Putting Natural in Natural Language Processing

Grzegorz Chrupała
Department of Cognitive Science and Artificial Intelligence
Tilburg University
grzegorz@chrupala.me

Abstract

Human language is firstly spoken and only secondarily written. Text, however, is a very convenient and efficient representation of language, and modern civilization has made it ubiquitous. Thus the field of NLP has overwhelmingly focused on processing written rather than spoken language. Work on spoken language, on the other hand, has been siloed off within the largely separate speech processing community which has been inordinately preoccupied with transcribing speech into text. Recent advances in deep learning have led to a fortuitous convergence in methods between speech processing and mainstream NLP. Arguably, the time is ripe for a unification of these two fields, and for starting to take spoken language seriously as the primary mode of human communication. Truly natural language processing could lead to better integration with the rest of language science and could lead to systems which are more data-efficient and more human-like, and which can communicate beyond the textual modality.

1 Introduction

The ACL 2023 theme track urges the community to check the reality of the progress in NLP. This position paper adopts an expansive interpretation of this question. It is definitely worth inquiring into the apparent advances of current NLP in their own terms. Here, however, I question these terms and argue that our field has focused on only a limited subset of human language which happens to be convenient to work with, and thus misses major aspects of human communication.

1.1 Human Language is Primarily Spoken

Humans are an exceptional species in many ways, and out of these, human language is one of the most salient. Unlike communication systems used by other organisms, human language is open-ended, capable of expressing abstract concepts, and of reference to events displaced in time and space. While the capacity to acquire language is universal and largely innate (Darwin, 1874; Pinker and Bloom, 1990) it also is culturally mediated and likely arose via gene-culture co-evolution (Deacon, 1998; Richerson and Boyd, 2010).

One revolutionary technology which turbocharged human language was writing, which was invented a handful of times in the most recent few thousand years of the human story (Fischer, 2003). Writing, followed by the printing press, followed by the Internet, have made written text ubiquitous to the extent that it is easy to forget that the primary and universal modality for most human communication throughout history has been spoken.[1]

Even today many of the world's languages do not have a standardized written form. For those that do, the written modality originated as a compressed, symbolic representation of the spoken form.

Children acquire a spoken language (and not infrequently two or more) within the first few years of their life with no or little explicit instruction, largely relying on weak, noisy supervision via social interaction and perceptual grounding. In contrast, they require hundreds of hours of explicit instruction and arduous conscious practice to learn to read and write, and most are only able to learn the written modality a couple of years at best after becoming fluent communicators in one or more spoken languages.

1.2 Reality check

Thus, arguably, the natural language for which we are biologically equipped is spoken. Written language is a secondary development, which happens to be very useful and widespread, but is nevertheless derivative of speech. This appears to be the

[1] I am using *spoken language* in the broad sense here, including both the oral and gestural (signed) modes of expression, and opposing these to the written modality.

Findings of the Association for Computational Linguistics: ACL 2023, pages 7820–7827
July 9-14, 2023

consensus view in linguistics going back at least a century (de Saussure, 1916; Bloomfield, 1933).[2]

Given these facts, is then the field of Natural Language Processing (NLP) a misnomer? Are we making less progress with getting machines to communicate via human language than current advances with processing written text would have us believe?

2 NLP is Written Language Processing

To anyone with experience reading, reviewing and publishing papers in NLP conferences and journals (such the ACL conferences and TACL) it is evident that the field is very strongly focused on processing written language. While this is evident to practitioners, it is also largely tacit and implicit.

2.1 Unstated assumptions

The fact that a paper is concerned with written as opposed to spoken oral or sign language is almost invariably assumed to be the default and not explicitly stated. Furthermore, even if there is some interest in tackling a dataset of originally spoken language (for example in much work on dialog and child language acquisition), the usual approach is to use a written transcription of this data rather than the actual audio. This is partly a matter of convenience, but partly due to the assumption that the written form of language is the canonical one while the audio modality is just a weird, cumbersome encoding of it.

To some extent such an implicit belief also lurks in much work within the speech community: the main thrust of speech research has always been on so called Automatic Speech Recognition (ASR), by which is meant automatically transcribing spoken language into a written form. Written text is treated as an interface and an abstraction barrier between the field of speech processing and NLP. In Sections 3 and 4 I address problems arising from the above assumptions, as well as the challenges and opportunities we have once we discard them. Firstly, however, it will be instructive to briefly quantify the assertion that NLP is Written Language Processing. by looking at historical publication patterns.

