How to Create Large Annotated Databases from Public Audio Data

Logan Walls, Natasha Vernooij, Eric Martell, Natalie Robbins
University of Michigan, USA

Presented at X-PPL 2023, Nov. 6 2023

A generic procedure with a concrete example

We provide concrete examples of each step from **ES COCO**: an English-Spanish code-switching corpus

Code-switching

- Code-switching
  - When a multilingual switches from one language to another (Poplack, 1980)
  - Still stigmatized but increased research, funding, and pop-science coverage is reducing stigma
- Research on code-switching can answer questions relating to:
  - language acquisition/learning, contact/transfer
  - sociopragmatics, social identity
  - cognitive strategies, mental representations of grammar

Problem: lack of usable data

- Collecting and preparing data is expensive
  - Collecting data:
    - trade-off between highly controlled experiments and naturalistic data
    - equipment, field work, community building
  - Preparing data: transcription and annotation

Limit the amount and type of data we can collect

- Available data…
  - does not focus on code-switching
  - focuses on high resource language pairings
  - is not in an easily usable format (ex. .mp3, .pdf files)

Limit the usability of existing data

Solution

Access publicly available audio, use machine learning to make data usable, and make data publicly available

- Tips for accessing publicly available audio
- Machine learning to automate both transcription and annotation
- Resources for creating an open access database

Finding available speech data

Rule of thumb: Ask for permission

Privately held recordings

Google/Google Scholar; academic listservs; related articles

Podcasts

Google/Apple Podcasts; Spotify

Be prepared to explain your work to non-academics; engage in community building

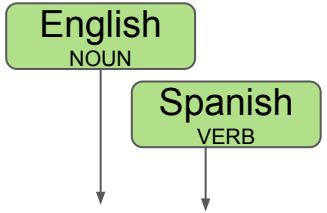
TV/Film

Local stations; YouTube

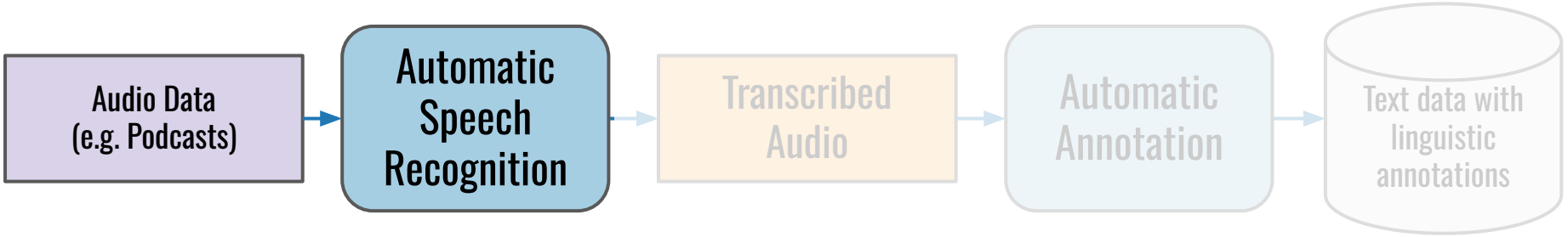
Note: Check local copyright laws- most information on the internet can be used for personal/academic use but redistribution may be prohibited

Automating transcription & annotation

The boy puso la caja on the table.



The boy puso la caja on the table.



Encoder-Decoder Model
(e.g. **Whisper**)

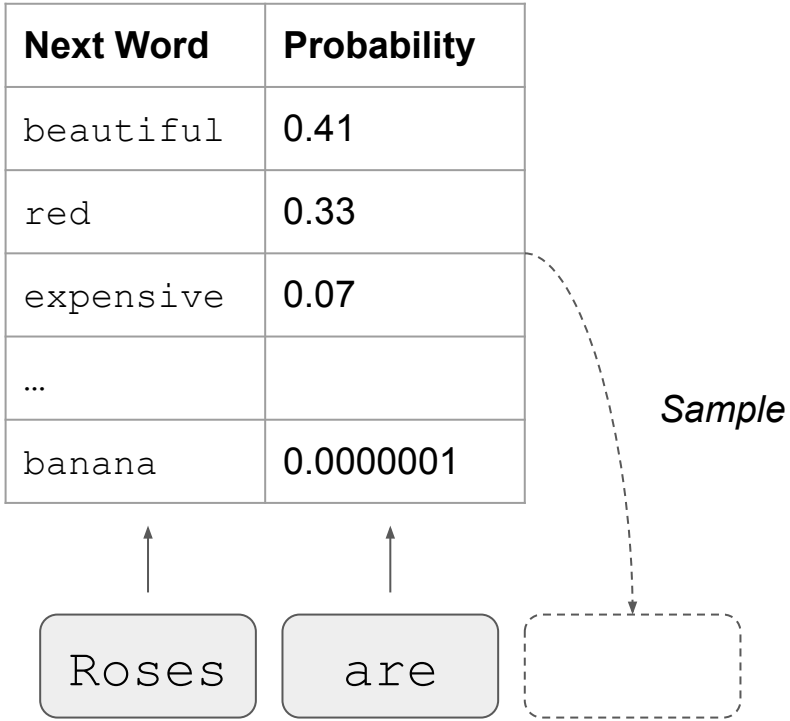
The boy puso la caja on the table.

The boy puso la caja on the table.

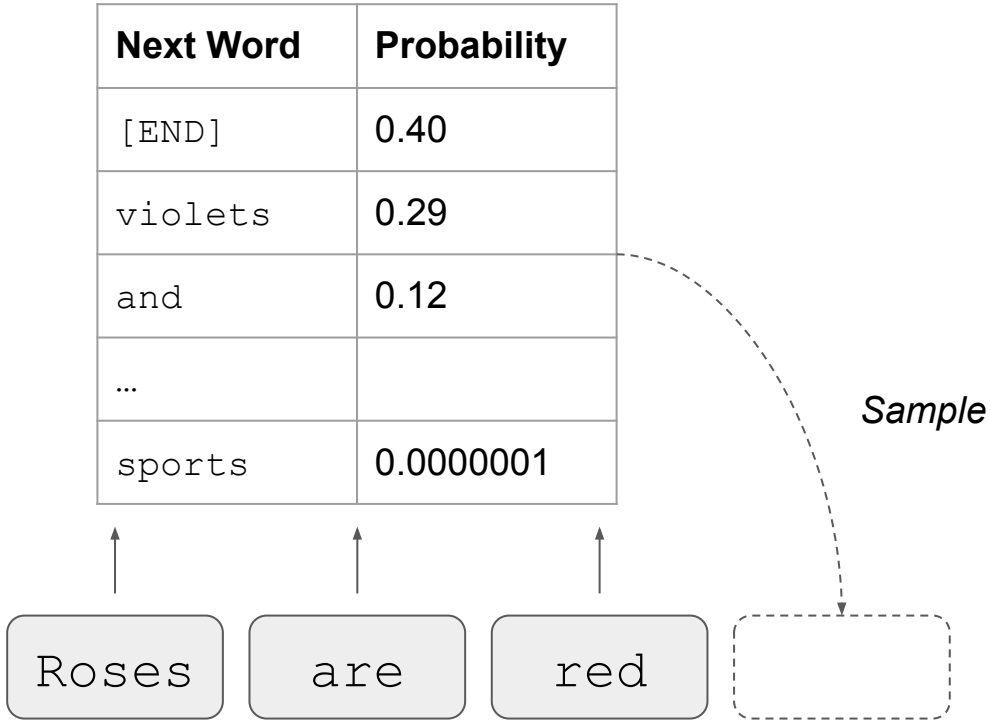
What is an **Encoder-Decoder** model?

