

# Putting *Natural* in NLP

Grzegorz Chrupała

*ML<sup>2</sup>*

Multimodal Language Learning Lab



# Reality Check

- Computational Linguistics (aka NLP) misses crucial aspects of human communication.
- We work on what is convenient, rather than on what is important.

# Human language is primarily spoken\*

- Capacity to acquire spoken (including sign) language is **universal** and **innate**.
- Writing was invented only a few times in the last **few thousand years**.
- Children learn to read and write the **hard way**.
- Most languages **lack** a standard and widely used written form.

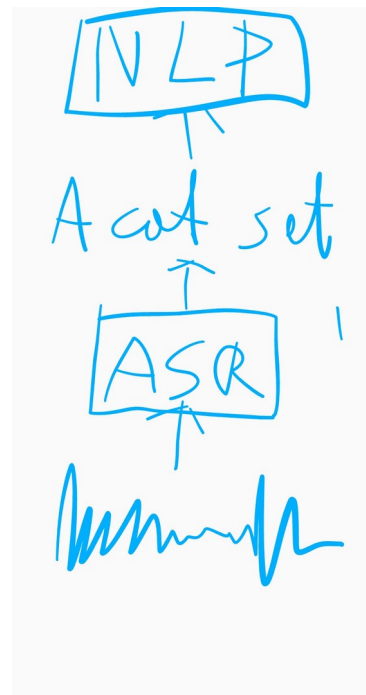


\* oral or signed

# NLP is Written Language Processing

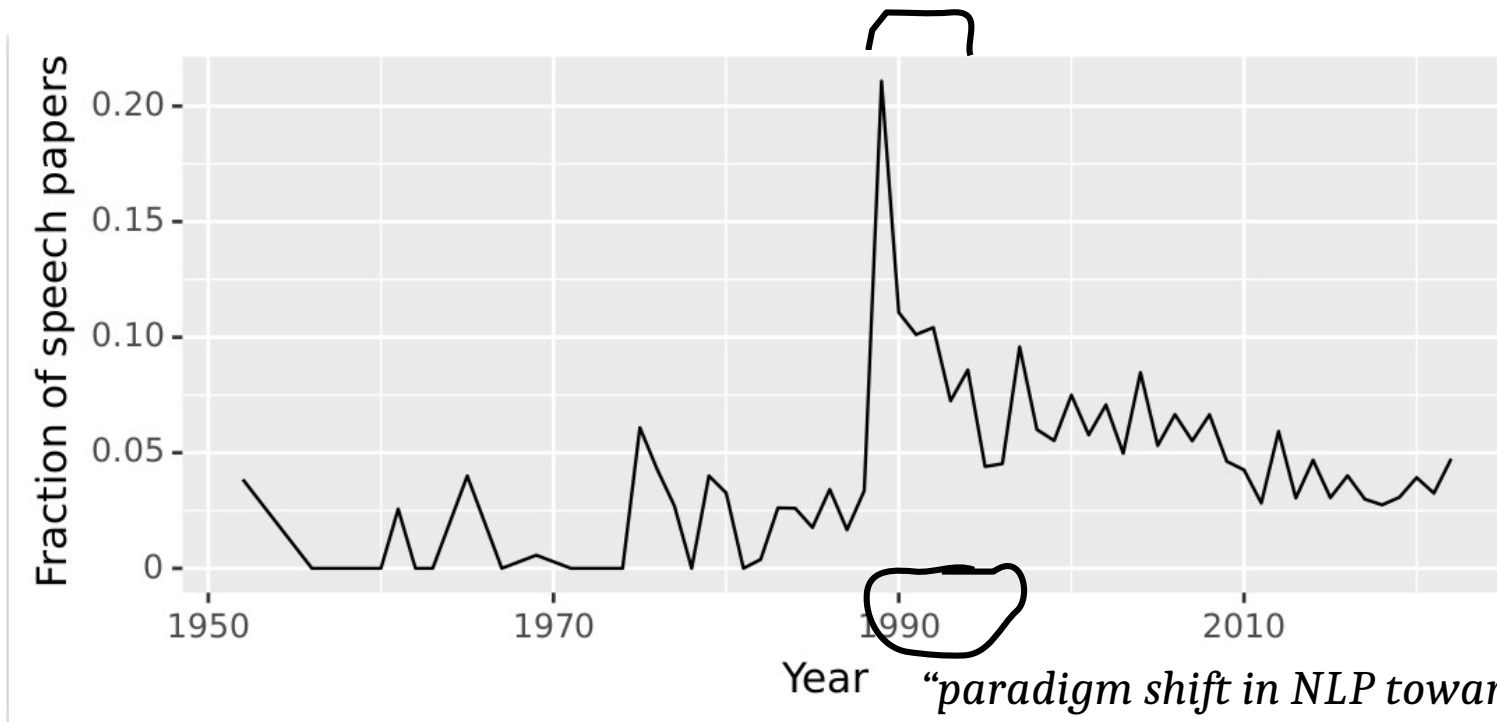
Unstated assumptions about nature of language within NLP and speech processing:

- Text is the default, canonical form of language.
- Speech is a cumbersome encoding which needs to be converted to text to be useful.



# Publication patterns in NLP

US DARPA Speech and Natural Language Workshop



*“paradigm shift in NLP towards empirical, corpus based methods” (Marcus, 1992)*

# Key features of spoken languages contrasted with text

# Information carried by speech

- Semantic and pragmatic content
  - Partly encoded in suprasegmental phonology (intonation)
- Speaker identity and characteristics
  - Age, sex, regional dialect, sociolect
- Speaker emotional state, attitude
  - Partly encoded in the visual modality (gestures, facial expression)



# Features of spoken communication

- Channel noise:  
overlapping speech,  
environmental sounds
- Fillers, hesitations, false  
starts, repairs
- Dialog: turn taking  
behavior





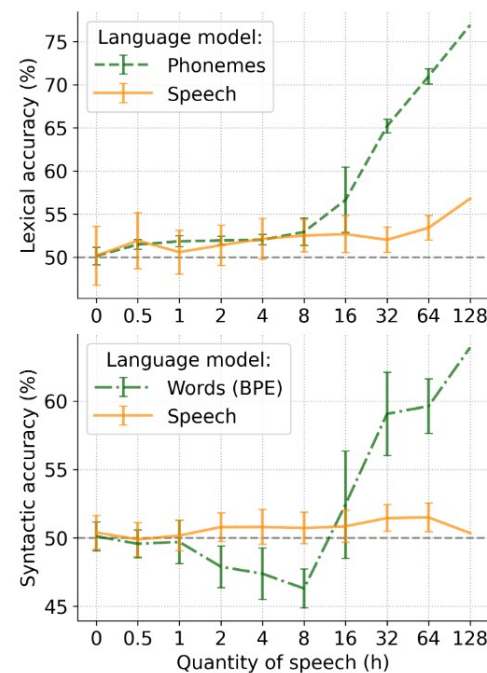
# Spoken vs written communication

- Spoken language carries much more information
- It is also more variable, complex and noisy
- **Knowledge** learned by studying **text** does not carry over to **talk**
- **Applications** for **text** and for **talk** have different challenges

# Challenges of speech

- From **LSA** to **ChatGPT**
  - self supervision for text just works.
- Self-supervision for speech – limited progress.

BabySLM





**Marvin Lavechin** @LavechinMarvin · Jun 5



Besides, turns can overlap across speakers, and people may under-articulate, mumble, shout, whisper, sing, or laugh while speaking!

▶ Naturalistic speech is highly variable. Unlike humans, LMs struggle to normalize the signal across non-linguistic information.



