – which is proportionally more likely for shorter utterances, one study on Swedish found (Cumbal et al., 2021). A recent error analysis of ASR performance across types of English speech shows that it fares worst for informal conversation. Furthermore, among function words, content words, and conversational words, it is the latter that cause the biggest drop in performance (Mansfield et al., 2021).

As ASR systems are stress-tested and the limitations of WER become more apparent, the need for complementary evaluation methods arises. Here, we build on the work reviewed above and provide two novel contributions. First, where most prior work has focused on English, we add two other languages. This baby step towards taking more of the world's linguistic diversity into account allows us to see to what extent prior findings generalize (Besacier et al., 2014). Second, we focus on error analysis not at the level of word classes but at the level of interactionally relevant phenomena: conversational words, self-repairs, and phonetic reductions. Both contributions are in line with our larger aim to improve human language technology through looking at linguistically diverse and ecologically valid conversational data (Bird, 2020; Birhane and Guest, 2021).

2 Data and Methods

To investigate how an ASR system processes conversational speech, we use data from English, Dutch, and German – three languages for which there are available corpora along with ASR solutions.

Human Transcripts. Human transcripts were obtained from three different conversational corpora, all of which capture natural conversations. For English, we use CallHome American English (Canavan et al., 1997), a corpus of informal telephone conversations between native speakers of American English from various places in the United States. A total of 140 recordings were used that ranged from 5 to 10 minutes in length. For Dutch, we use the IFA Dialog Video Corpus (van Son et al., 2008) of informal conversations between native Dutch speakers from different parts of the Netherlands. Transcripts follow the the Spoken Dutch Corpus format (Oostdijk, 2000). We used a total of 20 sound files with an average length of 15 minutes. For German, we use the Forschungs- und Lehrkorpus Gesprochenes (FOLK) Deutsch (Reineke and Schmidt, 2022), including 7 files of 10 to 30 minutes long. One sound file was excluded due to poor audio quality. Transcripts in all three corpora mark interjections, phonetically reduced forms, word fragments due to self-repairs and nonverbal conduct like coughs and lip smacks. We unified transcription formats to time-aligned utterance-level annotations, with nonverbal conduct and untranscribed stretches marked in "[]" and not included in our comparisons.

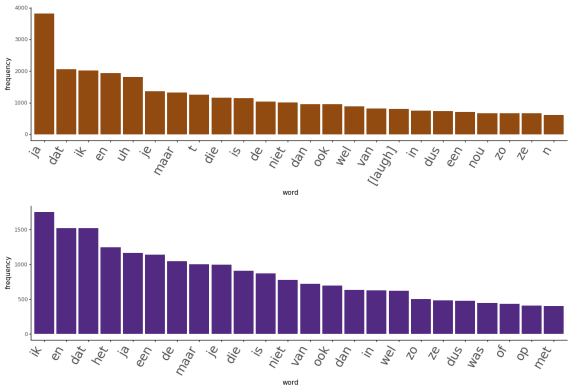

Figure 1: Most frequent words in Dutch human and ASR transcripts of conversational speech. See Appendix B for more details as well as English and German data.

ASR Transcripts. To generate ASR transcripts, we used three general purpose speech recognition engines made available through the Bavarian Archive for Speech Signals' CLARIN Transcription Portal (Draxler et al., 2020).[1] We picked these engines as examples of a class of widely available ASR solutions that are trained on large amounts of written language and that are designed to behave in a roughly comparable way: (i) emphasising textual representations over speech, and (ii) habitually removing some elements of language labeled as disfluencies. While specialist ASR solutions do exist, these general purpose engines are used in many applications and products that deal with conversational speech, such as voice assistants and social robots like Furhat (Al Moubayed et al., 2012) and Pepper (Pandey and Gelin, 2018).[2]