Transformer Language Models - **Decoders**

Transformer Language Models - **Decoders**

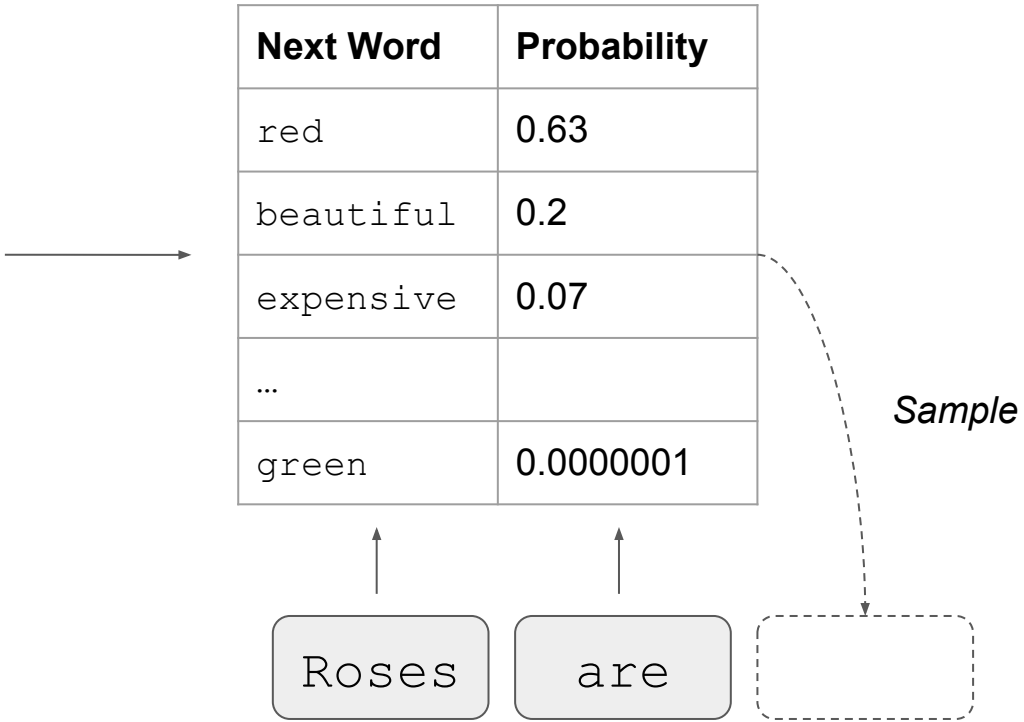


Transformer Language Models - **Decoders**

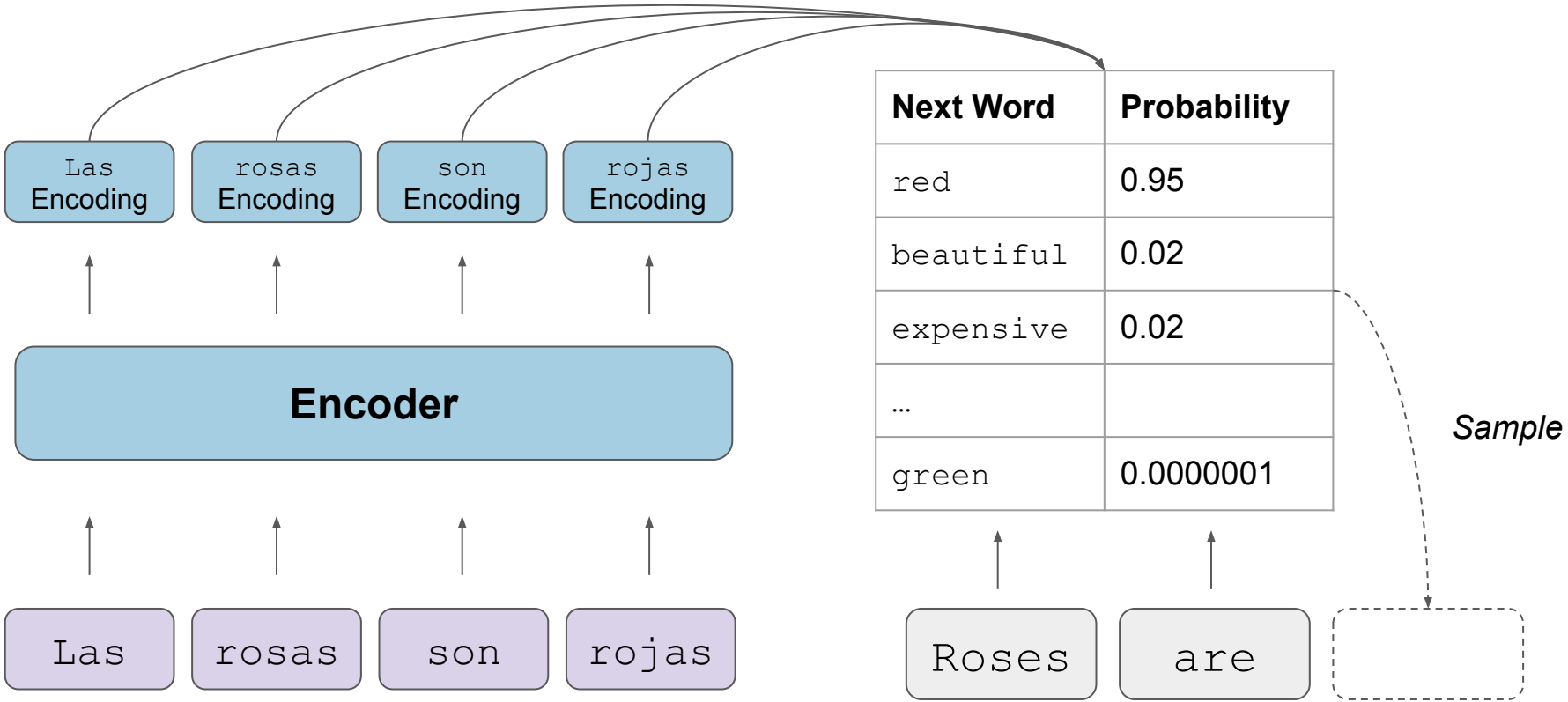


Transformer Language Models - **Encoder**-Decoders

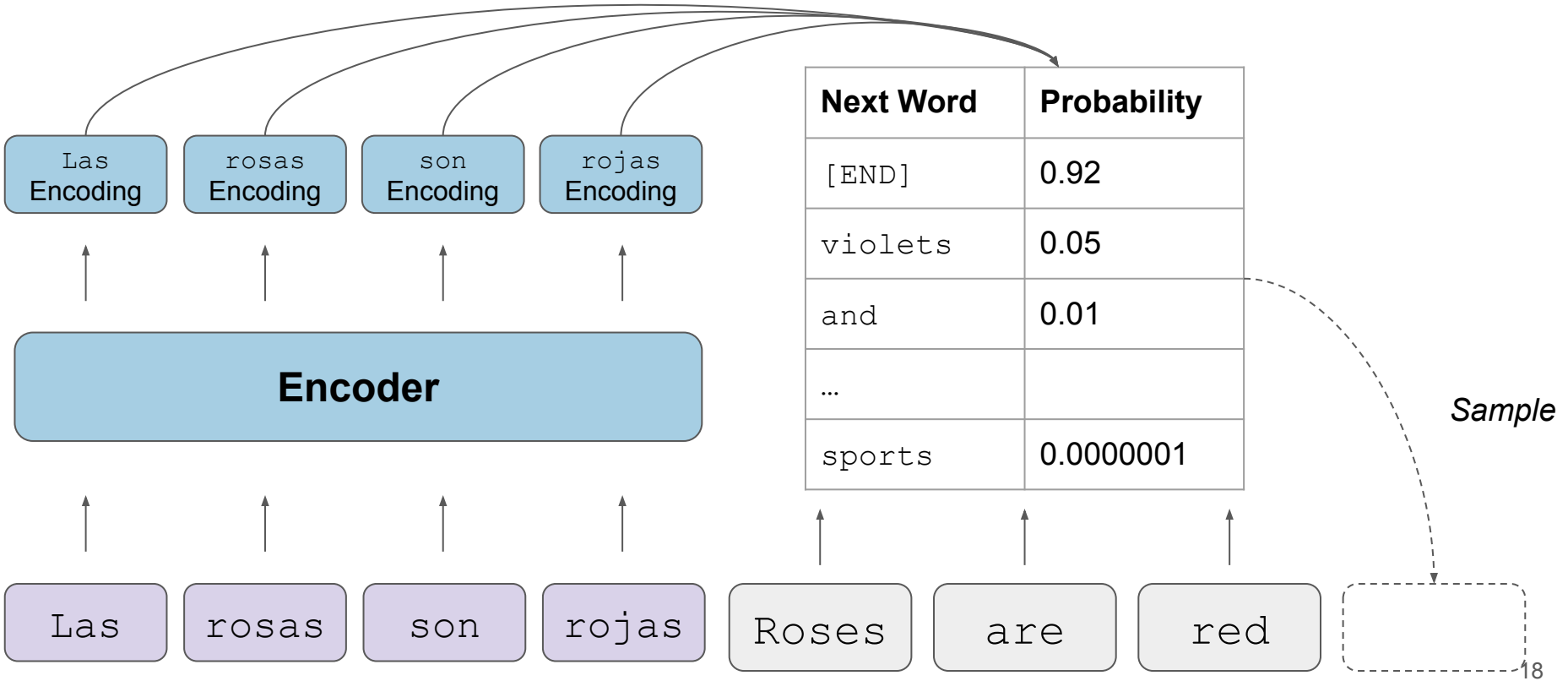
