Phylogenetic linguistic inference from acoustic speech data: Ideas for a novel research paradigm

Gerhard Jäger
gerhard.jaeger@uni-tuebingen.de

May 8, 2024

University of Zurich

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Objectives and Overview

- Objectives:
  - Provide insights into phylogenetic linguistic inference
  - Explore feature extraction with deep learning
  - Investigate potential inferences from acoustic speech data
- Overview:
  - Introduction to Phylogenetic Linguistics
  - State-of-the-Art Methods
  - Novel Research Paradigm with Acoustic Data
  - Project Steps & Future Directions

- main goal: infer **phylogenetic tree** from **lexical data**

Showing 1 to 38 of 38 entries

← Previous 1 Next →

No.	Meaning	Concepticon	Word	Loan
Search	Search		Search	--any--
1	I	I	n3k	False
2	you	THOU	k3j	False
3	we	WE	nucna	False
11	one	ONE	yan	False
12	two	TWO	zuZ	True
18	person	PERSON	insan	True
19	fish	FISH	amal3h	True
21	dog	DOG	ayda	False
23	tree	TREE	tagig3t	False
25	leaf	LEAF	afraw	False
28	skin	SKIN	IZ3ld	True
30	blood	BLOOD	a83m	False
31	bone	BONE	ax3s	False
34	horn	HORN (ANATOMY)	as3kaw	False
39	ear	EAR	am3zux	False
40	eye	EYE	tit	False
41	nose	NOSE	tax3nfurt	True
43	tooth	TOOTH	asan	False
44	tongue	TONGUE	il3s	False
47	knee	KNEE	afud	False
51	breast	BREAST	sd3r	True
53	liver	LIVER	l3fwad	True
54	drink	DRINK	su	False
57	see	SEE	zar	False
58	hear	HEAR	s3l	False
61	die	DIE	mu8	False

Coordinates WGS84	35°19'N, 4°58'W 35.31, -4.96
number of speakers	10,000
status	alive

Classification

WALS
AA > Berber

Glottolog
Afro Asiatic > Berber > Kabyle Atlasberber > Atlasberber > Northwesternmoroccanberber

Ethnologue
Afro Asiatic > Berber > Northern > Zenati > Ghomara

Sources

El Hannouche 2010
Arabic influence in Ghomara Berber. M.A. thesis, Leiden University.

Phylogenetic linguistics: state of the art

- main goal: infer **phylogenetic tree** from **lexical data**
- **input:** manual cognate classification
- example (*dunnielex* from Lexibank)

Row	Language_ID	Parameter_ID	Segments	Cognateset_ID
	String15?	String15?	String?	Int64?
1	urdu	180_tooth	d̪ ɑ̃ t̪	328
2	catalan	180_tooth	d e n	328
3	armenianmod	180_tooth	ɑ t ɑ m	328
4	bretonst	180_tooth	d ã n t	328
5	czech	180_tooth	z ʊ p	502
6	german	180_tooth	ts aː n	328
7	italian	180_tooth	d ɛ n t e	328
8	swedish	180_tooth	t a n d	328
9	greekmod	180_tooth	ð o̞ n d i	328
10	marathi	180_tooth	d a t	328
11	polish	180_tooth	z ɔ̃ p	502
12	portuguesest	180_tooth	d ẽ t ɨ	328
13	russian	180_tooth	z u b	502
14	spanish	180_tooth	d j e n t e	328
15	danish	180_tooth	d̥ʰ/dʰ a n	328
16	dutchlist	180_tooth	t ɑ n t	328
17	english	180_tooth	t uː θ	328
18	french	180_tooth	d ɑ̃	328
19	russian	180_tooth	d e s n a	328
20	bihari	180_tooth	d ã t	328
21	oriya	180_tooth	d a n t ɔ	328

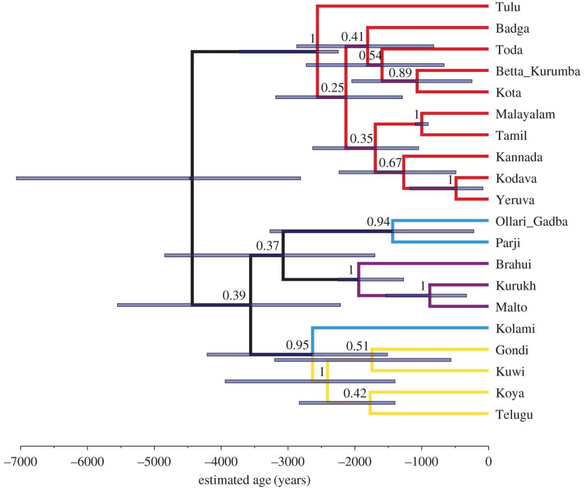
Phylogenetic linguistics: state of the art

