



Multilinguality & Speech Translation

Antonios Anastasopoulos

antonis@gmu.edu

<https://nlp.cs.gmu.edu/>

The Languages of the World

The Languages of the World

The Languages of the World

- More than **6000** languages:

The Languages of the World

- More than **6000** languages:
→ 45% oral

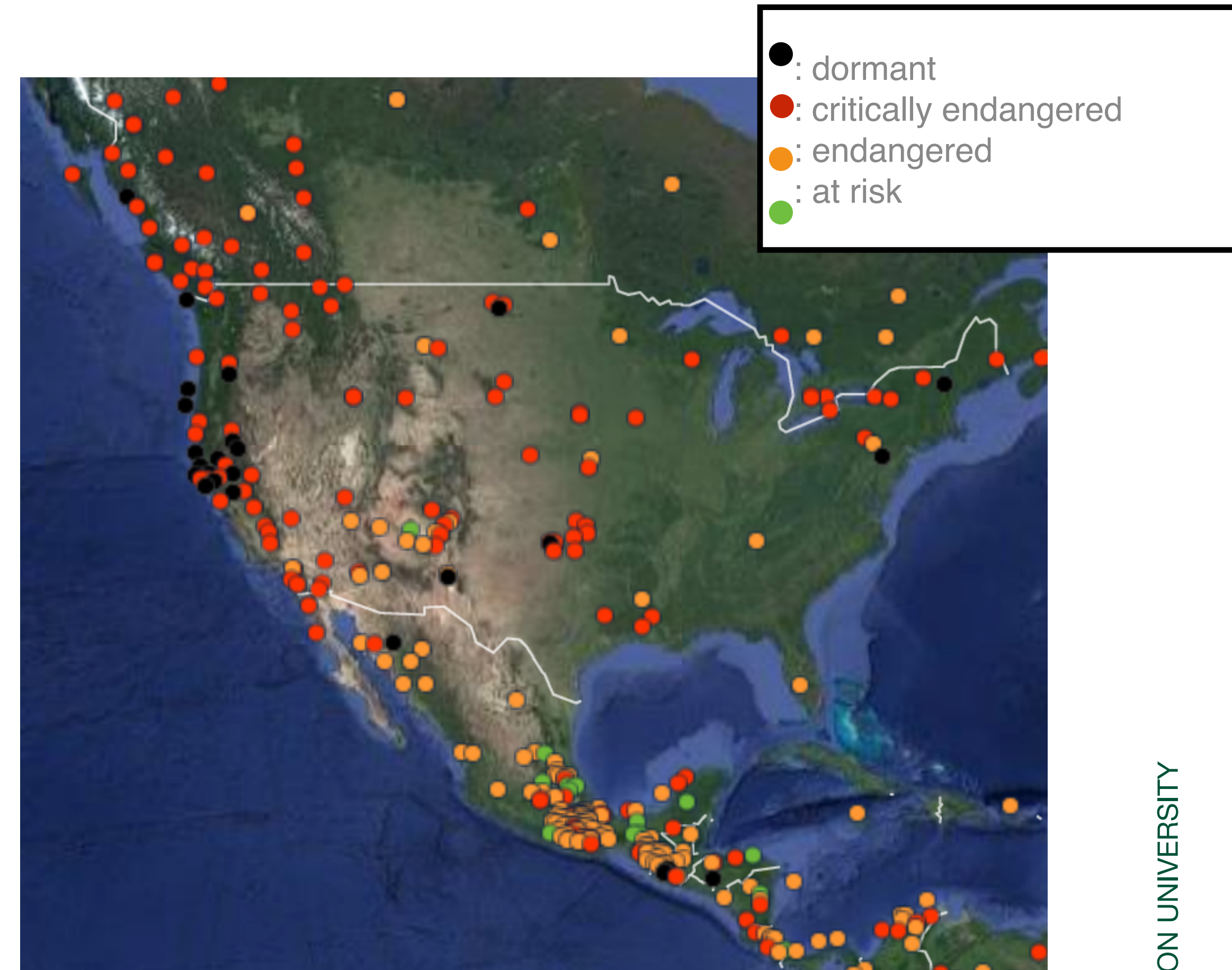


A traditional [Kyrgyz manaschi](#) performing part of the [Epic of Manas](#) at a [yurt](#) camp in [Karakol](#)

Image Source: Wikipedia

The Languages of the World

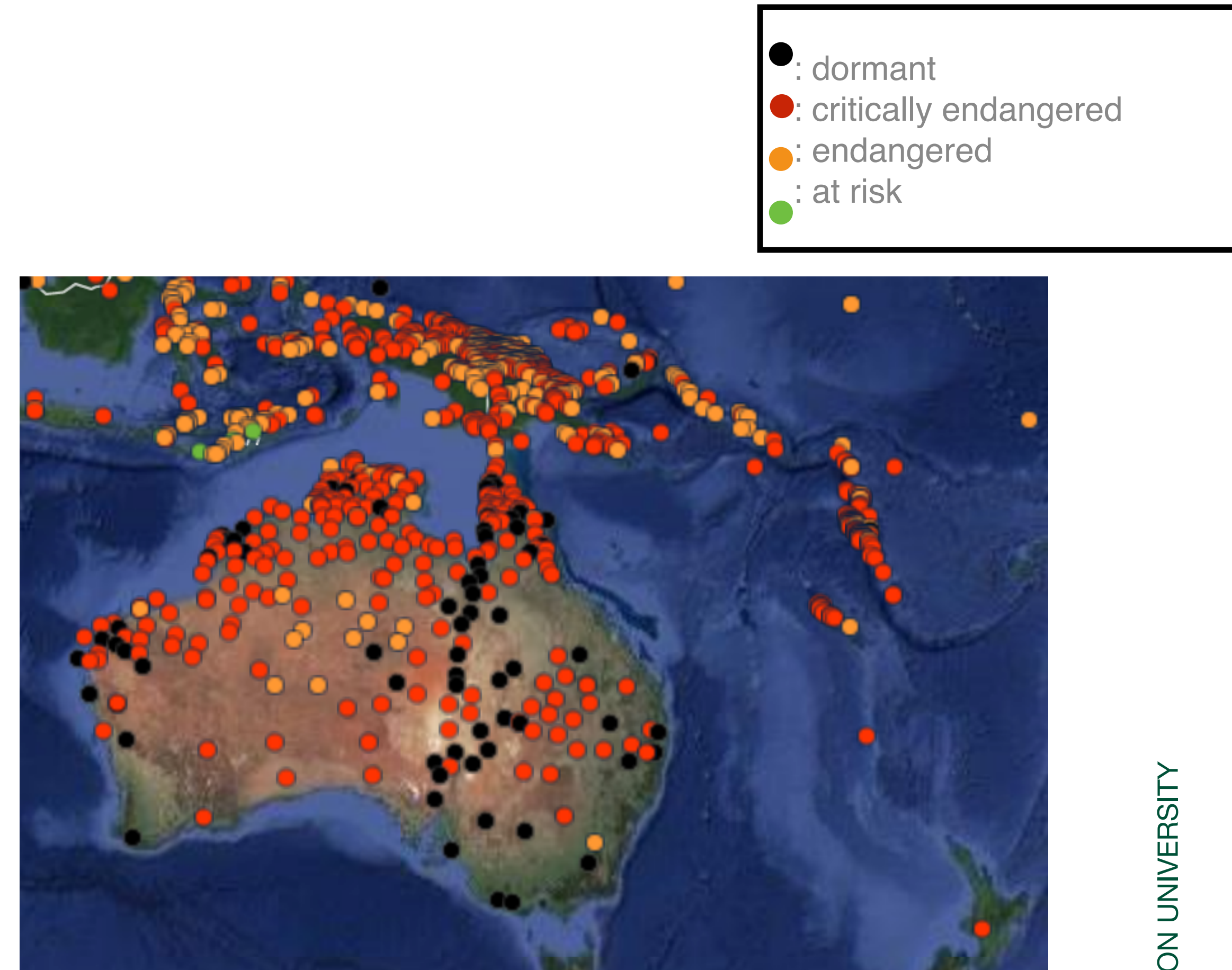
- More than **6000** languages:
 - 45% oral
 - 43% endangered or vulnerable



Source: the Endangered Languages Project

The Languages of the World

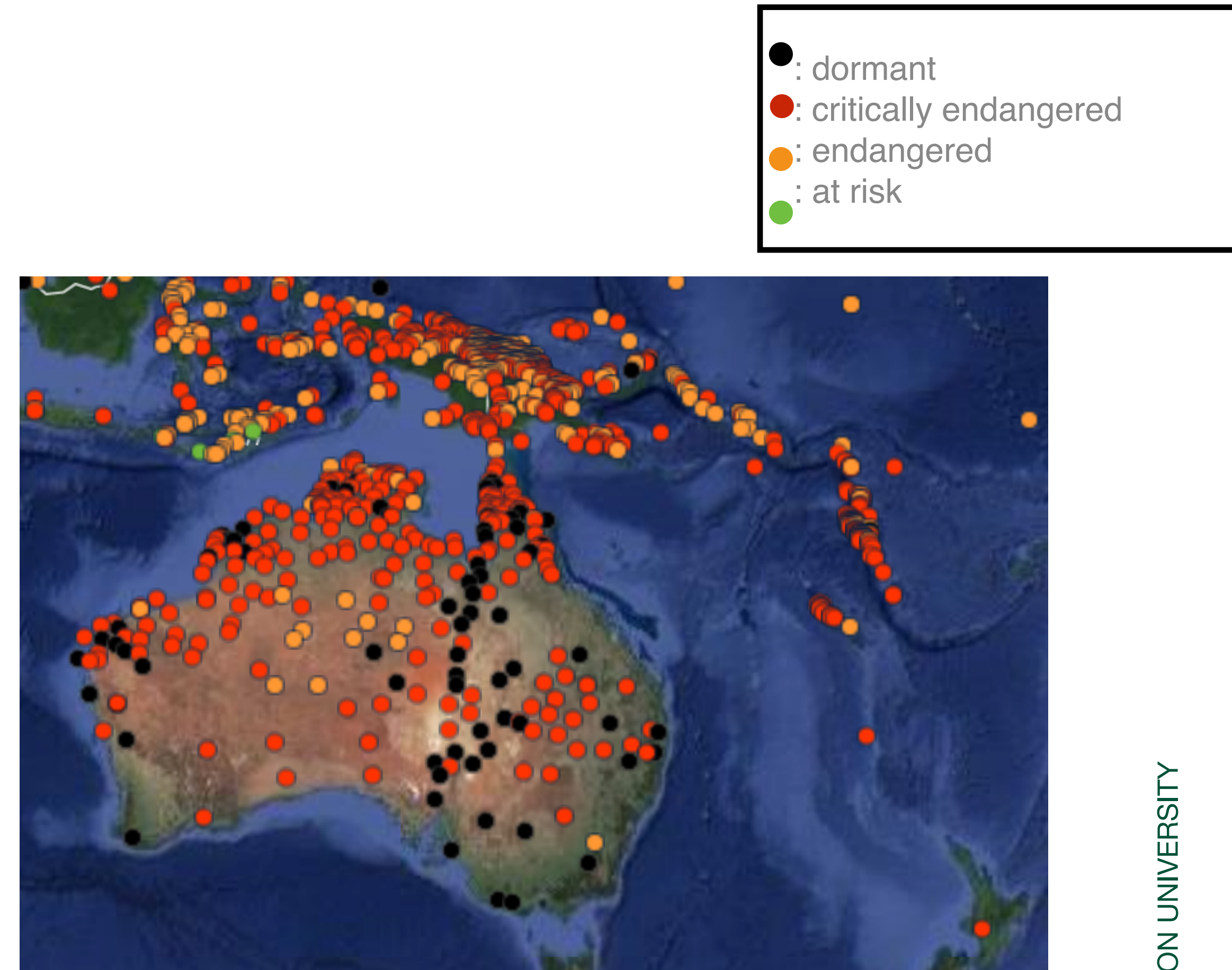
- More than **6000** languages:
 - 45% oral
 - 43% endangered or vulnerable
 - differences in culture, vocabulary



Source: the Endangered Languages Project

The Languages of the World

- More than **6000** languages:
 - 45% oral
 - 43% endangered or vulnerable
 - differences in culture, vocabulary
 - differences in morphological complexity, syntax, tonality, word order...



Source: the Endangered Languages Project

The Languages of the World

- More than **6000** languages:
 - 45% oral
 - 43% endangered or vulnerable
 - differences in culture, vocabulary
 - differences in morphological complexity, syntax, tonality, word order...

But also...



The Languages of the World

- More than **6000** languages:
 - 45% oral
 - 43% endangered or vulnerable
 - differences in culture, vocabulary
 - differences in morphological complexity, syntax, tonality, word order...
- But also...
- regional varieties (dialects)



The Languages of the World

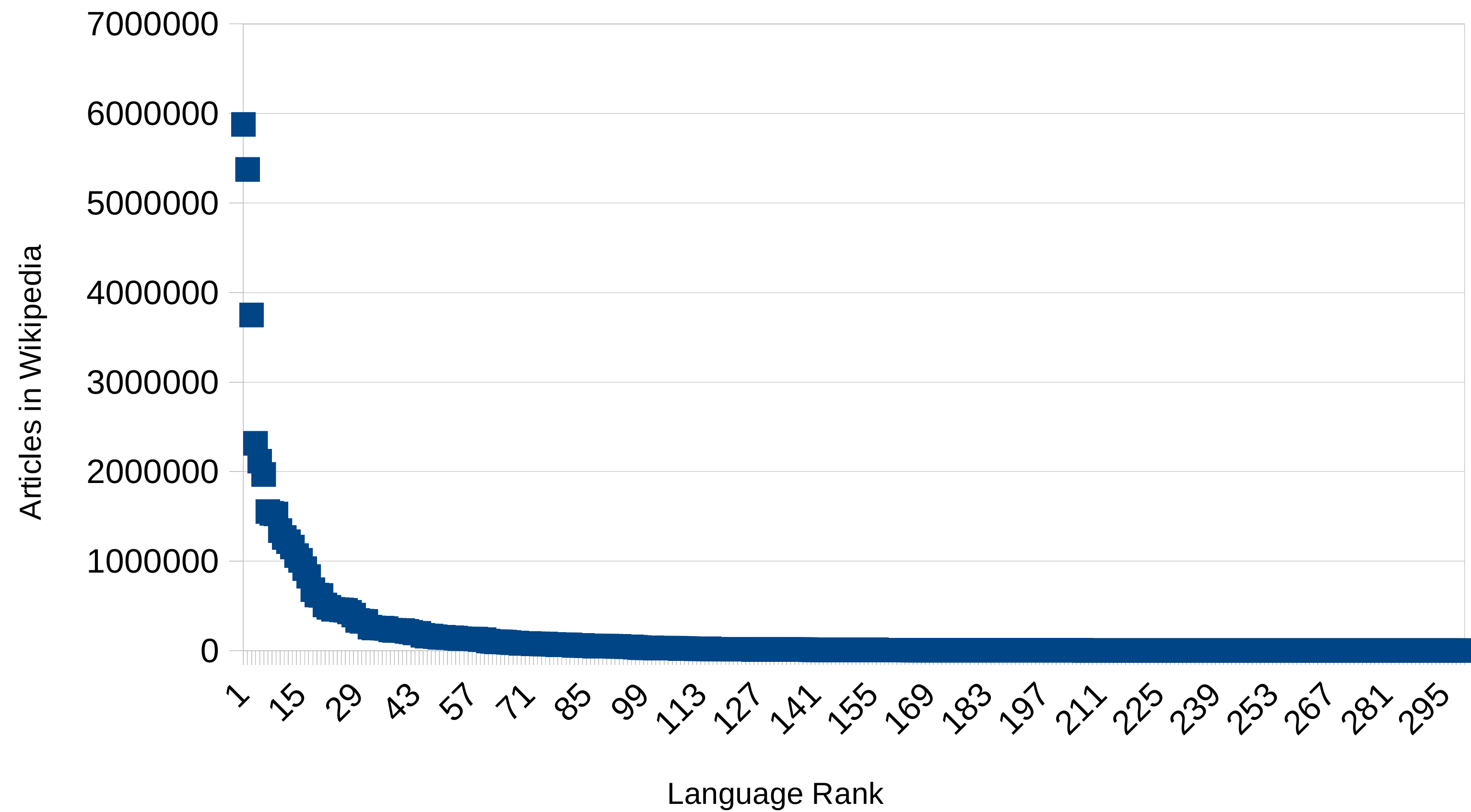
- More than **6000** languages:
 - 45% oral
 - 43% endangered or vulnerable
 - differences in culture, vocabulary
 - differences in morphological complexity, syntax, tonality, word order...

But also...

- regional varieties (dialects)
- L2 speakers
- sign languages



The Long Tail of Data



CASE STUDY: TRANSLATION BETWEEN SIMILAR LANGUAGES

Catalan: Què diu aquesta frase?

Spanish: ¿Qué dice esta oración?

Galician: Que di esta frase?

Portuguese: O que esta frase diz?

CASE STUDY: TRANSLATION BETWEEN SIMILAR LANGUAGES

Catalan: Què diu aquesta frase?

Spanish: ¿Qué dice esta oración?

Galician: Que di esta frase?

Portuguese: O que esta frase diz?

Many similarities to utilize

CASE STUDY: TRANSLATION BETWEEN SIMILAR LANGUAGES

Catalan: Què diu aquesta frase?

Spanish: ¿Qué dice esta oración?

Galician: Que di esta frase?

Portuguese: O que esta frase diz?

Many similarities to utilize

Team	Type	BLEU	TER
MLLPUPV	P	64.7	20.8
UPC-TALP	P	62.1	23.0
NICT	P	53.3	29.1
Uhelsinki	C	52.8	28.6
Uhelsinki	P	52.0	29.4
Uhelsinki	C	51.0	33.1
NICT	C	47.9	33.4
UBC-NLP	P	46.1	36.0
UBC-NLP	C	46.1	35.9
MLLPUPV	C	45.5	35.3
BSC	P	44.0	37.5

Table 27: Results for Spanish to Portuguese Translation

CASE STUDY: TRANSLATION BETWEEN SIMILAR LANGUAGES

Catalan: Què diu aquesta frase?

Spanish: ¿Qué dice esta oración?

Galician: Que di esta frase?

Portuguese: O que esta frase diz?

Team	Type	BLEU	TER
MLLPUPV	P	66.6	19.7
NICT	P	59.9	25.3
Uhelsinki	C	59.1	25.5
Uhelsinki	C	58.6	25.1
Uhelsinki	P	58.4	25.3
KYOTOUNIVERSITY	P	56.9	26.9
NICT	C	54.9	28.4
BSC	P	54.8	29.8
UBC-NLP	P	52.3	32.9
UBC-NLP	C	52.2	32.8
MLLPUPV	C	51.9	30.5
MLLPUPV	C	49.7	32.1
BSC	C	48.5	35.1

Table 26: Results for Portuguese to Spanish Translation

Team	Type	BLEU	TER
MLLPUPV	P	64.7	20.8
UPC-TALP	P	62.1	23.0
NICT	P	53.3	29.1
Uhelsinki	C	52.8	28.6
Uhelsinki	P	52.0	29.4
Uhelsinki	C	51.0	33.1
NICT	C	47.9	33.4
UBC-NLP	P	46.1	36.0
UBC-NLP	C	46.1	35.9
MLLPUPV	C	45.5	35.3
BSC	P	44.0	37.5

Table 27: Results for Spanish to Portuguese Translation

Many similarities to utilize

CASE STUDY: TRANSLATION BETWEEN SIMILAR LANGUAGES

Catalan: Què diu aquesta frase?

Spanish: ¿Qué dice esta oración?

Galician: Que di esta frase?

Portuguese: O que esta frase diz?

Many similarities to utilize

Team	Type	BLEU	TER
MLLPUPV	P	66.6	19.7
NICT	P	59.9	25.3
Uhelsinki	C	59.1	25.5
Uhelsinki	C	58.6	25.1
Uhelsinki	P	58.4	25.3
KYOTOUNIVERSITY	P	56.9	26.9
NICT	C	54.9	28.4
BSC	P	54.8	29.8
UBC-NLP	P	52.3	32.9
UBC-NLP	C	52.2	32.8
MLLPUPV	C	51.9	30.5
MLLPUPV	C	49.7	32.1
BSC	C	48.5	35.1

Table 26: Results for Portuguese to Spanish Translation

Team	Type	BLEU	TER
MLLPUPV	P	64.7	20.8
UPC-TALP	P	62.1	23.0
NICT	P	53.3	29.1
Uhelsinki	C	52.8	28.6
Uhelsinki	P	52.0	29.4
Uhelsinki	C	51.0	33.1
NICT	C	47.9	33.4
UBC-NLP	P	46.1	36.0
UBC-NLP	C	46.1	35.9
MLLPUPV	C	45.5	35.3
BSC	P	44.0	37.5

Table 27: Results for Spanish to Portuguese Translation

Team	Type	BLEU	TER
NITS-CNLP	C	53.7	36.3
Panlingua-KMI	P	11.5	79.1
CMUMEAN	P	11.1	79.7
UBC-NLP	P	08.2	77.1
UBC-NLP	C	08.2	77.2
NITS-CNLP	P	03.7	-
NITS-CNLP	C	03.6	-
CFILT_IITB	C	03.5	-
Panlingua-KMI	C	03.1	-
CFILT_IITB	P	02.8	-
CFILT_IITB	C	02.7	-
Panlingua-KMI	C	01.6	-
JUMT	P	01.4	-

Table 28: Results for Hindi to Nepali Translation

CASE STUDY: INDIAN SUBCONTINENT

এই বাক্যটি কী বলে? আ। वाक्य चूं कडे छे? ಈ ವಾಕ್ಯ ಏನು ಹೇಳುತ್ತದೆ? ਇਹ मन्ना की बहिंदी है?

ഇത് വാചകം എന്താണ് പറയുന്നത്? हि वाक्य क्या कहता है? हे वाक्य काय म्हणते?

ಈ ವಾಕ್ಯಂ ಎಮಿ చెಬುತ್ತುಂದಿ? यो वाक्यले के भन्छछुमम ಲಾಕಷ್ಯಃ ಠಲಿಱಿಱೆಱೆ ಕುಮಿವೆ?

- Phonetic and Orthographic Similarity
- Transliteration and Cognate mining
- Character-level translation

Issues: text normalization, tokenization

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Best WMT system: *The NiuTrans Machine Translation Systems for WMT19*, Li et al. 2019

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Very high resource, but:

Best WMT system: *The NiuTrans Machine Translation Systems for WMT19*, Li et al. 2019

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Very high resource, but:
logographic writing system —> huge vocabulary

Best WMT system: *The NiuTrans Machine Translation Systems for WMT19*, Li et al. 2019

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Very high resource, but:
logographic writing system —> huge vocabulary
tokenization?

Best WMT system: *The NiuTrans Machine Translation Systems for WMT19*, Li et al. 2019

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Very high resource, but:
logographic writing system —> huge vocabulary
tokenization?

Character-based decoding can help
when translating to Chinese (Bowden et al, 2019)

Best WMT system: *The NiuTrans Machine Translation Systems for WMT19*, Li et al. 2019

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Another idea: Modeling sub-character information

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Another idea: Modeling sub-character information

Neural Machine Translation of Logographic Languages
Using Sub-character Level Information, Zhang and Komachi, 2019.

Character	Semantic ideograph	Phonetic ideograph	Pinyin
驰 run	马 horse	也	chH
池 pool	水(氵) water	也	chH
施 impose	方 direction	也	sh
弛 loosen	弓 bow	也	chH
地 land	土 soil	也	dM
驱 drive	马 horse	区	q

Table 1: Examples of decomposed ideographs of Chinese characters. The composing ideographs of different functionality might be shared across different characters.

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Another idea: Modeling sub-character information

CASE STUDY: ENGLISH-CHINESE

what does this sentence mean? 這句話是什麼意思?
这句话是什么意思?

Another idea: Modeling sub-character information

Character-level Chinese-English Translation through ASCII Encoding,
Nikolov et al., 2019.

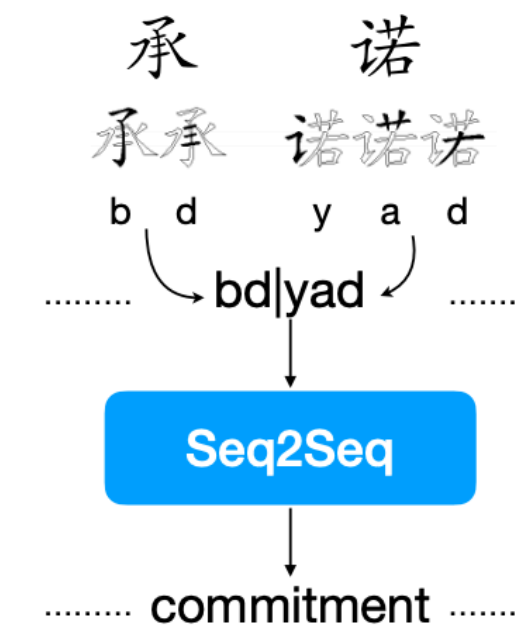


Figure 1: Overview of the **wubi2en** approach to Chinese-to-English translation. A raw Chinese word ('承诺') is encoded into ASCII characters ('bd|yad'), using the Wubi encoding method, before passing it to a Seq2Seq network. The network generates the English translation 'commitment', processing one ASCII character at a time.

CASE STUDY: ARABIC

what does this sentence mean? ماذا تعني هذه الجملة؟

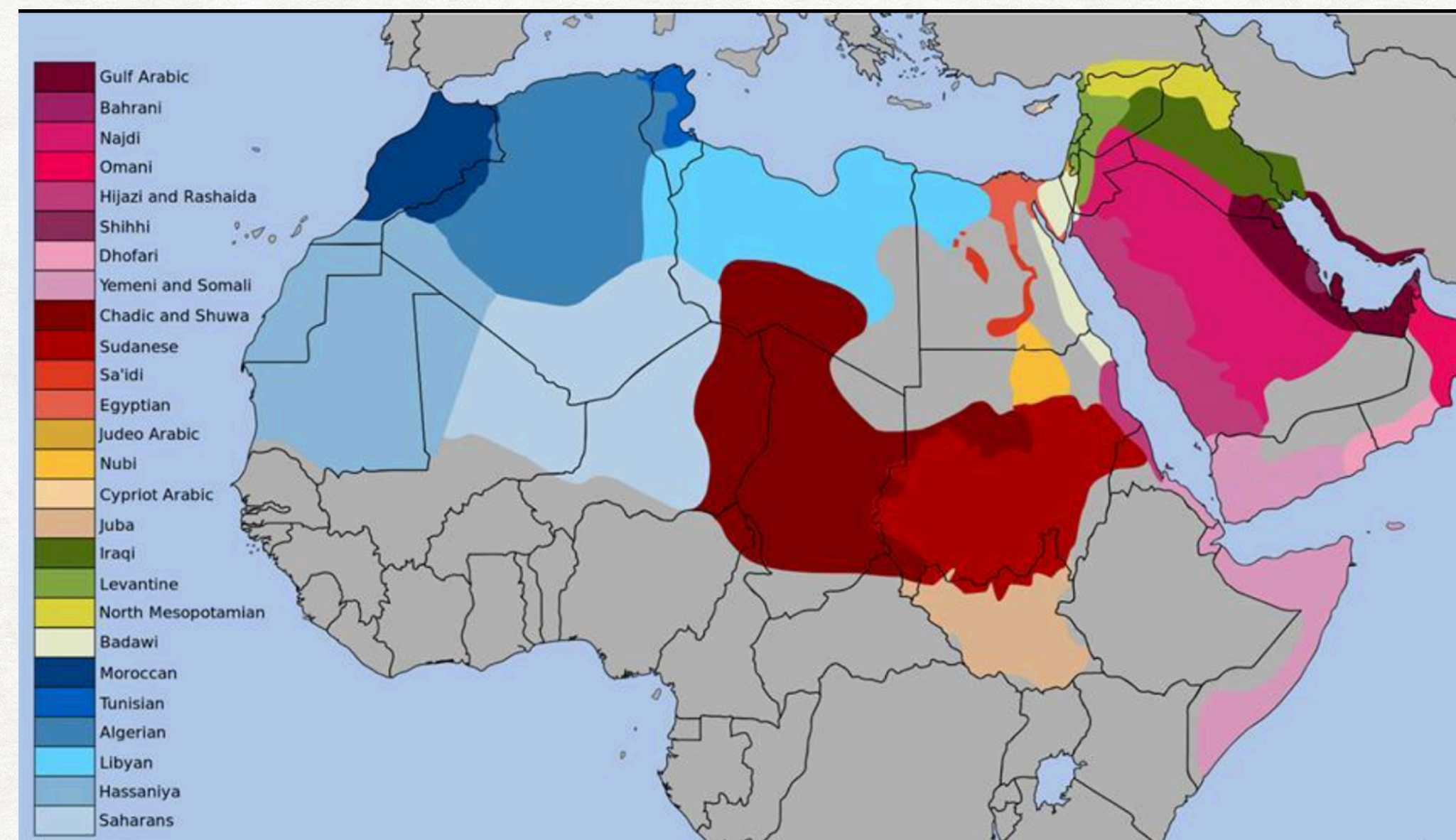
CASE STUDY: ARABIC

what does this sentence mean? ماذا تعني هذه الجملة؟



CASE STUDY: ARABIC

what does this sentence mean? ماذا تعني هذه الجملة؟



CASE STUDY: COMPLEX MORPHOLOGY (E.G. FINNISH, TURKISH)

What about linguistically-informed segmentation?

Words	He admits to shooting girlfriend
BPE	He admits to sho@@ oting gir@@ l@@ friend
Morfessor	He admit@@ s to shoot@@ ing girl@@ friend
Characters	H e _ a d m i t s _ t o _ s h o o t i n g _ g i r l f r i e n d

Table 2: Example with different segmentations.

USING RELATED LANGUAGES

USING RELATED LANGUAGES

How can you choose a related language
for cross-lingual transfer?

USING RELATED LANGUAGES

How can you choose a related language for cross-lingual transfer?

1. Intuition (maaaayyybe ok)

USING RELATED LANGUAGES

How can you choose a related language for cross-lingual transfer?

1. Intuition (maaaayyybe ok)
2. Geography (could be misleading)

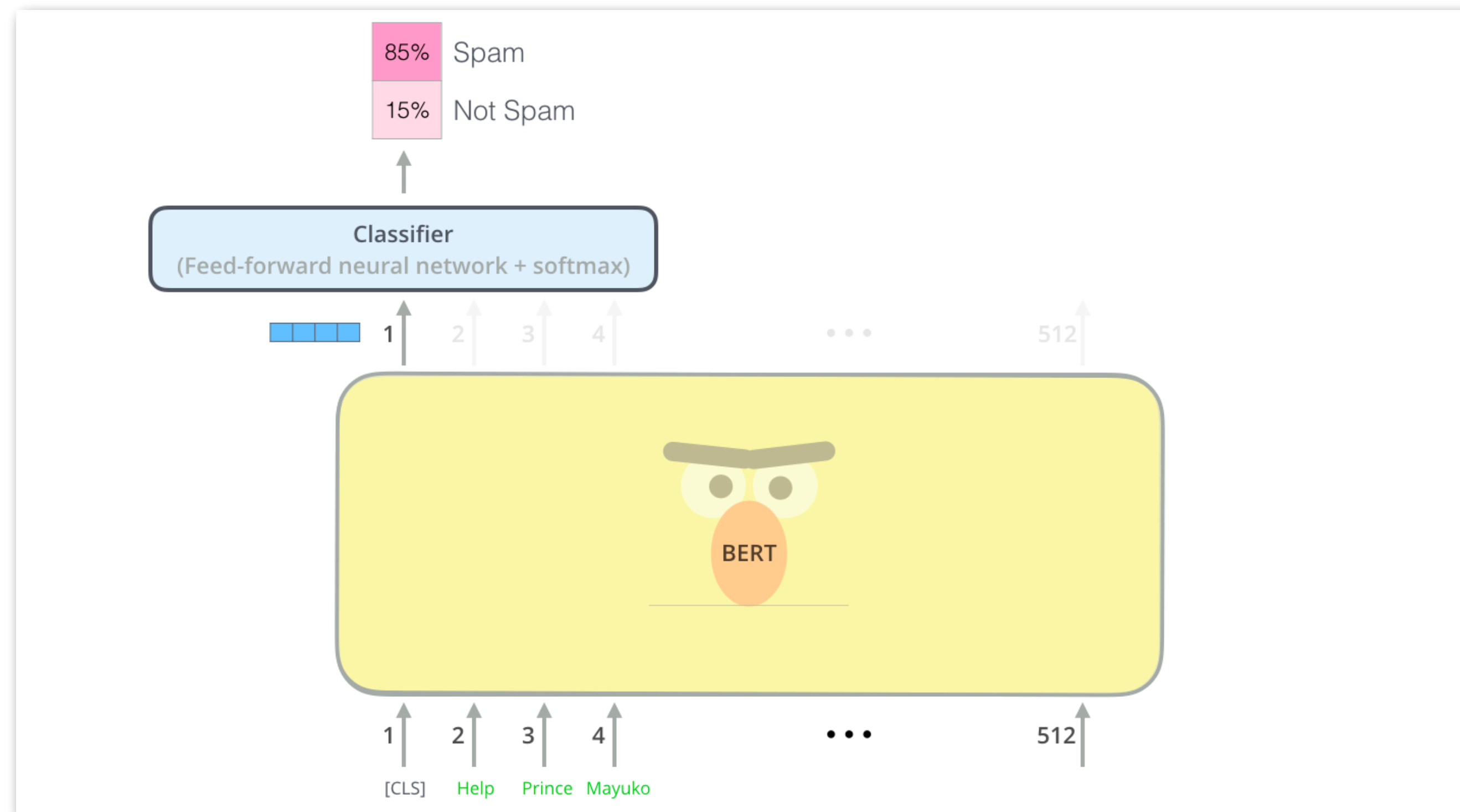


USING RELATED LANGUAGES

How can you choose a related language for cross-lingual transfer?