Additional Information



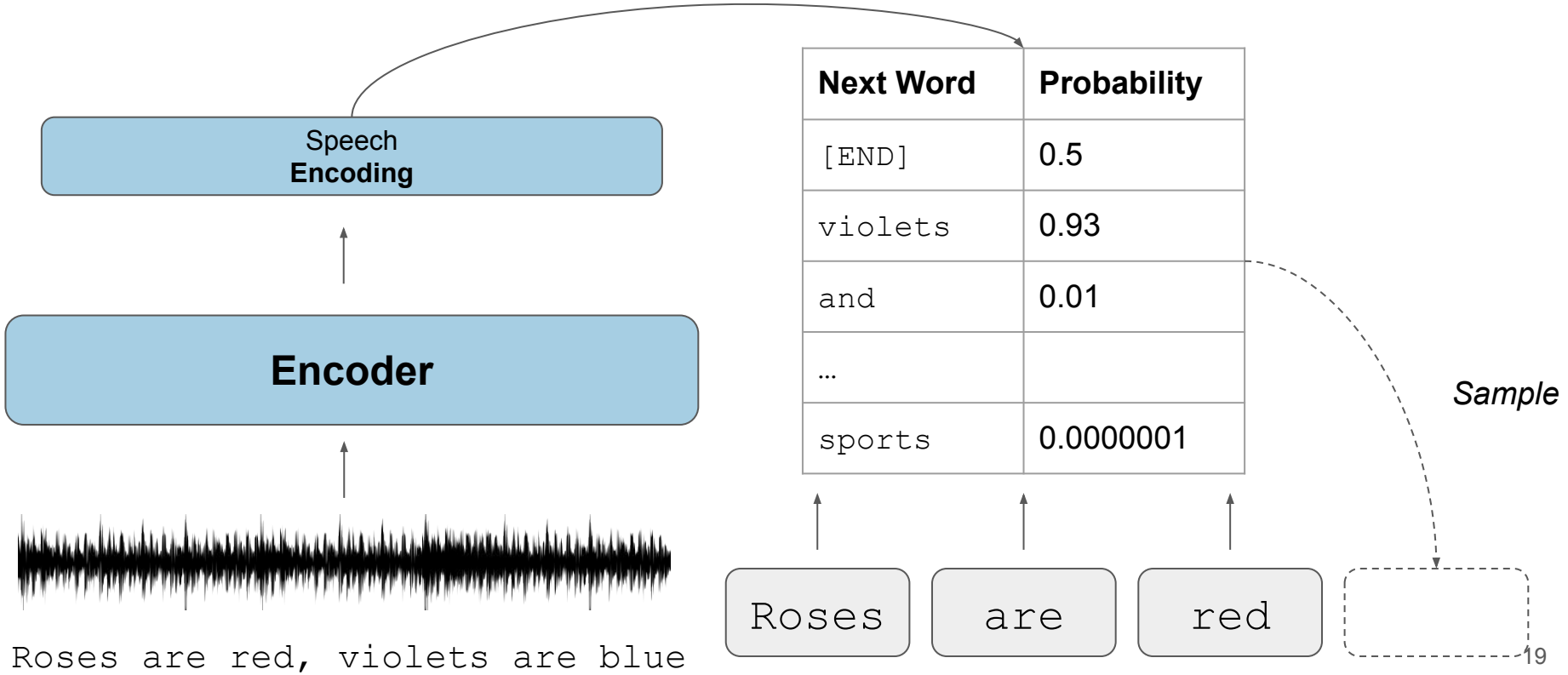
Transformer Language Models - **Encoder**-Decoders



Transformer Language Models - **Encoder**-Decoders



Transformer Language Models - **Encoder**-Decoders



How effective are these **A**utomatic **S**peech **R**ecognition (ASR) models?