- **intermediate step:** convert cognate classification into binary matrix

```
TNG.ENGAN.MAIBI                  10000000000000000000000000000000000000000000
TNG.ENGAN.POLE                   00000000000000000000000000000000000010000000
TNG.ENGAN.SAU                    00000000000000000000000000000000000000000000
TNG.ENGAN.YARIBA                 10000000000000000000000000000000000000000000
TNG.FASU.FASU                    00000000000000000000000000000000000010000000
TNG.FASU.NAMUMI                  00000000000000000000000000000000000000001000
TNG.FINISTERRE-HUON.AWARA        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BORONG       00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BURUM        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BURUM_MIND   00000000000000000000000000000000001010000000
TNG.FINISTERRE-HUON.DEDUA        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.HUBE         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KATE         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KOMBA        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KOSORONG     00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MAPE         00000000000000000000000000000000001000000000
TNG.FINISTERRE-HUON.MAPE_2       00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MIGABAC      00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MINDIK       00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MOMOLILI     00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.NABAK        00000000000000000000000000000000001000000000
TNG.FINISTERRE-HUON.NANKINA      00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.NEK          00000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.NUKNA        00000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.ONO          00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.SELEPET      00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.TIMBE        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.TOBO         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.WANTOAT      00000000000000000000000000000000000110000000
TNG.FINISTERRE-HUON.YOPNO        00000000000000000000000000000000000001000000
TNG.GOILALAN.AFOA                00000000000000000000000000000000000110000000
TNG.GOILALAN.KUNIMAIPA           00000000000000000000000000000000000010000000
TNG.GOILALAN.MAFULU              00000000000000000000000000000000000010000000
```

Phylogenetic linguistics: state of the art

- **output:** phylogenetic tree (here: Dravidian languages according to Kolipakam et al. 2018)



Phylogenetic linguistics: state of the art

Applications

- control for common ancestry in statistical models *(Jäger and Wahle 2021, ...)*
- estimate time depth and geographic location of ancestral populations (Bouckaert et al 2012)
- reconstruct properties of ancestral populations (Cathcart et al 2021, Carling and Cathcart 2021a,b, ...)
- statistic identification of patterns of language change *(Blasi et al. 2019)*
- ...

from word lists to trees

1. perform cognate classification (manual or automatic)
2. construct binary *character matrix*
3. let computer search the tree(s) that best explain(s) the distribution of 0s and 1s in the character matrix

Manual cognate detection

Cognate detection

Manual cognate detection

- labor intensive
- available data are geographically skewed
- requires tons of prior classical historical linguistics work

Automatic cognate detection

- lot of computational research over the past years to automate the process
- results are usable but far from perfect

Phylogenetic signal below cognacy

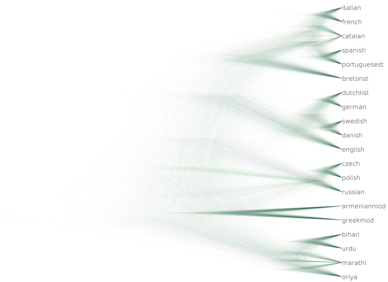
- sound change and morpological change contains relevant phylogenetic information

1	urdu	180_tooth	d̪ ɑ̃ t̪	328
2	catalan	180_tooth	d e n	328
3	armenianmod	180_tooth	ɑ t ɑ m	328
4	bretonst	180_tooth	d ã n t	328
5	german	180_tooth	ts aː n	328
6	italian	180_tooth	d ɛ n t e	328
7	swedish	180_tooth	t a n d	328
8	greekmod	180_tooth	ð o̞ n d i	328
9	marathi	180_tooth	d a t	328
10	portuguesest	180_tooth	d ẽ t ɨ	328
11	spanish	180_tooth	d j e n t e	328
12	danish	180_tooth	d̥^{h}/d^{h} a n	328
13	dutchlist	180_tooth	t ɑ n t	328
14	english	180_tooth	t uː θ	328
15	french	180_tooth	d ɑ̃	328
16	russian	180_tooth	d e s n a	328
17	bihari	180_tooth	d ã t	328
18	oriya	180_tooth	d a n t ɔ	328
19	czech	180_tooth	z ʊ p	502
20	polish	180_tooth	z ɔ̃ p	502
21	russian	180_tooth	z u b	502