1



5

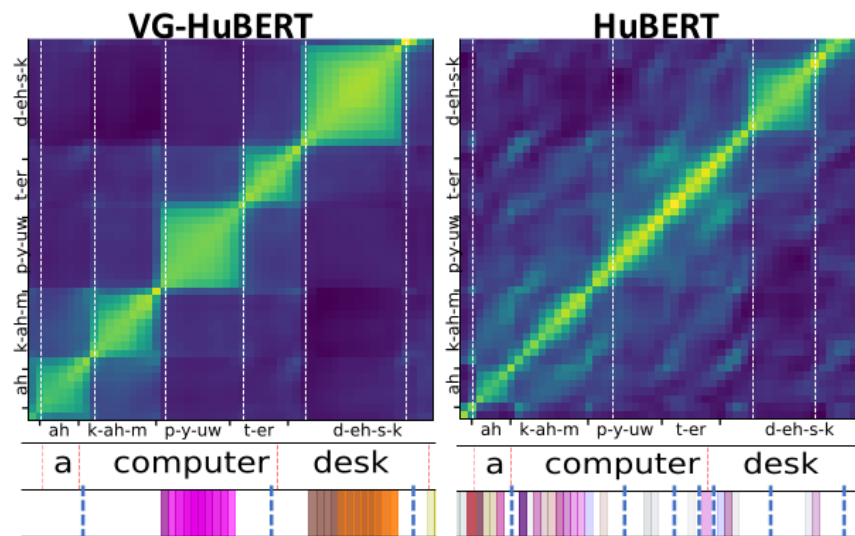


160



# Grounding

- Grounding speech in vision enables discovery words and syllables.
- Grounding text doesn't seem as crucial.