1. Intuition (maaaayyybe ok)
2. Geography (could be misleading)
3. Typological Features

Some recent trends



Some recent trends

Choose our own adventure

~~Gemini~~

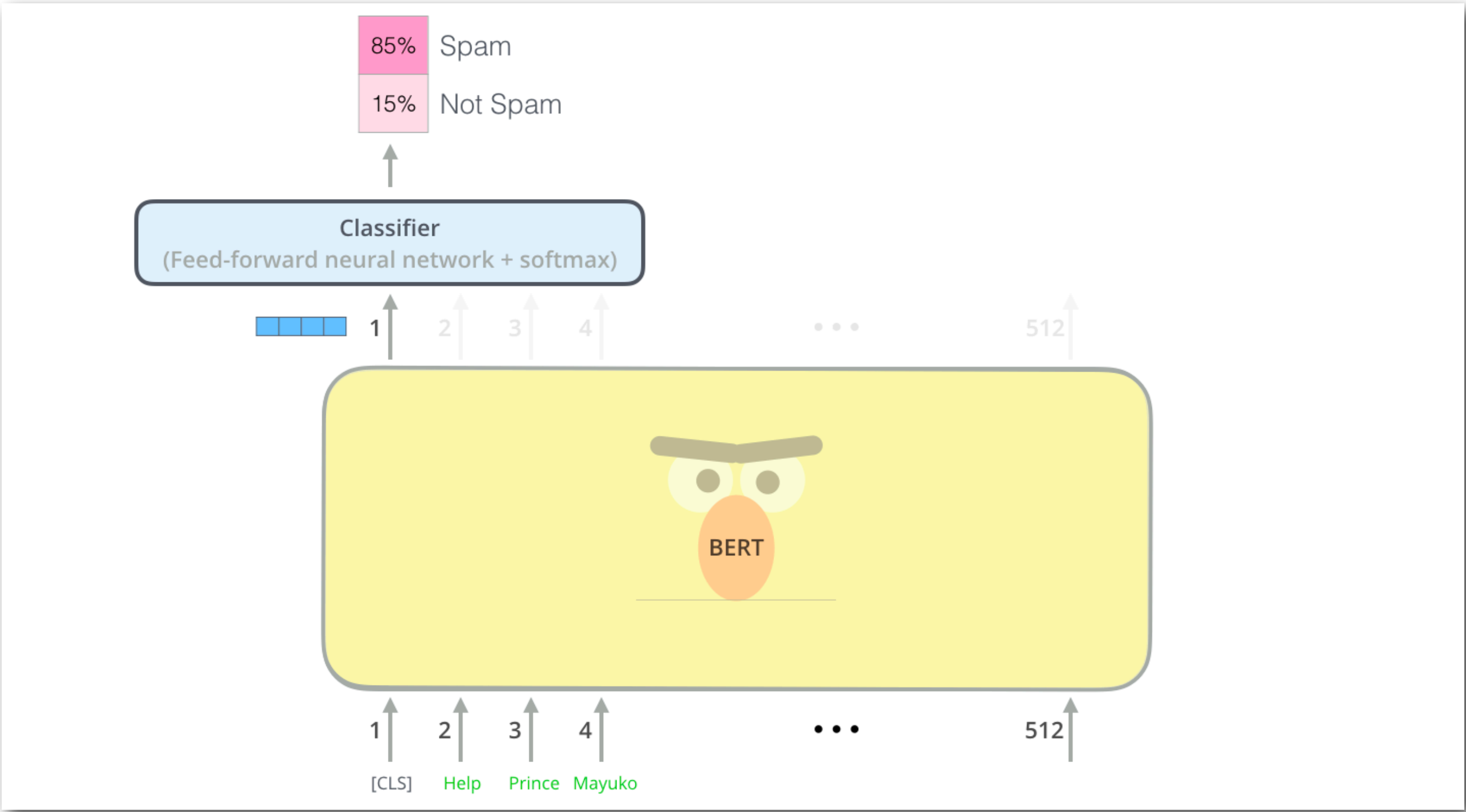
~~GPT 4~~

~~Claude~~

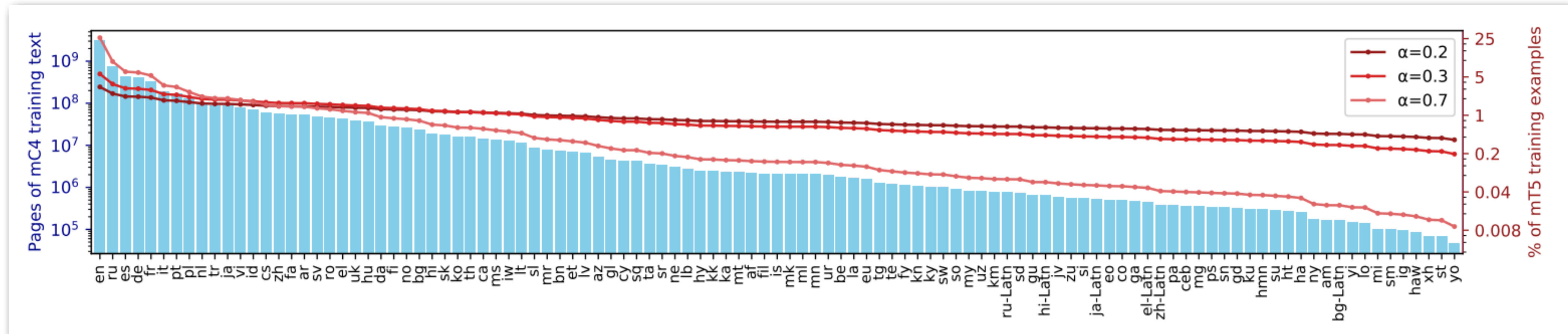
~~PaLM~~

~~GPT 2~~

~~XLNet~~



Make it multilingual!



mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer



Let's make a plan

NLP beyond the top-100 languages

Going Beyond the top-100 Languages



Going Beyond the top-100 Languages



Dominant
Written (Latin)
Standardized
high(ish)-resource

Going Beyond the top-100 Languages



Dominant
Written (Latin)
Standardized
high(ish)-resource

Local
Oral
non-Standardized
Very low-resource

Going Beyond the top-100 Languages

Going Beyond the top-100 Languages

Going Beyond the top-100 Languages

Train on all the internet (GPT-4?) → *incidental multilingualism*

Going Beyond the top-100 Languages

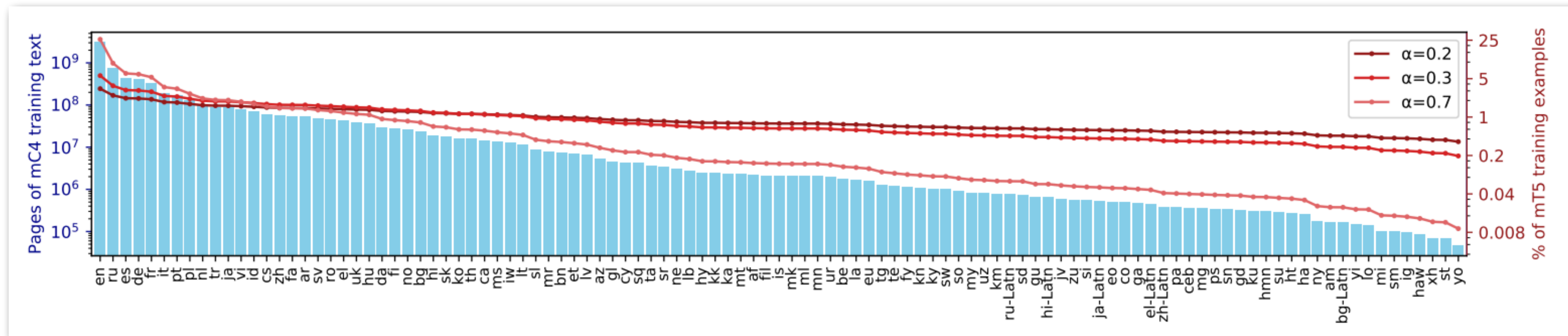
Train on all the internet (GPT-4?) → *incidental multilingualism*
or

Going Beyond the top-100 Languages

Train on all the internet (GPT-4?) → *incidental multilingualism*

or

Explicitly collect data in many languages and upsample low-resource ones



From incidental, to
Multilinguality by Design

Multilinguality *by Design*

We need interventions at all levels:

- data
- evaluation
- tokenization
- architecture
- post-training and alignment

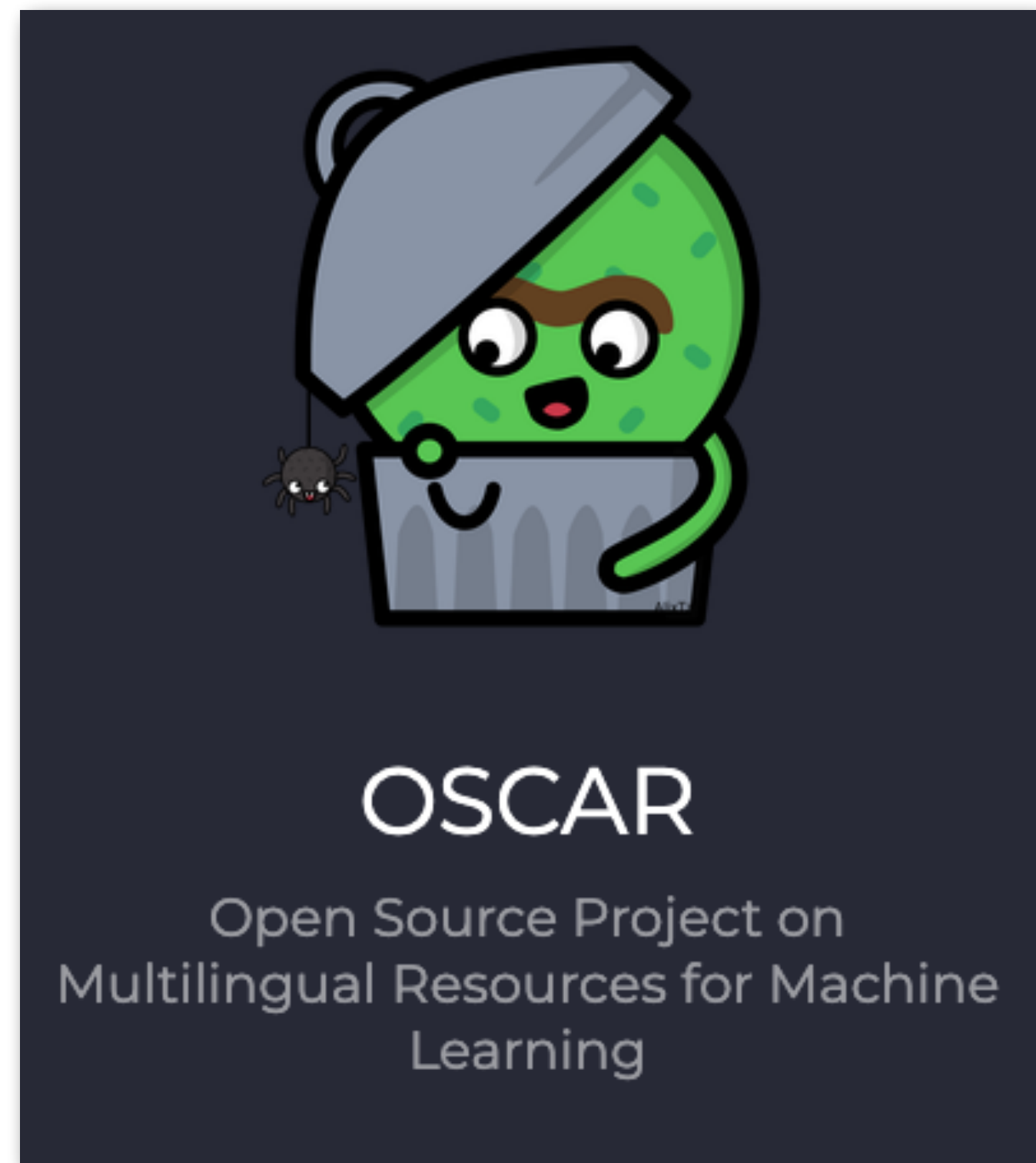
Getting Data - Internet Crawling

Getting Data - Internet Crawling

Getting Data - Internet Crawling

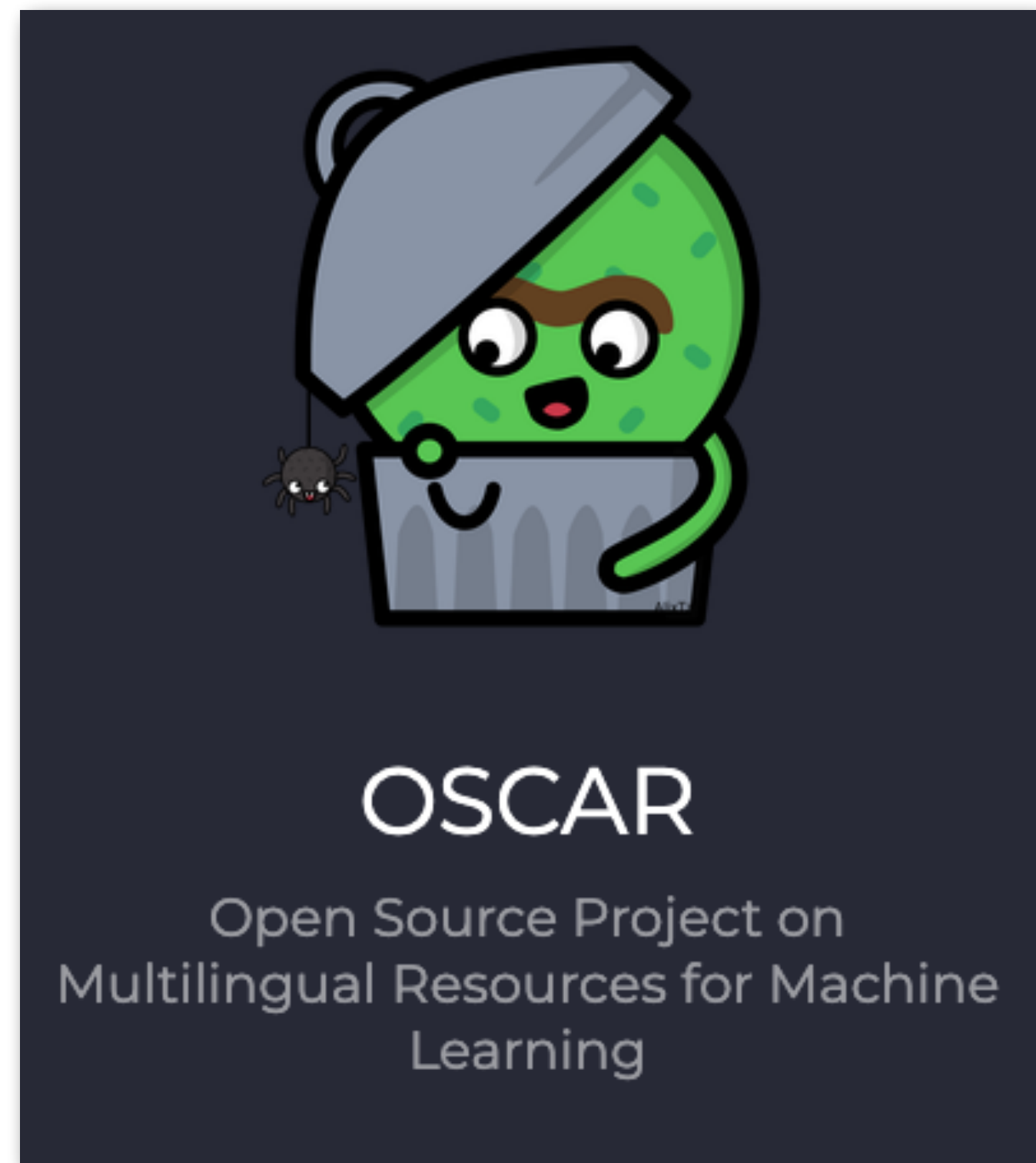


Getting Data - Internet Crawling



Crawling the internet → Language ID
Currently 166 languages

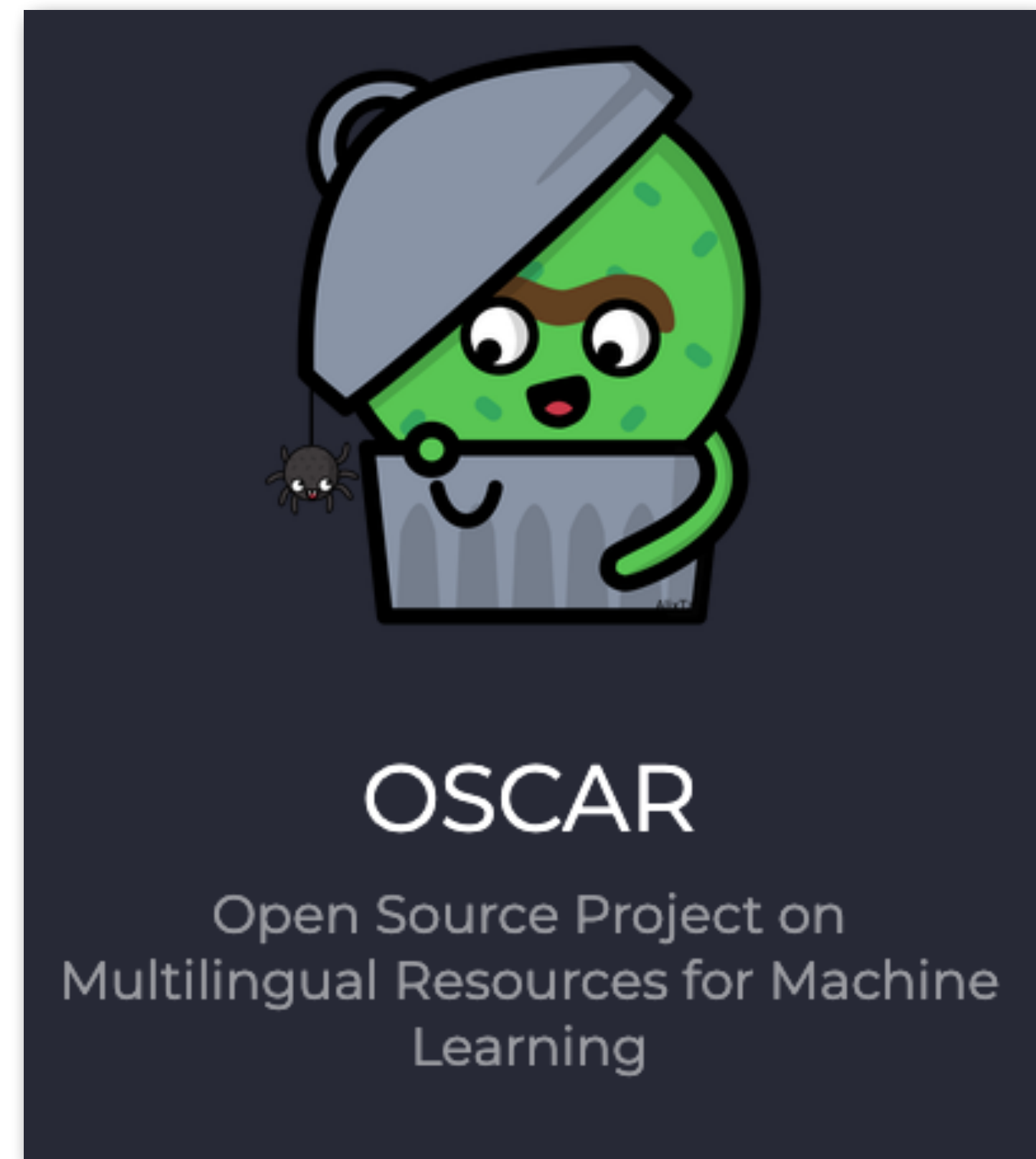
Getting Data - Internet Crawling



Crawling the internet → Language ID
Currently 166 languages

**Quality at a Glance:
An Audit of Web-Crawled Multilingual Datasets**

Getting Data - Internet Crawling

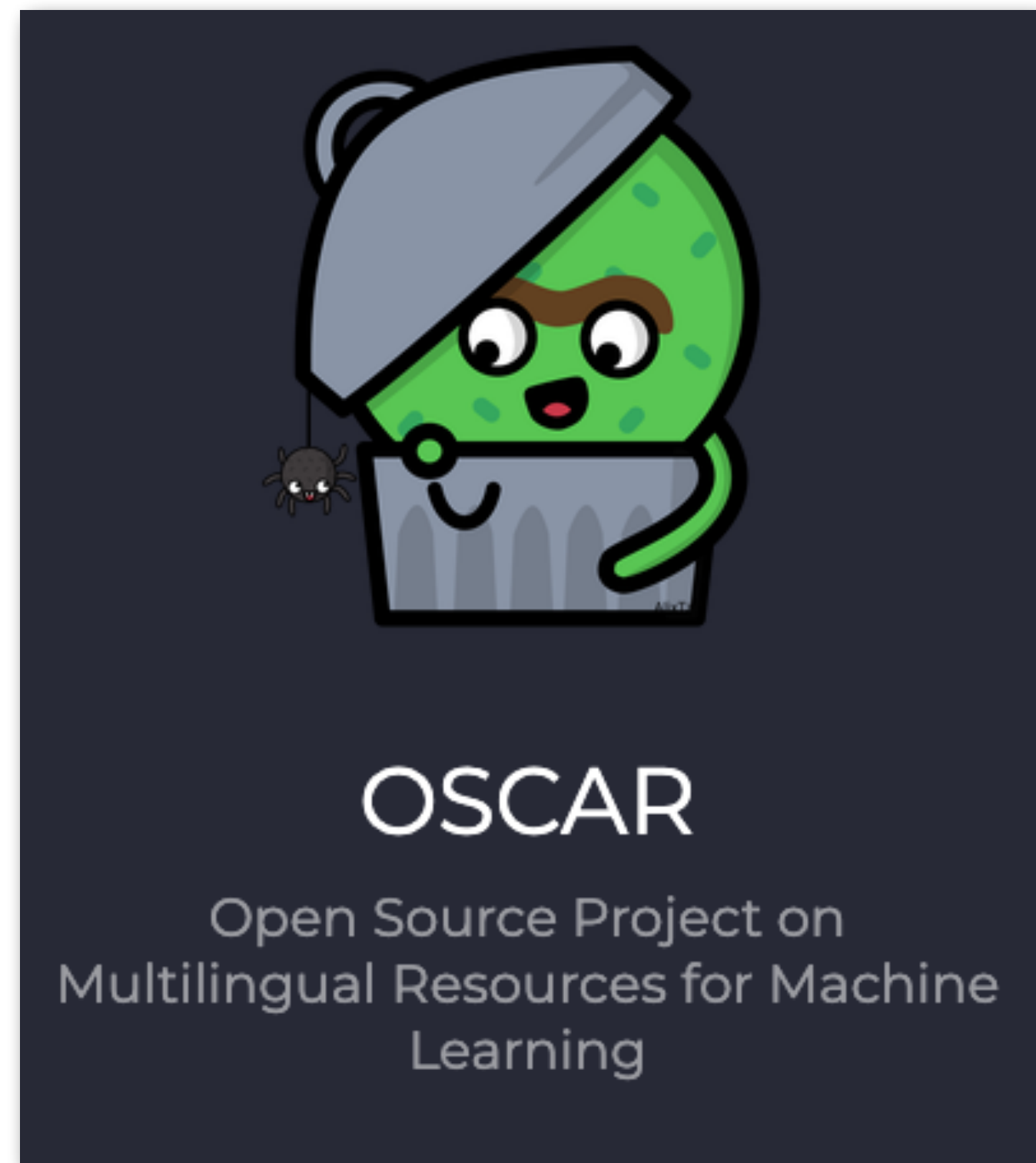


Crawling the internet → Language ID
Currently 166 languages

**Quality at a Glance:
An Audit of Web-Crawled Multilingual Datasets**

Very low quality for some languages
langID far from perfect

Getting Data - Internet Crawling



Crawling the internet → Language ID
Currently 166 languages

**Quality at a Glance:
An Audit of Web-Crawled Multilingual Datasets**

Very low quality for some languages
langID far from perfect

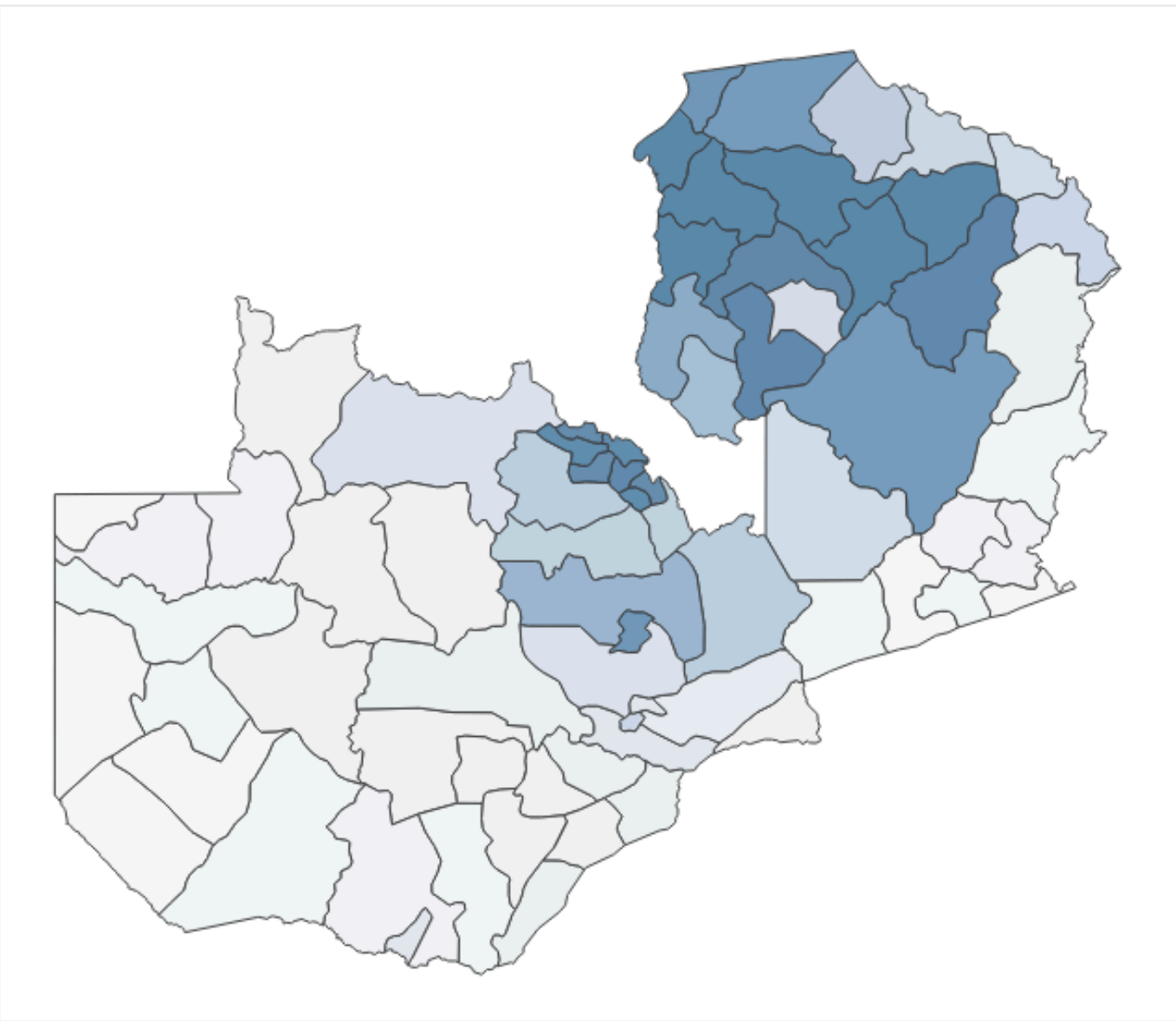


Our Solution: Work with Communities

Our Solution: Work with Communities

Language map of Zambia

Select a language from the menu to see where it's spoken as people's first language
Bemba



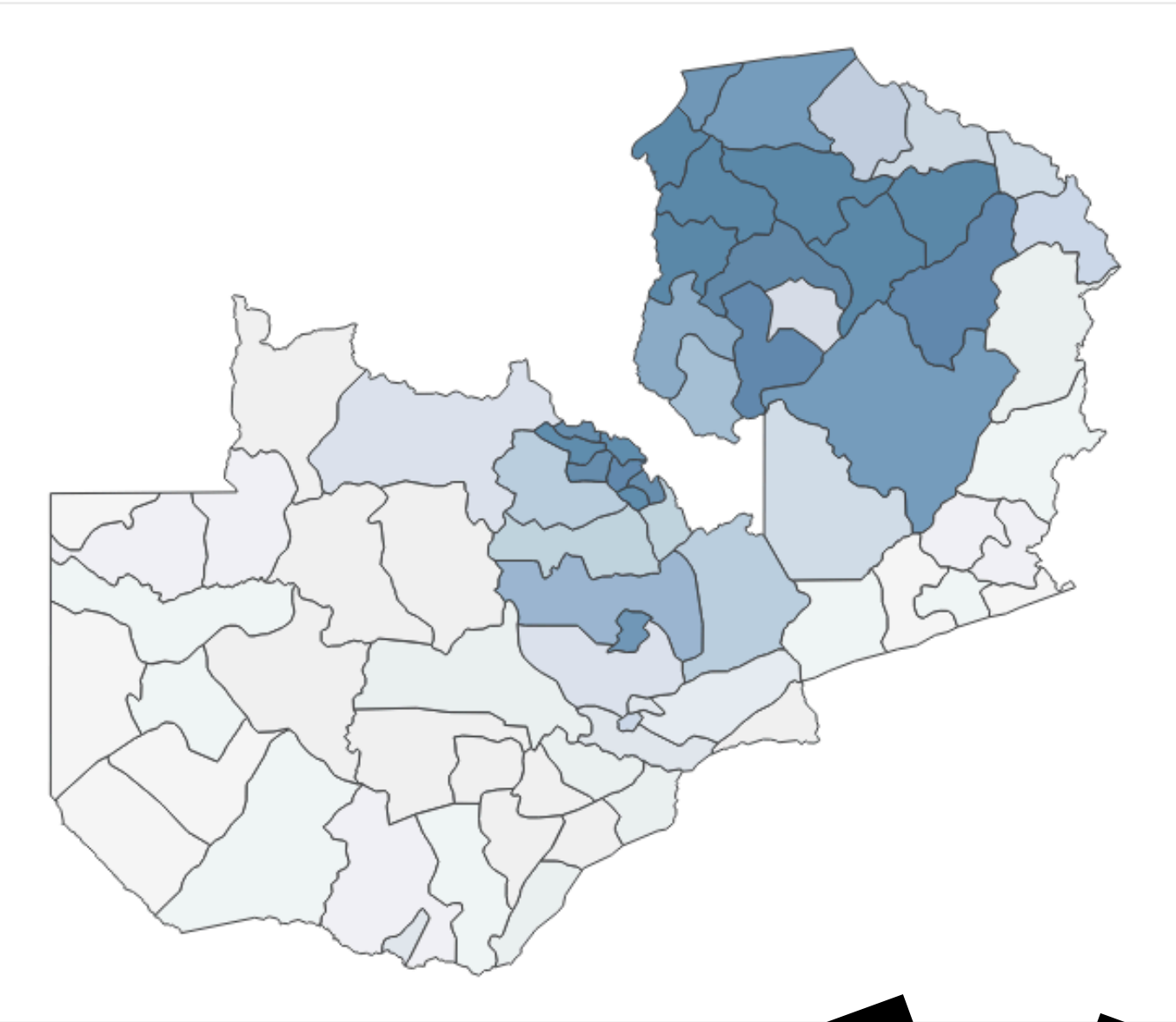
Select a district from the menu to see which languages are spoken as people's first language
All

Language	# of speakers	% of population
Bemba	3,727,677	28.9%
Tonga	1,585,877	12.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanj..	1,305,434	10.1%
Lozi	741,755	5.7%
Lunda	520,643	4.0%
Other langua..	464,474	3.6%
Luvale	416,725	3.2%
Lala-Bisa	379,548	2.9%
Nyamwanga	299,337	2.3%
Nsenga	292,814	2.3%
Mambwe-Lun..	276,006	2.1%
Kaonde	189,173	1.5%
Lamba	189,059	1.5%
Kunda	172,360	1.3%
Ila	150,976	1.2%
Ushi (Aushi)	132,887	1.0%
Soli	88,383	0.7%
Mbunda	77,004	0.6%
English	67,818	0.5%
Taabwa	62,831	0.5%
Lenje	56,770	0.4%
Bwile	49,996	0.4%
Ngoni	44,625	0.3%
Simaa	40,963	0.3%
Nkoya	29,116	0.2%

Our Solution: Work with Communities

Language map of Zambia

Select a language from the menu to see where it's spoken as people's first language
Bemba



Select a district from the menu to see which languages are spoken as people's first language
All

Language	# of speakers	% of population
Bemba	3,727,677	28.9%
Tonga	1,585,877	12.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanj..	1,305,434	10.1%
Lozi	741,755	5.7%
Lunda	520,643	4.0%
Other langua..	464,474	3.6%
Luvale	416,725	3.2%
Lala-Bisa	379,548	2.9%
Nyamwanga	299,337	2.3%
Nsenga	292,814	2.3%
Mambwe-Lun..	276,006	2.1%
Kaonde	189,173	1.5%
Lamba	189,059	1.5%
Kunda	172,360	1.3%
Ila	150,976	1.2%
Ushi (Aushi)	132,887	1.0%
Soli	88,383	0.7%
Mbunda	77,004	0.6%
English	67,818	0.5%
Taabwa	62,831	0.5%
Lenje	56,770	0.4%
Bwile	49,996	0.4%
Ngoni	44,625	0.3%
Simaa	40,963	0.3%
Nkoya	29,116	0.2%

BIG-C: a Multimodal Multi-Purpose Dataset for Bemba

Claytone Sikasote¹, Eunice Mukonde², Md Mahfuz Ibn Alam³, Antonios Anastasopoulos³

¹Department of Computer Science, University of Zambia, Zambia
²Department of Literature and Languages, University of Zambia, Zambia
³Department of Computer Science, George Mason University, USA

claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract

the BIG-C (Bemba Image Grounded Corpus) is a large multimodal dataset for the most populous language in Zambia. It contains a large number of images and videos of people speaking Bemba, which is the most widely spoken language in Zambia. The dataset is designed to support research in natural language processing, computer vision, and human-computer interaction. It is a valuable resource for the development of language processing systems for Bemba and other languages in Zambia.

BembaSpeech: A Speech Recognition Corpus for the Bemba Language

Claytone Sikasote¹ and Antonios Anastasopoulos²

¹Department of Computer Science, University of Zambia, Zambia
²Department of Computer Science, George Mason University, USA

claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract

We present a preprocessed, ready-to-use automatic speech recognition corpus, BembaSpeech, consisting over 24 hours of read speech in the Bemba language, a written but low-resourced language spoken by over 30% of the population in Zambia. To assess its usefulness for training and testing ASR systems for Bemba, we explored different approaches; supervised pre-training (training from scratch), cross-lingual transfer learning from a monolingual English pre-trained model using DeepSpeech on the portion of the dataset, and fine-tuning large scale self-supervised Wav2Vec2.0 based multilingual pre-trained models on the portion of the dataset. From our experiments, the 1 billion XLS-R parameter model gives the best results. The pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained models. The pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained models. The pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained models.