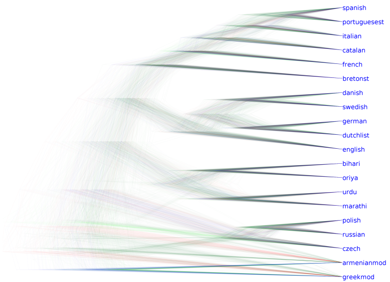
Feature extraction with deep learning

- **idea:** use deep learning to extract features from word lists rather than doing manual annotation or traditional machine learning
- pilot studies are quite encouraging

manual cognate classification

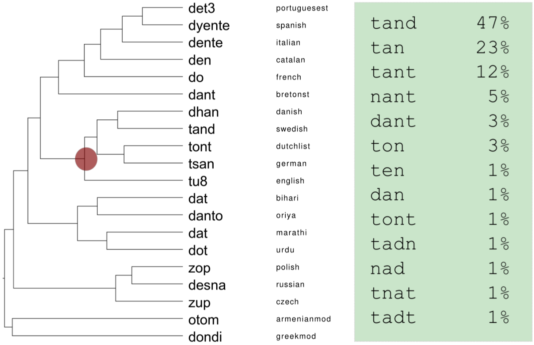


deep learning



Feature extraction with deep learning

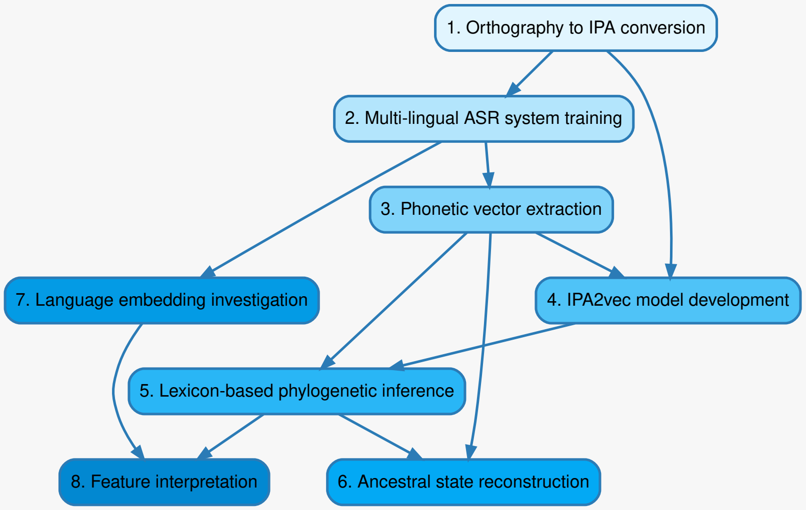
- **idea:** use deep learning to extract features from word lists rather than doing manual annotation or traditional machine learning
- pilot studies are quite encouraging



Working with speech data

- **idea:** start with speech data rather than word lists
- **advantages:**
  - sidesteps all the human decision-making involved in compiling dictionaries
  - applicable to low-resource languages
  - potentially accesses phylogenetically relevant information not easily accessible to introspection
- **challenges:**
  - get hold of a sufficient amount of data
  - develop methods to extract features from speech data
  - if it works: understand what the machine is doing