Word Error Rate (WER)

$$WER = \frac{S + D + I}{N}$$

- **S**ubstitutions `The boy puso la ~~caja~~ **casa** on the table.`
- **D**eletions `The boy puso la ~~caja~~ on the table.`
- **I**nsertions `The boy puso la **casa** caja on the table.`
- **N**umber (of total words)

Word Error Rates for Multilingual ASR Benchmarks

Multilingual LibriSpeech

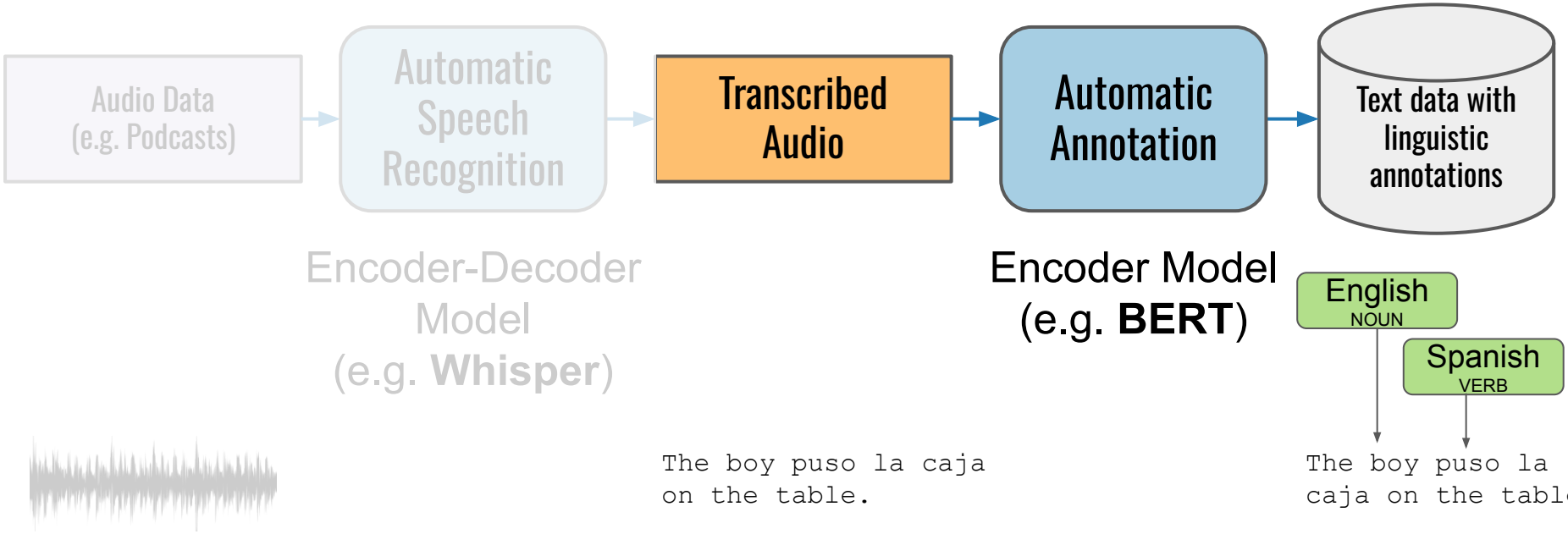
Model	English	Spanish
Whisper tiny	15.7	19.2
Whisper base	11.7	12.8
Whisper small	8.3	7.8
Whisper medium	6.8	5.3
Whisper large	6.3	5.4
Whisper large-v2	6.2	4.2

Common Voice 9

Model	English	Spanish
Whisper tiny	28.8	30.3
Whisper base	21.9	19.6
Whisper small	14.5	10.3
Whisper medium	11.2	6.9
Whisper large	10.1	6.4
Whisper large-v2	9.4	5.6