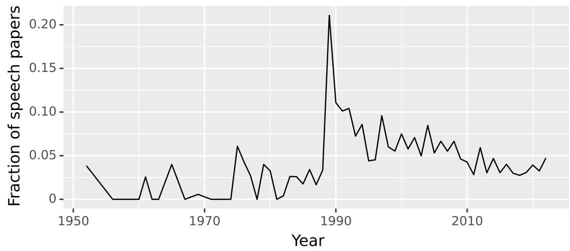

Figure 1: The proportion of papers in the ACL anthology up to year 2022 which mention the words *speech, spoken* or *audio* in the title, excluding those with *part(s)-of-speech* or *speech act(s)*.

2.2 Publication patterns

Figure 1 shows the proportion of NLP papers explicitly mentioning speech-related terms in their title over the years covered by the ACL anthology (1950 through 2022), which is a comprehensive database of NLP papers from a wide variety of relevant conferences, workshops and journals.[3] The fraction of speech-focused NLP papers varies quite a bit over the years, but mostly stays below 10%. There is a large peak going to 20% in 1989, followed by three years with around 10% of speech papers. A look at the underlying data reveals that the 1989 peak is associated with the inclusion in the anthology of the proceedings of the Speech and Natural Language Workshop (Hirshman, 1989) organized by the US Defense Advanced Research Projects Agency (DARPA), and featuring 79 papers. This workshop ran until 1992 and is thus largely responsible for the four-year run of sizable representation of spoken language research in the ACL anthology.

The overview of the last edition of this event notes the then ongoing *"paradigm shift in natural language processing towards empirical, corpus based methods"* (Marcus, 1992). It is likely that this shift in NLP methodology was at least partly driven by this workshop, the associated DARPA program, and the resulting increased interaction between researchers working on spoken and written language.

In recent years (since 2010) the proportion of NLP papers explicitly mentioning spoken language has resolutely stayed below 6%. While the major ACL events typically include speech processing as a topic in their calls for papers, as well as a

[2]However see Aaron and Joshi (2006) for a dissenting view.

[3]https://aclanthology.org/

track including the term *speech* in its name, such as *Speech and Multimodality*, processing of spoken language it clearly a rather minor concern of these conferences. Instead, speech work is published in different venues organized by a separate speech processing community.

3 Spoken Language is Richer

While the primacy of the spoken modality as means of communication is the consensus view in linguistics, Section 2.1 identifies unstated assumptions among NLP practitioners which amount to the opposite view. Here I outline why these assumptions contradicting the scientific view are not only incorrect but also detrimental to progress on understanding and processing real human language.

3.1 Key features of spoken language

Speech and writing are two different modalities with different affordances, and there is no straightforward mapping between them. Some writing systems such as those used for English, Arabic or Chinese do not even represent the phonology of the spoken language in a direct way. More crucially, writing only captures a small proportion of the information carried in the equivalent audio signal. Writing discards most of the information falling within the general category of paralinguistic phenomena, such as that related to speaker identity, speaker emotional state and attitude; likewise, information conveyed by speech tempo and amplitude, including most of suprasegmental phonology such as intonation and rhythm is typically not present in writing. In addition to the auditory signal, oral spoken language can also feature visual clues in the form of accompanying gestures, facial expressions and body posture. Sign languages rely on the visual channel exclusively, and in fact there are no widely used writing systems for any of them (Grushkin, 2017). Unlike most text, speech also typically contains a variable amount of channel noise (Shannon, 1948) such as environmental sounds.

Natural spontaneous speech contains fillers, hesitations, false starts, repairs and other disfluencies (Dinkar et al., 2023) which are usually edited out in the written form of language. Even more critically, spontaneous speech typically takes the form of a dialog between two or more participants. Dialog is unlike common written genres: crucially it features turn-taking behavior which is governed by complex and incompletely understood rules (Skantze, 2021). These features of natural dialog also mean that the traditional cascaded approach of ASR followed by NLP faces serious limitations, not least due to low ASR performance in this regime (Szymański et al., 2020), but also due to its inherently interactive nature.

For all these reasons, spoken language is more informationally rich than written language;[4] the same factors also make it more variable, complex and noisy, and consequently more challenging for automated processing (Shriberg, 2005). Thus any understanding of language as a human faculty gained via the written modality does not necessarily generalize to the spoken modality. The same is also the case about language applications: for example the successes and shortcomings of state-of-the-art text chatbot systems (e.g. Stiennon et al., 2020) are likely to be substantially different from those of spoken dialog systems.