[1] https://clarin.phonetik.uni-muenchen.de/apps/TranscriptionPortal/

[2] *Cobalt Speech* is an example of specialist ASR engine for

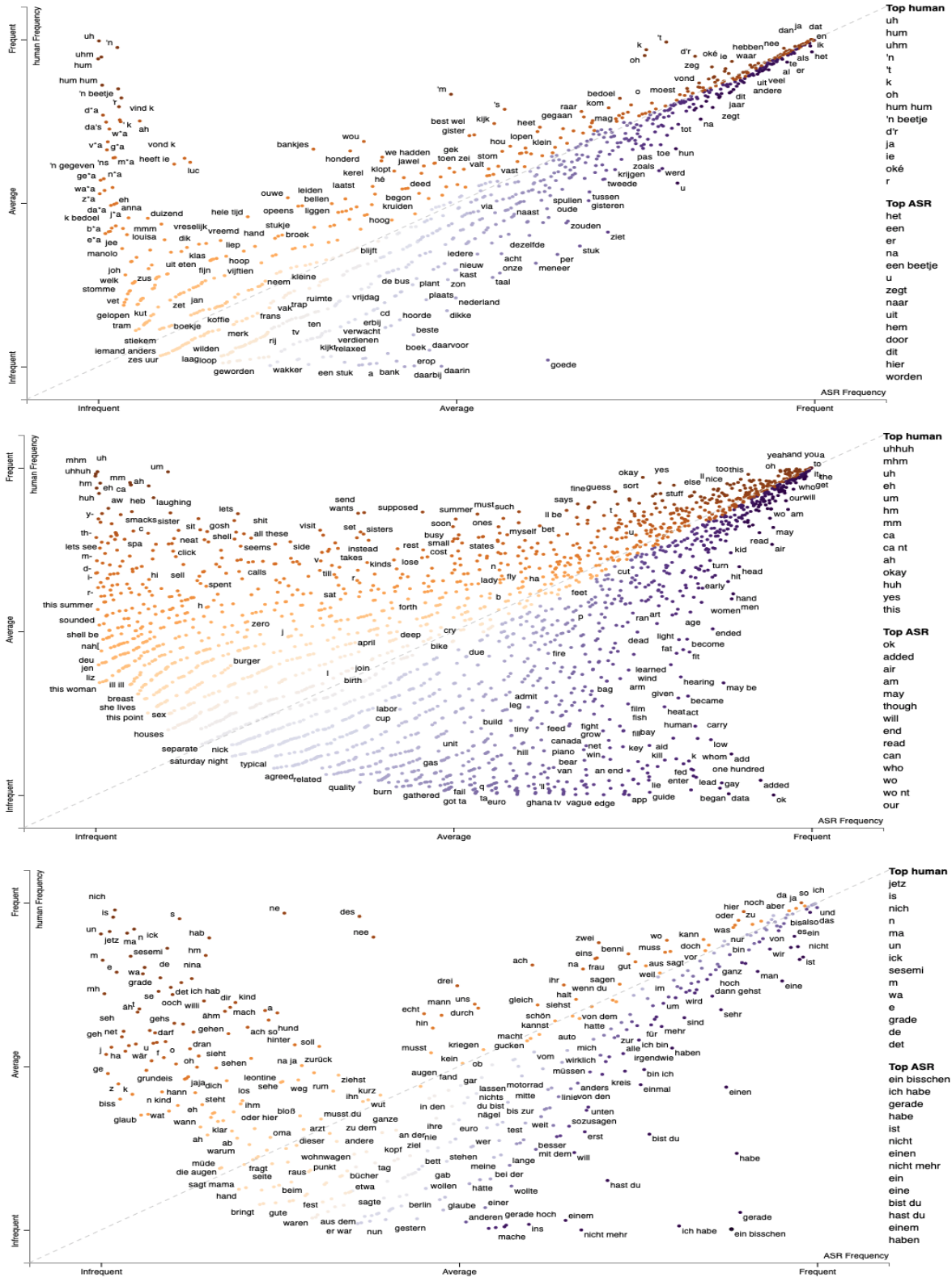

Figure 2: Most characteristic elements in human-transcribed (orange) and ASR transcribed (purple) conversational speech in Dutch, English and German, with right panels showing the top 10 most distinctive items for each type. Plotted using `scaled F score` metric using `scattertext` (Kessler, 2017).