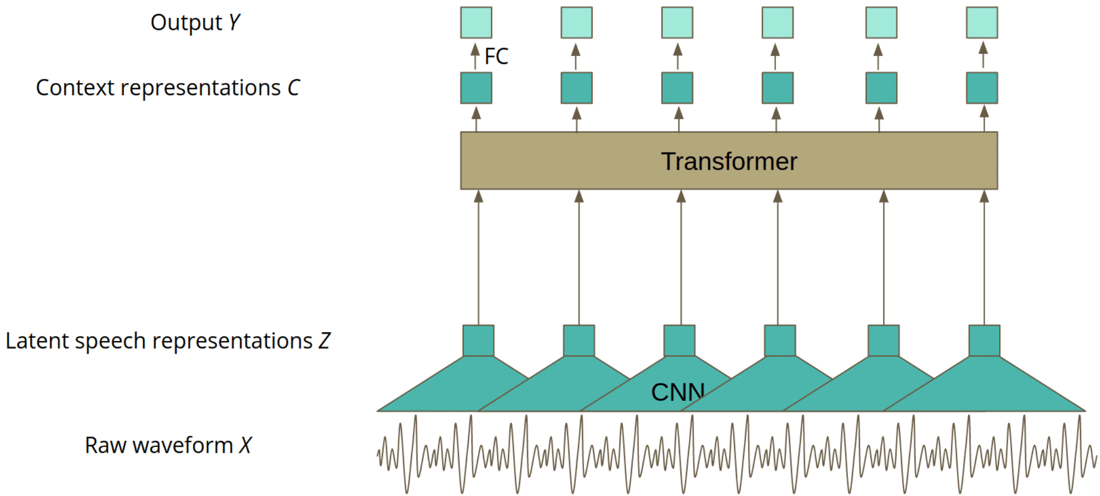
Workflow


1. Orthography to IPA conversion
2. Multi-lingual ASR system training
3. Phonetic vector extraction
7. Language embedding investigation
4. IPA2vec model development
5. Lexicon-based phylogenetic inference
8. Feature interpretation
6. Ancestral state reconstruction

Automatic speech recognition

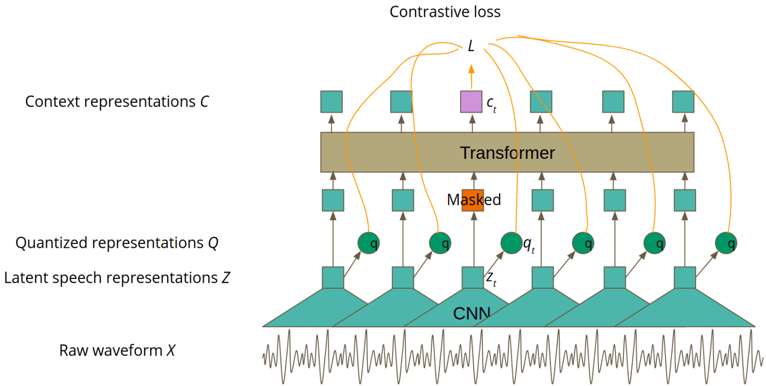
- major advances in the past years
- important steps:
  - *wav2vec* (Schneider et al., 2019)
  - *wav2vec 2.0* (Baevski et al., 2020)
  - *wav2vec-u* (Baevski et al., 2021)

Wav2vec 2.0

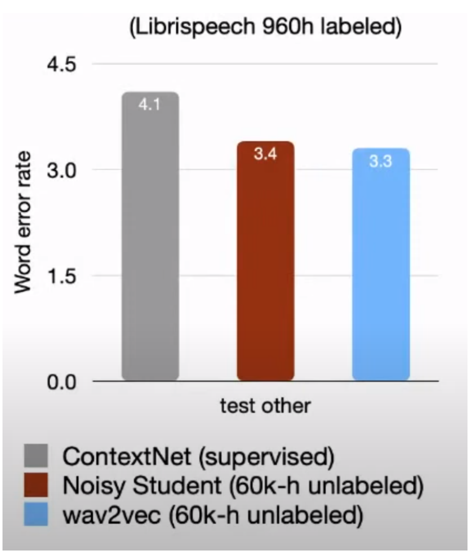

Output Y
FC
Context representations C
Transformer
Latent speech representations Z
CNN
Raw waveform X

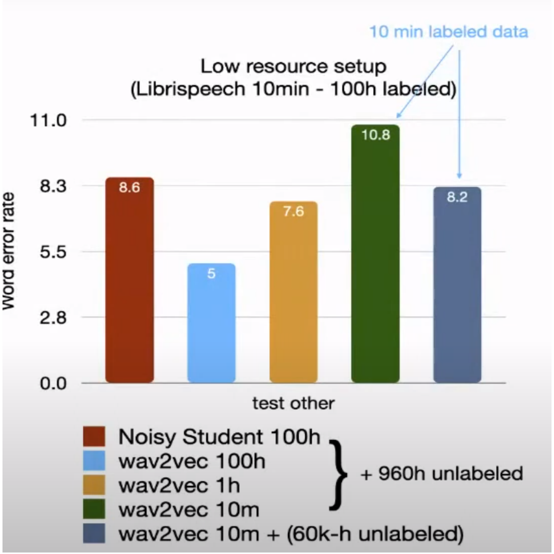
Wav2vec 2.0

First phase: unsupervised: pretrain a model to predict the context of a speech segment