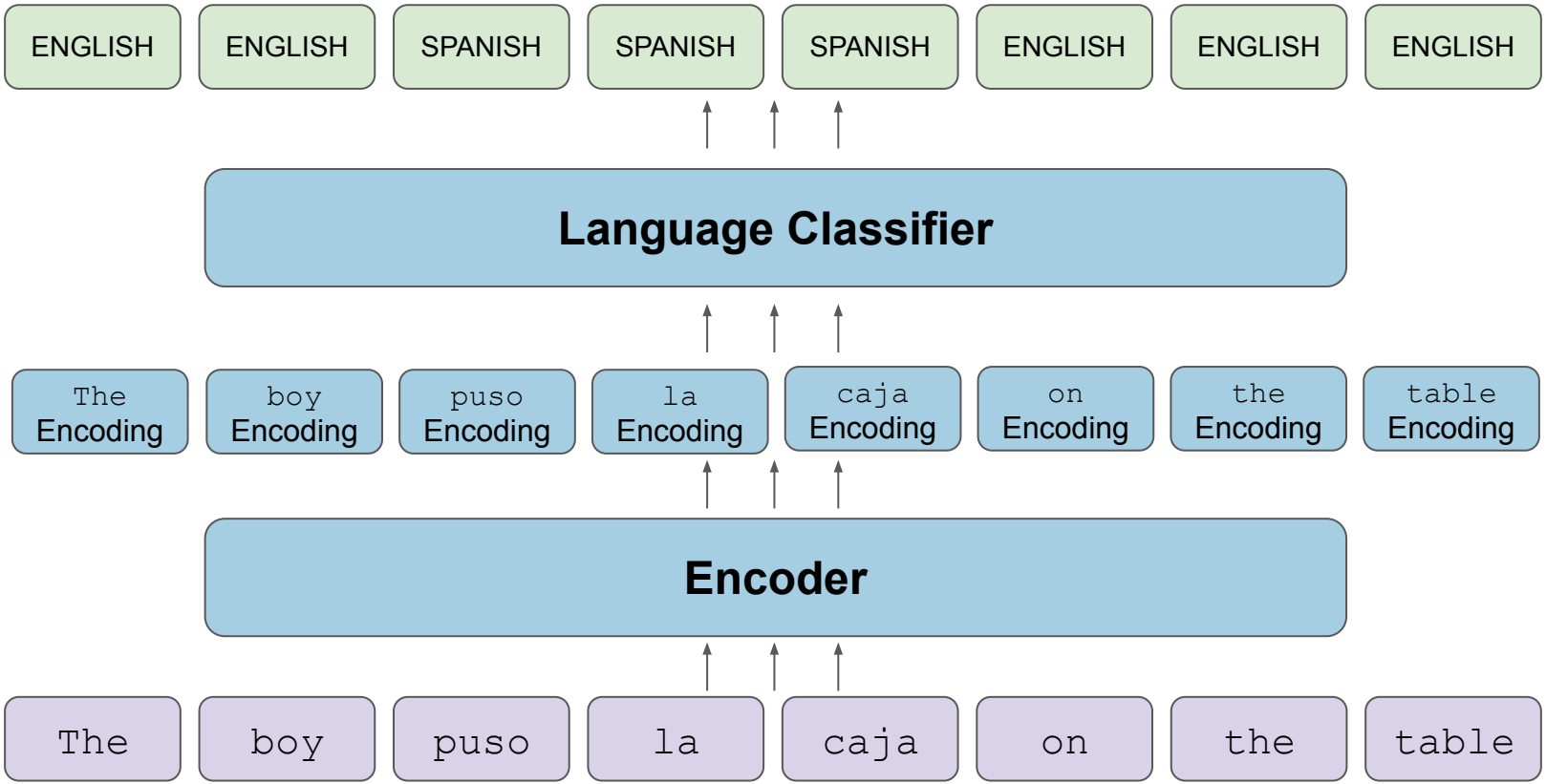
FLEURS

Model	English	Spanish
Whisper tiny	12.4	15.9
Whisper base	8.9	9.9
Whisper small	6.1	5.6
Whisper medium	4.4	3.6
Whisper large	4.5	3.5
Whisper large-v2	4.2	3.0

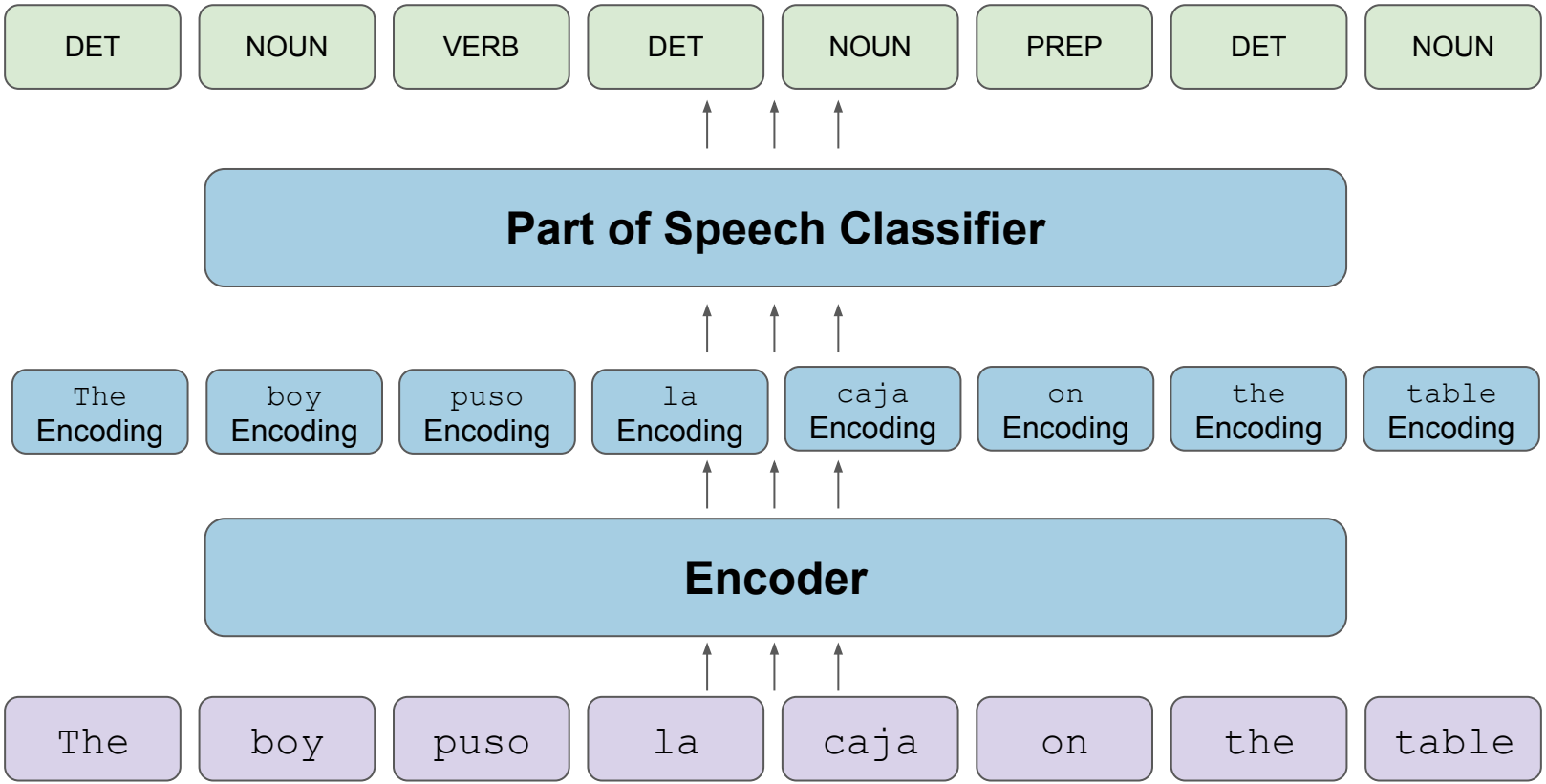

Audio Data
(e.g. Podcasts)
Automatic Speech Recognition
Transcribed Audio
Automatic Annotation
Text data with linguistic annotations
Encoder-Decoder Model
(e.g. Whisper)
Encoder Model
(e.g. BERT)
English
NOUN
Spanish
VERB
The boy puso la caja on the table.
The boy puso la caja on the table.

What is an **Encoder** model?

Transformer Language Models - **Encoders**



Transformer Language Models - **Encoders**



How effective are these annotation models?