3.2 Challenges of speech

As an illustrative example, let us consider the effectiveness of self-supervision: inducing representations of words and phrases from just listening to speech or reading text. For text, this general family of methods has been successful since around the time of Latent Semantic Analysis (Dumais, 2004), and currently large written language models exhibit a constantly expanding range of abilities (Wei et al.). In contrast, self-supervision with spoken language has met with a limited amount of success only in the last few years (e.g. Baevski et al., 2020; Hsu et al., 2021), and these models as of now are usually only fine-tuned on the task of ASR. One obvious difference is that items such as words and morphemes are either explicitly delimited or easily discovered in text, but finding them is an unsolved research problem in speech, due to the inherent variability of this modality.

On the other hand, learning spoken language becomes much more tractable when self-supervision is augmented with grounding in perception. The cross-modal correlations, though unreliable and noisy, are often sufficient to substantially facilitate the discovery and representation of words (Peng and Harwath, 2022; Nikolaus et al., 2022) and syllables (Peng et al., 2023) in spoken language. For written language, grounding in the visual modality

[4] One exception to this general pattern is the presence of two spatial dimensions in written language, and the role of 2D layout in textual publications.

has also been found to help in some cases (e.g. Tan and Bansal, 2020) but it does not appear crucial, as the dominance of text-only language models demonstrates.

Since spoken language is richer in information content, it should in principle be possible to exploit this extra signal for improving performance. One obstacle to such developments is the increased variability and channel noise. Perhaps less obviously, a second obstacle is that widely used benchmarks are often designed in a way which obstructs obtaining such gains. For example the 2021 Zerospeech challenge (Dunbar et al., 2021) which aimed to benchmark spoken language modeling, evaluates systems according to the following criteria: phoneme discrimination, word recognition, syntactic acceptability and correlation to human judgments of word similarities. None of these metrics would benefit much from modeling speaker characteristics, speech tempo, pitch, loudness or even suprasegmental phonology. Except for the first one, these metrics would be very well suited for models trained exclusively on written language. The combined effect of these two obstacles was evident in the results of Zerospeech 2021 where written-language toplines, such as RoBERTa (Liu et al., 2019), outperformed spoken language models on the latter three metrics, often by large margins.

4 Unifying Speech Processing and NLP

As evident from the examples highlighted above, spoken language is in some ways quite different from written language and presents a distinct set of challenges and potentials. In order to understand how much progress the fields of speech and NLP are making in understanding and implementing human language, we need to take speech seriously *qua* language, not just a cumbersome modality, and measure our progress accordingly.

4.1 Converging methodology

The time is ripe for a closer integration of the speech and NLP communities and for a unified computational science of language. The set of methodologies used in speech and text processing used to be quite distinct in the past. Since the adoption of deep learning both fields have converged to a large extent: currently the state-of-the-art models for both spoken and written language rely on transformer architectures (Vaswani et al., 2017) self-trained on large amounts of minimally preprocessed data, with optional fine-tuning. The technical communication barriers across disciplinary boundaries are thus much lower. The recent emergence of the concept of *textless NLP* (Lakhotia et al., 2021) exemplifies the potential of unifying these two fields.

4.2 Opportunities

The following paragraphs outline the most important benefits of making NLP more natural, ranging from basic science to practical applications.

Modeling language acquisition. An increased attention to spoken language within NLP has the potential to lead to a more realistic understanding of how well our current methods can replicate key human language abilities. Acquiring language under constraints that human babies face is the big one. There is a large amount of work on modeling human language acquisition which uses exclusively written data (at best transcribed from the original audio). Hopefully by this point the reader will be convinced that the relevance of this work to the actual issue under consideration is highly questionable. We stand a much better chance of figuring out human language acquisition if we refocus attention on spoken language.

Data efficiency. Linzen (2020) argues convincingly for language models which are human-like in their data-efficiency and generalization capabilities. It is, however, unclear whether these properties can even be properly evaluated via the medium of written language. Since the informational density and the signal-to-noise ratio in written vs spoken language are so very different, it makes little sense to compare human children with language models trained on text. Furthermore, the challenges of pure self-supervision may motivate us to take seriously the impact of grounding in perception and interaction, which humans use universally as a learning signal.

Unwritten languages. Many modes of human communication lack standard written representation. These range from major languages spoken by millions of people such as Hokkien (Mair, 2003), to small or non-standard language varieties, to sign languages. Shifting the emphasis of NLP research from text to the primary, natural oral and gestural modalities will benefit the communities using these varieties.