	Dutch	English	German
Conversational Words	*uh, hum, uhm, hum hum, oh, ja*	*uhhuh, mhm, uh, eh, um, hm, mm, ah, huh, okay*	*hm, mh*
Reductions	*d'r (haar), , 'n (een), 'n beetje (een beetje), 't (het), ie (hij)*		*'n (ein), wa (wir), grade (gerade), det (das)*
Self-repairs	*k-, r-*	*m-, e-*	*se-*

Table 1: Top elements that are underrepresented (or missing) in the ASR versus human-produced transcripts. Three interactional phenomena make up most of the top 20 salient tokens by Scaled F score: short *conversational words* (this includes backchannels, response tokens, continuers, non-lexical utterances), phonetic *reductions* (including contractions), and *self-repairs* (also known as word fragments or truncated words).

2.1 Pre-Processing

Transcripts were processed to bring them to a more comparable format. This entailed removing punctuation, correcting the spelling for proper names, and removing capitalization. For the English ASR transcript, the inconsistent formats for contractions were changed to match the human transcript (i.e *can' t* to *cant*). Word fragments and shortened forms were left untouched. To further enhance comparability, tags and other special characters from the human transcripts were removed. All transcripts were then tokenized using `spaCy`'s "Core web" language models.[3]

2.2 Error Analysis

We investigate systematic differences between human-produced and ASR transcripts in the three languages. Which elements are underrepresented in ASR transcripts, and which elements go missing completely? We adopt the `scaled F-Score` introduced by Kessler (2017) as a metric of n-gram salience scoring to compare the two types of transcripts (see appendix A for details). We make the processing and error analysis pipeline available via an OSF repository as part of this paper.[4]

3 Results and Analysis

Across all languages, we find three systematic differences between human and ASR transcripts. This shows that there are indeed certain elements in conversational speech that are incongruously represented.

Shorter output text: In all cases, the ASR transcripts contained fewer words than their human counterparts with a 33% difference for Dutch, 37% for English, and 57% for German. This indicates a significant gap between how humans and ASR engines transcribe conversational speech (Scharenborg, 2007; Mansfield et al., 2021).

Skewed frequency distributions: Furthermore, the frequency distributions of the human transcripts are skewed differently from the ASR transcripts (see Figure 1).

Missing elements: The ngram salience score-based error analysis, visualized in Figure 2, revealed that the words missed by the ASR are notably similar in all the languages studied. First, the lack of conversational words in the ASR transcript indicate that current systems have difficulties picking up these short but important utterances regardless of the language. For reductions, only those in English were well detected by the ASR. This may be because Dutch and German reductions are more exclusive to conversational speech; thus occurring less frequently in written language than their English counterpart. On instances when these reductions are actually detected, the ASR then tends to transcribe them in their expanded form instead of how they were actually said. Lastly, self-repairs are completely missed too. Aside from these self-repairs also being short, they are often omitted from speech datasets as well due to their "incompleteness". However, these word fragments were nonetheless uttered and consequently still carry meaning in conversations.

These findings indicate that current general-purpose ASR engines tend to struggle with three interactional phenomena: short conversational words, reductions, and self-repair (see Table 1).

4 Limitations

We are aware of several limitations. First, the examined corpora are too small to provide a com-

conversational speech. Such products are not only few and far between, but also proprietary and expensive.