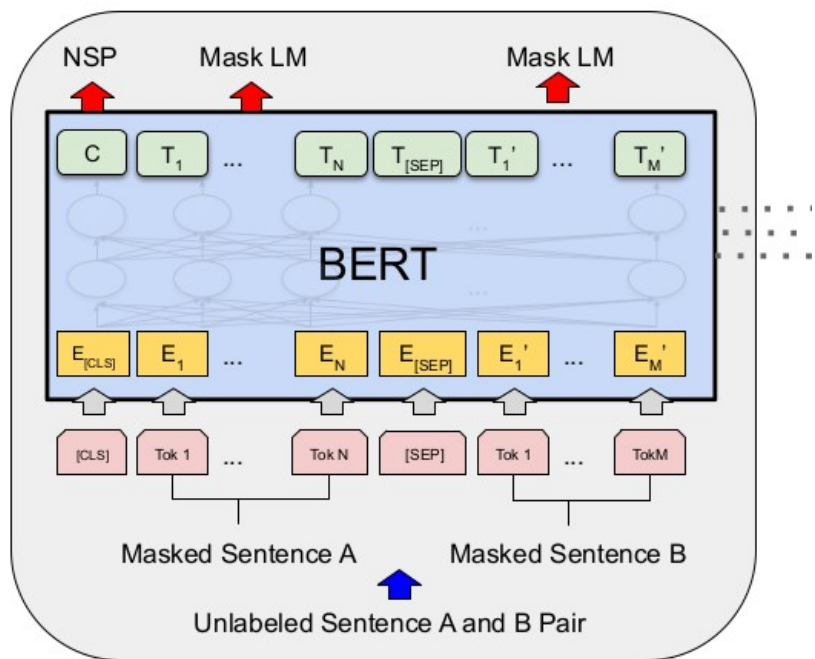
Peng et al 2023

<https://arxiv.org/abs/2305.11435>

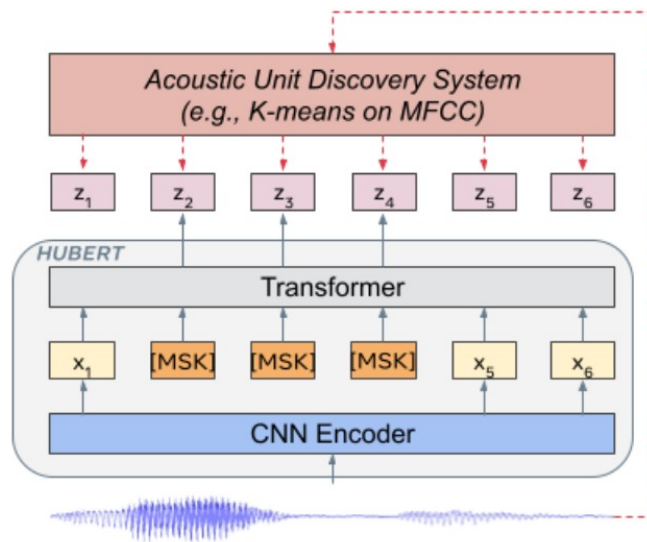
# Harnessing the richness of speech

- We'd like to make use of all the **extra information** conveyed in the speech signal.
- Noise and variability are obstacles
- But, **proper** evaluation also crucial.
  - Zerospeech 2021: phoneme discrimination, word recognition, syntactic acceptability, correlation to human judgments of word similarities
  - **Not clear** how speech-specific info can help with the above.

# Methodological convergence



BERT



HuBERT

# Opportunities of unifying NLP and Speech Processing

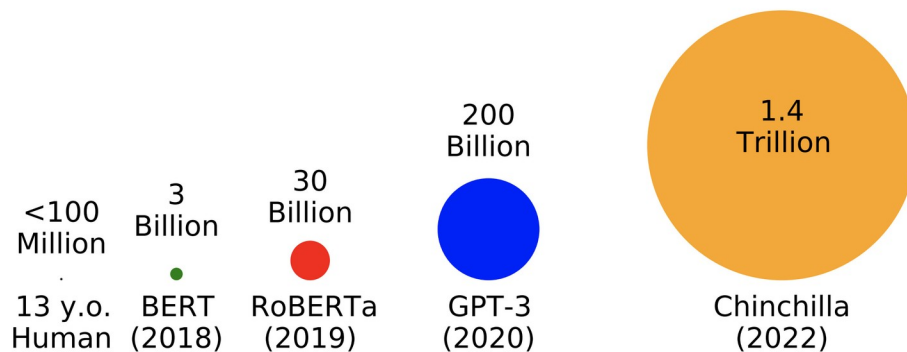
# (1) Modeling Language Acquisition

- Most work on this problem uses exclusively **textual data**.
- Hopefully, by now, you think that this is modeling the **wrong thing**.
- Not least because of ...



## (2) Data Efficiency

- Linzen (2020) argues for LM with **human-like** data-efficiency and generalization.
- BabyLM Challenge: Sample-efficient pretraining on a **developmentally plausible** corpus





**Grzegorz Chrupala** 🇪🇺 🇺🇦 @gchrupala · May 8



Comparing the number of tokens used to train an LM to the number of words spoken to a child learning a language is like comparing the number of synapses in a human brain to the number of parameters in an ML model. Borderline meaningless.



4



1



25

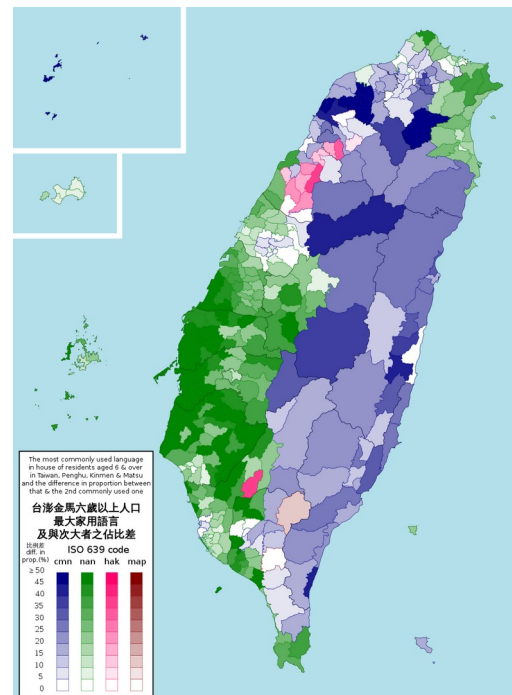


3,519



# (3) Unwritten Languages

- Most languages or are unwritten.
- All sign languages are unwritten.
- Some are used by a lot of people.
  - Taiwanese Hokkien spoken by tens of millions of people. Usually written in Chinese characters, but this doesn't really work. Latin script not widely used.
  - Indo-Pakistani Sign language: over six million speakers.