Keywords:

Automatic Speech Recognition, ASR corpus, Low-resource language, Bemba, Zambia

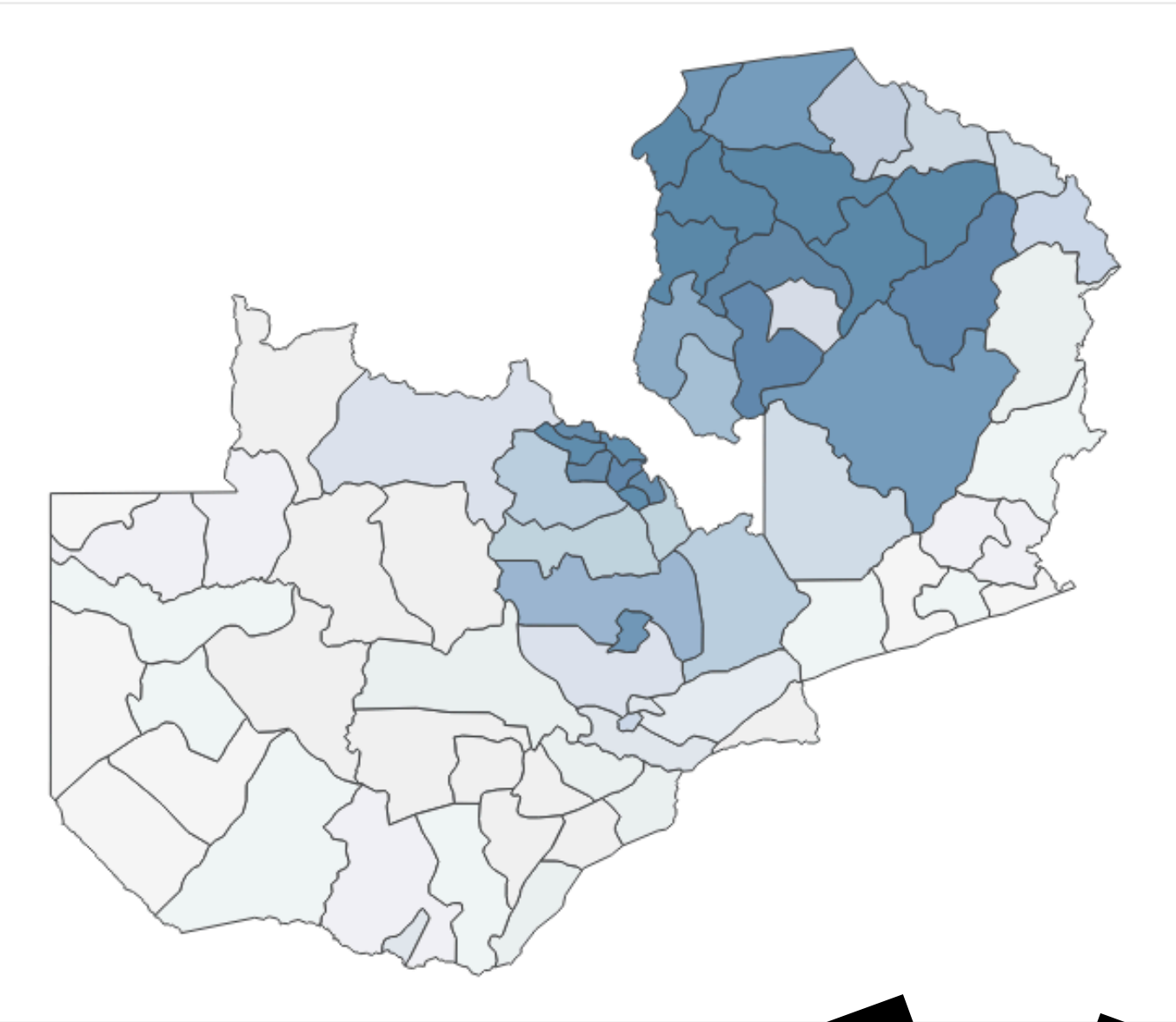
1. Introduction

Speech-to-Text, also known as Automatic Speech Recognition (ASR), is a technology that converts spoken language into written text. It is a key component of many modern communication systems, such as voice search, virtual assistants, and transcription services. However, ASR systems are typically designed for high-resource languages like English, and they often perform poorly for low-resource languages like Bemba. This is because ASR systems rely on large amounts of training data to learn the patterns of a language, and Bemba has very little training data available. In this paper, we present BembaSpeech, a new ASR corpus for the Bemba language. BembaSpeech consists of over 24 hours of read speech in Bemba, which is a written but low-resourced language spoken by over 30% of the population in Zambia. We explore different approaches to train ASR systems for Bemba, including supervised pre-training, cross-lingual transfer learning, and fine-tuning large scale self-supervised models. Our experiments show that the 1 billion XLS-R parameter model gives the best results. The pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained models. The pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained models.

Our Solution: Work with Communities

Language map of Zambia

Select a language from the menu to see where it's spoken as people's first language
Bemba



Select a district from the menu to see which languages are spoken as people's first language
All

Language	# of speakers	% of population
Bemba	3,727,677	28.9%
Tonga	1,585,877	12.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanj..	1,305,434	10.1%
Lozi	741,755	5.7%
Lunda	520,643	4.0%
Other langua..	464,474	3.6%
Luvala	416,725	3.2%
Lala-Bisa	379,548	2.9%
Nyamwanga	299,337	2.3%
Nsenga	292,814	2.3%
Mambwe-Lun..	276,006	2.1%
Kaonde	189,173	1.5%
Lamba	189,059	1.5%
Kunda	172,360	1.3%
Ila	150,976	1.2%
Ushi (Aushi)	132,887	1.0%
Soli	88,383	0.7%
Mbunda	77,004	0.6%
English	67,818	0.5%
Taabwa	62,831	0.5%
Lenje	56,770	0.4%
Bwile	49,996	0.4%
Ngoni	44,625	0.3%
Simaa	40,963	0.3%
Nkoya	29,116	0.2%



BIG-C: a Multimodal Multi-Purpose Dataset for Bemba
Claytone Sikasote¹, Eunice Mukonde², Md Mahfuz Ibn Alam³, Antonios Anastasopoulos³
¹Department of Computer Science, University of Zambia, Zambia
²Department of Literature and Languages, University of Zambia, Zambia
³Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract
BIG-C (Bemba Image Grounded Corpus) is a large multimodal dataset for the Bemba language, the most populous language in Zambia. It contains a wide variety of images and audio recordings, capturing the rich cultural and linguistic diversity of the community. This dataset is designed to support research in natural language processing, computer vision, and human-computer interaction, particularly in the context of low-resource languages and communities.

BembaSpeech: A Speech Recognition Corpus for the Bemba Language
Claytone Sikasote¹ and Antonios Anastasopoulos²
¹Department of Computer Science, University of Zambia, Zambia
²Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract
We present a preprocessed, ready-to-use automatic speech recognition corpus, BembaSpeech, consisting of over 24 hours of read speech in the Bemba language, a written but low-resourced language spoken by over 30% of the population in Zambia. To assess its usefulness for training and testing ASR systems for Bemba, we explored different approaches: supervised pre-training (training from scratch), cross-lingual transfer learning from a monolingual English pre-trained model using DeepSpeech on the portion of the dataset, and fine-tuning large scale self-supervised Wav2Vec2.0 based multilingual pre-trained models on the portion of BembaSpeech corpus. From our experiments, the 1 billion XLS-R parameter model gives the best results. The pre-trained models transfers cross-lingual acoustic representation better than monolingual representations. The pre-trained models for Bemba ASR. Lastly, results also show that the corpus can be used for other tasks such as speech-to-text, text-to-speech, and machine translation.

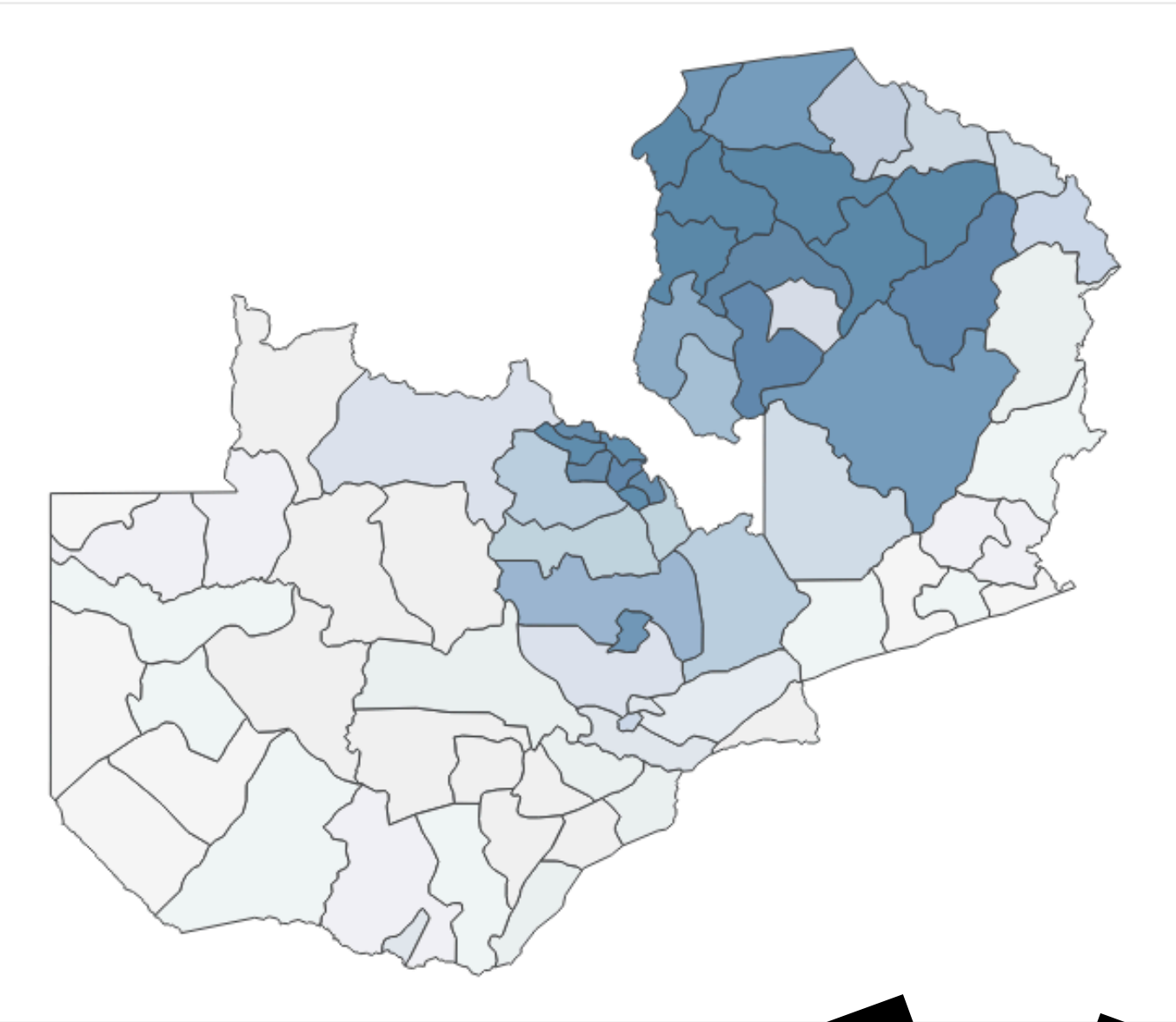
Keywords: Automatic Speech Recognition, ASR corpus, Low-resource language, Bemba, Zambia

1. Introduction
Speech-to-Text, also known as Automatic Speech Recognition (ASR), is a technology that converts spoken language into written text. It is a crucial component in many applications, including voice search, virtual assistants, and transcription services. However, ASR systems are typically trained on large amounts of data from high-resource languages, leaving low-resource languages like Bemba with limited support. This paper introduces BembaSpeech, a new ASR corpus for the Bemba language, and discusses its construction and potential applications.

Our Solution: Work with Communities

Language map of Zambia

Select a language from the menu to see where it's spoken as people's first language
Bemba



Select a district from the menu to see which languages are spoken as people's first language
All

Language	# of speakers	% of population
Bemba	3,727,677	28.9%
Tonga	1,585,877	12.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanj..	1,305,434	10.1%
Lozi	741,755	5.7%
Lunda	520,643	4.0%
Other langua..	464,474	3.6%
Luvale	416,725	3.2%
Lala-Bisa	379,548	2.9%
Nyamwanga	299,337	2.3%
Nsenga	292,814	2.3%
Mambwe-Lun..	276,006	2.1%
Kaonde	189,173	1.5%
Lamba	189,059	1.5%
Kunda	172,360	1.3%
Ila	150,976	1.2%
Ushi (Aushi)	132,887	1.0%
Soli	88,383	0.7%
Mbunda	77,004	0.6%
English	67,818	0.5%
Taabwa	62,831	0.5%
Lenje	56,770	0.4%
Bwile	49,996	0.4%
Ngoni	44,625	0.3%
Simaa	40,963	0.3%
Nkoya	29,116	0.2%



BIG-C: a Multimodal Multi-Purpose Dataset for Bemba

Claytone Sikasote¹, Eunice Mukonde², Md Mahfuz Ibn Alam³, Antonios Anastasopoulos³

¹Department of Computer Science, University of Zambia, Zambia
²Department of Literature and Languages, University of Zambia, Zambia
³Department of Computer Science, George Mason University, USA

claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract

BIG-C (Bemba Image Grounded Corpus) is a large multimodal dataset for the Bemba language, the most populous language in Zambia. It contains a wide variety of audio, video, and text data, including spontaneous speech, formal speeches, and educational materials. The dataset is designed to support research in natural language processing, speech recognition, and machine learning for the Bemba language.

BembaSpeech: A Speech Recognition Corpus for the Bemba Language

Claytone Sikasote¹ and Antonios Anastasopoulos²

¹Department of Computer Science, University of Zambia, Zambia
²Department of Computer Science, George Mason University, USA

claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract

We present a preprocessed, ready-to-use automatic speech recognition corpus, BembaSpeech, consisting of over 24 hours of read speech in the Bemba language, a written but low-resourced language spoken by over 30% of the population in Zambia. To assess its usefulness for training and testing ASR systems for Bemba, we explored different approaches: supervised pre-training (training from scratch), cross-lingual transfer learning from a monolingual English pre-trained model using DeepSpeech on the portion of the dataset, and fine-tuning large scale self-supervised Wav2Vec2.0 based multilingual pre-trained models on the portion of BembaSpeech corpus. From our experiments, the 1 billion XLS-R parameter model gives the best results. The pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained models. The pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained models. The pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained models.

Keywords:

Automatic Speech Recognition, ASR corpus, Low-resource language, Bemba

1. Introduction

Speech-to-Text, also known as Automatic Speech Recognition (ASR), is a technology that converts spoken language into written text. It is a key component of many modern communication systems, including voice search, virtual assistants, and transcription services. However, ASR systems are typically trained on large amounts of data from high-resource languages, leaving low-resource languages like Bemba with limited access to these technologies. In this work, we develop a tool for educational

Educational Tools for Mapuzugun

Cristian Ahumada¹ Claudio Gutierrez¹ Antonios Anastasopoulos²

¹Department of Computer Science, Universidad de Chile
²Computer Science Department, George Mason University

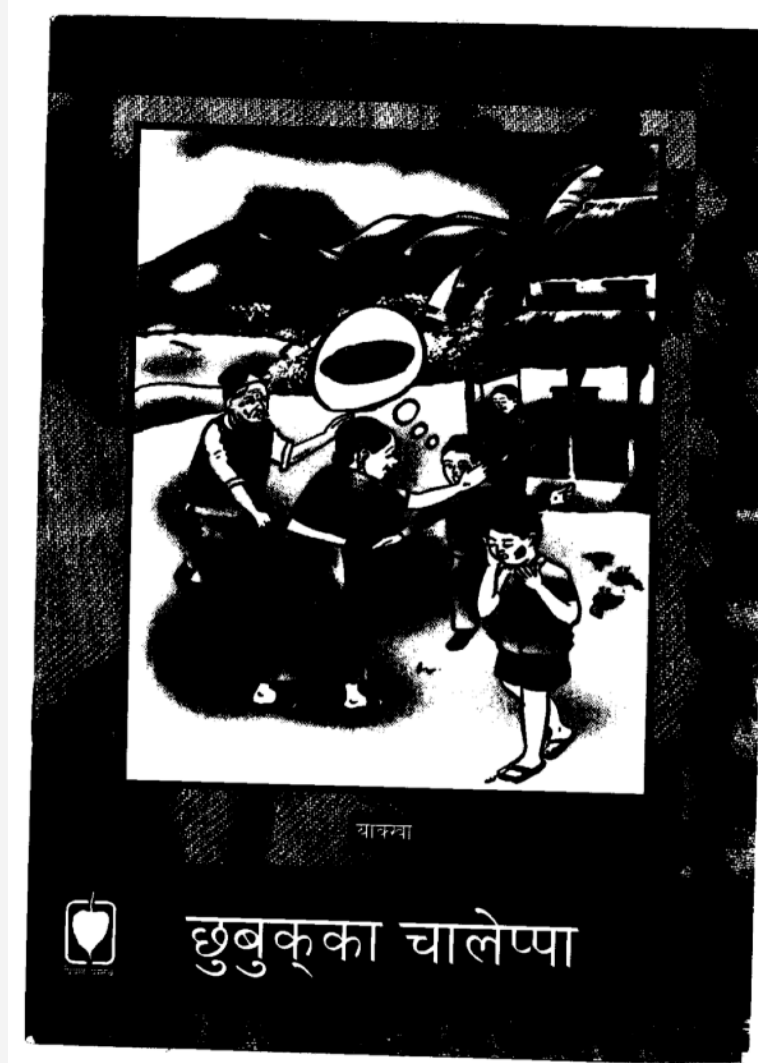
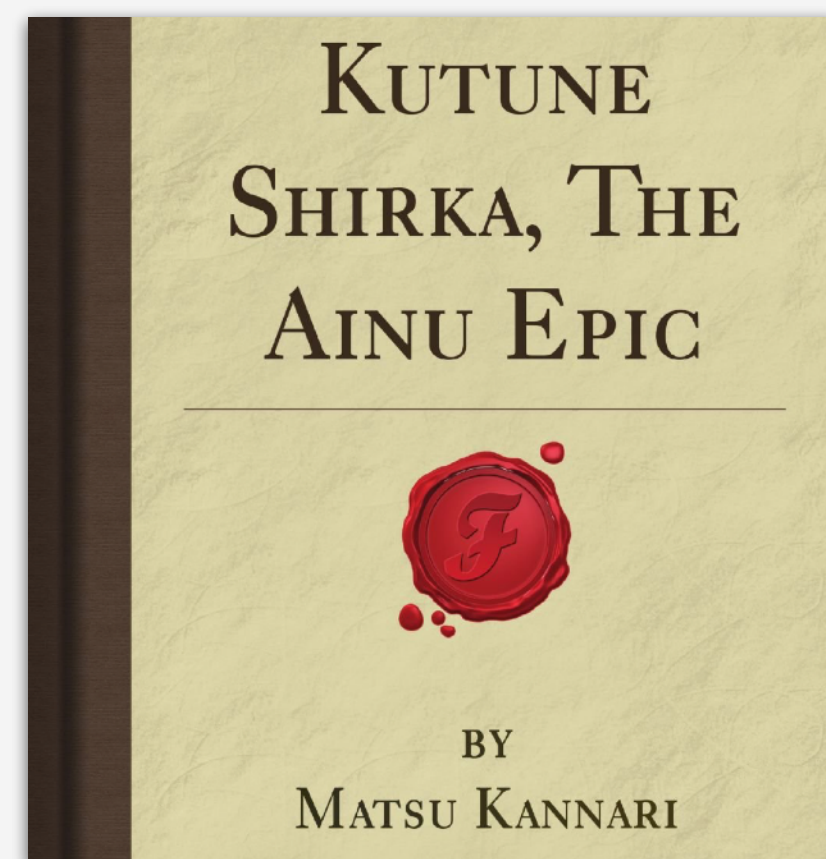
ahumada.860@gmail.com cgutierrez@dcc.uchile.cl antonis@gmu.edu

Abstract

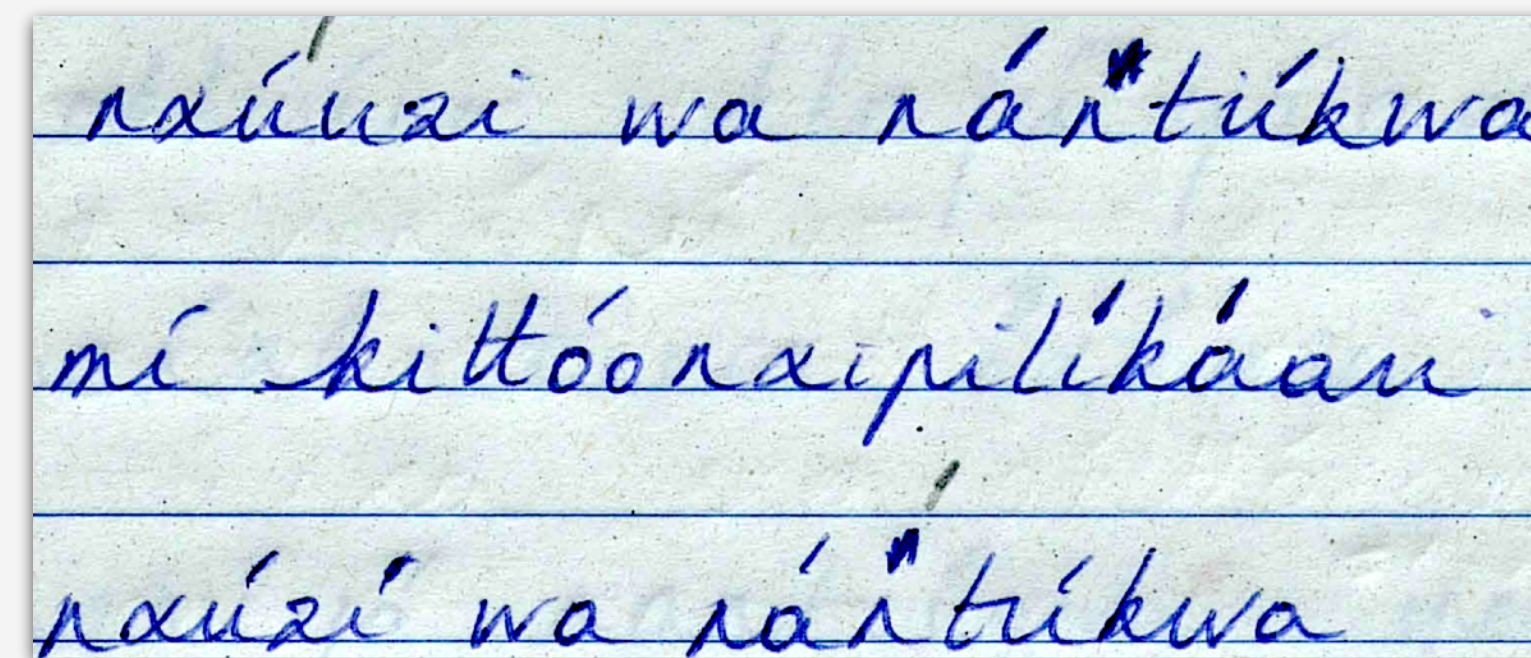
Mapuzugun is the language of the Mapuche people. Due to political and historical reasons, its number of speakers has decreased and the language has been excluded from the educational system in Chile and Argentina. For this reason, it is very important to support the revitalization of the language. In this work, we develop a tool for educational

Guaraní for Spanish speakers,² learning resources for Indigenous languages are hard to come by, let alone ones that incorporate language technologies in the educational setting in order to aid learners. In particular, it is undeniable that the development of NLP tools that reach the users lags further behind that NLP research itself (Blasi et al., 2021).

Our Solution: Make Existing Data ML-Usable

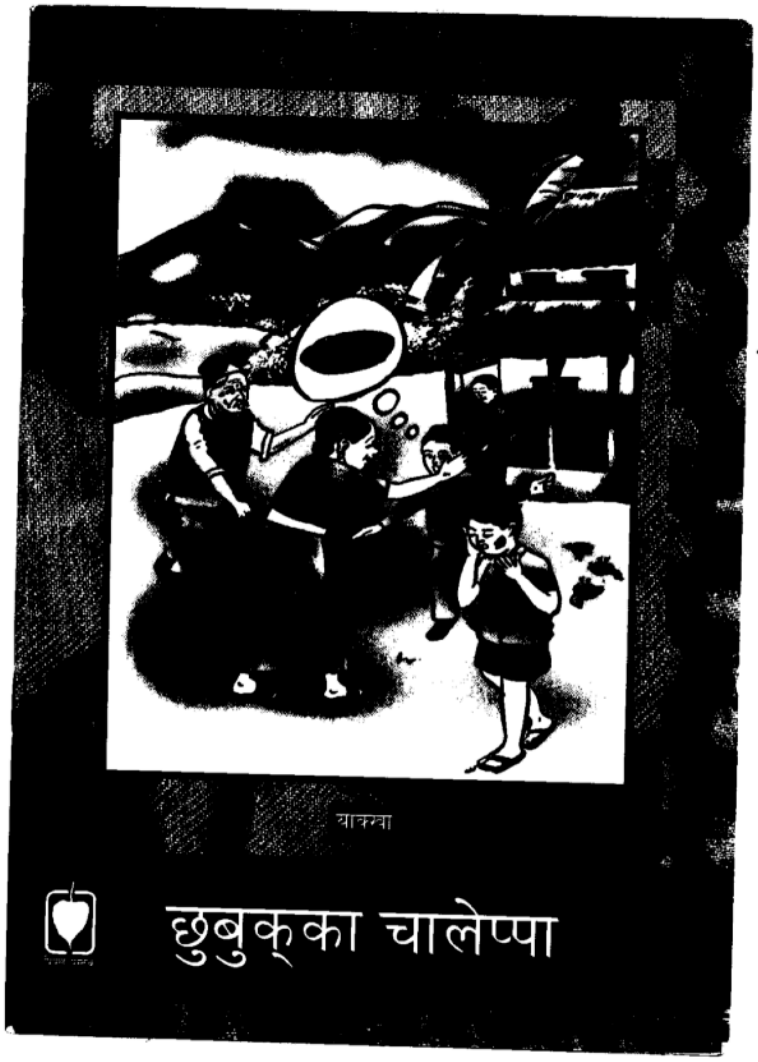
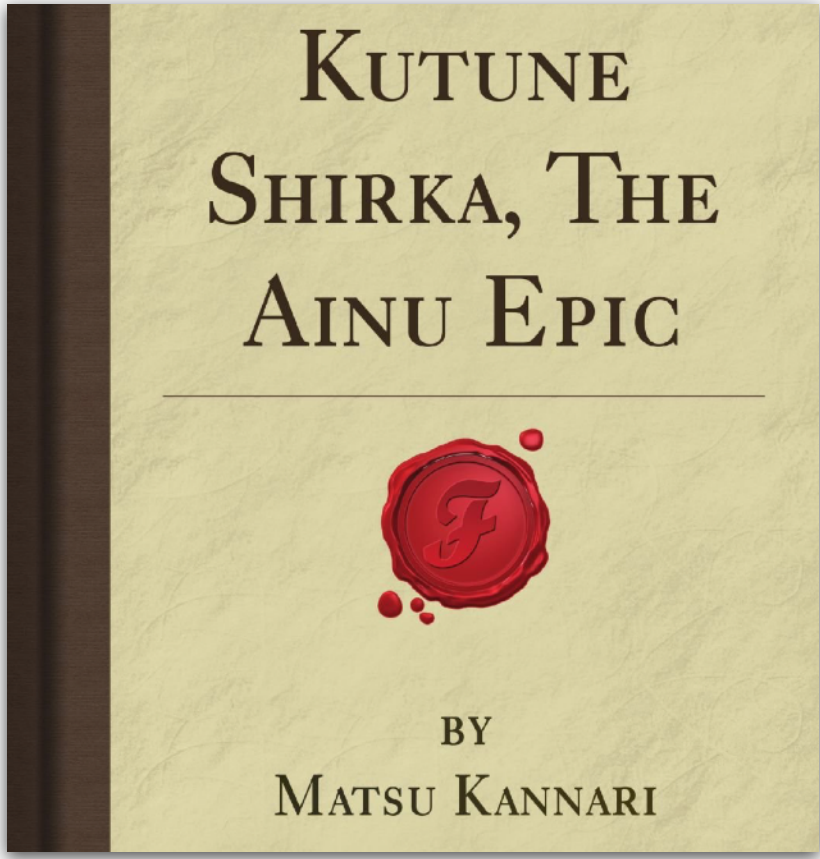


Printed books

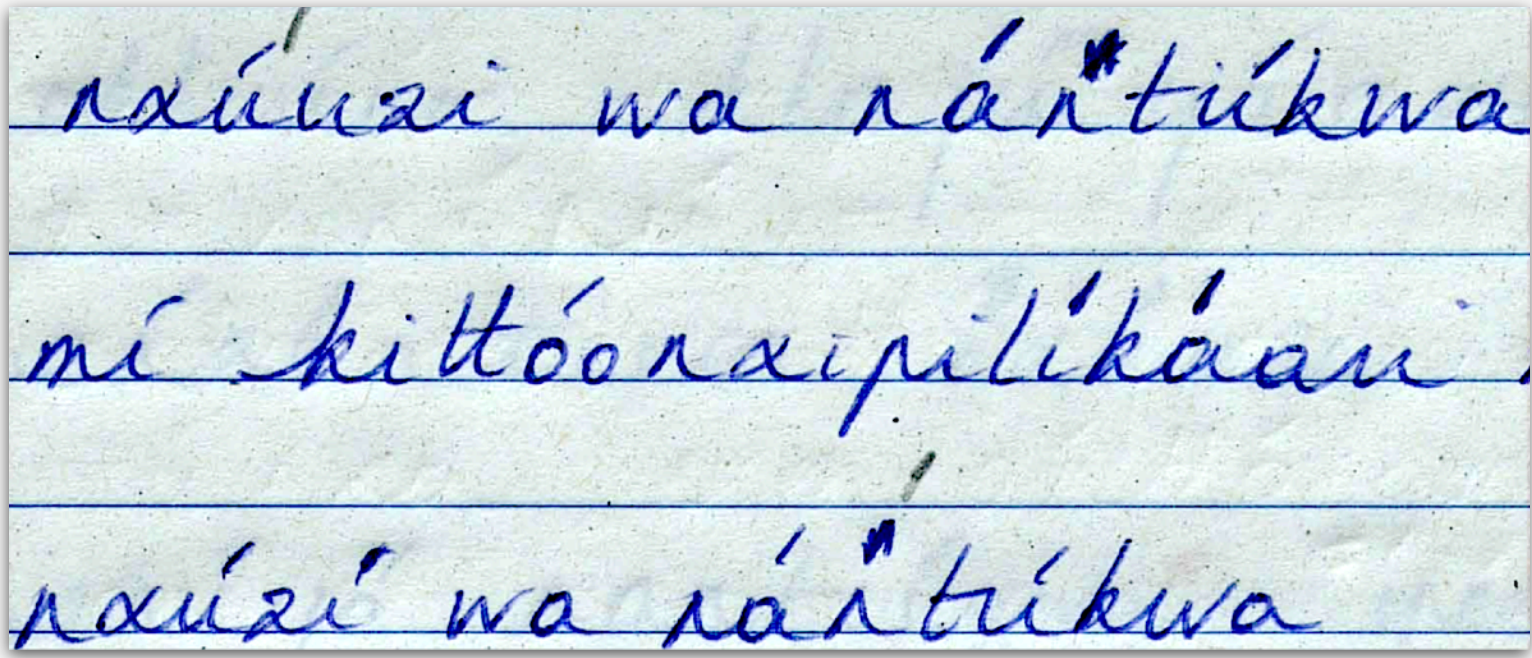


Handwritten notes

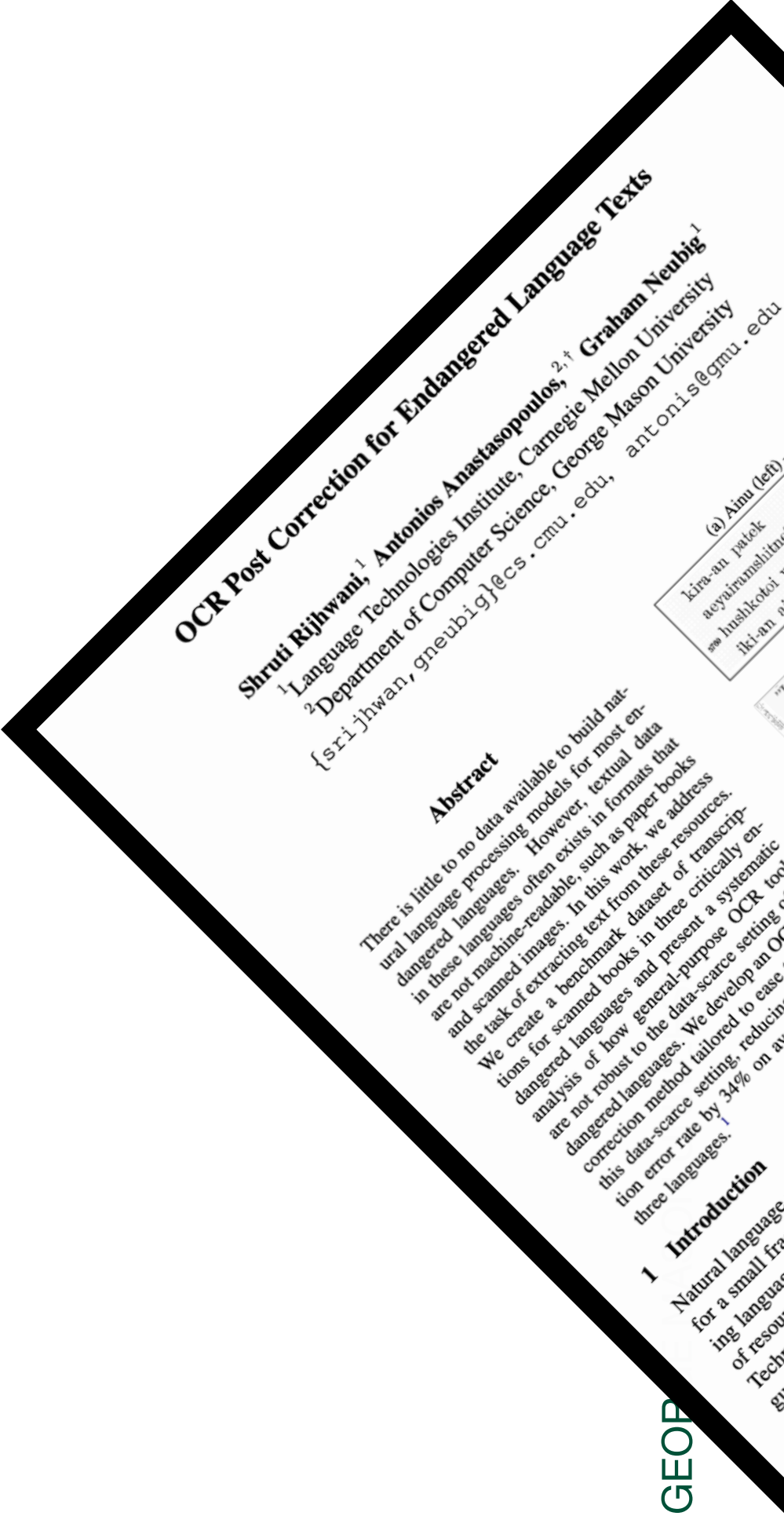
Our Solution: Make Existing Data ML-Usable