Percentage of tokens predicted correctly (LinCE benchmark)

Language Identification

Model	SPA-ENG
XLMR_multi-labels	98.64
Char2subword mBERT	98.33
mBERT	98.36
BERT base, cased	98.35
ELMo small	97.93

Part of Speech

Model	Avg	SPA-ENG
XLMR_multi-labels	**94.48**	97.22
XLM-R Large	**94.38**	97.18
XLM-R Base	**93.98**	96.96
XLM-MLM-100	**93.07**	97.04
HME-Ensemble	**93.04**	96.78
HME	**92.60**	96.66
Char2subword mBERT	**92.55**	96.88
BERT base, cased	**91.97**	96.92

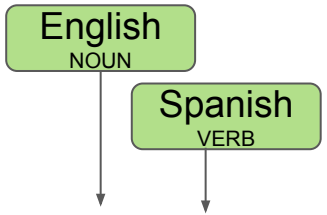
Considerations for ASR and annotation

- Efficiency vs. Accuracy
  - Larger models are more accurate, but slower to run / require more powerful computers
- Specialization
  - Consider how compatible the model's training data is with your target data
    - What languages?
    - Multilingual vs. code switched speech

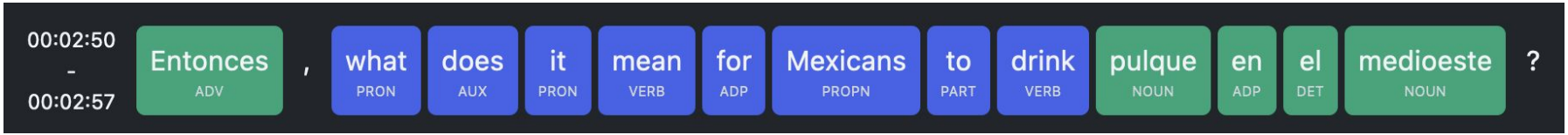
Encoder-Decoder Model (e.g. **Whisper**)

Encoder Model (e.g. **BERT**)

```
The boy puso la caja
on the table.
```




```
The boy puso la
caja on the table.
```



Sharing your data

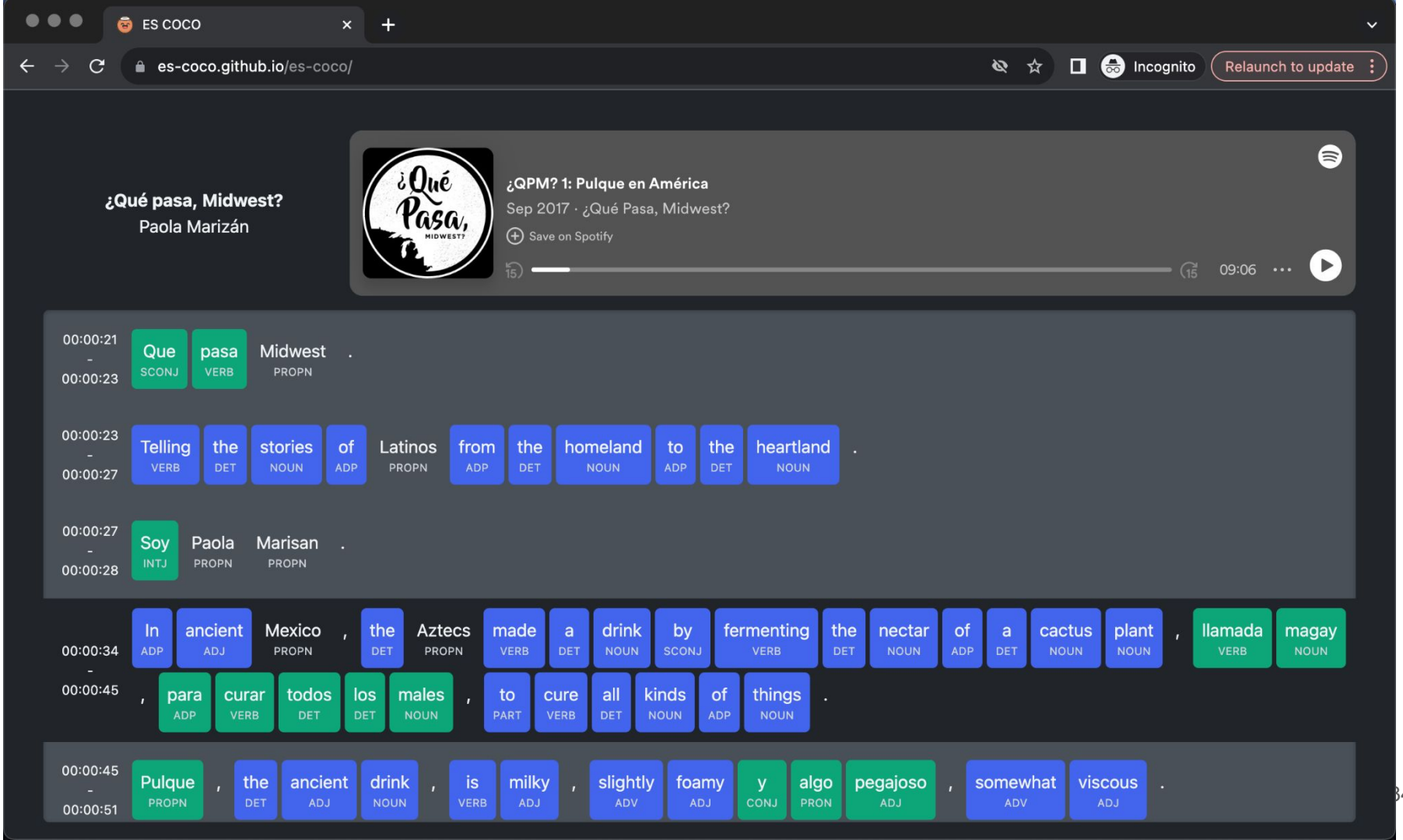
How to share your data

- Distribute as SQLite database
  - One-file format, easy to backup and share
  - Widely supported in popular data analysis languages (R, Python)

name	url	creator	segment	segment_start	segment_end	id	surface_form	segment_id	word_index
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	1	From	1	0
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	2	WNIN	1	1
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	3	and	1	2
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	4	PRX	1	3
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	5	,	1	4
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	6	this	1	5
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	7	is	1	6
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	8	Que	1	7
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	9	pasa	1	8
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	10	Midwest	1	9
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	1	12520	17180	11	.	1	10
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	2	17180	18120	12	Que	2	0
¿Qué pasa, Midwest?	https://www.quepasapodcast.com/	Paola Marizán	2	17180	18120	13	pasa	2	1

How to share your data

- Create & distribute a self-contained web-interface
  - Can be hosted / maintained very cheaply (even for free)
  - Provides a higher-level interface that doesn't require programming expertise to interact with the data
    - Ex. GitHub pages, GitLab pages


ES COCO
es-coco.github.io/es-coco/
Incognito
Relaunch to update
¿Qué pasa, Midwest?
Paola Marizán
¿QPM? 1: Pulque en América
Sep 2017 · ¿Qué Pasa, Midwest?
Save on Spotify
09:06
00:00:21 - 00:00:23
Que SCONJ
pasa VERB
Midwest PROPN
.
00:00:23 - 00:00:27
Telling VERB
the DET
stories NOUN
of ADP
Latinos PROPN
from ADP
the DET
homeland NOUN
to ADP
the DET
heartland NOUN
.
00:00:27 - 00:00:28
Soy INTJ
Paola PROPN
Marisan PROPN
.
00:00:34 - 00:00:45
In ADP
ancient ADJ
Mexico PROPN
,
the DET
Aztecs PROPN
made VERB
a DET
drink NOUN
by SCONJ
fermenting VERB
the DET
nectar NOUN
of ADP
a DET
cactus NOUN
plant NOUN
,
llamada VERB
magay NOUN
,
para ADP
curar VERB
todos DET
los DET
males NOUN
,
to PART
cure VERB
all DET
kinds NOUN
of ADP
things NOUN
.
00:00:45 - 00:00:51
Pulque PROPN
,
the DET
ancient ADJ
drink NOUN
,
is VERB
milky ADJ
,
slightly ADV
foamy ADJ
y CONJ
algo PRON
pegajoso ADJ
,
somewhat ADV
viscous ADJ
.