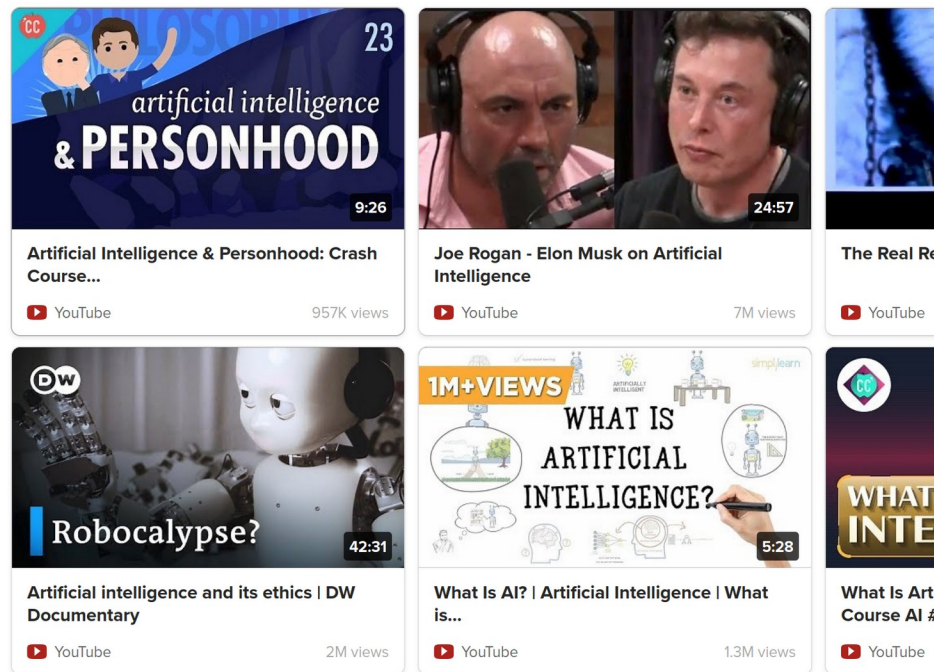


# (4) Spoken Dialog Systems

- Language technology should transition from text to talk (Dingemanse and Liesenfeld, 2022)
- Dialog systems which understand humans better and interact with them in a natural way.
- ASR + ChatGPT is not likely to work great.

# (5) Non-textual content


- Voice chat, podcasts, videos – non textual language content on the rise.
- Again, ASR + NLP is a lossy, suboptimal approach.
- Better to process this data in its native modality.



# Recommendations

- If you work in NLP, CL, or Psycholinguistics
  - Speech is not outside the scope of your field.
  - Technical obstacles are there, but much less now than 20 years ago.
- If you work in speech processing
  - Think beyond ASR and Text-to-Speech
- Move from sound to meaning and back, **textfree**.

# Selected works from $ML^2$

- Chrupała G. 2023. Putting Natural in Natural Language Processing. In Findings of the Association for Computational Linguistics.
- Shen G., Alishahi A., Bisazza A. & Chrupała G. 2023. Wave to Syntax: Probing spoken language for syntax. Proc. Interspeech 2023
- Nikolaus, M., Alishahi, A. & Chrupała, G. (2022). Learning English with Peppa Pig. *Transactions of the Association for Computational Linguistics*, 10, 922–936. 
- Chrupała, G. (2022). Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*, 73, 673-707.
- Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 613-622).