Printed books



Handwritten notes



Our Solution: Curation at Scale

Our Solution: Curation at Scale

Let's get *small, but high quality* data

Our Solution: Curation at Scale

Let's get *small, but high quality* data



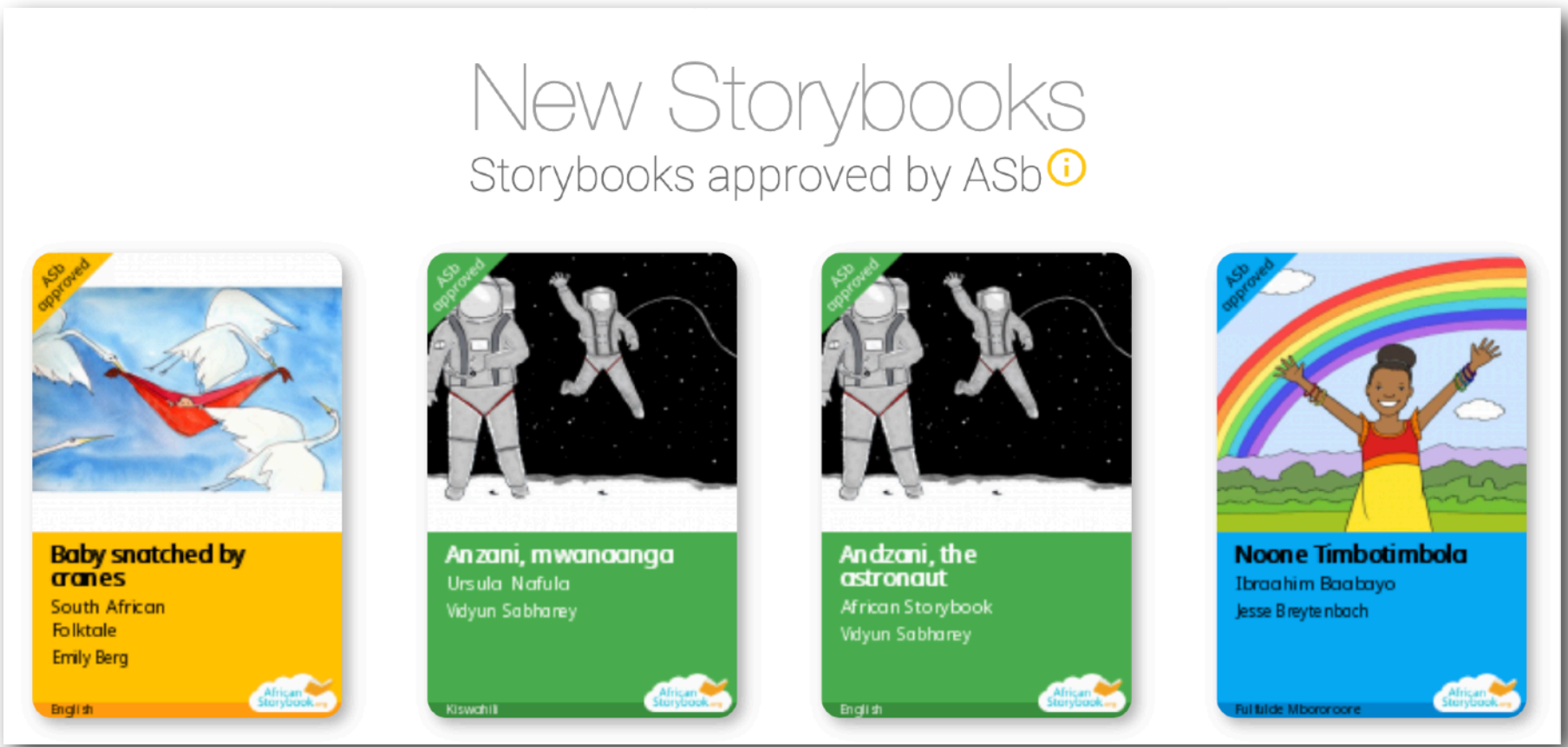
Our Solution: Curation at Scale

Let's get *small, but high quality* data

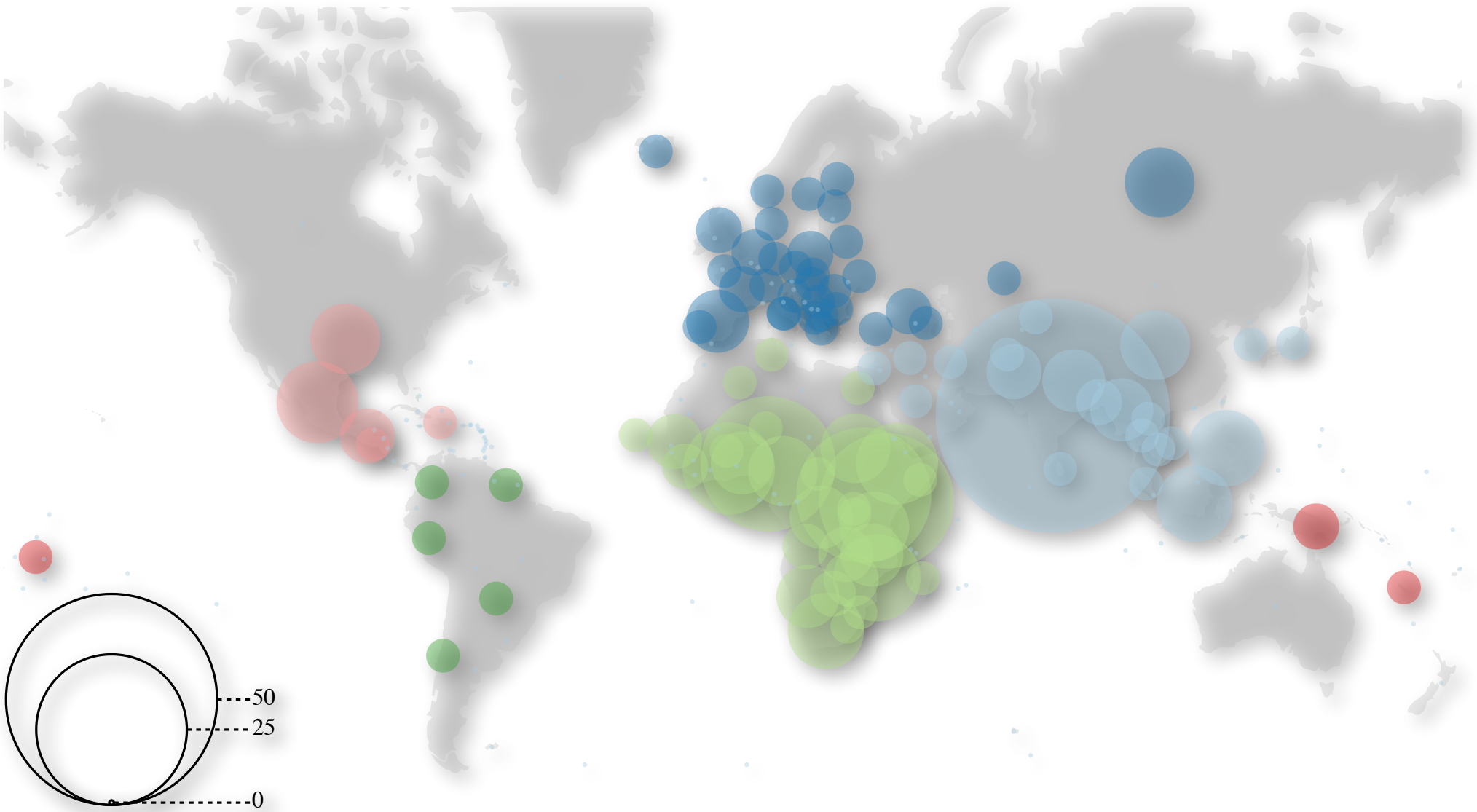


Our Solution: Curation at Scale

Let's get *small, but high quality* data



>350 languages

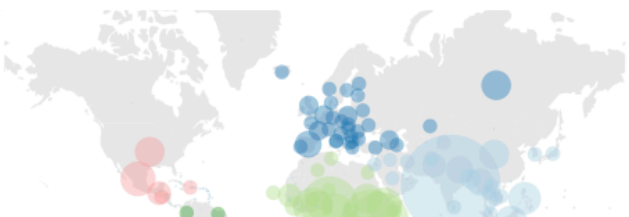


LIMIT: Language Identification, Misidentification, and Translation using Hierarchical Models in 350+ Languages

Milind Agarwal Md Mahfuz Ibn Alam Antonios Anastasopoulos
Department of Computer Science, George Mason University
{magarwa, malam21, antonis}@gmu.edu

Abstract

Knowing the language of an input text/audio is a necessary first step for using almost every NLP tool such as taggers, parsers, or translation systems. Language identification is a well-



Learning Under Privileged Information

Training

Inference

Convergence Rate

Supervised learning:

Learning Under Privileged Information

Training

Inference

Convergence Rate

Supervised learning:

$$x \xrightarrow{f} y$$

$$y = f(x)$$

Learning Under Privileged Information

Training

Inference

Convergence Rate

Supervised learning:

$$x \xrightarrow{f} y$$

$$y = f(x)$$

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Learning Under Privileged Information

Training

Inference

Convergence Rate

Supervised learning:

$$x \xrightarrow{f} y$$

$$y = f(x)$$

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Learning under
Privileged Information:

Learning Under Privileged Information

Training

Inference

Convergence Rate

Supervised learning:

$$x \xrightarrow{f} y$$

$$y = f(x)$$

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Learning under
Privileged Information:

$$(x, x^*) \xrightarrow{f} y$$

Learning Under Privileged Information

Training

Inference

Convergence Rate

Supervised learning:

$$x \xrightarrow{f} y$$

$$y = f(x)$$

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Learning under
Privileged Information:

$$(x, x^*) \xrightarrow{f} y$$

$$y = f(x)$$

Learning Under Privileged Information

Training

Inference

Convergence Rate

Supervised learning:

$$x \xrightarrow{f} y$$

$$y = f(x)$$

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Learning under
Privileged Information:

$$(x, x^*) \xrightarrow{f} y$$

$$y = f(x)$$

$$\mathcal{O}\left(\frac{1}{n}\right)$$

Learning Under Privileged Information

Training

Inference

Convergence Rate

Supervised learning:

$$x \xrightarrow{f} y$$

$$y = f(x)$$

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Learning under
Privileged Information:

$$(x, x^*) \xrightarrow{f} y$$

$$y = f(x)$$

$$\mathcal{O}\left(\frac{1}{n}\right)$$

What is good privileged information for language tasks?

- 1) Context
- 2) Meta-language

Meta-data Augmented LLMs

news article

news article

TITLE: Spread The Love With NY Country 94.7 And Kelly Ford's 'Valentines For The Troops'

CONTENT: The team at 94.7, led by Kelly Ford, are asking people across the Tri-State area to send Valentine's Day cards to troops. "It's really all about the message because you're away from home on a holiday that's all about love and you need to know that people back here love ya," Ford tells Wayne Cabot and Paul Murnane. "It's about saying, 'Hey, we love and appreciate you.'"

Metadata augmented news article

URL: <https://wcbs880.radio.com/articles/news/ny-country-947s-kelly-ford-seeks-valentines-for-troops>

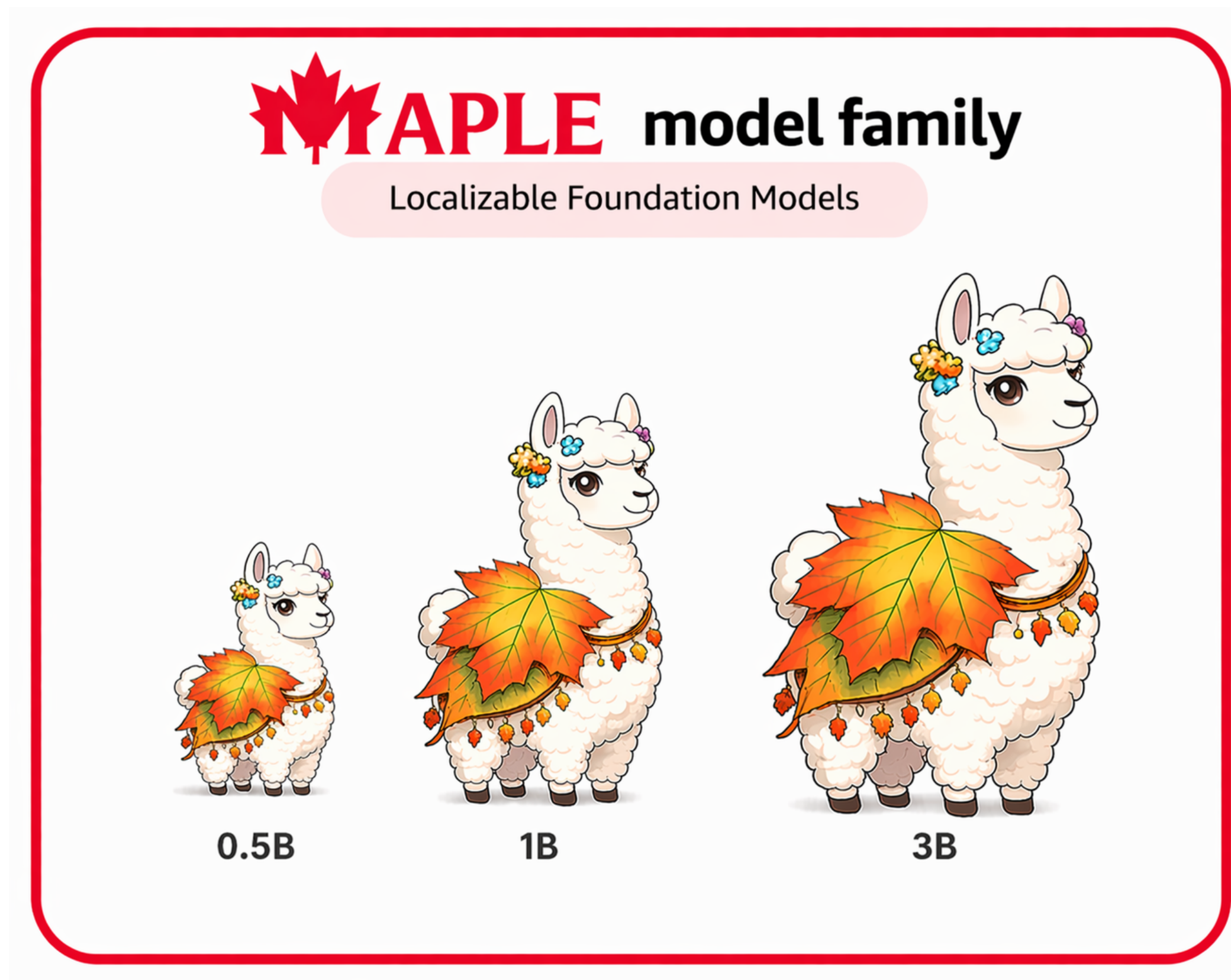
COUNTRY: us

CONTINENT: America

TITLE: Spread The Love With NY Country 94.7 And Kelly Ford's 'Valentines For The Troops'

CONTENT: The team at 94.7, led by Kelly Ford, are asking people across the Tri-State area to send Valentine's Day cards to troops. "It's really all about the message because you're away from home on a holiday that's all about love and you need to know that people back here love ya," Ford tells Wayne Cabot and Paul Murnane. "It's about saying, 'Hey, we love and appreciate you.'"

The MAPLE model family



**MAPLE: Metadata Aware Pre-training
for Localized Evaluation**

The MAPLE model family

 **MAPLE** model family
Localizable Foundation Models



Standard decoder transformer LMs

**MAPLE: Metadata Aware Pre-training
for Localized Evaluation**

The MAPLE model family

 **MAPLE** model family

Localizable Foundation Models



0.5B



1B



3B

Standard decoder transformer LMs

Model sizes: .5B, 1B, 3B

MAPLE: Metadata **A**ware **P**re-training
for **L**ocalized **E**valuation

The MAPLE model family

 **MAPLE** model family

Localizable Foundation Models



Standard decoder transformer LMs

Model sizes: .5B, 1B, 3B

Pretrained on the NOW corpus:

Scraped English news from across the world:
US, Canada, UK, Nigeria, Australia, India, etc

**MAPLE: Metadata Aware Pre-training
for Localized Evaluation**

The MAPLE model family

 **MAPLE** model family

Localizable Foundation Models



Standard decoder transformer LMs

Model sizes: .5B, 1B, 3B

Pretrained on the NOW corpus:

Scraped English news from across the world:
US, Canada, UK, Nigeria, Australia, India, etc

Metadata: URL, Country, Continent

**MAPLE: Metadata Aware Pre-training
for Localized Evaluation**

The MAPLE model family

 **MAPLE** model family

Localizable Foundation Models



**MAPLE: Metadata Aware Pre-training
for Localized Evaluation**

Standard decoder transformer LMs

Model sizes: .5B, 1B, 3B

Pretrained on the NOW corpus:

Scraped English news from across the world:
US, Canada, UK, Nigeria, Australia, India, etc

Metadata: URL, Country, Continent

Ablations: Global model, Continent-localized model



Significantly improved **token efficiency**

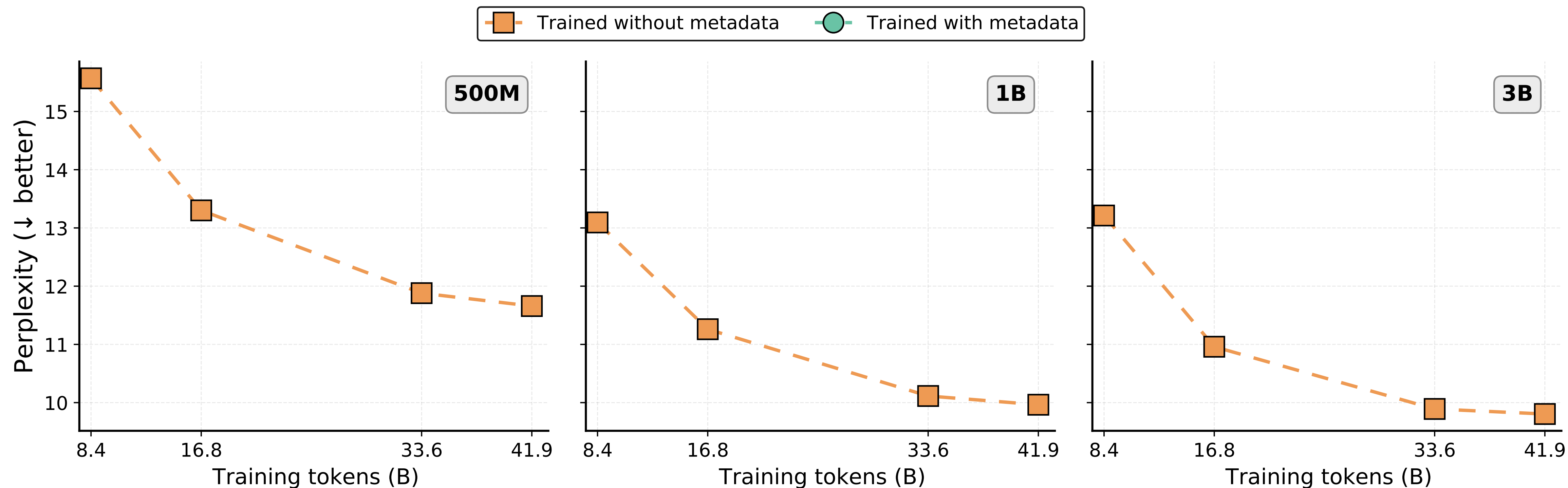


Significantly improved **token efficiency**

Without metadata

$$p(\text{text}) = \sum_{\ell} p(\ell) p(\text{text} \mid \ell)$$

mixture across locales





Significantly improved **token efficiency**

Without metadata

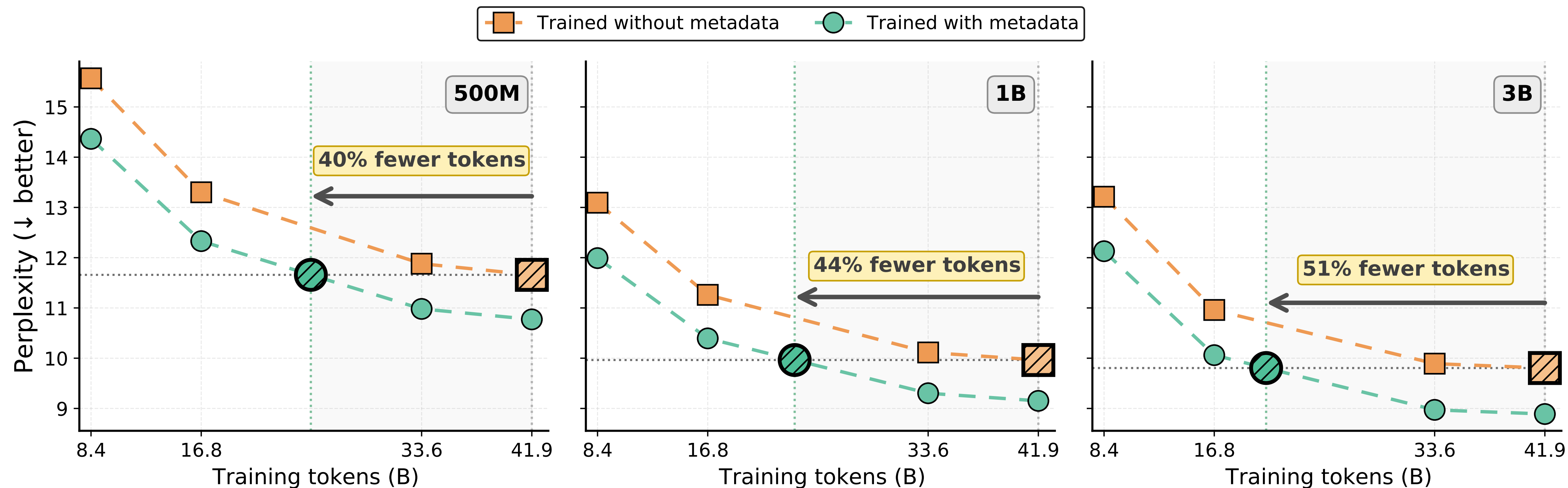
$$p(\text{text}) = \sum_{\ell} p(\ell) p(\text{text} \mid \ell)$$

mixture across locales

With metadata

$$p(\text{text} \mid \textcolor{teal}{m}) \approx p(\text{text} \mid \ell)$$

condition on locale signal





Local Models cannot generalize to unseen regions



Local Models cannot generalize to unseen regions

		Local(T+,I+)			
Train Region	Africa	9.1			
	America		8.3		
	Asia			7.5	
	Europe				8.5
		Africa	America	Asia	Europe

Showing perplexity on corresponding test sets



Local Models cannot generalize to unseen regions

		Local(T+,I+)			
Train Region	Africa	9.1	22.2	20.2	22.5
	America		8.3		
	Asia			7.5	
	Europe				8.5
		Africa	America	Asia	Europe

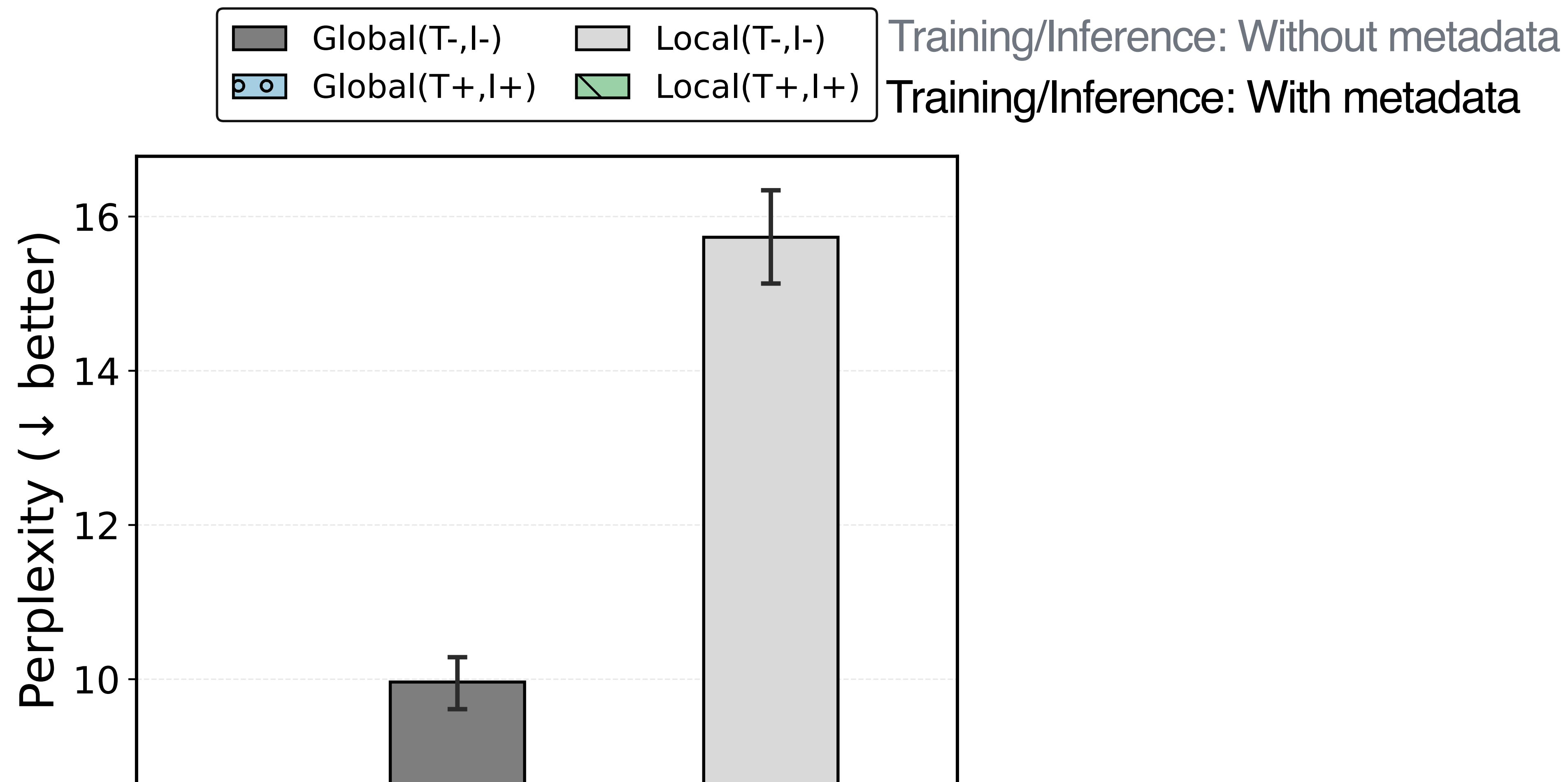
Showing perplexity on corresponding test sets



We can localize our global models

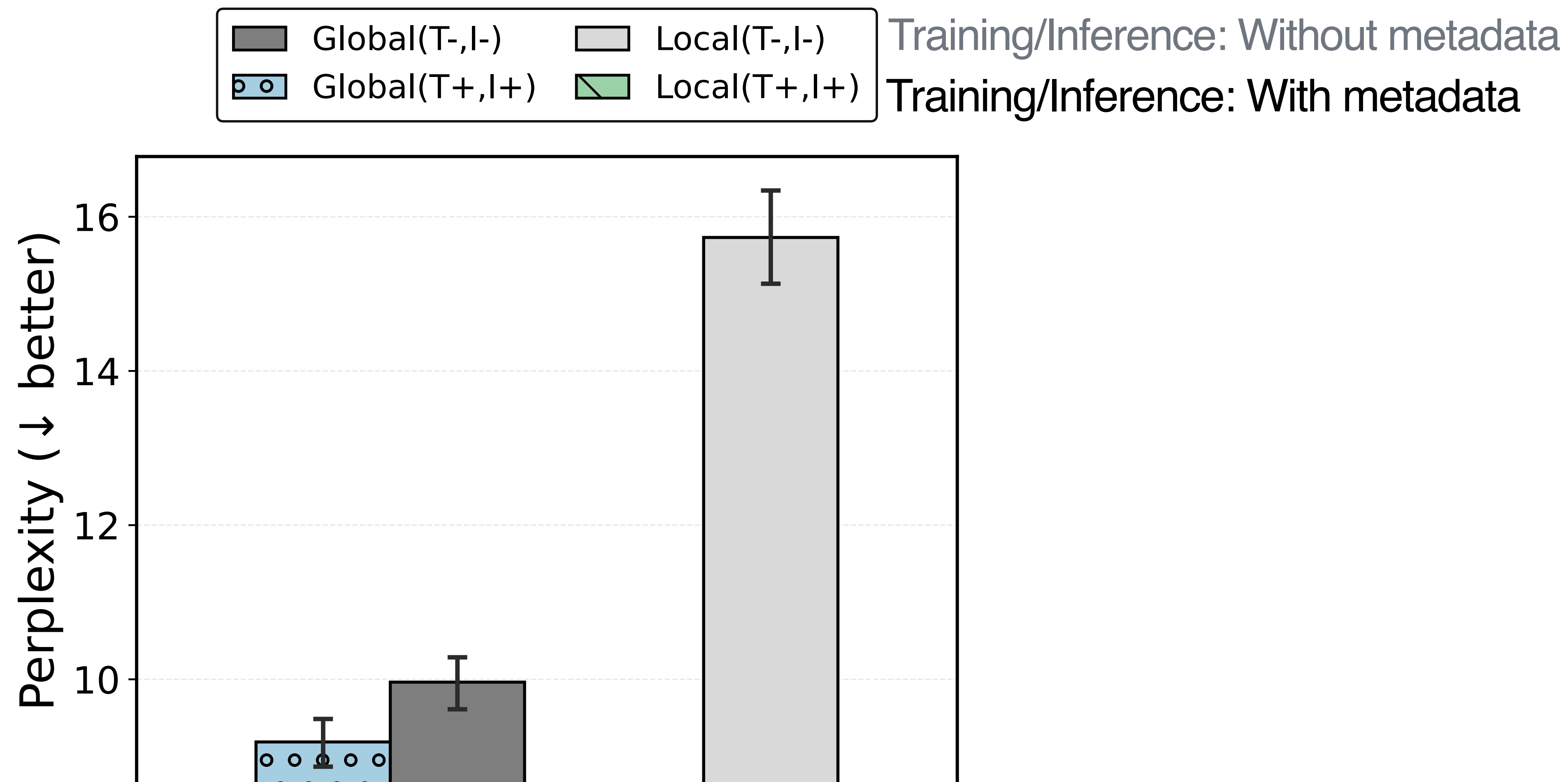


We can localize our global models



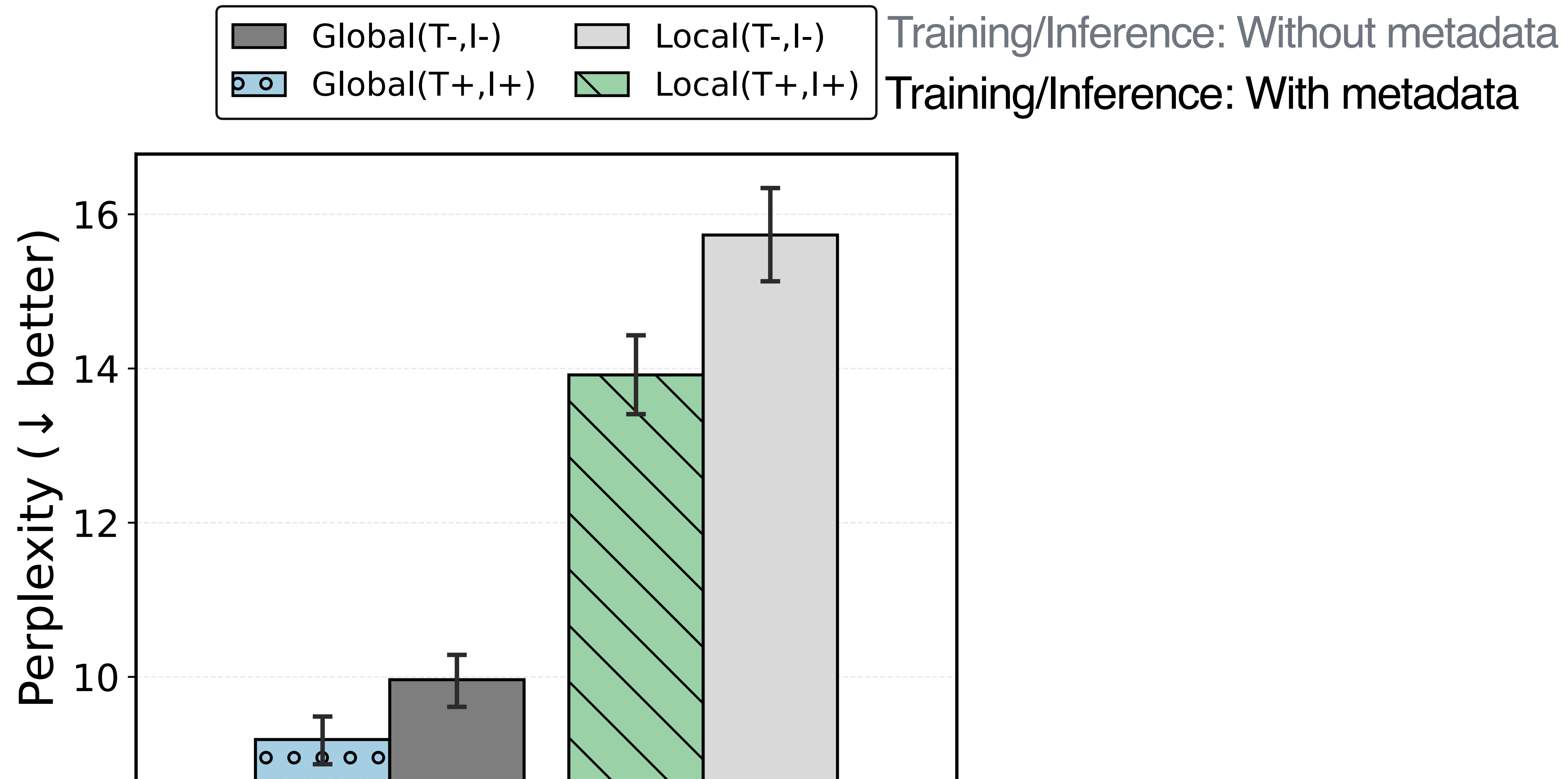


We can localize our global models



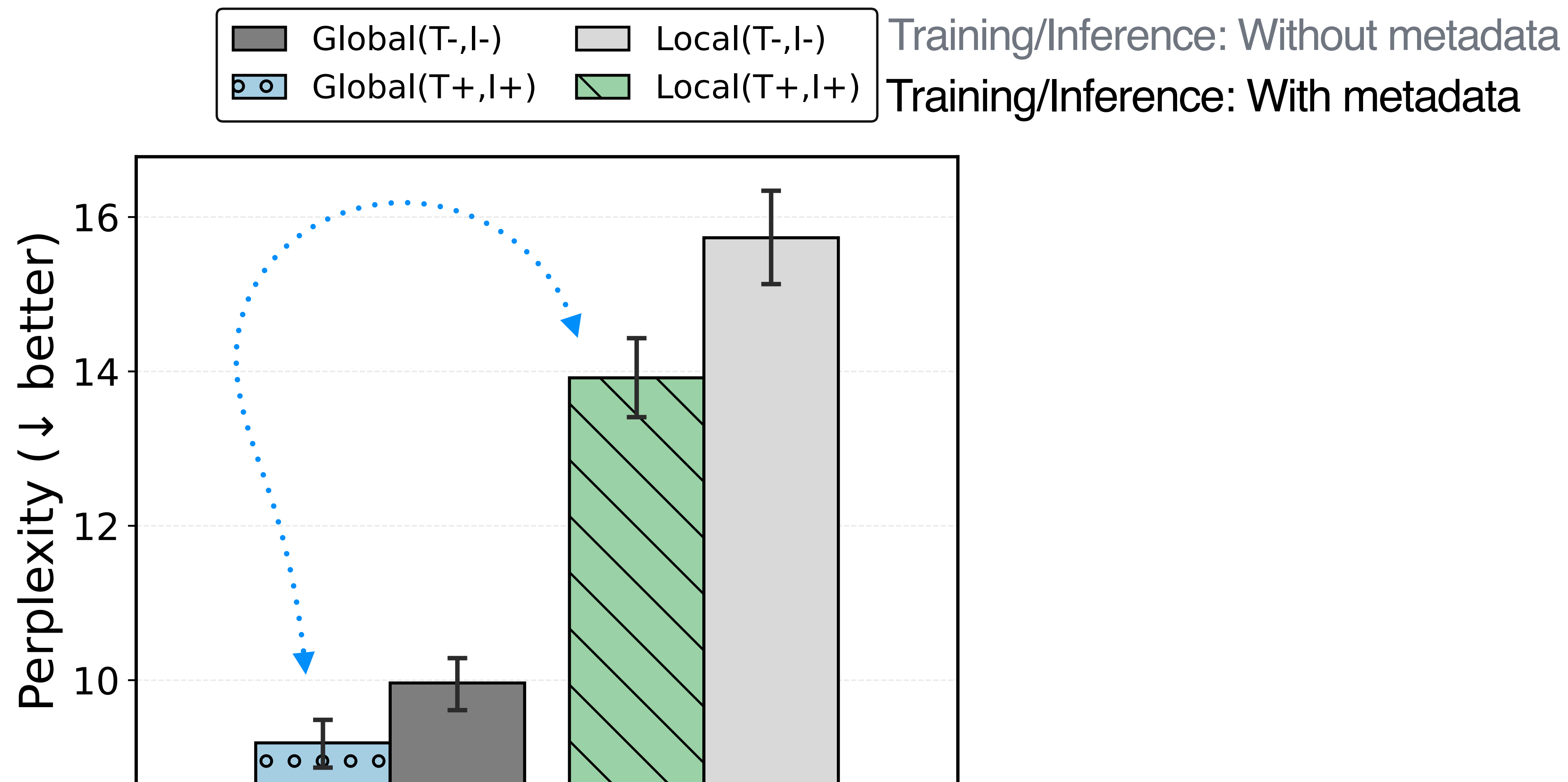


We can localize our global models





We can localize our global models



Models trained on all continents together
can perform as well as **local models**.