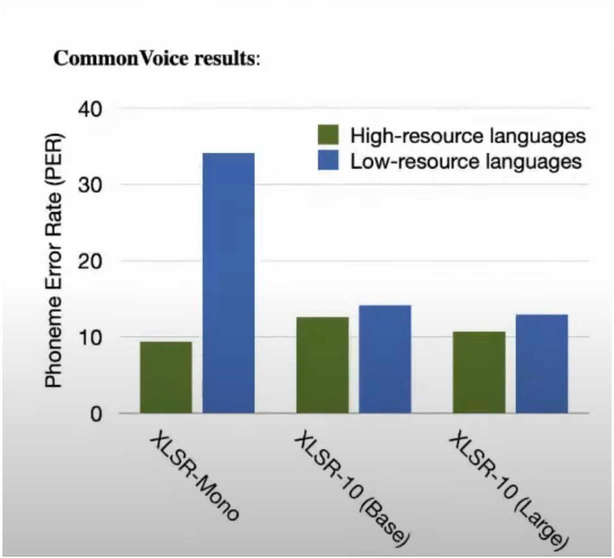
Second phase: supervised: fine-tune the model to predict the phonemes of a speech segment





(image from https://www.youtube.com/watch?v=EQOBE7sJSJY)

Wav2vec 2.0: Cross-linguistic transfer

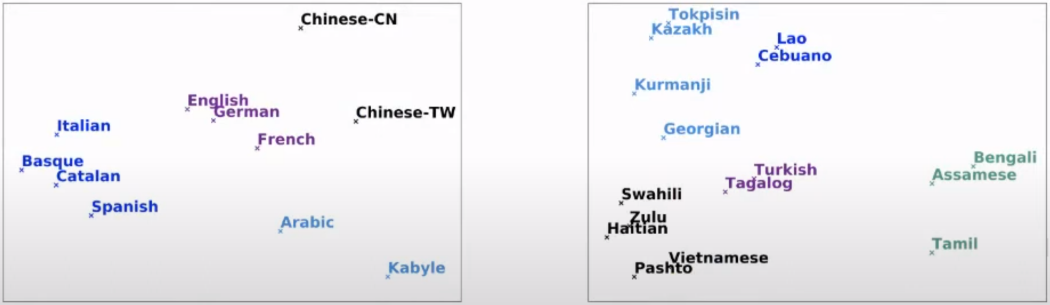


(image from https://www.youtube.com/watch?v=EQOBE7sJSJY)

Wav2vec 2.0: Cross-linguistic transfer

PCA visualization of latent discrete representations from the multilingual codebook

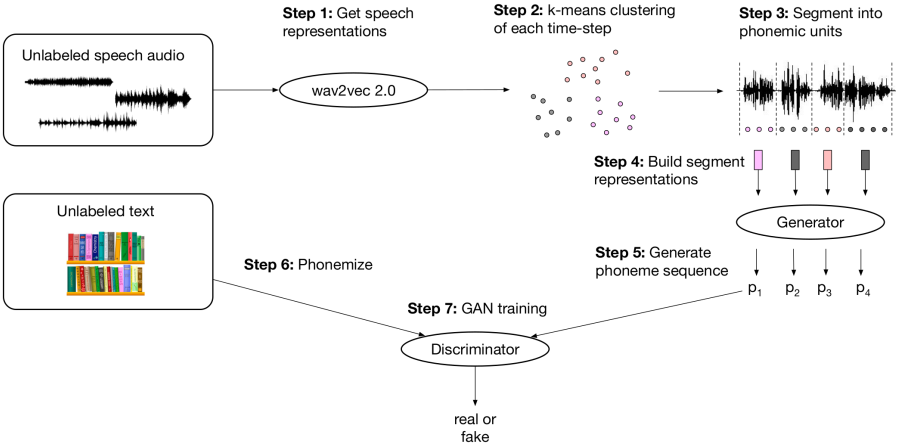
Similar languages tend to share discrete tokens and thus cluster together