[3] https://spacy.io

[4] https://osf.io/7ts3y

prehensive overview of the missing interactional elements. It is likely that a larger dataset will help to discover even more elements that this study has missed. Next, while our analysis revealed the disparity in the representation of certain elements between human and ASR transcripts, an analysis at the utterance level will provide more insight on how and why this disparity exists (Cumbal et al., 2021). An accurate representations of conversational speech has to not only take into account what is being say, but also how it is said, which makes the task a lot harder. This may require a whole new ASR processing pipeline design (Faruqui and Hakkani-Tür, 2022; Merz and Scrivner, 2022; Wepner et al., 2022). Finally, we have not computed WER and similar measures – making it harder to relate such measures to our results (cf. Georgila et al. 2020).

5 Conclusion

Conversation is the primary ecology of natural language use (Schegloff, 2006). ASR systems are an integral part of conversational agents and any technology that deals with speech input, and they are increasingly exposed to conversational settings (Baumann et al., 2017). However, they are far from able to handle free-flowing conversations (Addlesee et al., 2020), a major cause of interactional turbulence and user dissatisfaction (Hoegen et al., 2019; Clark et al., 2019). Here we have shown that across three languages, off-the-shelf ASR solutions have trouble with quintessentially interactional phenomena like conversational words (backchannels, delay markers, and other interjections) and word fragments resulting from self-repair. Yet, it is precisely these items that people use to streamline interaction. Dealing with these items as interactional tools, rather than indiscriminately erasing them, represents the next frontier in the development of voice-driven human language technologies.

Acknowledgments

This work was supported by NWO Vidi 016.vidi.185.205.

References

Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020. A Comprehensive Evaluation of Incremental Speech Recognition and Diarization for Conversational AI. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3492–3503, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How Might We Create Better Benchmarks for Speech Recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34, Online. Association for Computational Linguistics.

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction. In *Cognitive Behavioural Systems*, Lecture Notes in Computer Science, pages 114–130, Berlin, Heidelberg. Springer.

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.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th language resources and evaluation conference*, pages 4218–4222.

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. Recognising Conversational Speech: What an Incremental ASR Should Do for a Dialogue System and How to Get There. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Lecture Notes in Electrical Engineering, pages 421–432. Springer, Singapore.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

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.

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

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. 1997. CALLHOME American English Speech. Artwork Size: 1830160 KB Pages: 1830160 KB Type: dataset.

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

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84:73–111.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA. Association for Computing Machinery.

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.

Mark Dingemanse and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin. Association for Computational Linguistics.

Tanvi Dinkar. 2022. *Computational models of disfluencies : fillers and discourse markers in spoken language understanding*. These de doctorat, Institut polytechnique de Paris.

Christoph Draxler, Henk van den Heuvel, Arjan van Hessen, Silvia Calamai, and Louise Corti. 2020. A CLARIN Transcription Portal for Interview Data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3353–3359, Marseille, France. European Language Resources Association.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science*, 128:32–37.

Manaal Faruqui and Dilek Hakkani-Tür. 2022. Revisiting the Boundary between ASR and NLU in the Age of Conversational Dialog Systems. *Computational Linguistics*, 48(1):221–232.

Donna T. Fujimoto. 2007. Listener responses in interaction: A case for abandoning the term, backchannel. *Bulletin paper of Osaka Jogakuin College*, 9(28):35–54.

Kallirroi Georgila, Anton Leuski, Volodymyr Yanov, and David Traum. 2020. Evaluation of Off-the-shelf Speech Recognizers Across Diverse Dialogue Domains. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6469–6476, Marseille, France. European Language Resources Association.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7.

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.

Julian Hough and David Schlangen. 2017. Joint, Incremental Disfluency Detection and Utterance Segmentation from Speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 326–336. Association for Computational Linguistics.

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.

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

Loulou Kosmala. 2020. Euh le saviez-vous ? le rôle des (dis)fluences en contexte interactionnel : étude exploratoire et qualitative. *SHS Web of Conferences*, 78:01018. Publisher: EDP Sciences.