Example advantage: Underspecified user queries

35k+ questions

Example advantage: Underspecified user queries

35k+ questions



MAPLE

LocalNewsQA Explicit

Who was the Prime Minister in the UK in 2015?

(A) Gordon Brown 

(B) David Cameron 

(C) Narendra Modi 

(D) Angela Merkel 

✓ **(B) David Cameron**



Example advantage: Underspecified user queries

35k+ questions



MAPLE

LocalNewsQA Explicit

Who was the Prime Minister in the UK in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 

✓ (B) David Cameron



LocalNewsQA Ambiguous

Who was the Prime Minister in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 



LocalNewsQA Ambiguous

Who was the Prime Minister in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 



Example advantage: Underspecified user queries

35k+ questions



MAPLE

LocalNewsQA Explicit

Who was the Prime Minister in the UK in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 

✓ (B) David Cameron



LocalNewsQA Ambiguous

URL: news.co/uk

Who was the Prime Minister in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 

✓ (B) David Cameron



LocalNewsQA Ambiguous

Who was the Prime Minister in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 



Example advantage: Underspecified user queries

35k+ questions



MAPLE

LocalNewsQA Explicit

Who was the Prime Minister in the UK in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 

✓ (B) David Cameron



LocalNewsQA Ambiguous

URL: news.co/uk

Who was the Prime Minister in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 

✓ (B) David Cameron



LocalNewsQA Ambiguous

URL: news.co/in

Who was the Prime Minister in 2015?

- (A) Gordon Brown 
- (B) David Cameron 
- (C) Narendra Modi 
- (D) Angela Merkel 

✓ (C) Narendra Modi

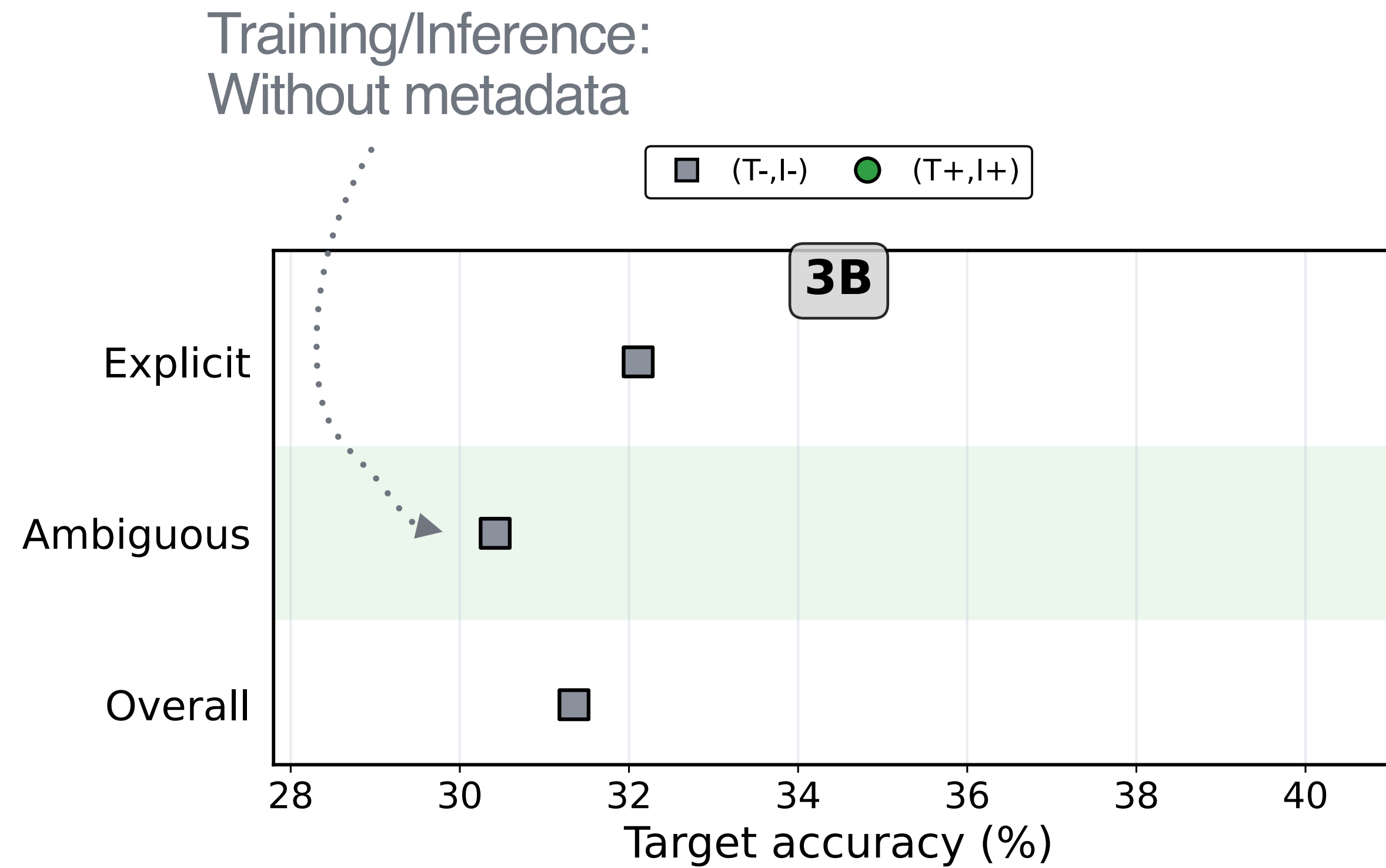




Disambiguation of under-specified user queries

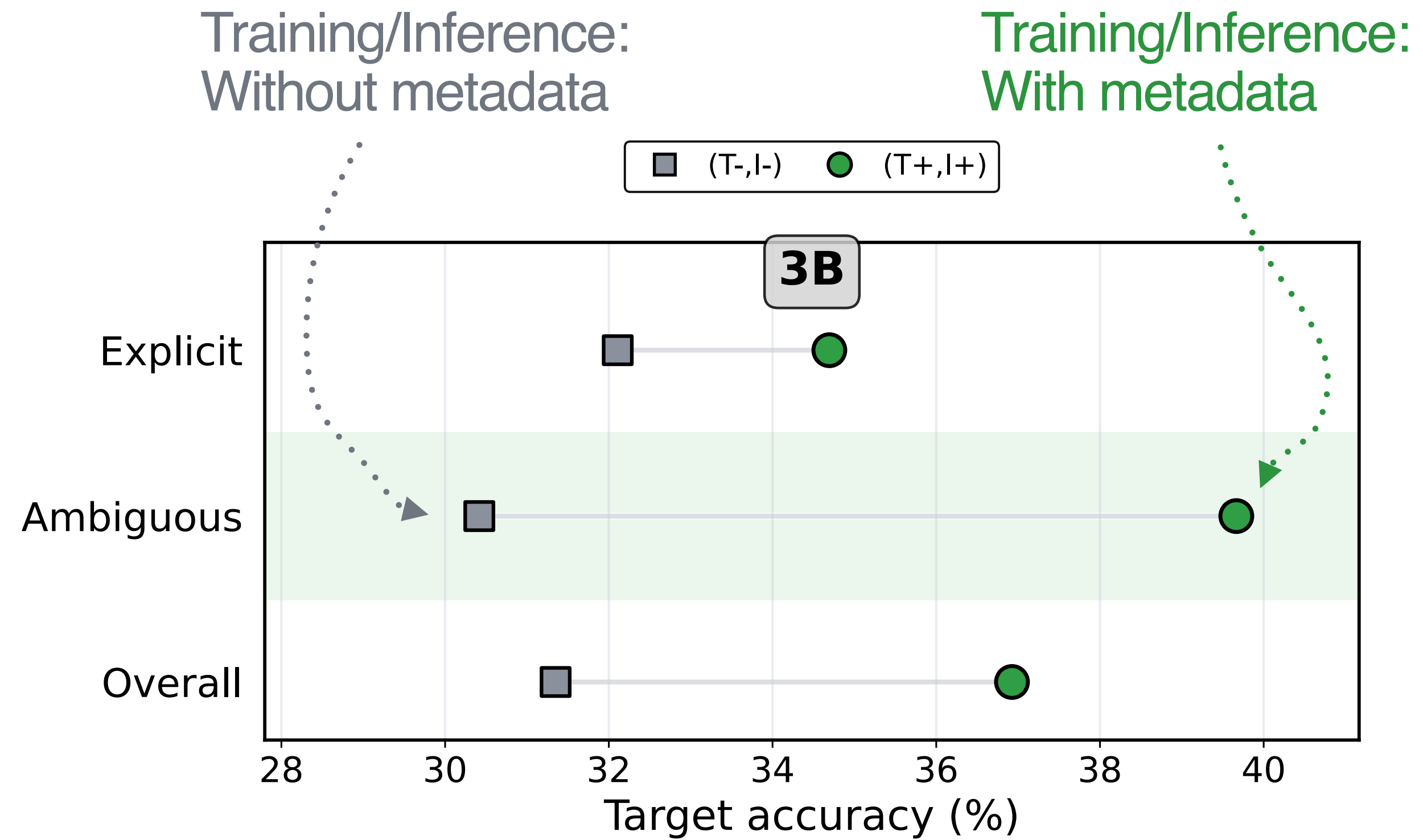


Disambiguation of under-specified user queries



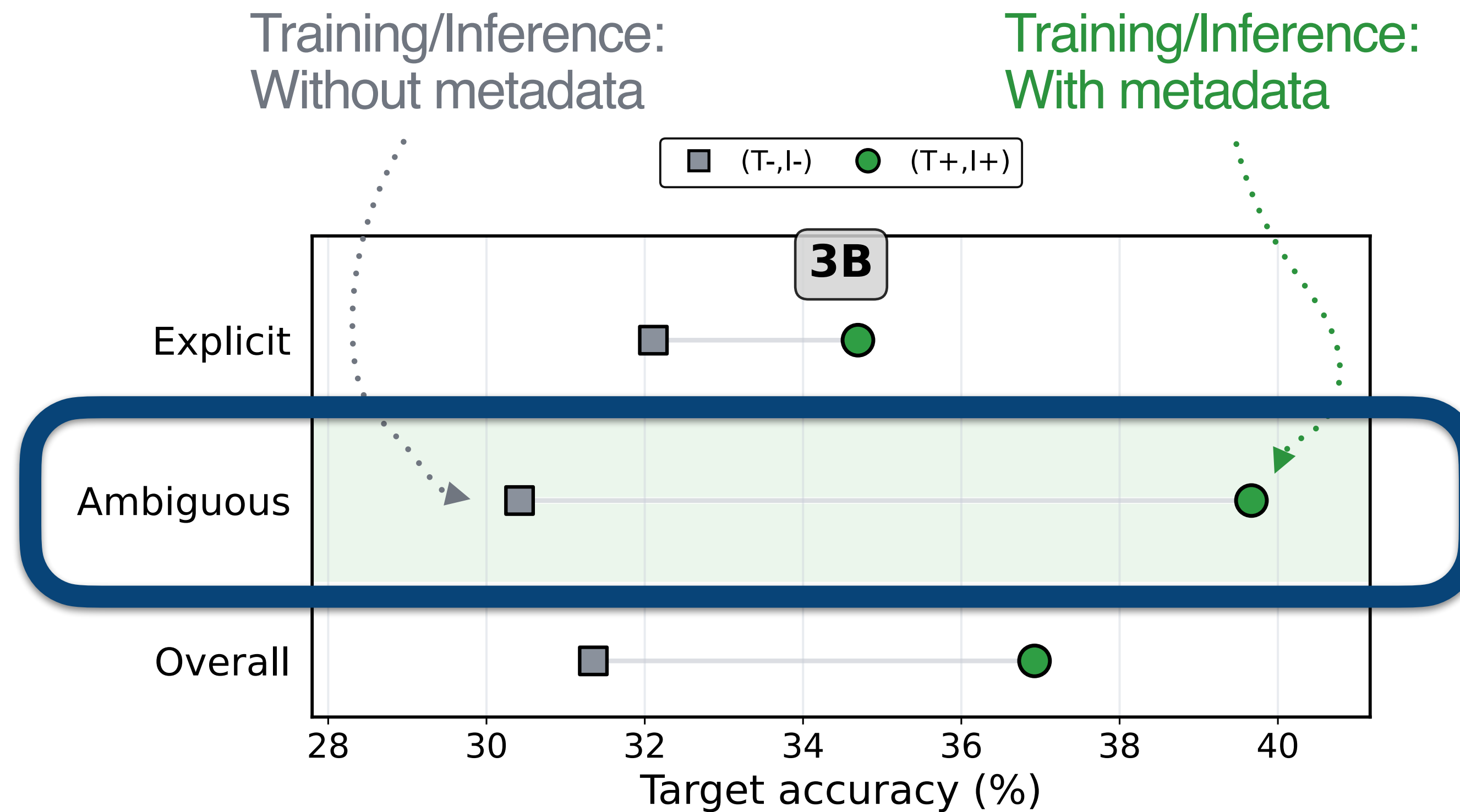


Disambiguation of under-specified user queries





Disambiguation of under-specified user queries



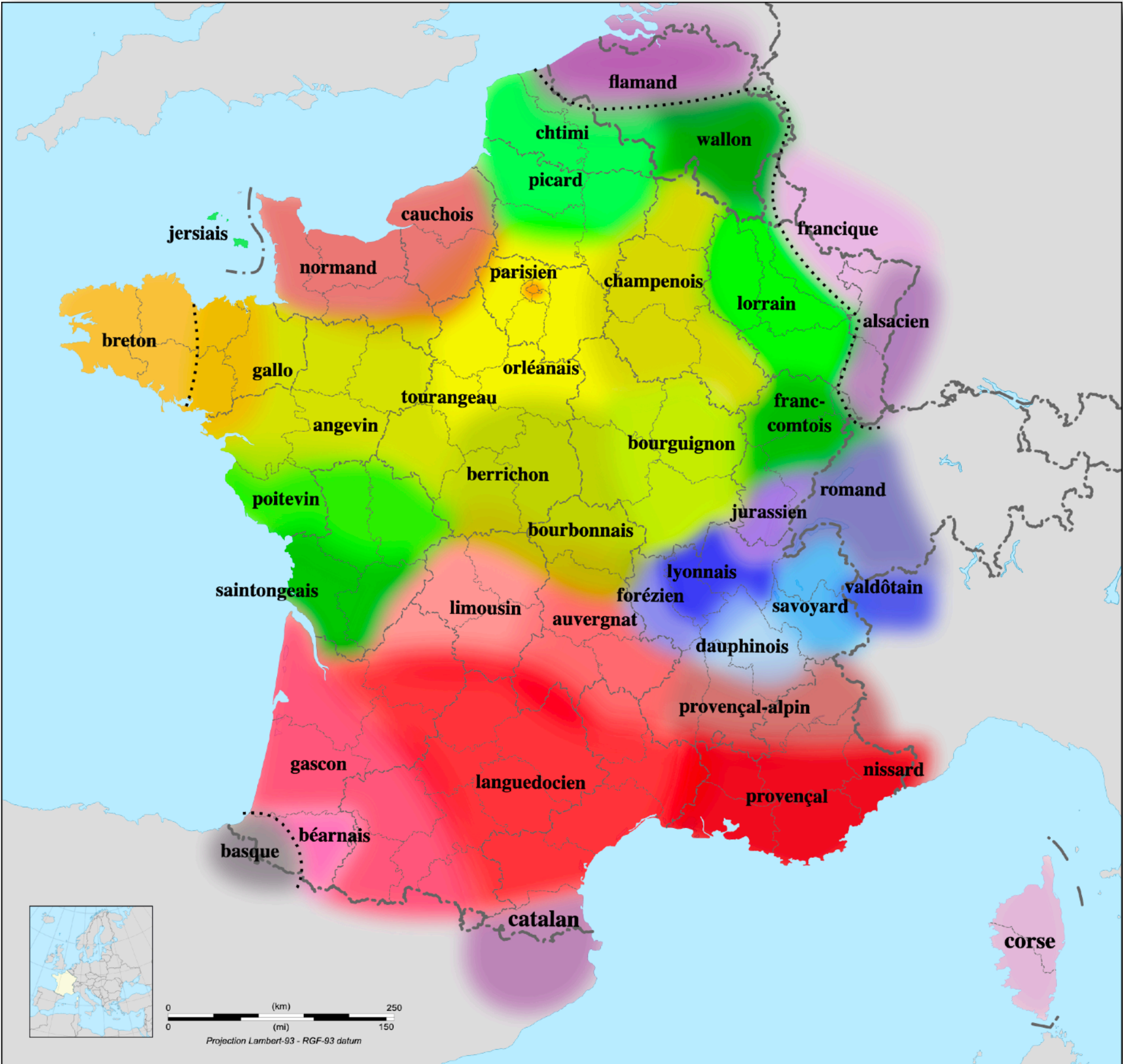
Dialects

Languages are not Monoliths

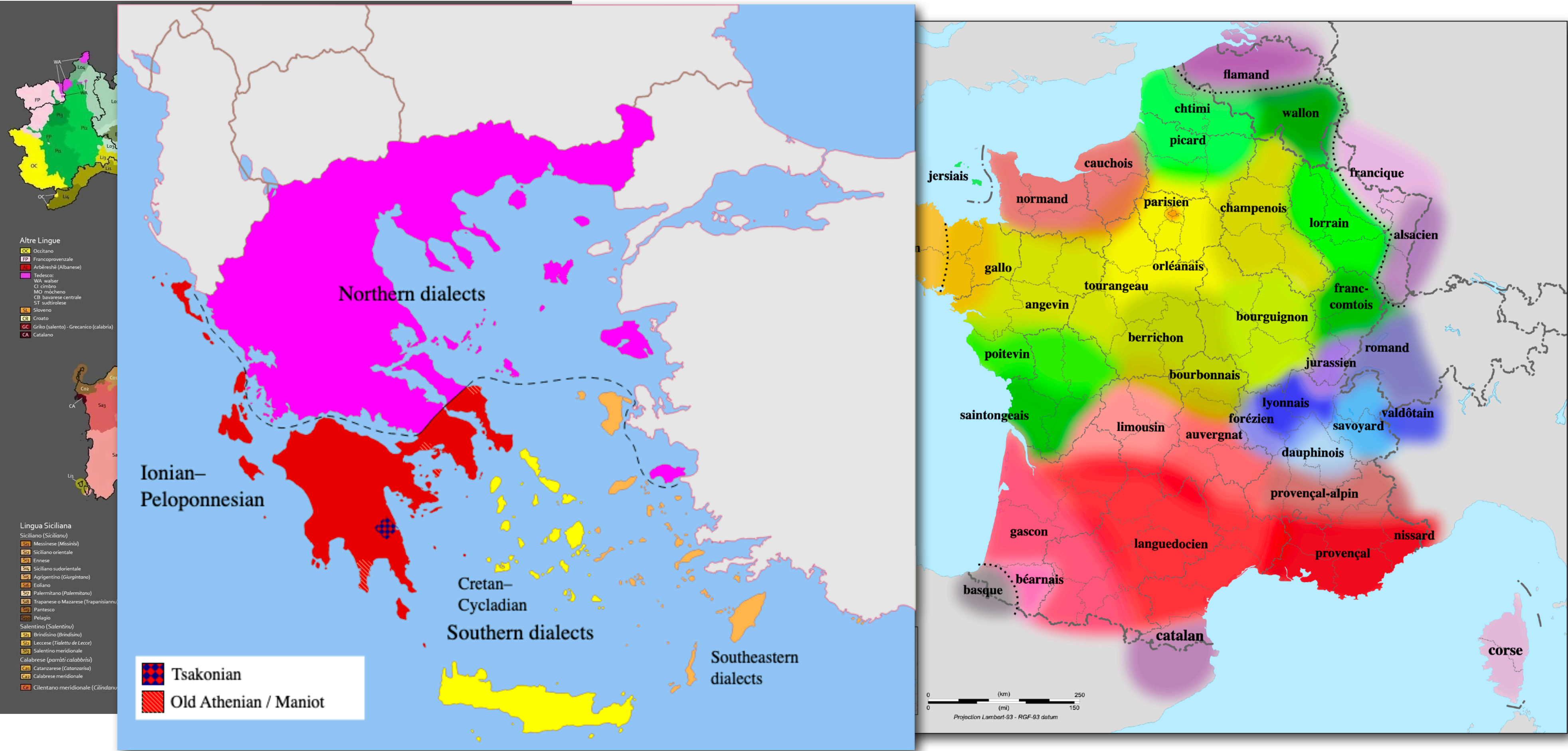
Languages are not Monoliths



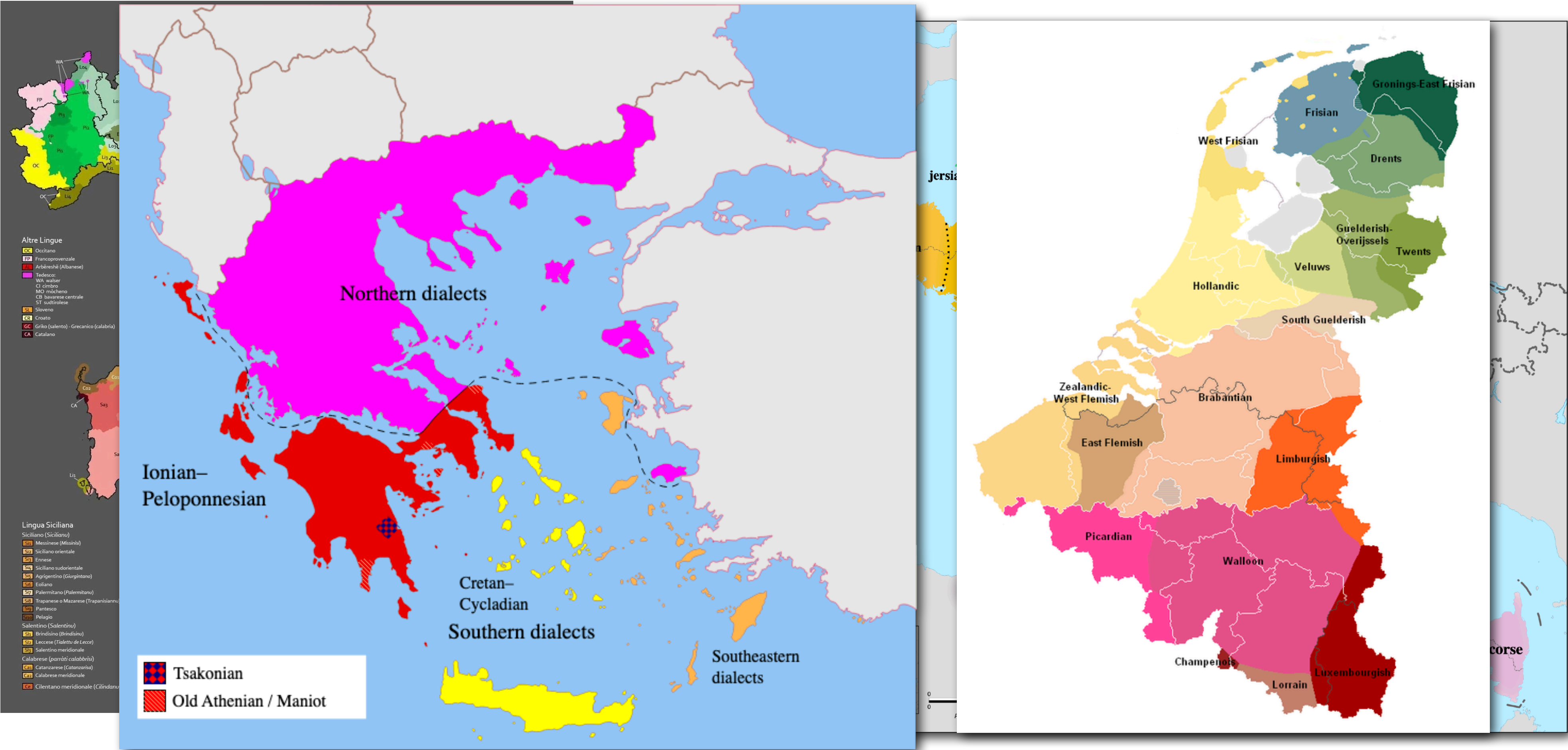
Languages are not Monoliths



Languages are not Monoliths



Languages are not Monoliths

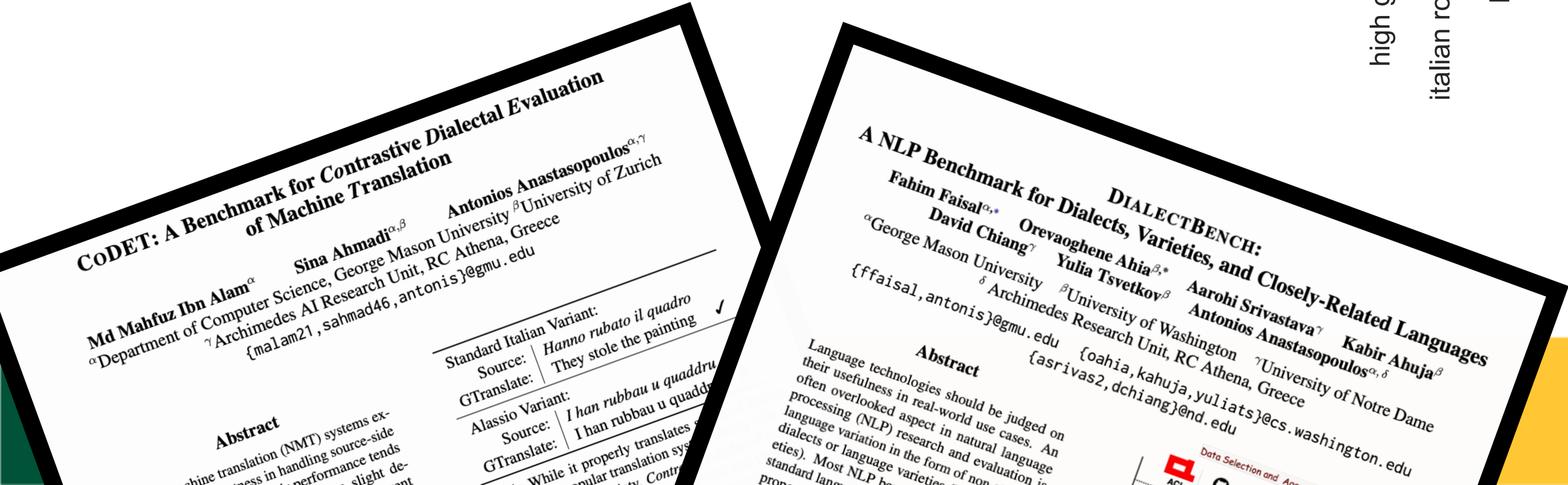


DialectBench

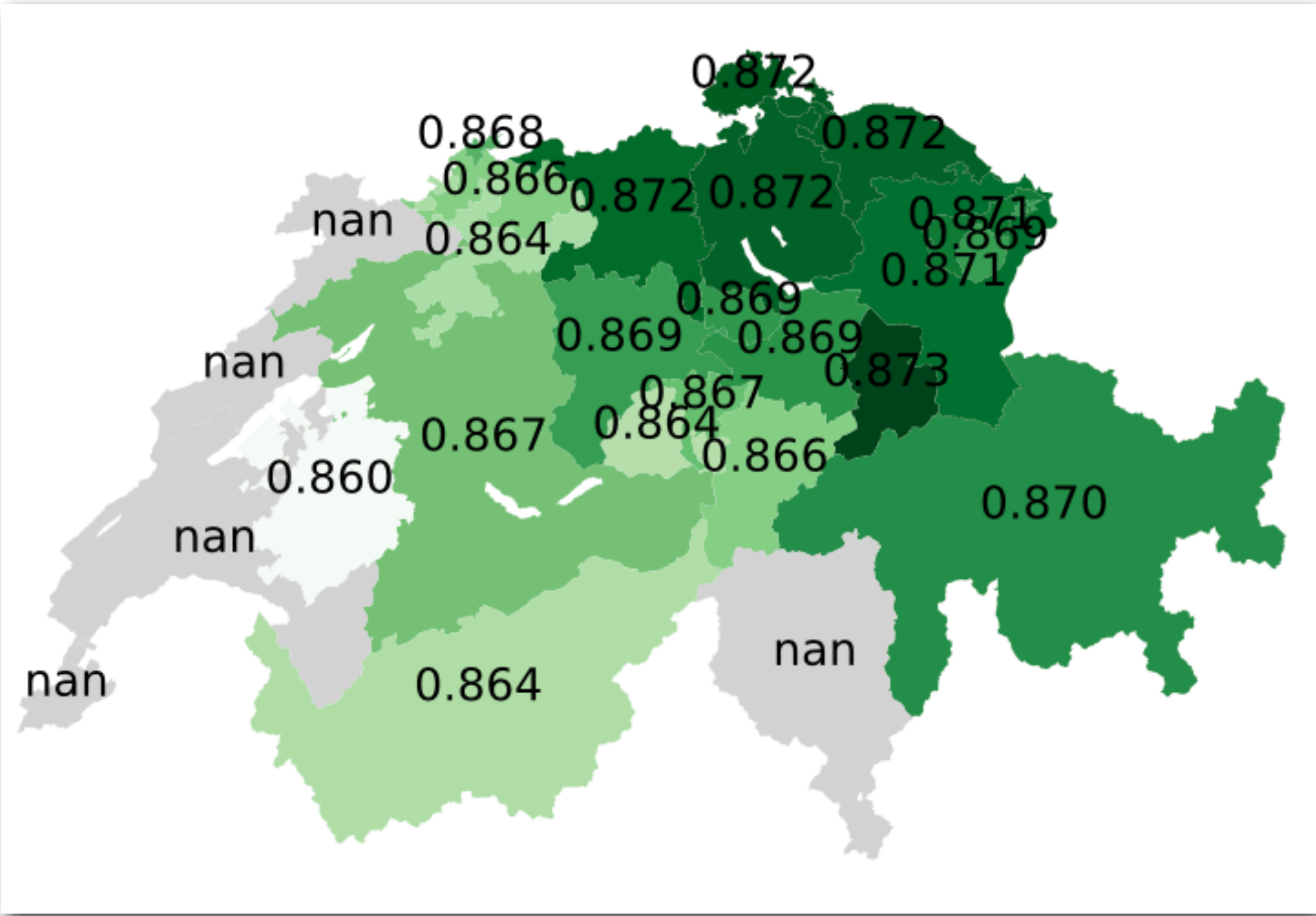
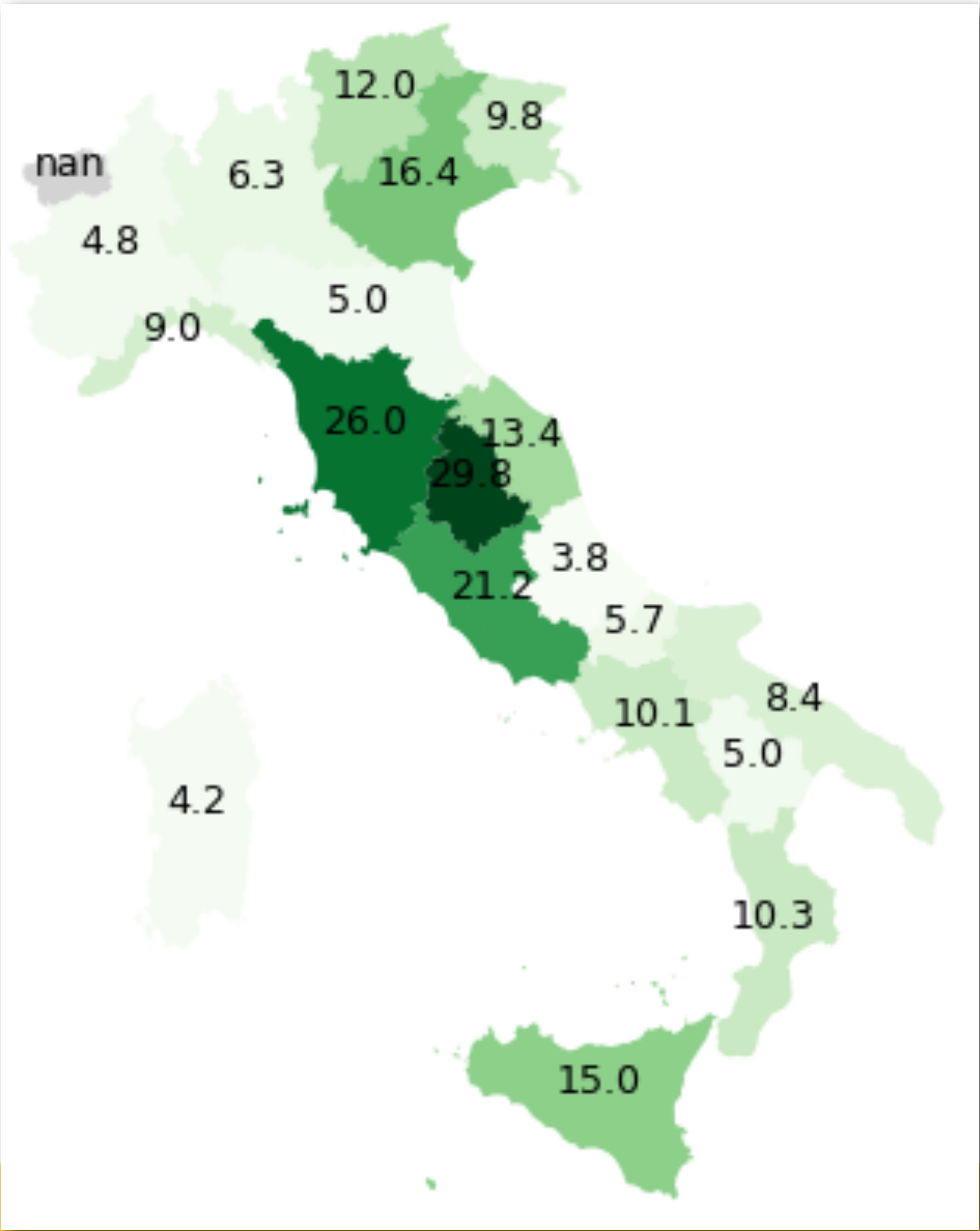
First large-scale benchmark
10 tasks, 40 continua, 281 varieties