(image from https://www.youtube.com/watch?v=EQOBE7sJSJY)

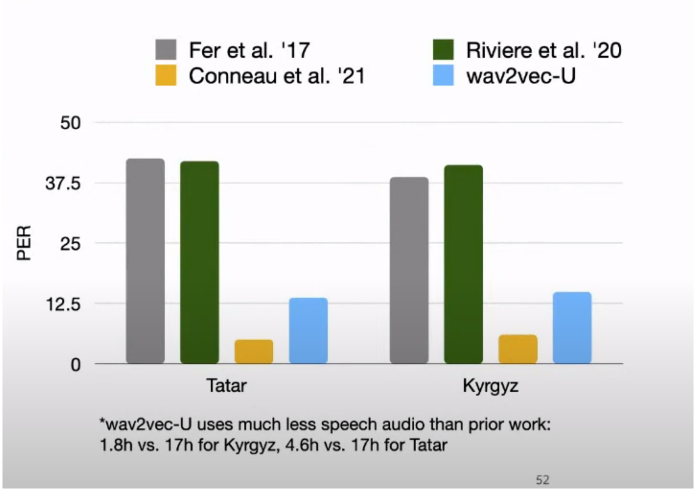
Wav2vec-u

- *u* stands for *unsupervised*



(image from Baevski et al. 2021)

Wav2vec-u: Low-resource setting



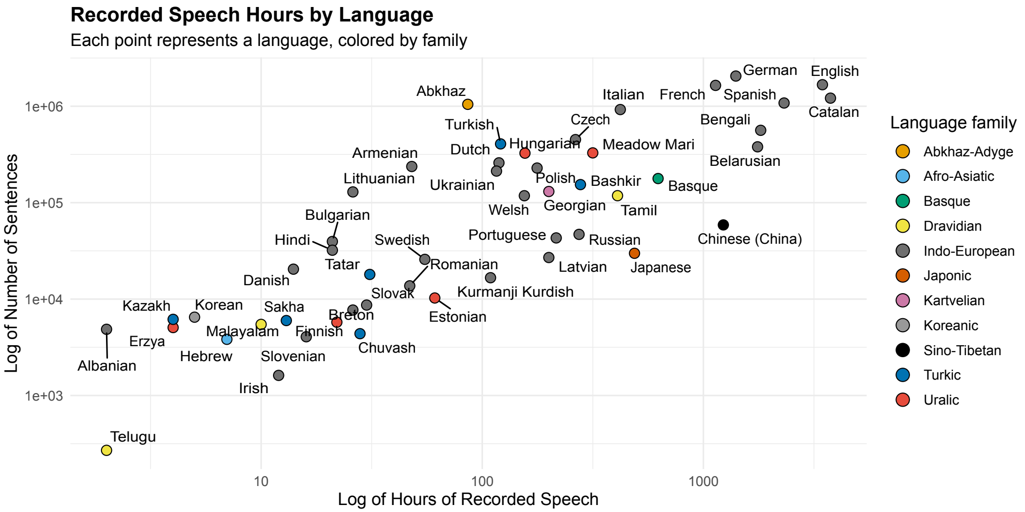
(image from https://www.youtube.com/watch?v=EQOBE7sJSJY)