Conclusion

Takeaways

- Code-switching is an important aspect of multilingual language usage and should be studied
  - Problem: Lack of usable data
- Our procedure:
  - reduces the barriers to accessing and analyzing audio data
  - is not specific to code-switching: different stages can be adopted for any audio data, especially for under-studied languages or linguistic features
  - can also apply this to audio data collected in the lab

Future work

- Speaker diarization
- Add metadata to database
- Add search / filter + download capabilities to website
- Make detailed procedure publicly available

ES COCO GitHub

Please Submit Your Feedback!

What would you like to do with this type of data?

How could we improve this procedure to better suit your work?

Thank you!

References

Aguilar, G., Kar, S., & Solorio, T. (2020). Lince: A centralized benchmark for linguistic code-switching evaluation. arXiv preprint arXiv:2005.04322. https://doi.org/10.48550/arXiv.2005.04322

Conneau, A., et al. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.https://doi.org/10.48550/arXiv.1911.02116

MacWhinney, B. (2019). TalkBank and SLA. In N. Tracy-Ventura & M. Paquot (Eds.), *The Handbook of SLA and Corpora*. Routledge.

Maher, J. C., (2017). *Multilingualism: A very short introduction*. Oxford University Press. https://doi.org/10.1093/actrade/9780198724995.003.0001.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. Linguistics, 18(7–8). https://doi.org/10.1515/ling.1980.18.7-8.581

Pratap, V., et al. (2023). Scaling speech technology to 1,000+ languages [Preprint]. arXiv. https://arxiv.org/abs/2305.13516

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. https://openai.com/blog/whisper/

Limitations & considerations

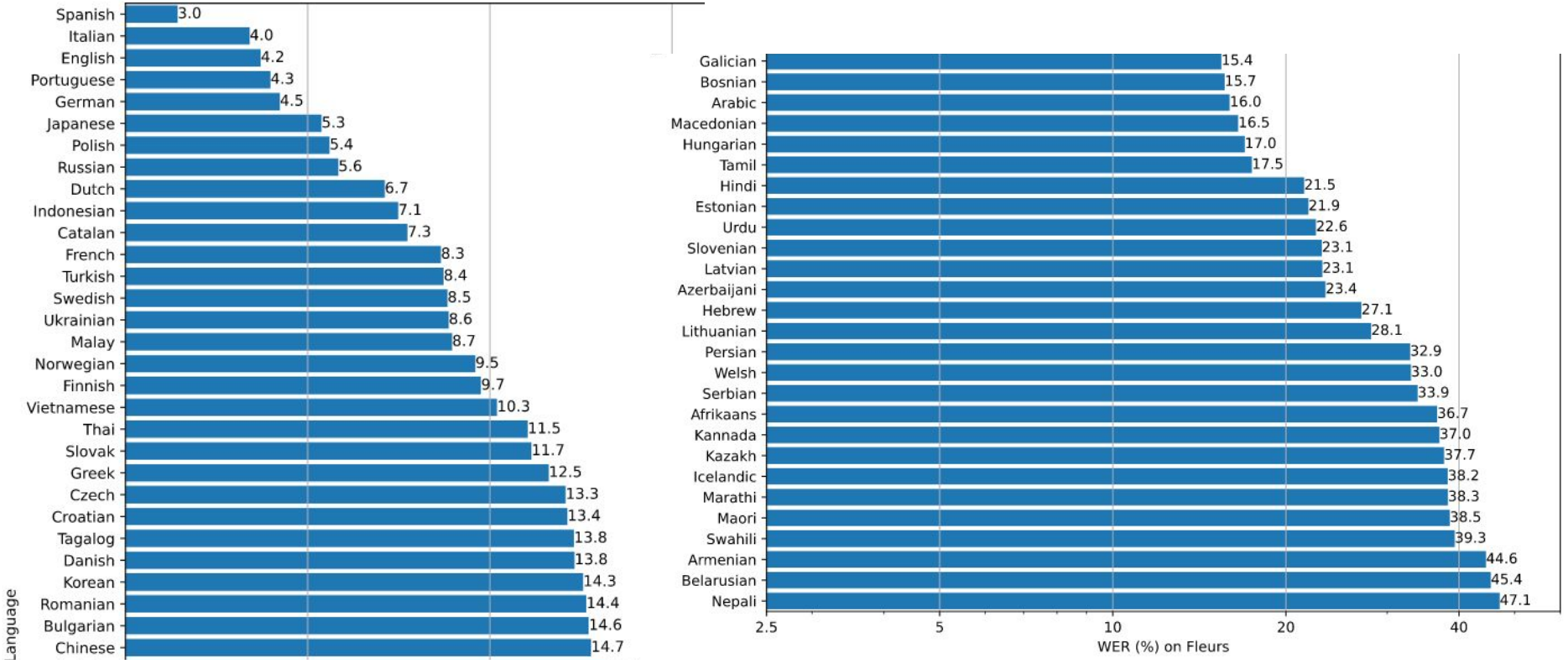
Limitations / considerations and how to address them

- No easy solution to the lack of publicly available content
  - Can address by collecting + hand-annotating more data
- Model accuracy is good but not perfect
  - Can measure via cross-validation
  - Can improve by collecting + hand-annotating more data
- Conversational contexts / content domains may be a biased sample
  - Can measure by analyzing metadata
  - Can address by supplementing with new data

Hand Annotation Reliability: Bangor-Miami Corpus

“For 10% of the transcripts an independent transcription was done, in which a member of the transcription team transcribed one (randomly selected) minute of the recording independently from the original transcriber of that particular transcript. Transcripts were then compared and a rate of similarity was calculated. The average reliability score for independent transcriptions was 83%. Furthermore, all the transcripts were proofread by another member of the transcription team and corrections made accordingly. An additional team of transcribers and checkers included the following researchers in addition to the original transcription team: Margaret Deuchar, Sarah Fairchild, Marika Fusser, Lara Gil Vallejo, Guillermo Montero Melis, Esther Nuñez, Susana Sabin-Fernández, and Jonathan Stammers.”