Task	DEP.	40	3	2	4		3	3		4		3	3			1		3		3					8
	POS.	51	6	2	4		2	3		5		3	3	8		1		3		3					8
	NER	85	2	8	4		4	6	4	5	2	6	3			5	2	2	4		3	3	3		19
	EQA	24	7				11								2						3		2	2	
	MRC	11	6				1	2																2	
	NLI	38	9	2	2		1	3	3	4		1	2			3	2				1				5
	TC	38	9	2	2		1	3	3	4		1	2			3	2				1				5
	SA	9	9																						
	DId	49	26	4			3	4		6	6														
	MT	114	25	23	20	21			8	1	2		3		5		2								4
	Total	281	42	31	26	21	19	13	12	11	11	8	8	8	6	5	5	4	4	3	3	3	3	3	3
	Total	arabic	high german	italian romance	basque	anglic	sinitic	common turkic	sw shift. romance	greek	gallo-rhaetian	norwegian	neva	bengali	gallo-italian	kurdish	komi	serb.-croa.-bosnian	tupi-guarani.	modern dutch	eastern romance	frisian	swahili	Other	

Language Clusters



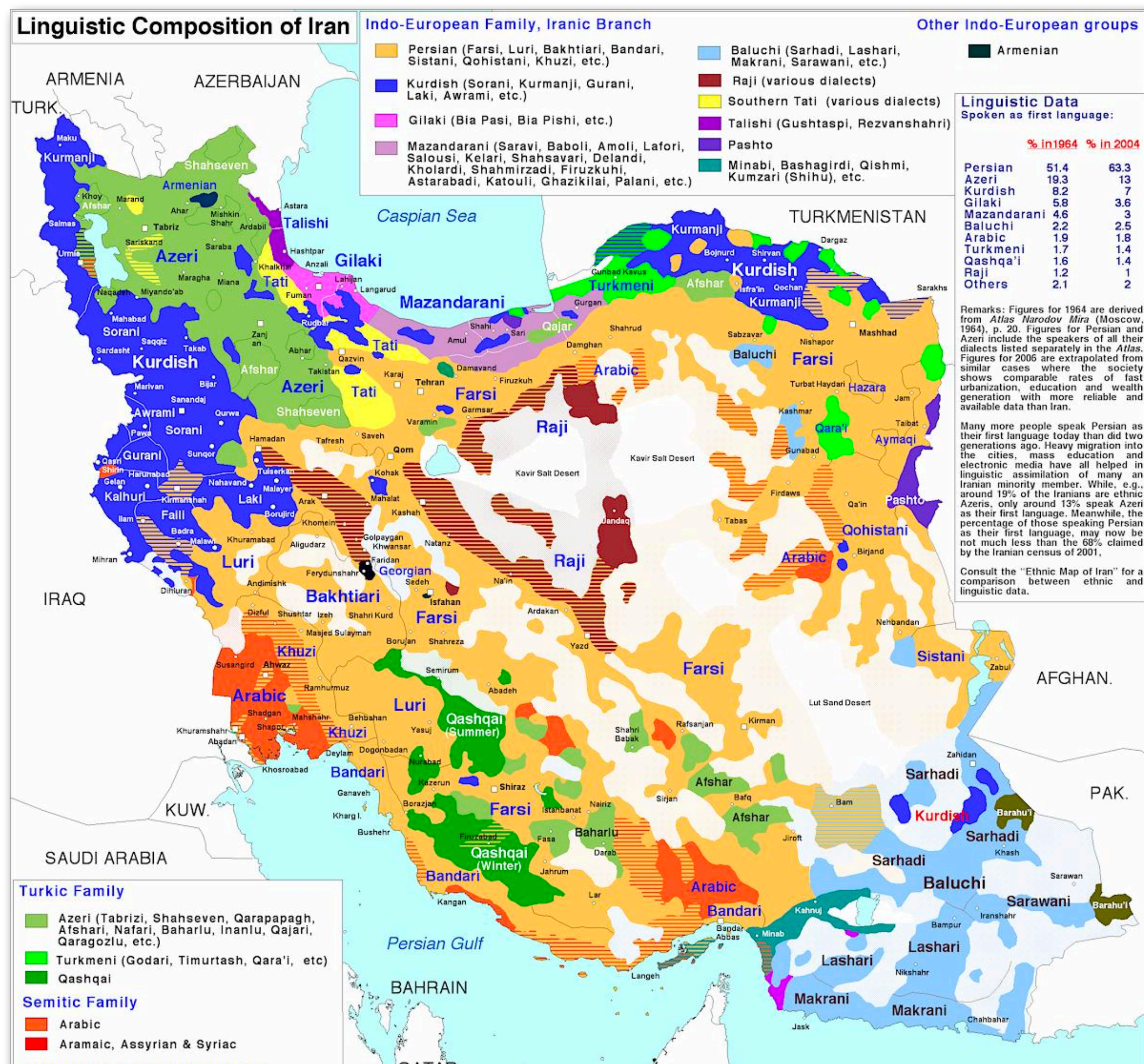
DialectBench Results



Minority Languages

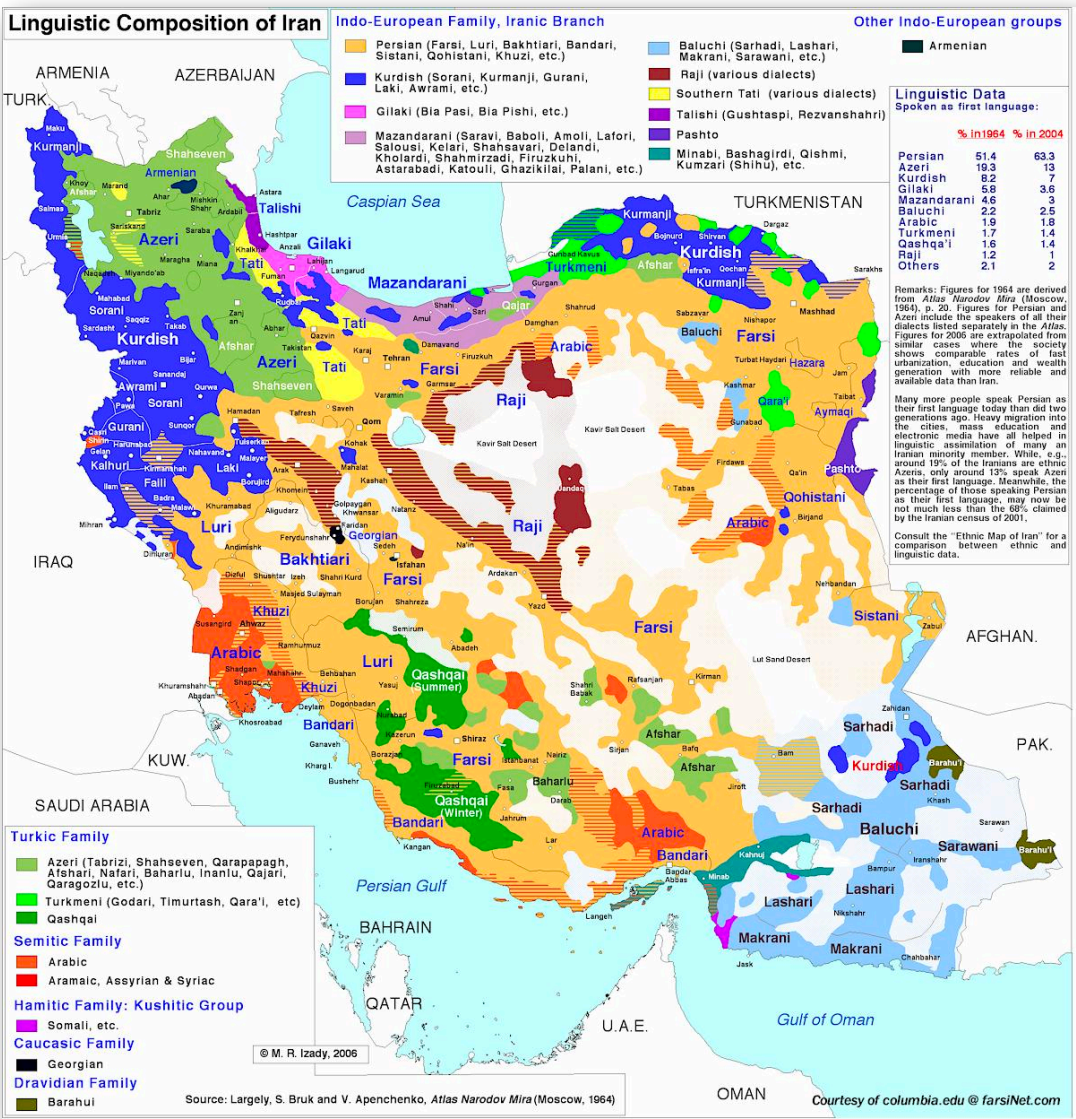
Minority Languages in X-lingual Communities

Minority Languages in X-lingual Communities

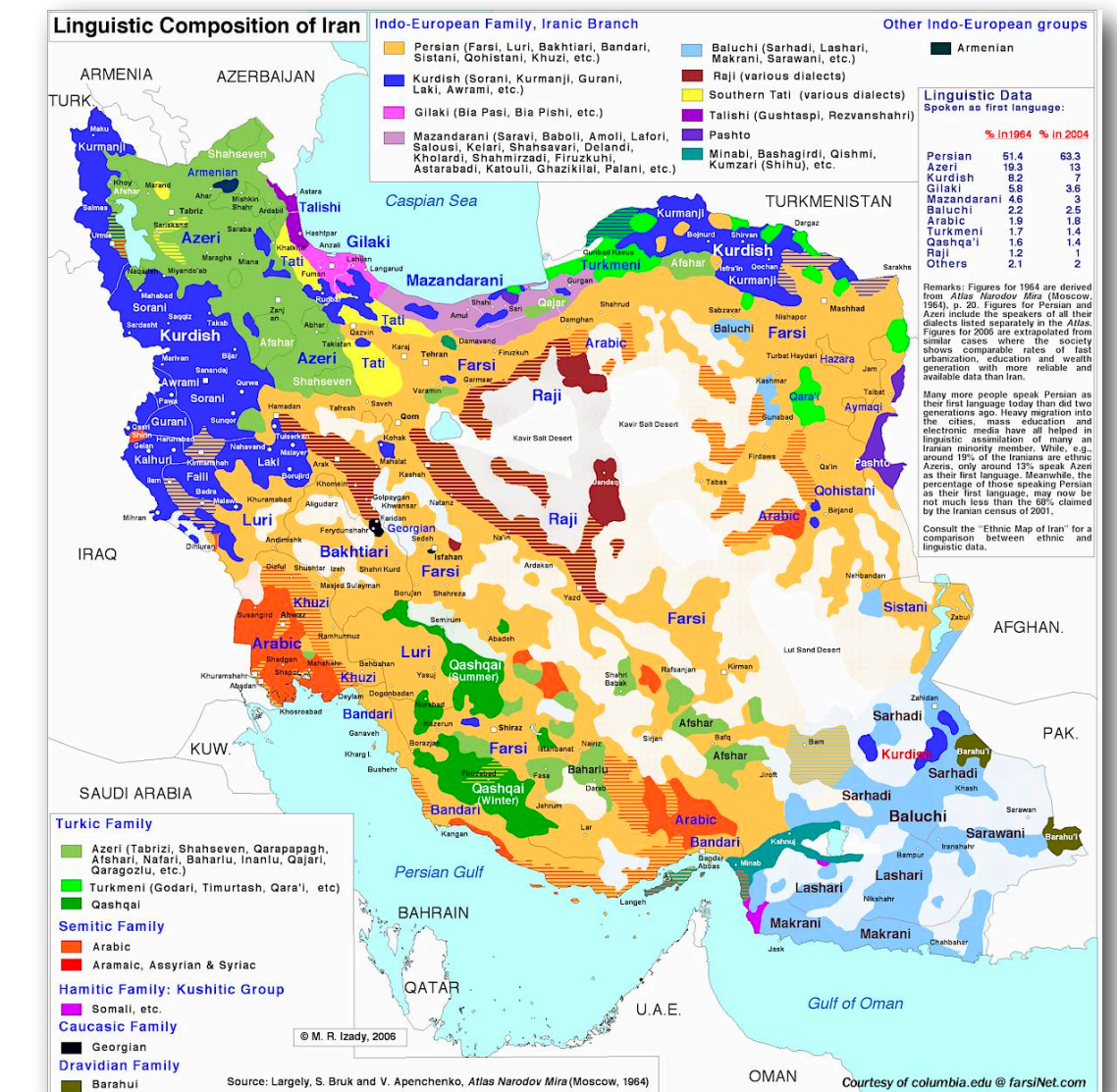
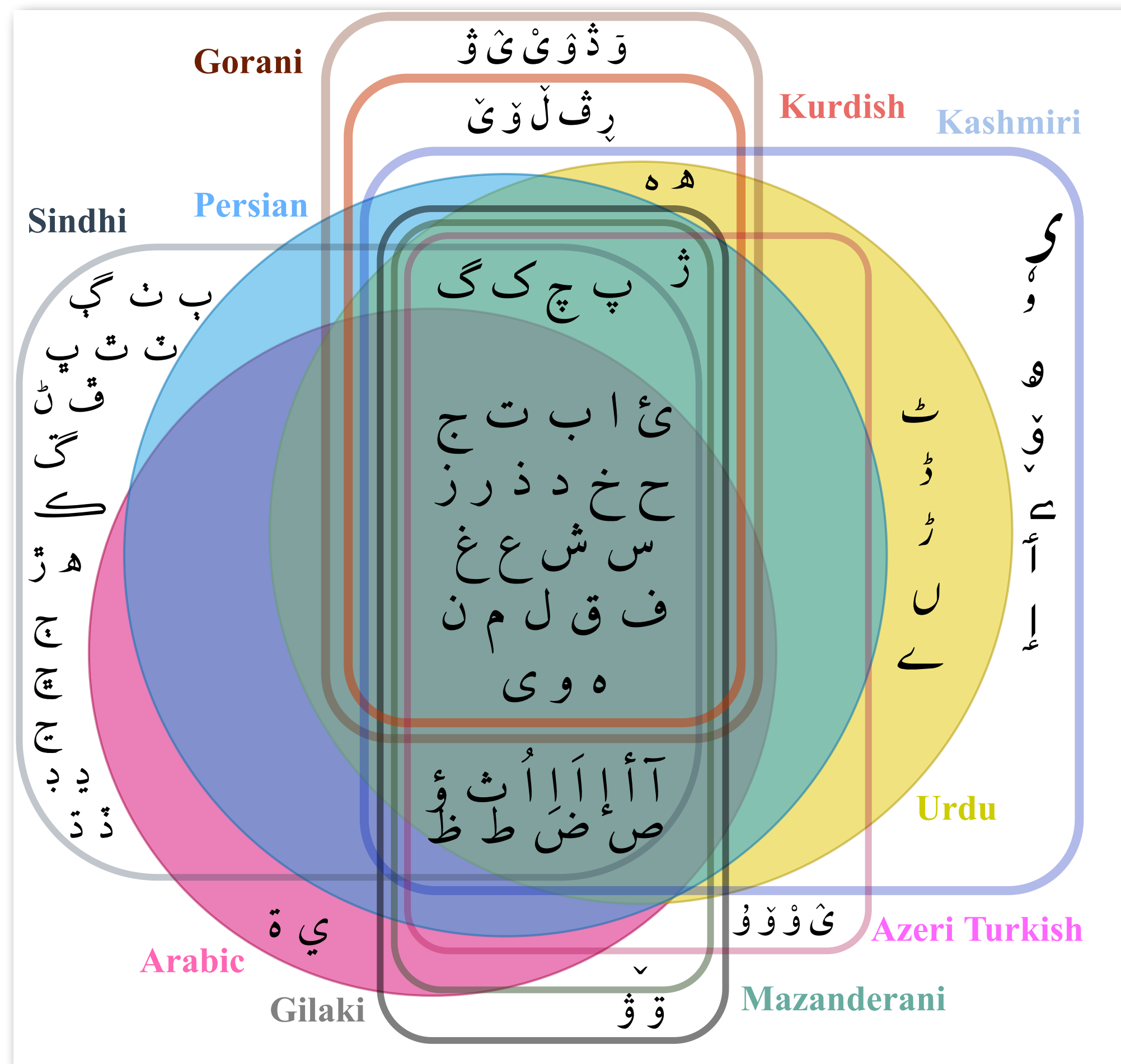


Source:

Minority Languages in X-lingual Communities

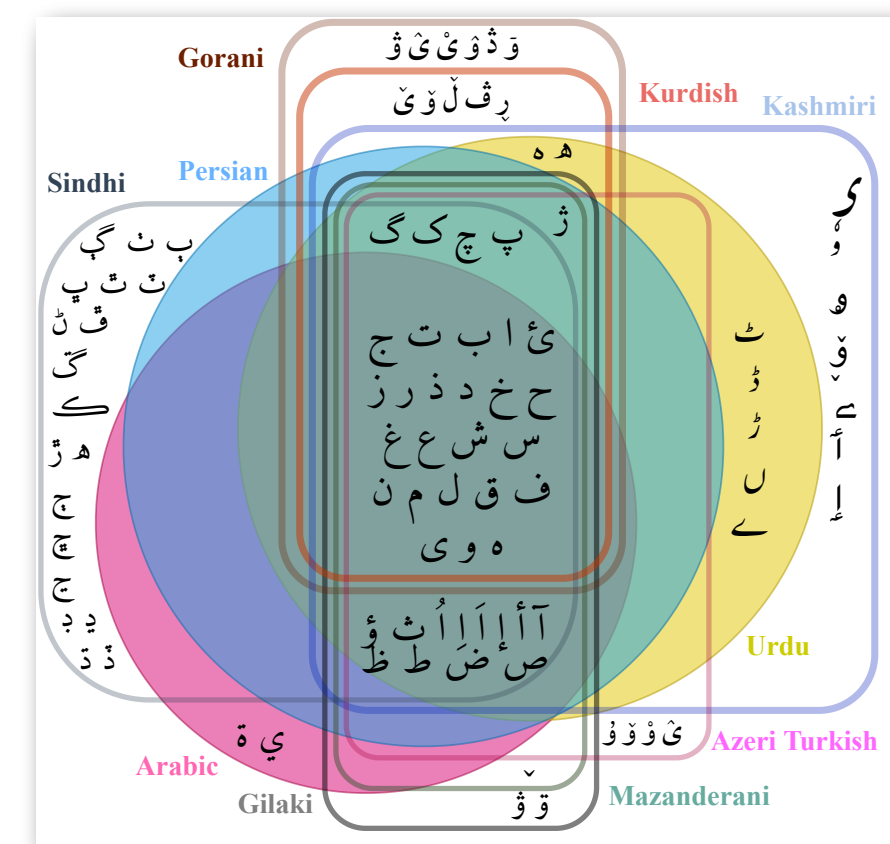
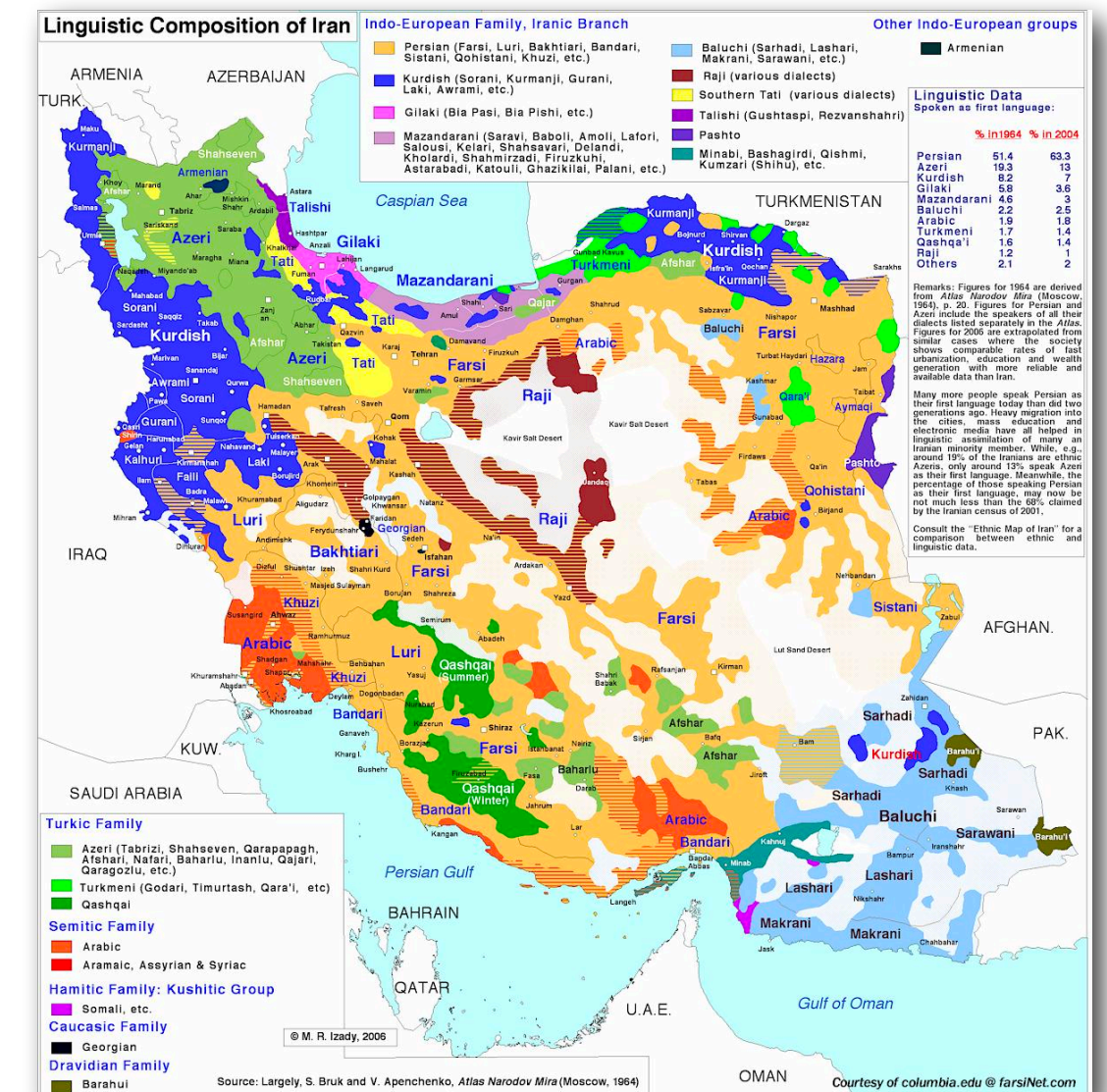


Minority Languages in X-lingual Communities



Minority Languages in X-lingual Communities

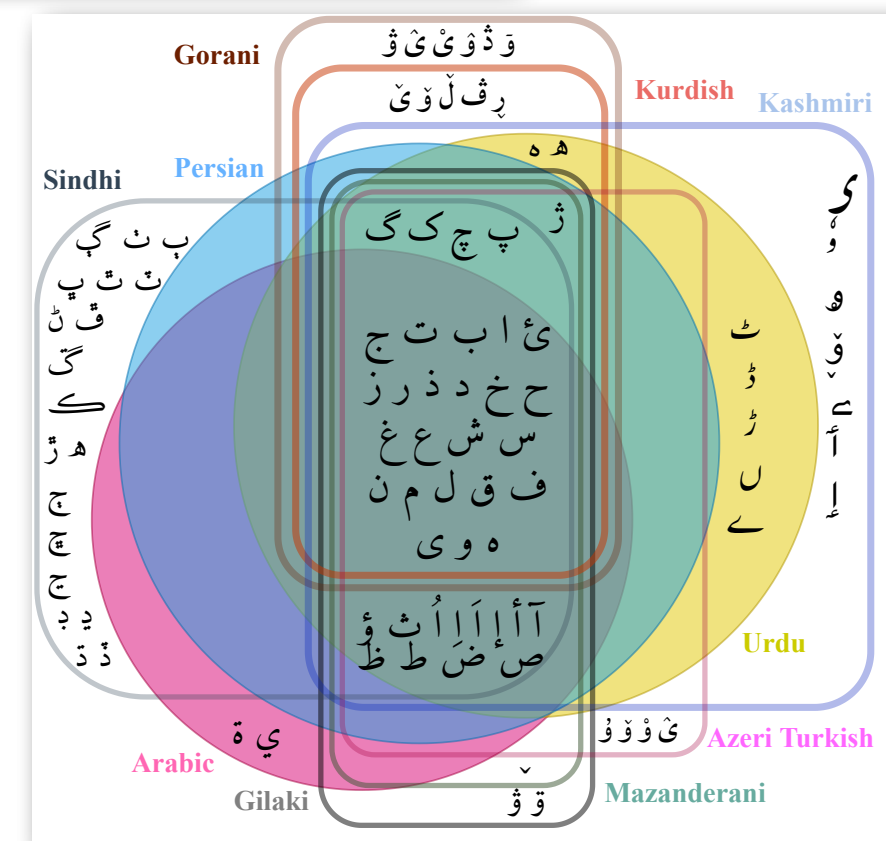
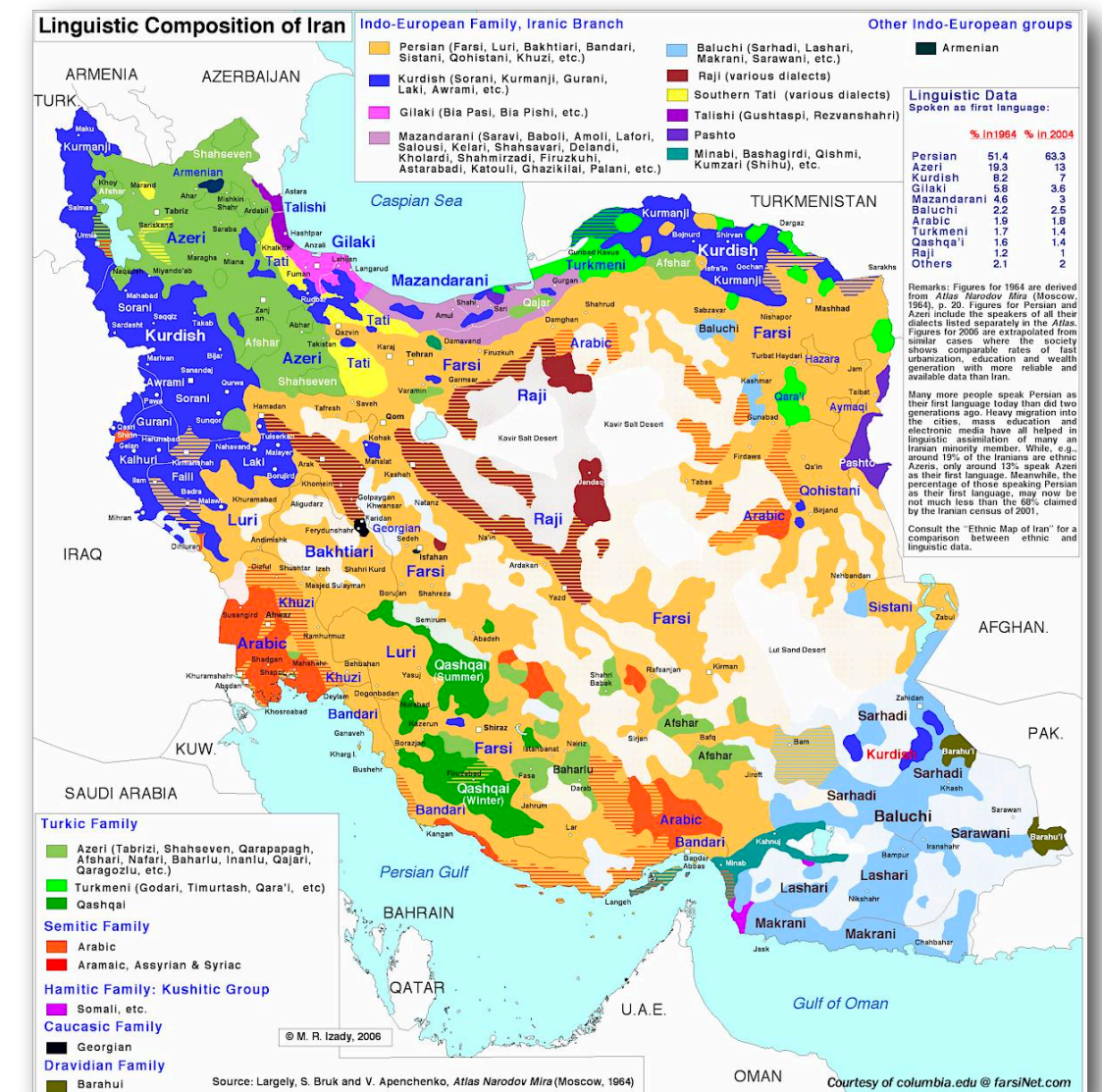
Dominant language (e.g. Farsi) influences the minority one:



Minority Languages in X-lingual Communities

Dominant language (e.g. Farsi) influences the minority one:

Language	Unconventional script	Unconventional writing	Conventional writing
Gilaki	Persian	یتہ زون نم ہیسه گه گیلکن اون جی گب زنن	یتہ زوؤن نۆم هیسه گه گیلکۆن اۆن جی گب زنن
Kashmiri	Urdu	برور چھ اکھ ورا سے جانور۔	برور چھُ اکھ وُرأسِ جانور۔
Kurmanji	Arabic	قایمقام الامدی بترثوا پارزکار دھوک دا	قایمقامی ئامیدی بهرسقا پاریزگاری دهوکی دا
Sorani	Arabic	هەر له یه کهم شانوو دياره فهه دیان دهویت	هەر له یه کهم شانوو دياره فهه دیان دهویت
Sindhi	Urdu	مدیني ذانهن هجرت وقت فقط هيء گھر واري سان گد هئي	مدیني ڈانهن هجرت وقت فقط هيء گهر واري ساڻن گڏ هئي