Data availability

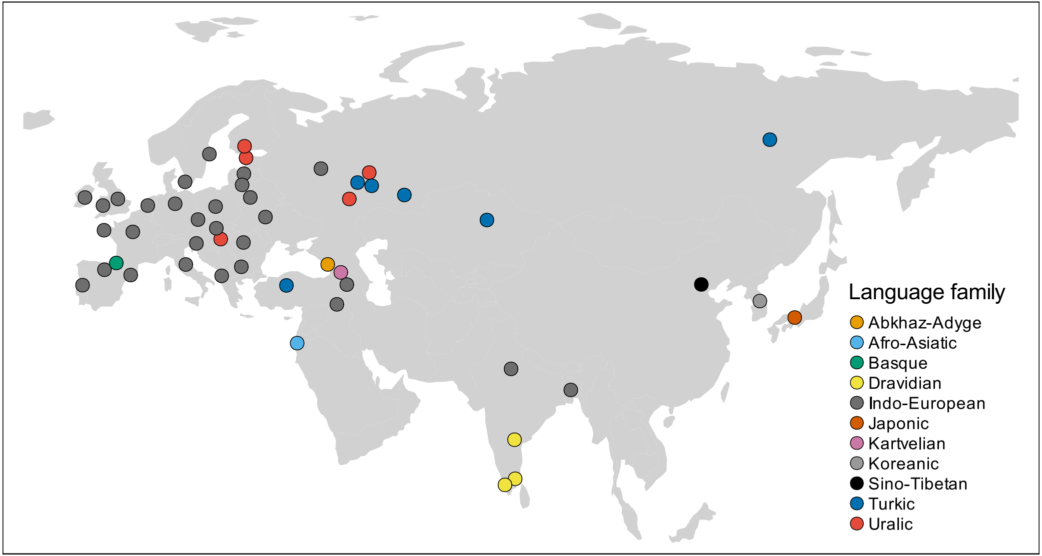
- **Common Voice** (Mozilla)
  - **Open Dataset**: Common Voice provides a freely accessible and diverse multilingual dataset, covering over 70 languages, for developing voice-enabled technologies.
  - **Community Contributions**: The project is community-driven, relying on global volunteers for voice donations and data validation.
  - **Ethical and Accessible**: Emphasizing privacy and accessibility, Mozilla aims to enhance voice technology inclusivity for underrepresented languages and dialects.

Data availability

Languages to be covered in *phylomilia*



Data availability



Project steps

Create IPA transcriptions

- chosen languages all belong to NorthEuraLex
- for these languages, orthography-to-IPA transducers are available

Project steps

Train multi-lingual ASR system

- **Model Implementation**: Implement and train a version of the wav2vec-u model that inputs an audio recording and a language identifier.
- **Output Format**: The model will output a sequence of IPA (International Phonetic Alphabet) symbols.
- **Training Approach**: Training will utilize transcriptions from the earlier phase as a guide.

Project steps

Phonetic vector representations of NorthEuralex

- **Concept Selection**: The 1,016 concepts in NorthEuralex are data-driven, selected for their clear reflexes in Northern Eurasian languages and diachronic stability.
- **Concept List Basis**: This concept list extends the Swadesh list (Swadesh 1955), a foundational tool in historical linguistics.
- **Data Extraction**: The next step involves extracting spoken counterparts of the entries from the NorthEuralex database.
- **Transcription Process**: This will be achieved by querying the database for IPA transcriptions of the words and using the wav2vec-u model to transcribe the audio recordings.

Project steps

IPA2vec model for missing data imputation

- **Model Training**: Train a model that inputs a sequence of IPA symbols and a language identifier, outputting a vector representation of the word.
- **Model Type**: The model will use a neural sequence autoencoder to mimic the embedding from a spoken version of the word.
- **Application of Model**: The model will be used to impute all NorthEuralex entries not covered by the previous extraction step.

Project steps

Lexicon-based phylogenetic inference

- **Vector Usage**: Use the vector representations of words from the NorthEuralex database as input for phylogenetic inference.
- **Model Training for Vector Conversion**: Train a straight-through autoencoder (Bengio et al. 2013) to convert phonetic vector embeddings into binary vectors.
- **Binary Vector Concatenation**: Concatenate the binary vectors of all words in a language (or a selected subset) to serve as input for phylogenetic analysis.
- **Phylogenetic Inference Tool**: Perform phylogenetic inference using a standard package such as BEAST or MrBayes.

Project steps

Ancestral state reconstruction

- perform ancestral state reconstruction with binary vectors
- convert reconstructed vectors back into phonetic vectors

Project steps

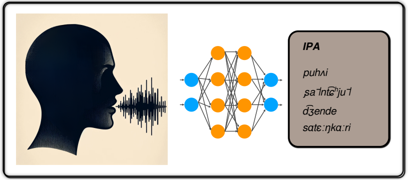
Phylogenetic inference from direct language embeddings

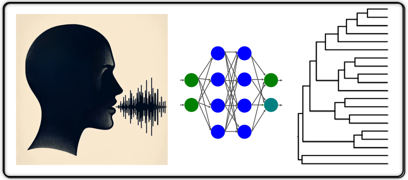
- **High-Risk Approach Exploration**: Investigate the feasibility of performing phylogenetic inference directly from representations of entire phonetic systems of languages.
- **Model Configuration**: Train a deep neural network that inputs a language identifier and the sound recording of an entire sentence.
- **Internal Mapping**: Internally map the language identifier onto a dense vector within the neural network.
- **Loss Function Design**: Utilize a classifier as the loss function to determine the familial origin of the sound recording, incorporating a BERT-style masking of parts of the audio input.