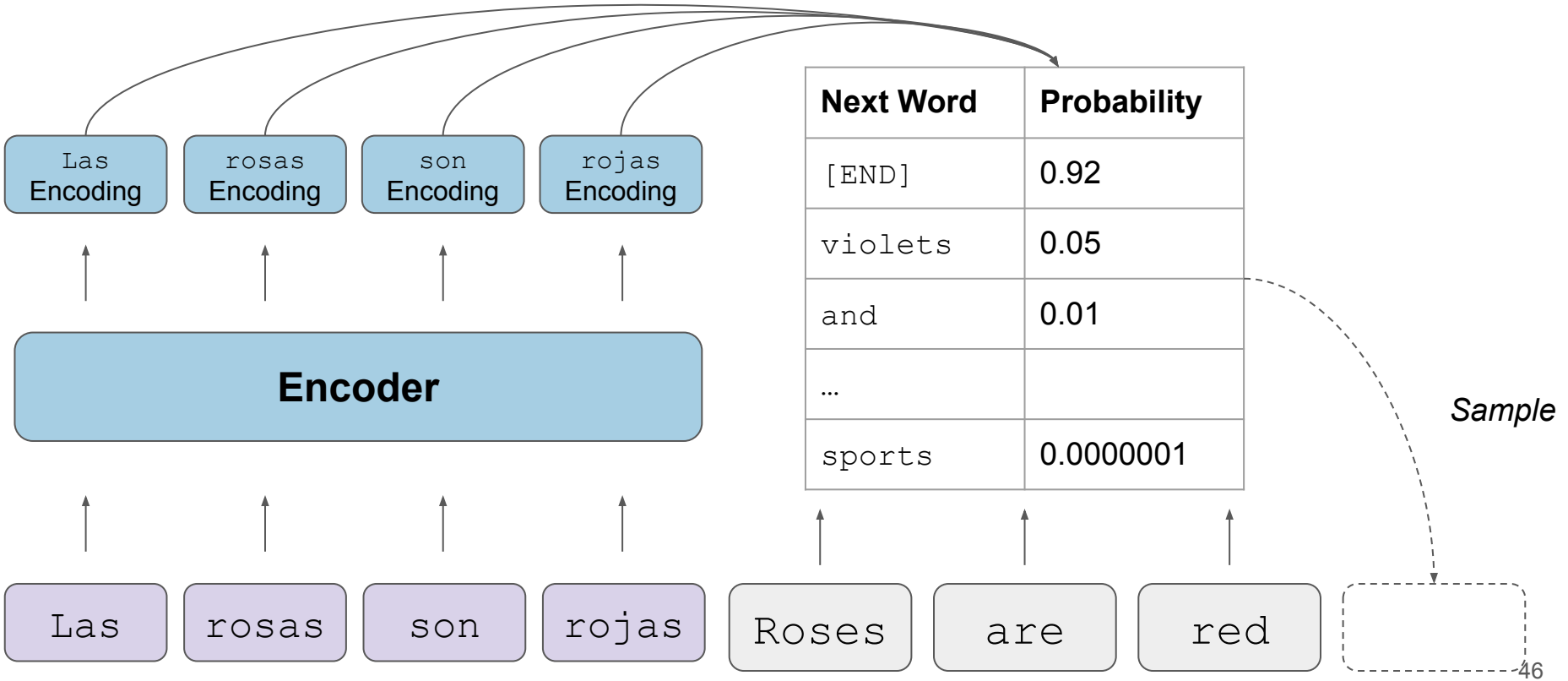
Whisper Word Error Rates (Fleurs dataset)



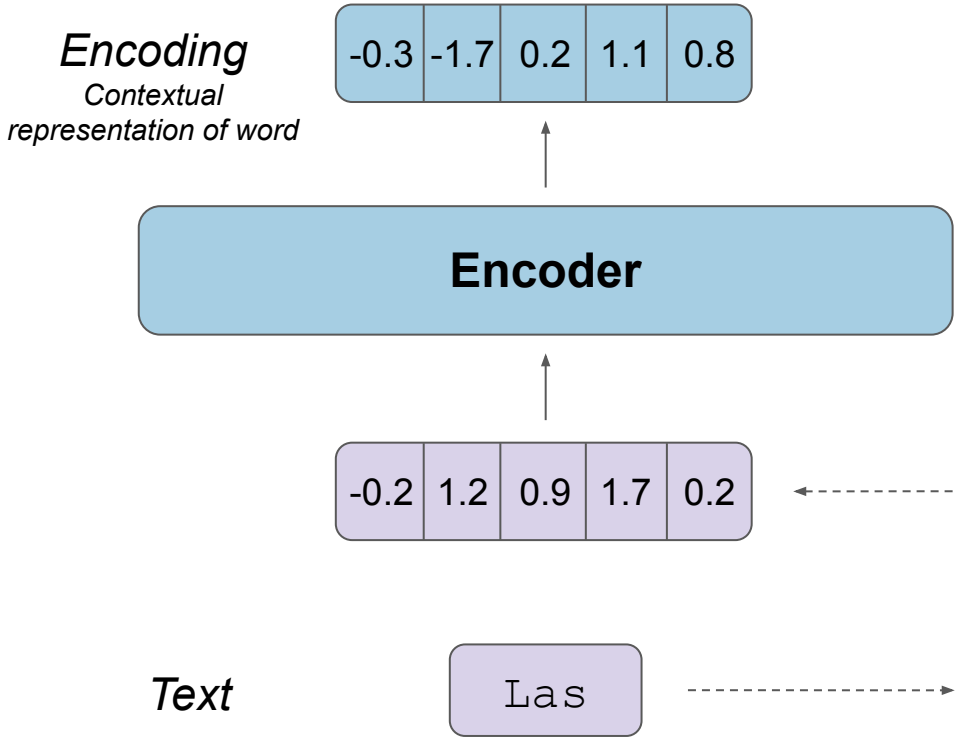
Alternative ASR model for low-resource languages: **MMS** (Pratap et al.)

	Whisper large-v2	MMS L-1107 CC LM LSAH
Amharic	140.3	31.1
Arabic	16.0	21.0
Assamese	106.2	19.2
Azerbaijani	23.4	19.1
Bengali	104.1	12.1
Bulgarian	14.6	13.5
Burmese	115.7	16.0
Catalan	7.3	10.8
Dutch	6.7	14.5
English	4.2	12.3
Filipino	13.8	12.4
Finnish	9.7	23.1
French	8.3	15.0
German	4.5	13.3

Transformer Language Models - **Encoder**-Decoders



Transformer Language Models - **Encodings**



Vector Table

Word	Vector
…	…
Las	
Last	
…	…
Zwickelbier	

Helpful links

GitHub: https://github.com/ES-COCO

Website: https://es-coco.github.io/es-coco/

Feedback form: https://forms.gle/wHKEMj5widQoY3Lk8