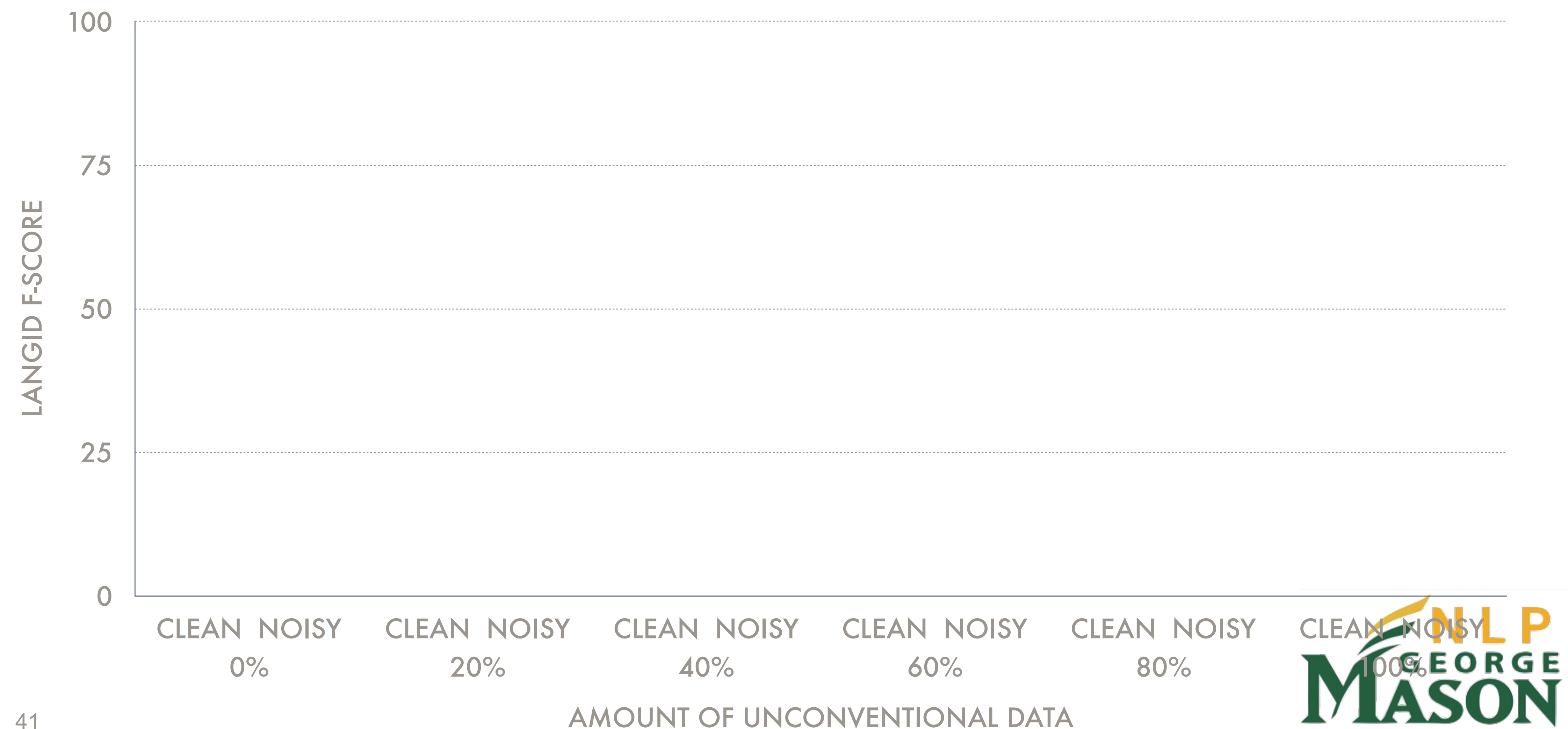
Case Study: Languages using Perso-Arabic Script

Language	639-3	WP	Script type	Diacritics ZWNJ		Dominant
Azeri Turkish	azb	azb	Abjad	✓	✓	Persian
Gilaki	glk	glk	Abjad	✓	✓	Persian
Mazanderani	mzn	mzn	Abjad	✓	✓	Persian
Pashto	pus	ps	Abjad	✓	✗	Persian
Gorani	hac	-	Alphabet	✗	✗	Persian, Arabic, Sorani
Northern Kurdish (Kurmanji)	kmr	-	Alphabet	✗	✗	Persian, Arabic
Central Kurdish (Sorani)	ckb	ckb	Alphabet	✗	✗	Persian, Arabic
Southern Kurdish	sdh	-	Alphabet	✗	✗	Persian, Arabic
Balochi	bal	-	Abjad	✓	✗	Persian, Urdu
Brahui	brh	-	Abjad	✓	✗	Urdu
Kashmiri	kas	ks	Alphabet	✓	✗	Urdu
Sindhi	snd	sd	Abjad	✓	✗	Urdu
Saraiki	skr	skr	Abjad	✓	✗	Urdu
Torwali	trw	-	Abjad	✓	✗	Urdu
Punjabi	pnb	pnb	Abjad	✓	✗	Urdu
Persian	fas	fa	Abjad	✓	✓	-
Arabic	arb	ar	Abjad	✓	✗	-
Urdu	urd	ur	Abjad	✓	✓	-
Uyghur	uig	ug	Alphabet	✗	✗	-

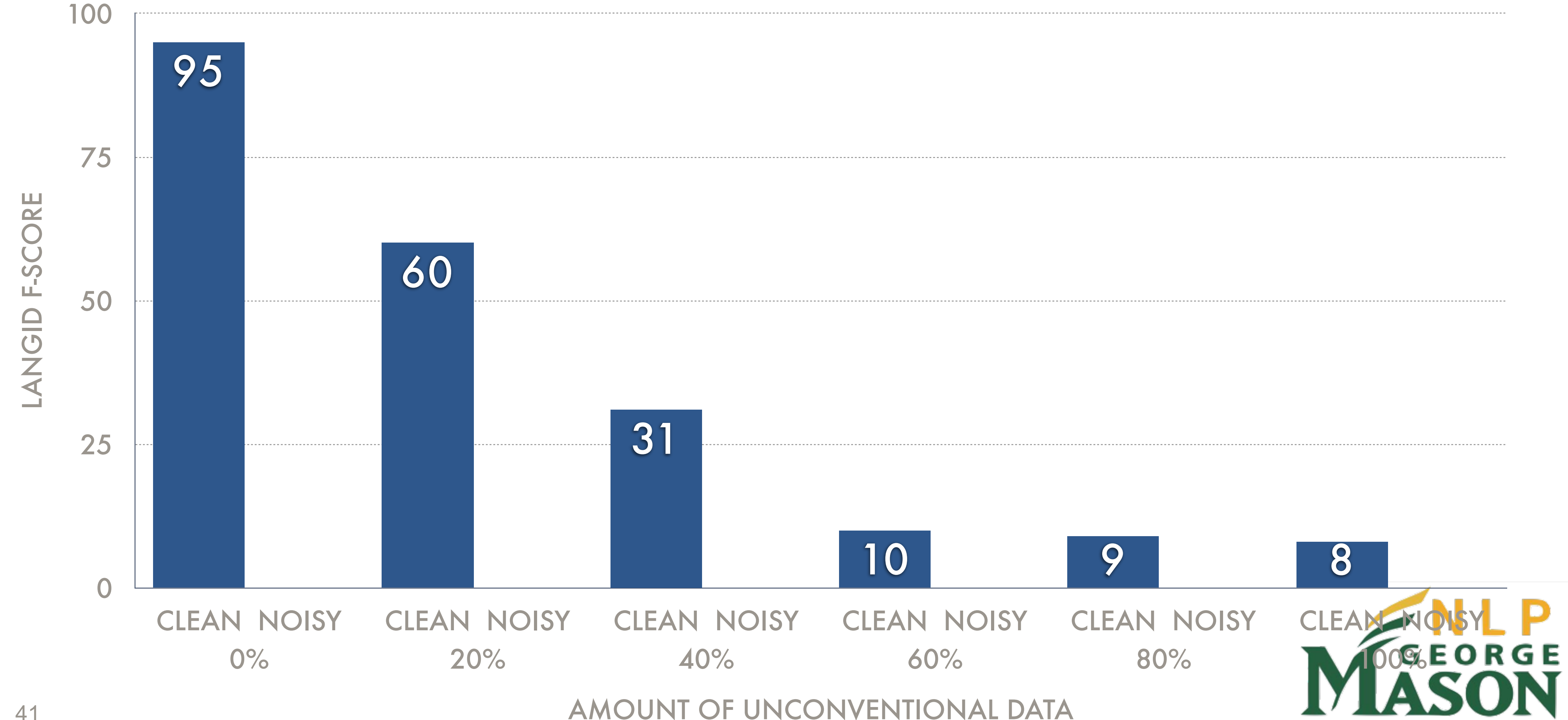
Table 1: Perso-Arabic scripts of the selected languages studied in this paper. Columns 2 and 3 show the codes of the languages in ISO 639-3 and on their specific Wikipedia (WP), if available. The diacritics and zero-width non-joiner (ZWNJ) columns refer to the usage of diacritics (*Harakat*) and ZWNJ as individual characters.

Effect of Unconventional Writing

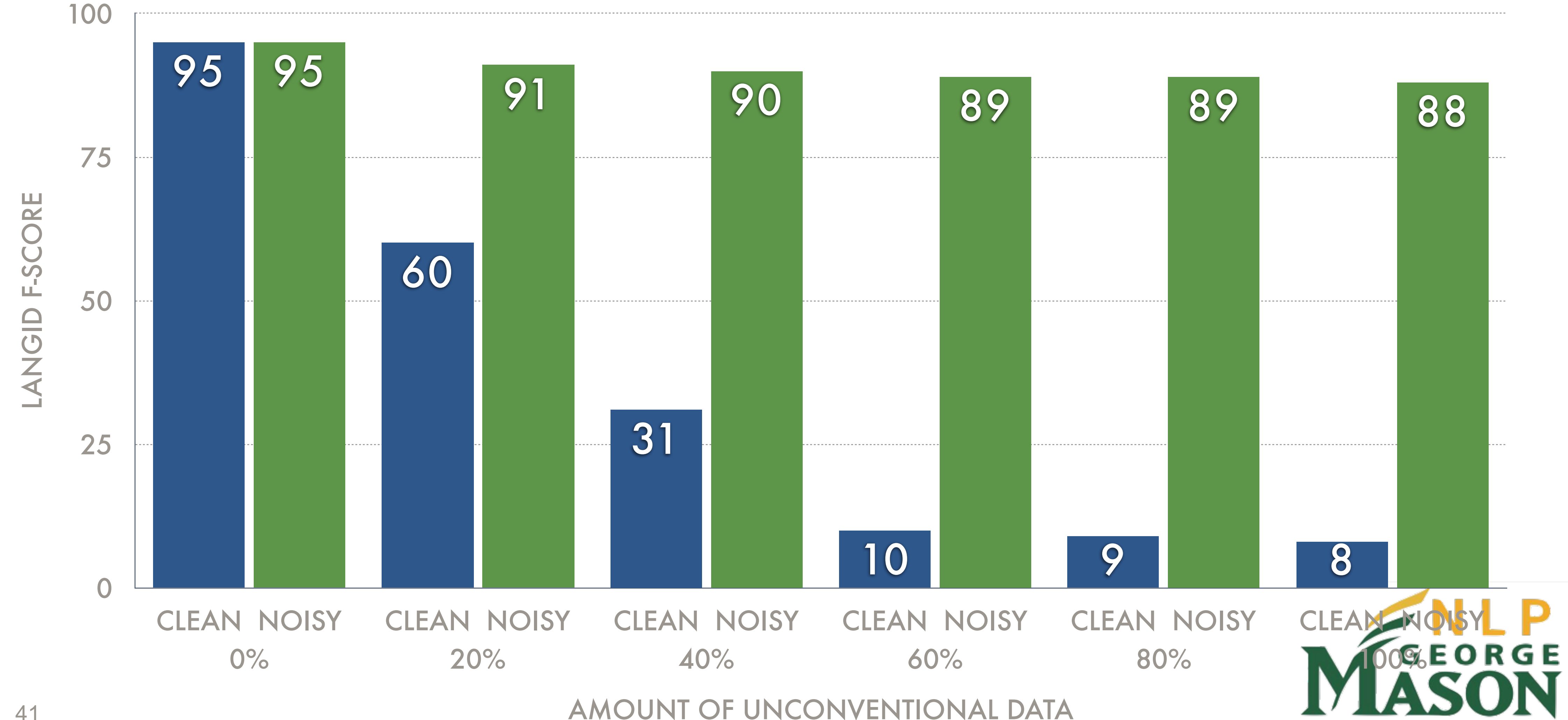
Effect of Unconventional Writing



Effect of Unconventional Writing



Effect of Unconventional Writing



Mitigating the Effect of Unconventional Writing

Mitigating the Effect of Unconventional Writing

Train a Normalization model
(Encoder-decoder, self-attention based)

Mitigating the Effect of Unconventional Writing

Train a Normalization model

(Encoder-decoder, self-attention based)

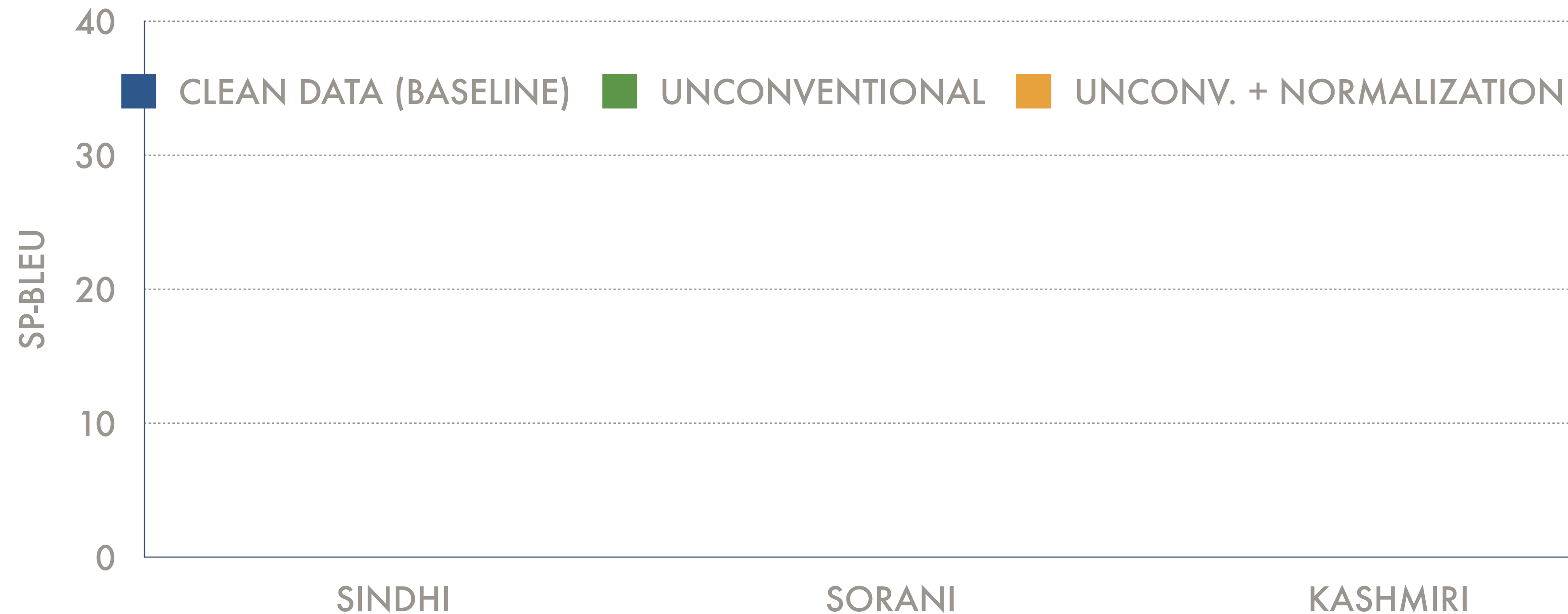
Evaluate its effect on Machine Translation

Mitigating the Effect of Unconventional Writing

Train a Normalization model

(Encoder-decoder, self-attention based)

Evaluate its effect on Machine Translation

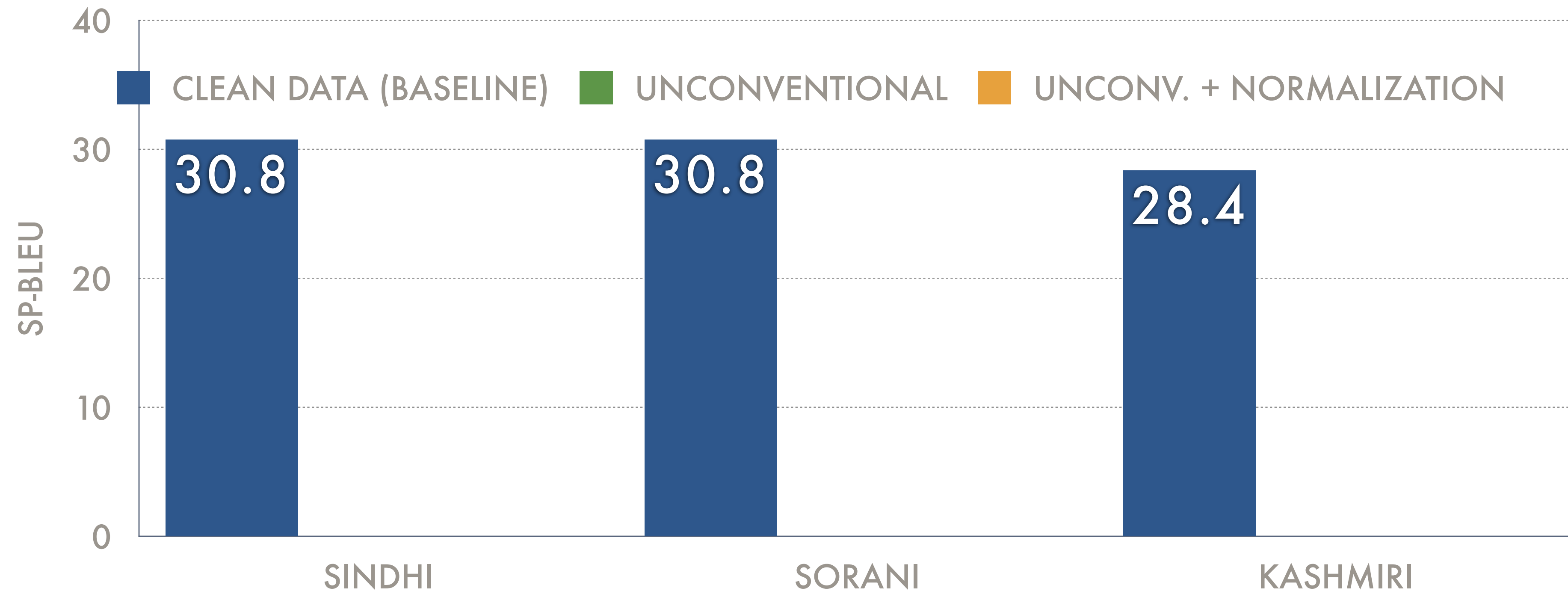


Mitigating the Effect of Unconventional Writing

Train a Normalization model

(Encoder-decoder, self-attention based)

Evaluate its effect on Machine Translation

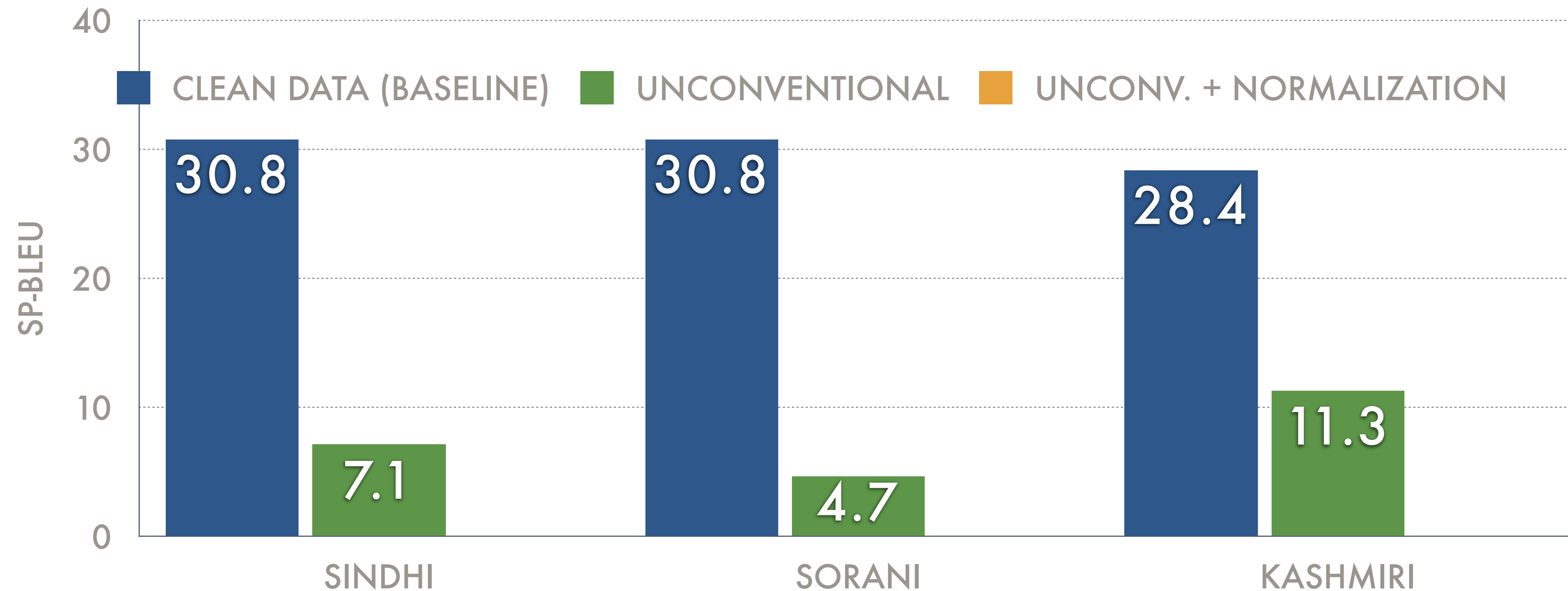


Mitigating the Effect of Unconventional Writing

Train a Normalization model

(Encoder-decoder, self-attention based)

Evaluate its effect on Machine Translation

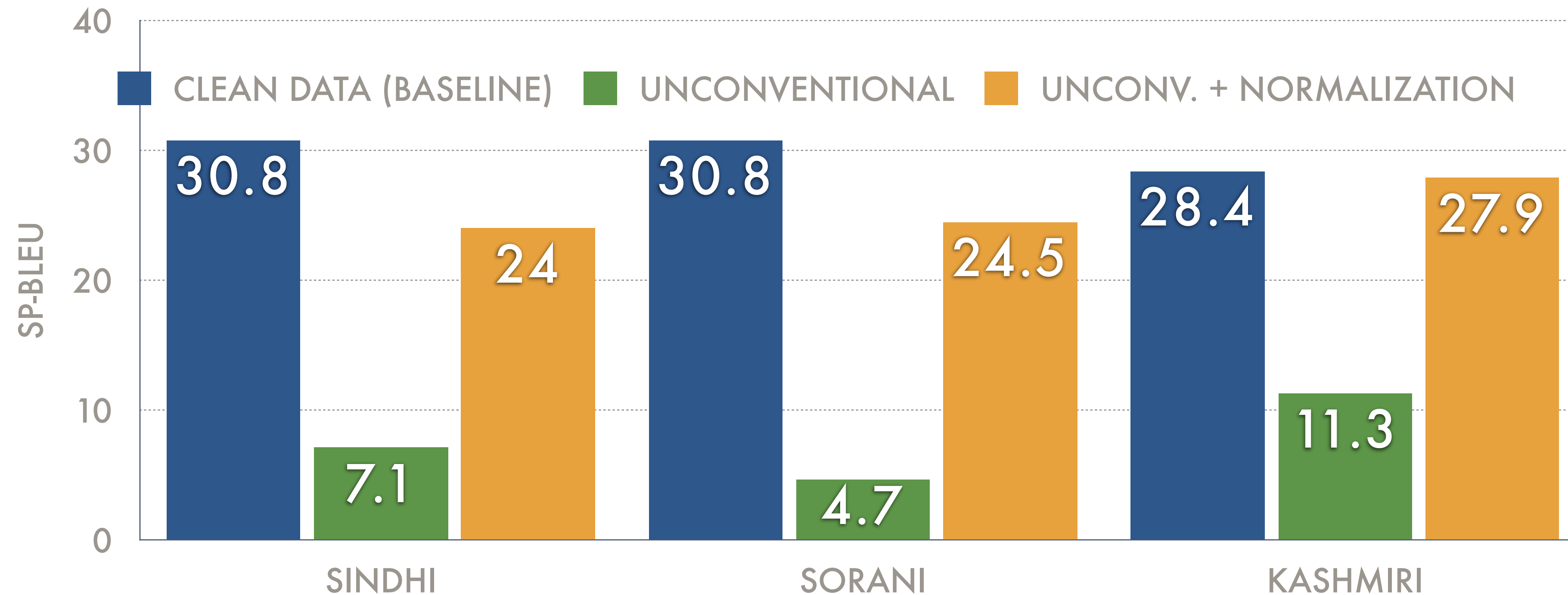


Mitigating the Effect of Unconventional Writing

Train a Normalization model

(Encoder-decoder, self-attention based)

Evaluate its effect on Machine Translation



We Need to Handle Speech Input

With many slides from the
“End-to-end-ST tutorial” at EACL 2021
by Jan Niehues, Liz Salesky, Marco
Turchi, and Matteo Negri

Speech Translation - History (before e2e)

Late '80s: first proofs of concept

Constraints to control language ambiguity (phonetics, syntax, semantics)

- Restricted vocabulary
- Controlled speaking style
- Narrow domain
- Offline processing

'90s: Less constraints (vocabulary, speaking style)

First spontaneous ST systems (C-STAR, Verbmobil, Nespole,...)

2003-2006: Less constraints (domain)

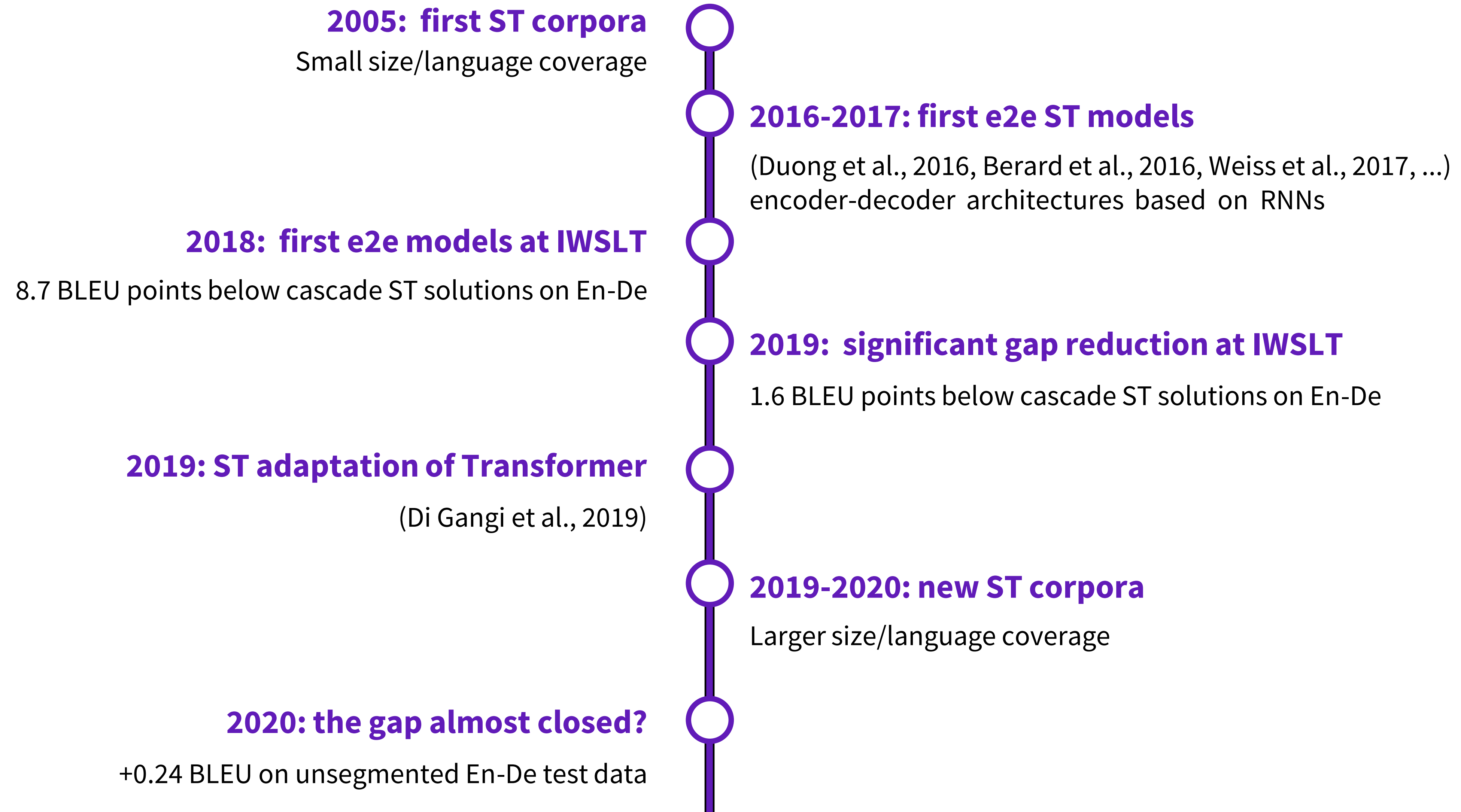
First open-domain ST systems (STR-DUST, TC-STAR, GALE)

- different scenarios (broadcast news, parliamentary speeches, academic lectures)
- different languages (Zh, Ar, Es)

2006: Less constraints (operating conditions)

First simultaneous translator
(real-time translation of spontaneous lectures and presentations)

Speech Translation - History (the e2e era)



Sec 1.2

Challenges in Translation of Speech

Challenges in translation of speech

- Audio challenges
 - Multiple speaker
 - e.g. Meetings
 - Challenges:
 - Overlapping voice
 - Background noise
 - Audio segmentation



Challenges in translation of speech

- Audio challenges
- Text-Speech mismatch
 - Disfluencies
 - Hesitations: “uh”, “uhm”, “hmm”,
 - Discourse markers: “you know”, “I mean”,...
 - Repetitions: “It had, it had been a good day”
 - Corrections: “no, it cannot, I cannot go there”
 - No punctuation
 - Let’s eat Grandpa !
 - Let’s eat, Grandpa !