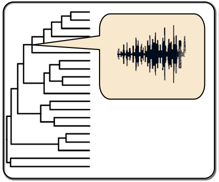
Project steps

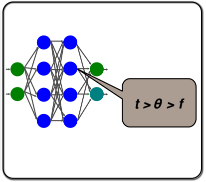
Feature interpretation

- **Feature Identification with LIME**: Utilize the LIME method to determine the most informative features for phylogenetic inference.
- **High-Dimensional Mapping**: Map the phylogenetic tree onto a high-dimensional vector space by applying multidimensional scaling to pairwise co-phenetic language distances.
- **Deep Network Training**: Train a deep network to predict the position of a language from its vector representation, which could be either a concatenation of word embeddings from its NorthEuralex entries or a direct language embedding.
- **Explainable AI Application**: Apply explainable-AI methods to identify and interpret the features most significant to the model's decision-making process.


IPA
puhʌi
ʂa˥ntɕ͡ʰju˥
d͡ʒende
satɛːŋkaːri






t > θ > f

Related work

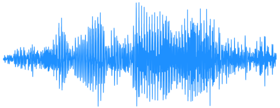
- came to my attention this morning (He et al., 2024):

WAV2GLOSS: Generating Interlinear Glossed Text from Speech

Taiqi He[1], Kwanghee Choi[1], Lindia Tjuatja[1], Nathaniel R. Robinson[2], Jiatong Shi[1], Shinji Watanabe[1], Graham Neubig[1], David R. Mortensen[1], Lori Levin[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Center for Language and Speech Processing, Johns Hopkins University

Abstract

Thousands of the world's languages are in danger of extinction—a tremendous threat to cultural identities and human language diversity. Interlinear Glossed Text (IGT) is a form of linguistic annotation that can support documentation and resource creation for these languages' communities. IGT typically consists of (1) transcriptions, (2) morphological segmentation, (3) glosses, and (4) free translations to a majority language. We propose WAV2GLOSS: a task to extract these four annotation components



wd:	n	sìginde		yan	de
sr:	n	sìgi	-nde	yan	de
ur:	n	sìgi	-len	yan	le
gl:	1.SG	sit	-PC.RES	that	FOC
tr:	"I live here."				

Thank You

Thank You for Your Attention!

Questions or feedback are welcome.

Contact

Gerhard Jäger

gerhard.jaeger@uni-tuebingen.de

References I

Baevski, A., W.-N. Hsu, A. Conneau, and M. Auli (2021). Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, **34**:27826–27839.

Baevski, A., Y. Zhou, A. Mohamed, and M. Auli (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**:12449–12460.

Blasi, D. E., S. Moran, S. R. Moisik, P. Widmer, D. Dediu, and B. Bickel (2019). Human sound systems are shaped by post-neolithic changes in bite configuration. *Science*, **363**(6432):eaav3218.

Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, **337**(6097):957–960.

Carling, G. and C. Cathcart (2021). Evolutionary dynamics of Indo-European alignment patterns. *Diachronica*, **38**(3):358–412.

References II

Dunn, M. (2012). Indo-European lexical cognacy database (IELex). URL: http://ielex.mpi.nl/.

He, T., K. Choi, L. Tjuatja, N. R. Robinson, J. Shi, S. Watanabe, G. Neubig, D. R. Mortensen, and L. Levin (2024). Wav2gloss: Generating interlinear glossed text from speech. *ArXiv*, **abs/2403.13169**.

Jäger, G. and J. Wahle (2021). Phylogenetic typology. *Frontiers in Psychology*, **12**:682132.

Kolipakam, V., F. M. Jordan, M. Dunn, S. J. Greenhill, R. Bouckaert, R. D. Gray, and A. Verkerk (2018). A Bayesian phylogenetic study of the Dravidian language family. *Royal Society open science*, **5**(3):171504.

Schneider, S., A. Baevski, R. Collobert, and M. Auli (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.