Spoken dialog systems. Dingemanse and Liesenfeld (2022) argue that language technology needs to transition from the text to talk, and provide a roadmap of how to harness conversational corpora in diverse languages to effect such a transition. Indeed, one of the most obvious benefits of spoken language NLP would be dialog systems that do not need to rely on ASR and are able to exploit the extra information lost when transcribing speech, enabling them to understand humans better and interact with them in a more natural way.

Non-textual language data. Finally, there is a large and increasing stream of non-textual language data such as podcasts, audio chat channels and video clips. Processing such content could also benefit from an end-to-end holistic treatment without the need of going through the lossy conversion to text.

4.3 Recommendations

If you are an NLP practitioner and view spoken language as outside the scope of your field, reconsider. Getting into speech processing does require understanding its specifics, but it is not as technically daunting as it used to. Conversely, if you are a speech researcher, consider that ASR and text-to-speech is not all there is: we can get from sound to meaning and back without going through the written word. Both fields would do well to consider the whole of human language as their purview. Increased collaboration would benefit both communities, and more importantly, would give us a chance of making real progress towards understanding and simulating natural language.

5 Limitations

The main limitation of this paper is the one applying to any opinion piece: it is subjective and personal, as the views of the authors are inherently limited by their expertise and experience. More specifically, this paper argues for an increased interaction between the speech and NLP communities, but the author is more strongly embedded in the latter, and thus addresses this audience primarily. Additionally, the short paper format imposes significant constraints on the amount of nuance, detail and discussion of relevant literature, and thus readers may find some of the claims to be less strongly supported and less hedged than would be ideal, or proper in a longer treatment of this topic.

Acknowledgements

I would like to thank Hosein Mohebbi, Afra Alishahi, Mark Dingemanse, Tanvi Dinkar, Piotr Szymański and three anonymous reviewers for their valuable feedback on this paper.

References

P. G. Aaron and R. Malatesha Joshi. 2006. Written language is as natural as spoken language: A biolinguistic perspective. *Reading Psychology*, 27(4):263–311.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Leonard Bloomfield. 1933. *Language*. Henry Holt, New York.

Charles Darwin. 1874. *The descent of man, and selection in relation to sex*. D. Appleton and Company, New York.

Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Paris.

Terrence W Deacon. 1998. *The Symbolic Species: The Co-evolution of Language and the Brain*. WW Norton & Company.

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, Ireland. Association for Computational Linguistics.

Tanvi Dinkar, Chloé Clavel, and Ioana Vasilescu. 2023. Fillers in spoken language understanding: Computational and psycholinguistic perspectives. *Traitement Automatique des Langues*, 63(3).

Susan T Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST)*, 38:189–230.

Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, and Emmanuel Dupoux. 2021. The Zero Resource Speech Challenge 2021: Spoken language modelling.

Steven Roger Fischer. 2003. *History of writing*. Reaktion books.

Donald A Grushkin. 2017. Writing signed languages: What for? What form? *American annals of the deaf*, 161(5):509–527.

Lynette Hirshman. 1989. Overview of the DARPA speech and natural language workshop. In *Proceedings of the workshop on Speech and Natural Language*, pages 1–2.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Victor H Mair. 2003. How to forget your mother tongue and remember your national language.

Mitchell P Marcus. 1992. Overview of the fifth DARPA speech and natural language workshop. In *Proceedings of the workshop on Speech and Natural Language*, pages 3–4.

Mitja Nikolaus, Afra Alishahi, and Grzegorz Chrupała. 2022. Learning English with Peppa Pig. *Transactions of the Association for Computational Linguistics*, 10:922–936.

Puyuan Peng and David Harwath. 2022. Word Discovery in Visually Grounded, Self-Supervised Speech Models. In *Proc. Interspeech 2022*, pages 2823–2827.

Puyuan Peng, Shang-Wen Li, Okko Räsänen, Abdelrahman Mohamed, and David Harwath. 2023. Syllable discovery and cross-lingual generalization in a visually grounded, self-supervised speech mode. In *Proc. Interspeech 2023*.

Steven Pinker and Paul Bloom. 1990. Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4):707–727.

Peter J Richerson and Robert Boyd. 2010. Why possibly language evolved. *Biolinguistics*, 4(2-3):289–306.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Elizabeth Shriberg. 2005. Spontaneous speech: how people really talk and why engineers should care. In *Proc. Interspeech 2005*, pages 1781–1784.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

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

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5

☒ A2. Did you discuss any potential risks of your work?
It's a position paper and does not propose or implement any particular method.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.