Loulou Kosmala and Aliyah Morgenstern. 2018. Should 'uh' and 'um' be categorized as markers of disfluency? The use of fillers in a challenging conversational context. In *Fluency and Disfluency across Languages and Language Varieties*.

Courtney Mansfield, Sara Ng, Gina-Anne Levow, Richard A. Wright, and Mari Ostendorf. 2021. Revisiting Parity of Human vs. Machine Conversational Speech Transcription. In *Interspeech 2021*, pages 1997–2001. ISCA.

Megan Merz and Olga Scrivner. 2022. Discourse on ASR Measurement: Introducing the ARPOCA Assessment Tool. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 366–372, Dublin, Ireland. Association for Computational Linguistics.

Roger K. Moore. 2017. Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Lecture Notes in Electrical Engineering, pages 281–291. Springer, Singapore.

Nelleke Oostdijk. 2000. The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings of the second international conference on language resources and evaluation (LREC'00)*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Amit Kumar Pandey and Rodolphe Gelin. 2018. A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind. *IEEE Robotics & Automation Magazine*, 25(3):40–48. Conference Name: IEEE Robotics & Automation Magazine.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A Dataset Of Primary Care Mock Consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Silke Reineke and Thomas Schmidt. 2022. Das Archiv für Gesprochenes Deutsch und das Forschungs-und Lehrkorpus für Gesprochenes Deutsch. In *Sprache in Politik und Gesellschaft*, pages 323–330. de Gruyter.

Odette Scharenborg. 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347.

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.

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2018. Multi-Task Learning for Domain-General Spoken Disfluency Detection in Dialogue Systems.

Rob van Son, Wieneke Wesseling, Eric Sanders, and Henk van den Heuvel. 2008. The IFADV corpus: A free dialog video corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

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

Saskia Wepner, Barbara Schuppler, and Gernot Kubin. 2022. How prosody affects ASR performance in conversational Austrian German. pages 195–199.

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

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.

6 Appendix

A `Scaled F-score`: measuring ngram salience by class

`Scaled F-score` is a modified version of the vanilla F-score calculated by taking the harmonic means of precision and frequency. Given a word $w_i \in W$ and a category $c_j \in C$, the precision of word w_i with respect to a category c_j is defined as the following:

$$\text{prec}(i,j) = \frac{\#(w_i, c_j)}{\sum_{c \in C} \#(w_i, c)}$$

The function $\#(w_i, c_j)$ represents either the number of times w_i occurs in an utterance labeled with the category c_j or the number of utterances labeled c_j which contain w_i. The frequency of a word within a category is defined as:

$$\text{freq}(i,j) = \frac{\#(w_i, c_j)}{\sum_{w \in W} \#(w, c_j)}$$

Then, the harmonic mean of these two values is defined as:

$$\mathcal{H}_\beta(i,j) = (1+\beta^2)\frac{\text{prec}(i,j) \cdot \text{freq}(i,j)}{\beta^2 \cdot \text{prec}(i,j) + \text{freq}(i,j)}$$

$\beta \in \mathcal{R}^+$ is a scaling factor where frequency is favored if $\beta < 1$, precision if $\beta > 1$, and both are equally weighted if $\beta = 1$. F-score is equivalent to the harmonic mean where $\beta = 1$.

This score is then modified in two ways to address two issues, namely that (1) harmonic means are dominated by precision, and that (2) low scores are "low-frequency brittle terms". In short, the Scaled F-Score aims to better take into account tokens of extremely high and low token frequencies and balances the score to this end. On a scale from -1 to 1, the score indicates whether an n-gram exhibits an association with a class (positive score) or not (negative score). For a more detailed explanation of these modification, see: https://github.com/JasonKessler/scattertext#understanding-scaled-f-score

B Word Frequency distributions in human versus ASR transcripts

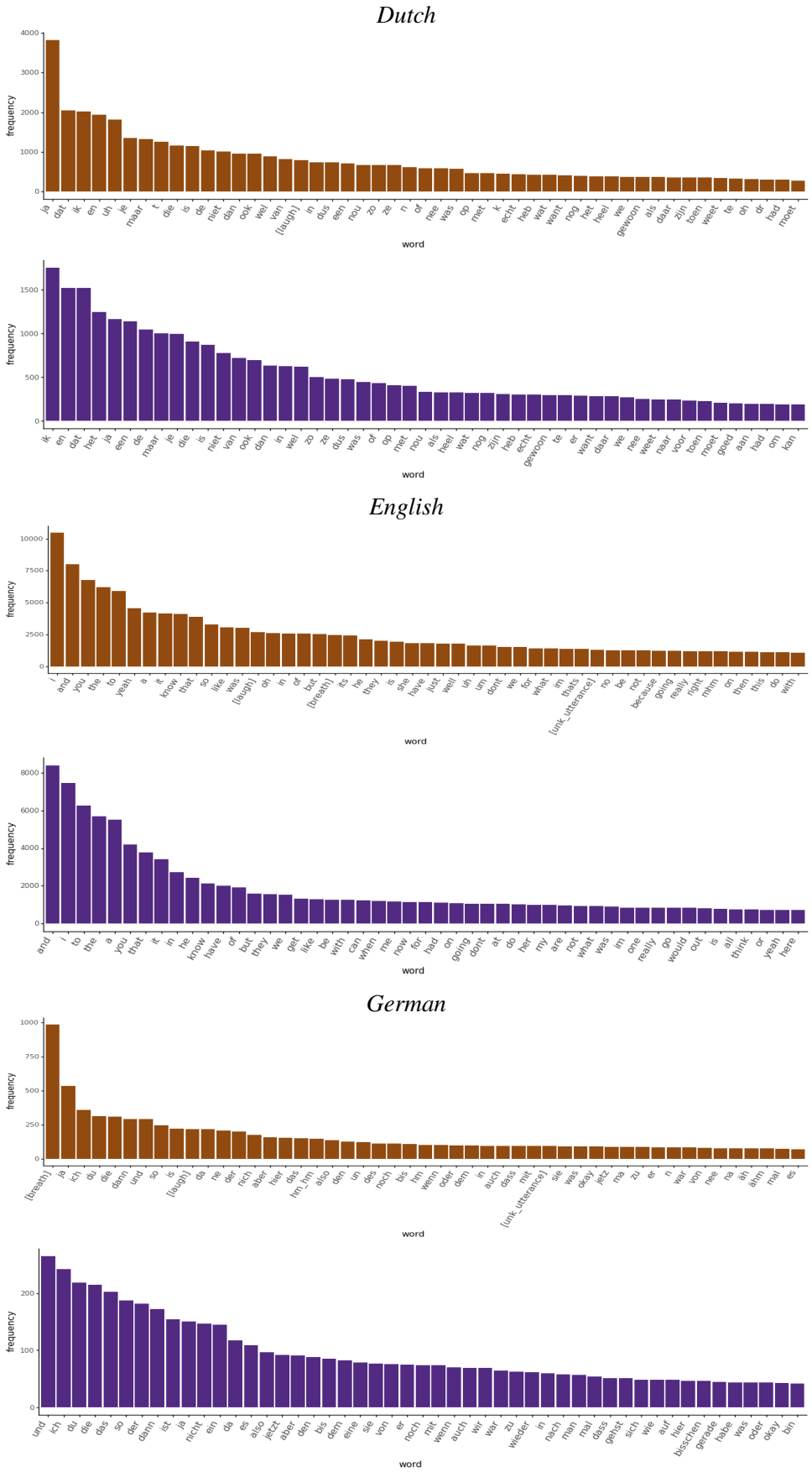

Figure 3: Most frequent words in Dutch, English, and German human (orange) and ASR (purple) transcripts of conversational speech.