Challenges in translation of speech

- Audio challenges
- Text-Speech mismatch
- Error propagation
 - ASR errors worse after translation
 - More difficult to compensate by human
 - MT adds additional errors



Re**d**en (engl. speeches)



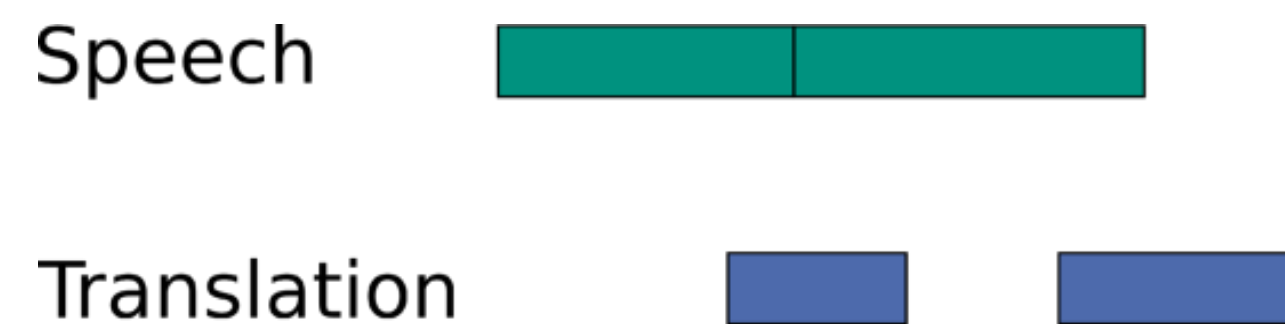
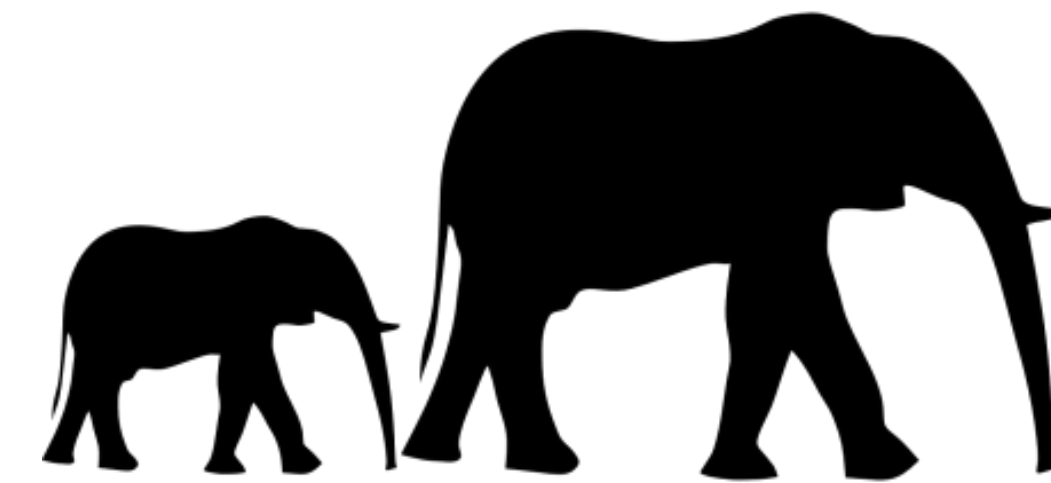
Re**b**en (engl. vines)

Challenges in translation of speech

- Audio challenges
- Text-Speech mismatch
- Error propagation
- Data
 - End-to-End data:
 - Growing amount but still limited
 - Integration of other data types
 - Speech transcripts
 - Parallel data

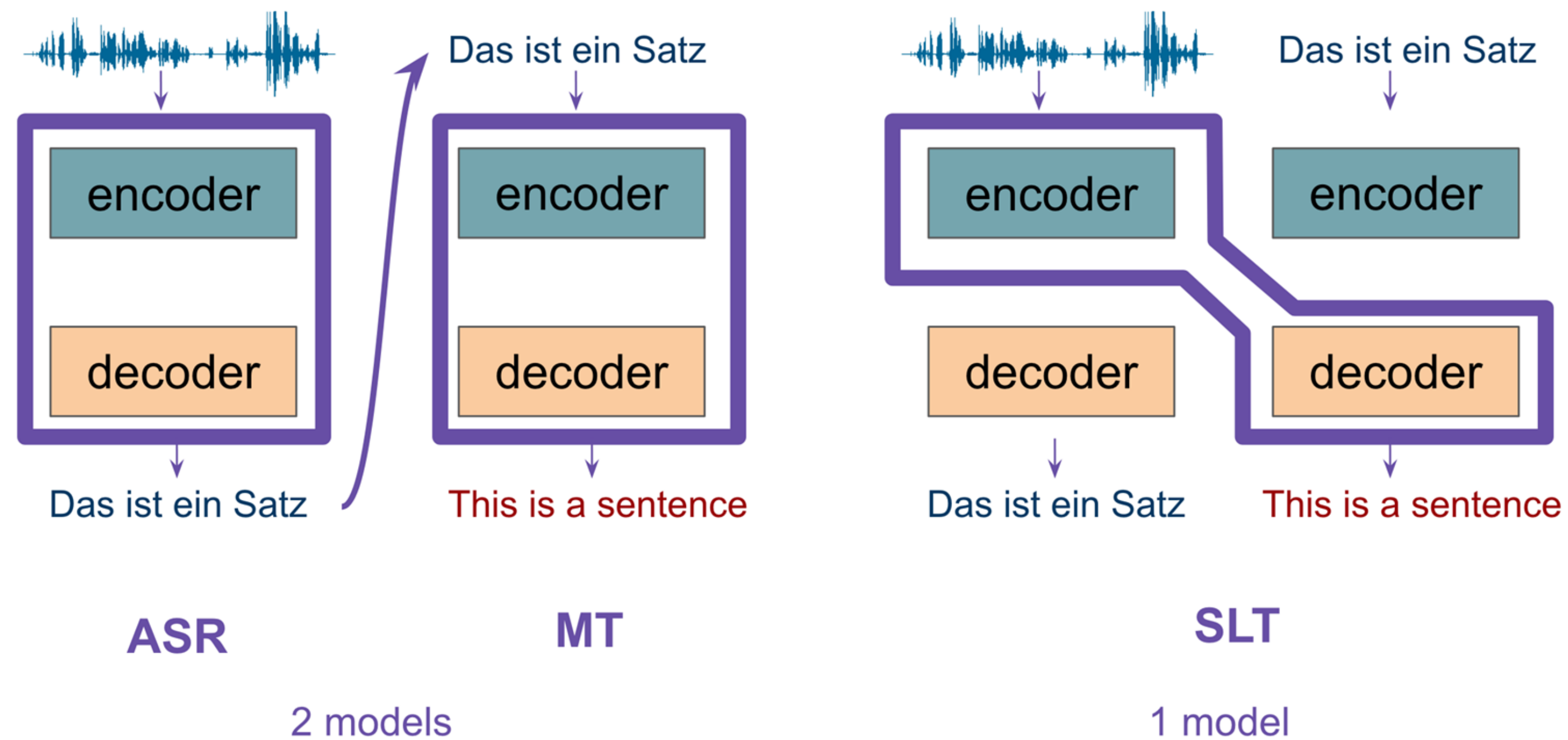
Challenges in translation of speech

- Audio challenges
- Text-Speech mismatch
- Error propagation
- Data
- Partial information
 - Online: Translate during production of speech
 - Generate translation before full sentence is known

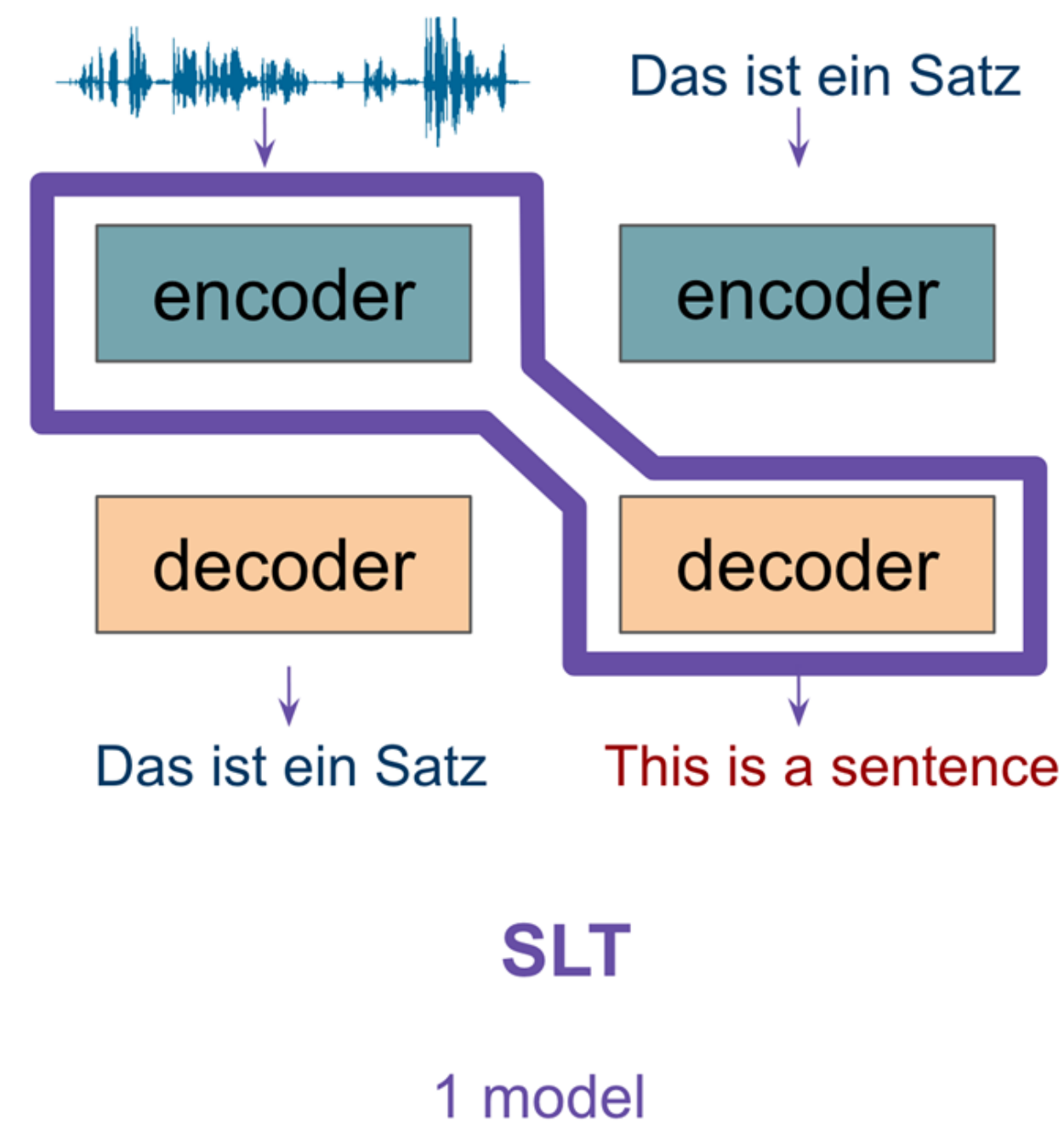
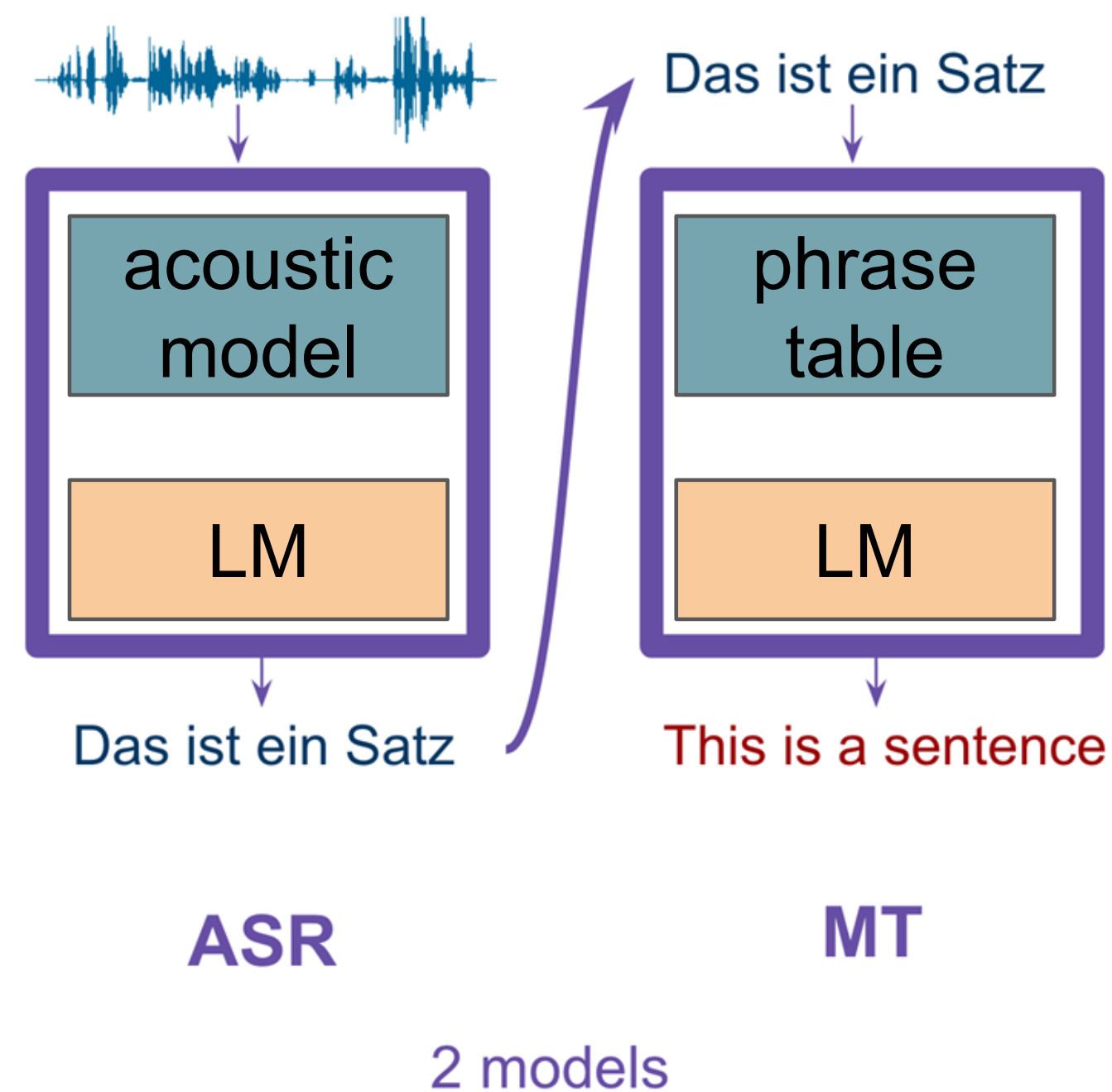


Traditional Cascade Approach

Traditional cascade approach



Traditional cascade approach



Modular, pipeline approach
ASR, MT: isolated objectives

(Waibel et al. 1991; Vidal, 1997; Ney, 1999; Saleem et al. 2004; Matusov et al. 2005; Bertoldi and Federico, 2005; Quan et al. 2005; Kumar et al. 2014; IWSLT Eval Campaigns 2004—)

End-to-End ST



Encoder-Decoder with Attention

the cat sat on the mat





Encoder-Decoder with Attention

the

cat

sat

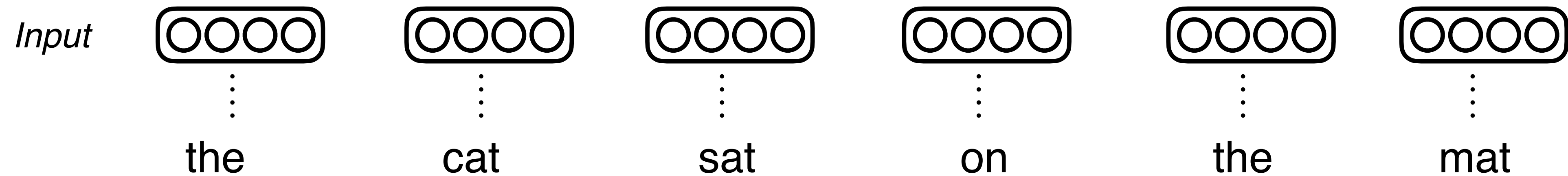
on

the

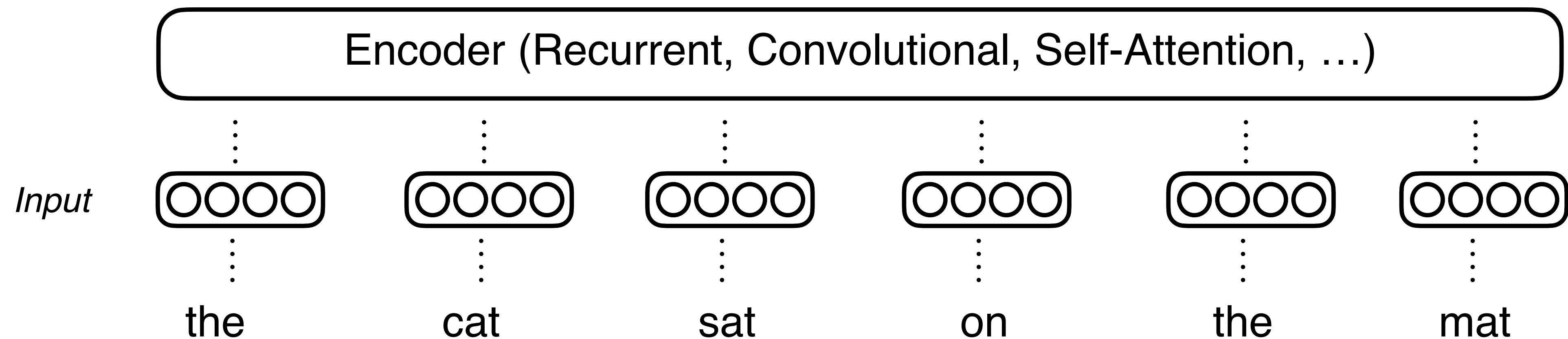
mat



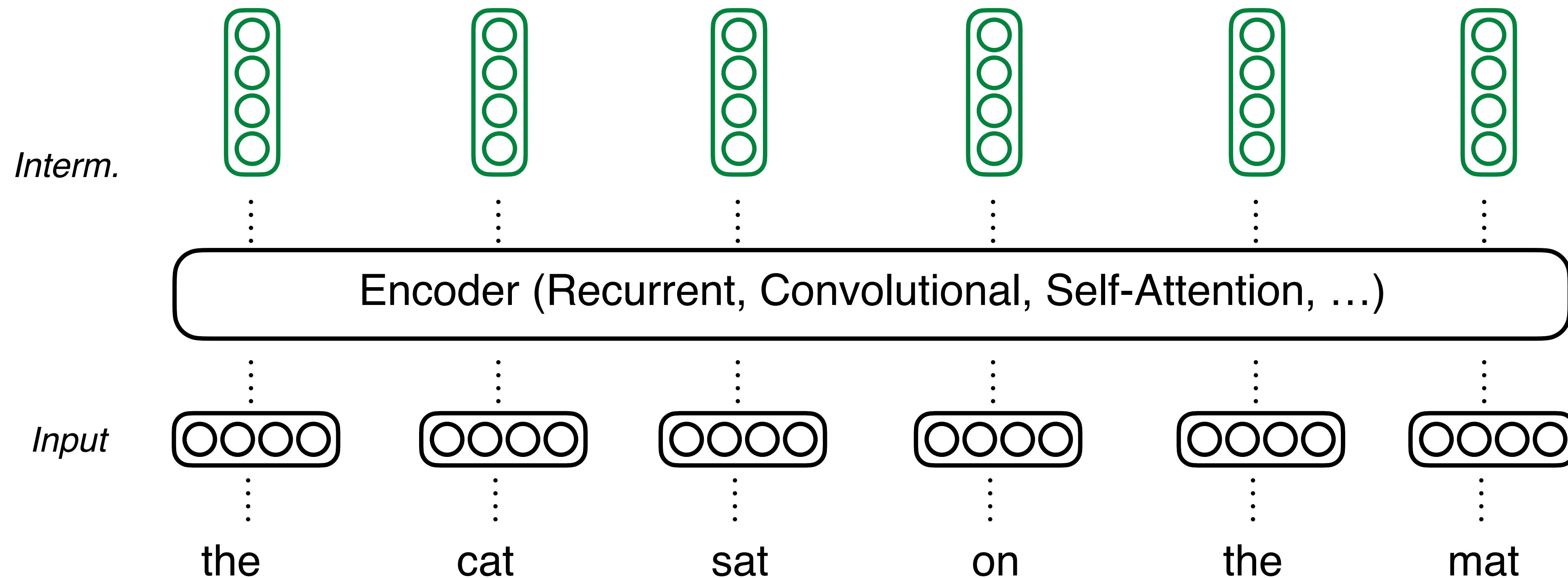
Encoder-Decoder with Attention



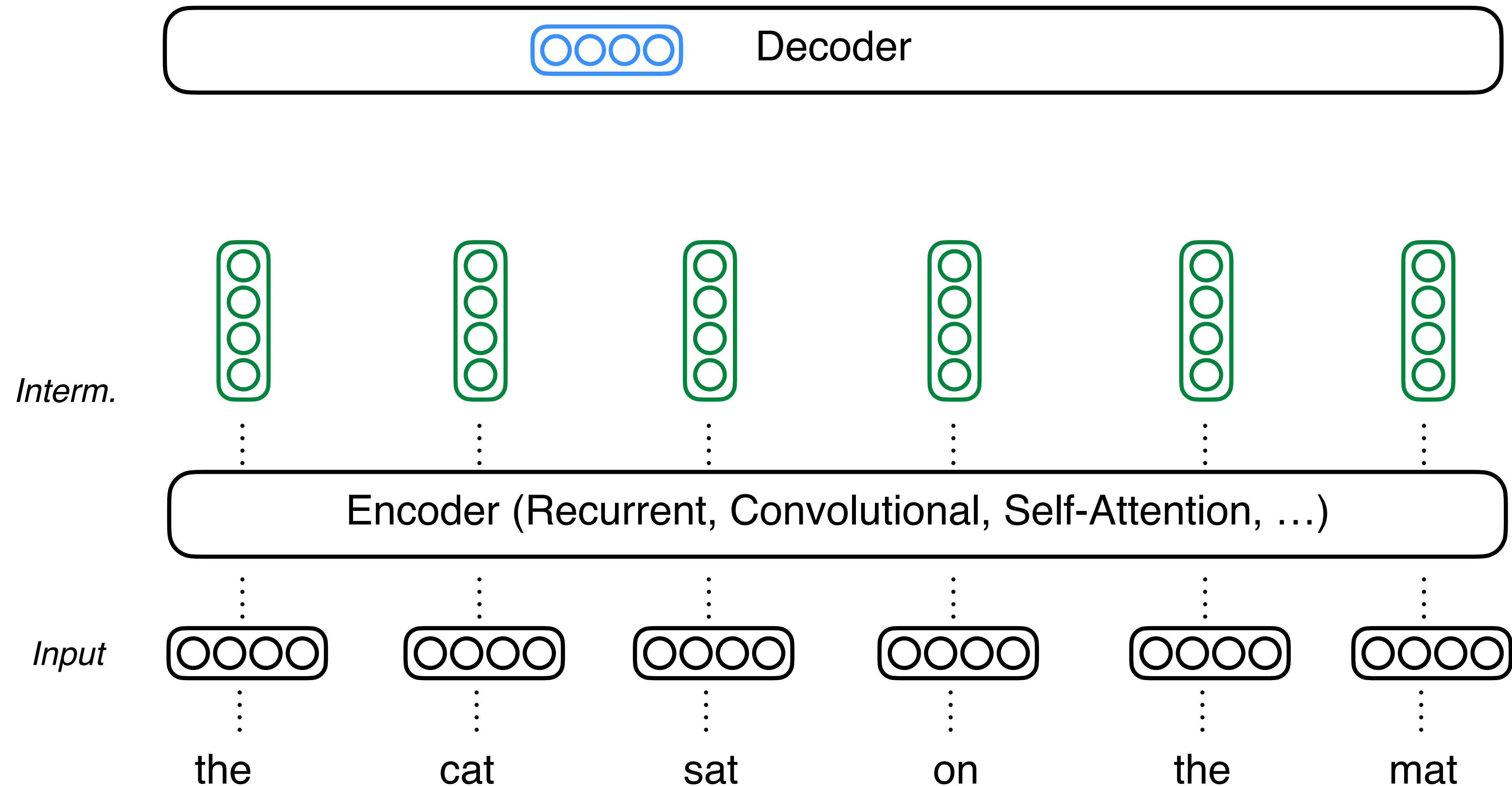
Encoder-Decoder with Attention



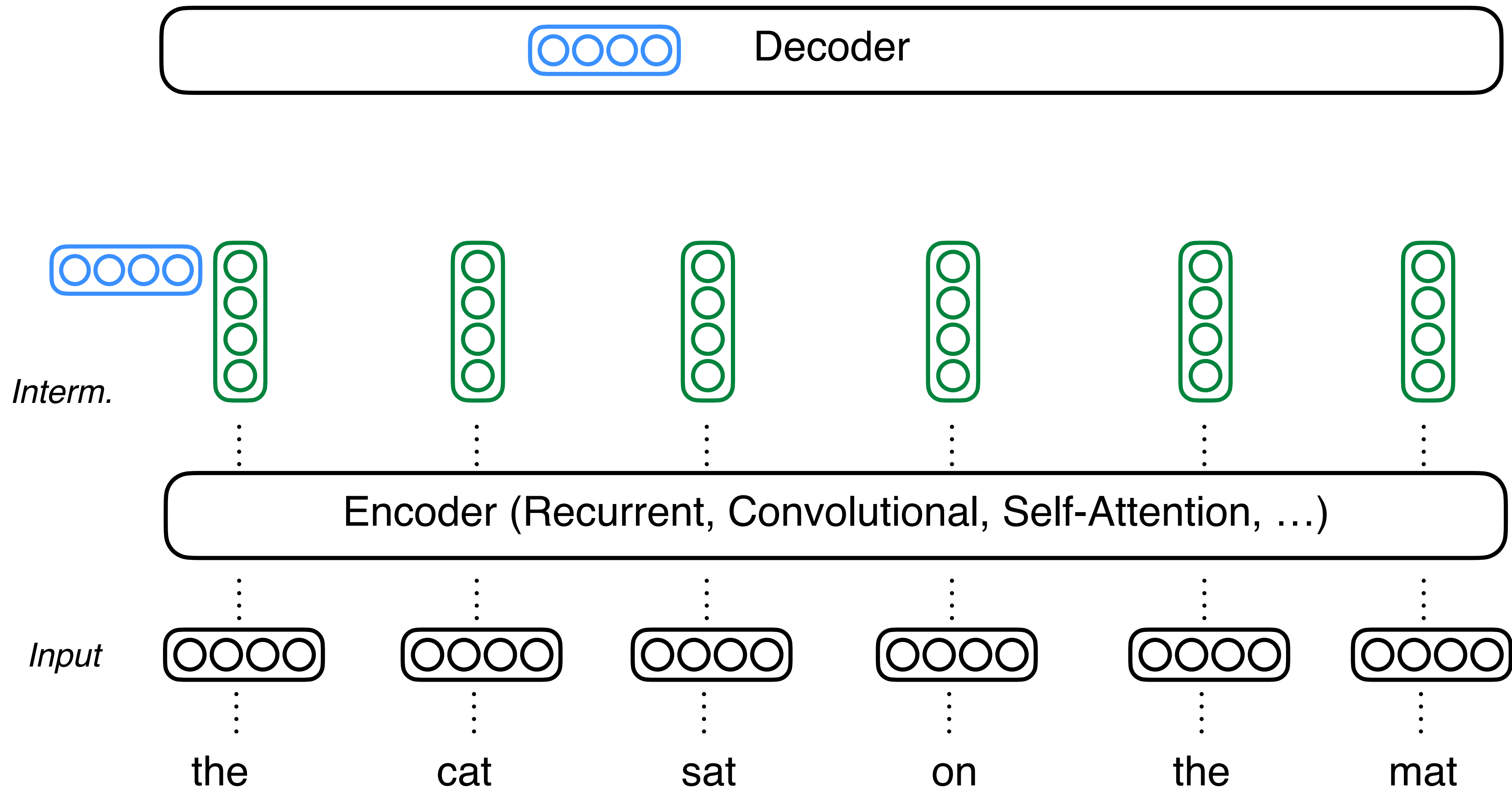
Encoder-Decoder with Attention

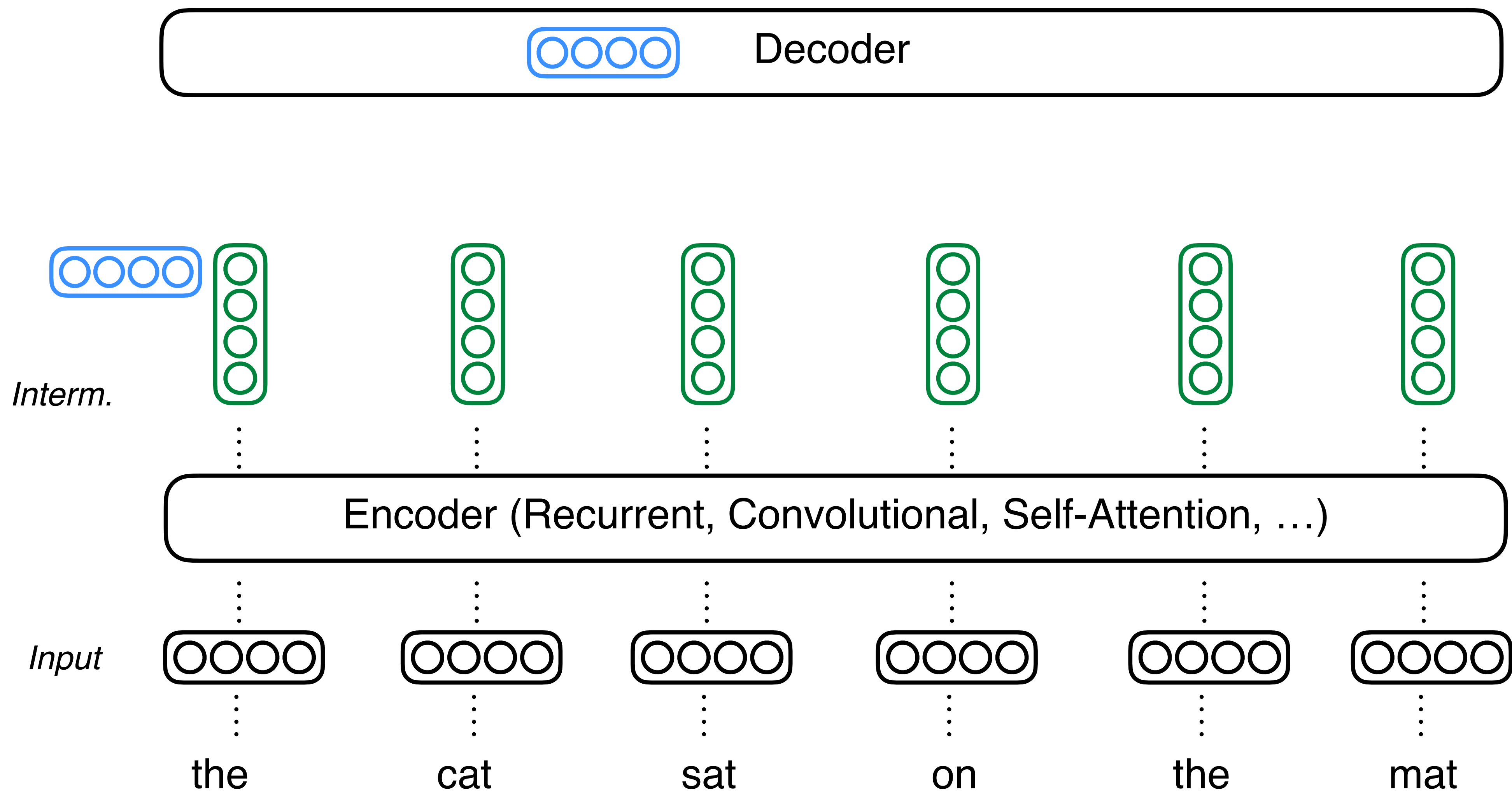


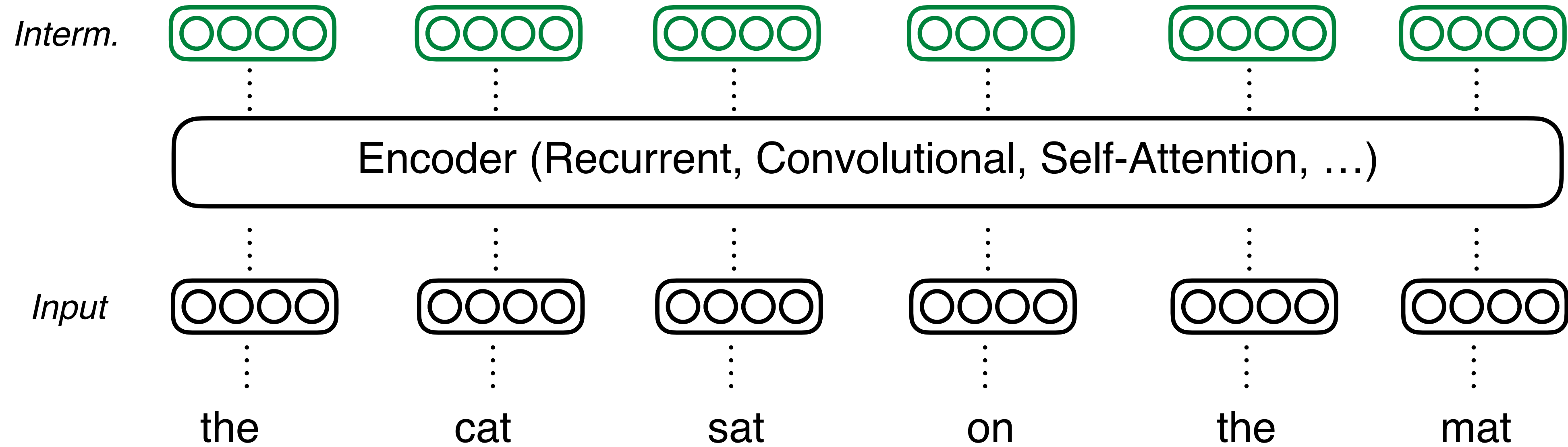
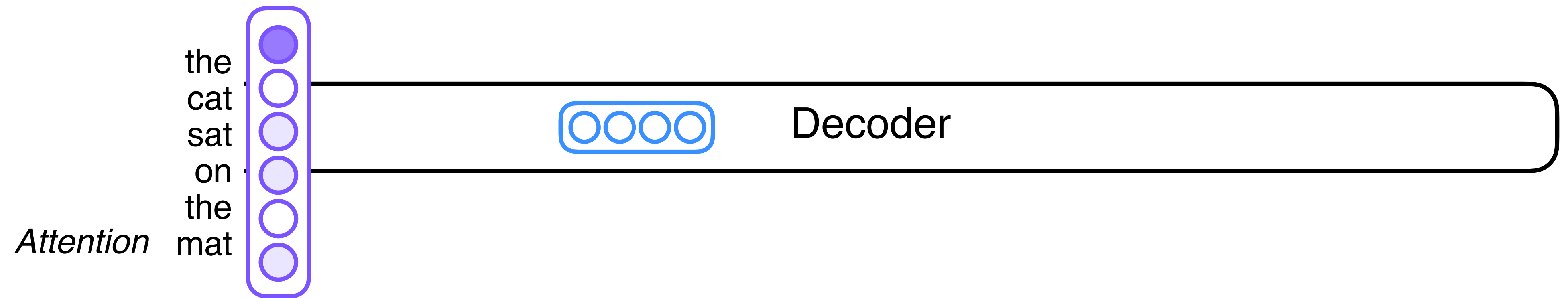
Encoder-Decoder with Attention

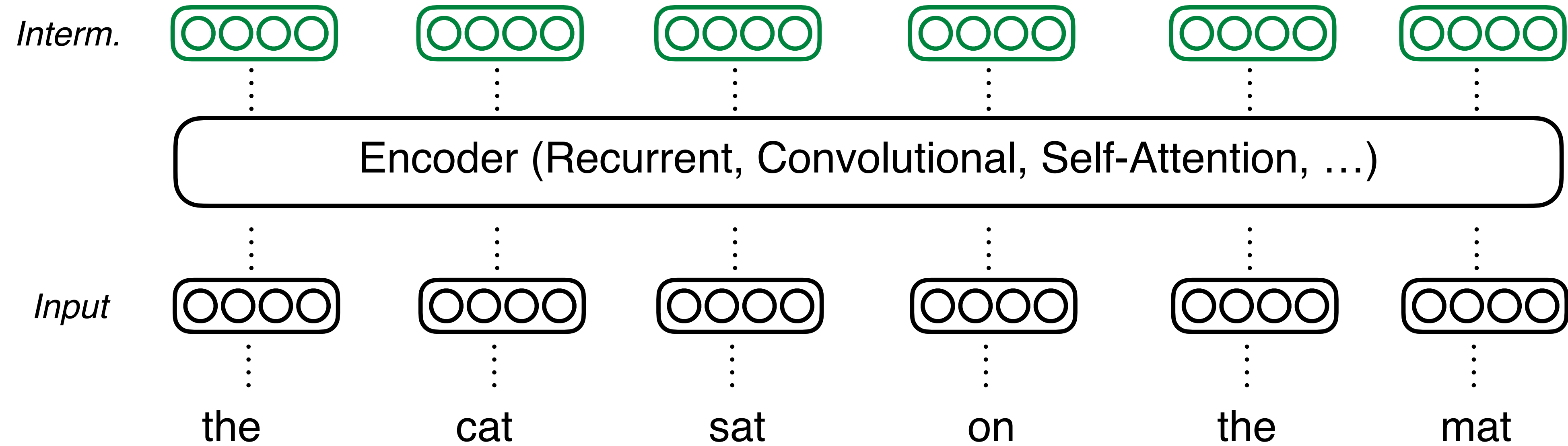
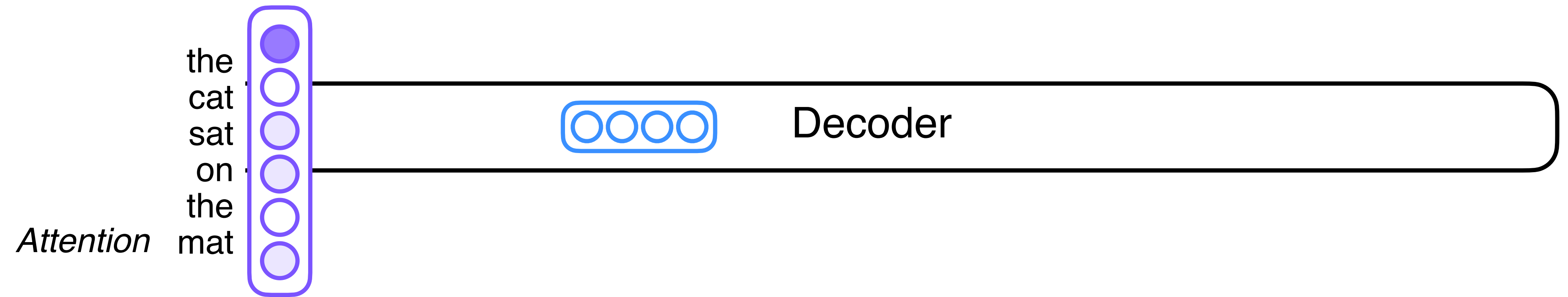


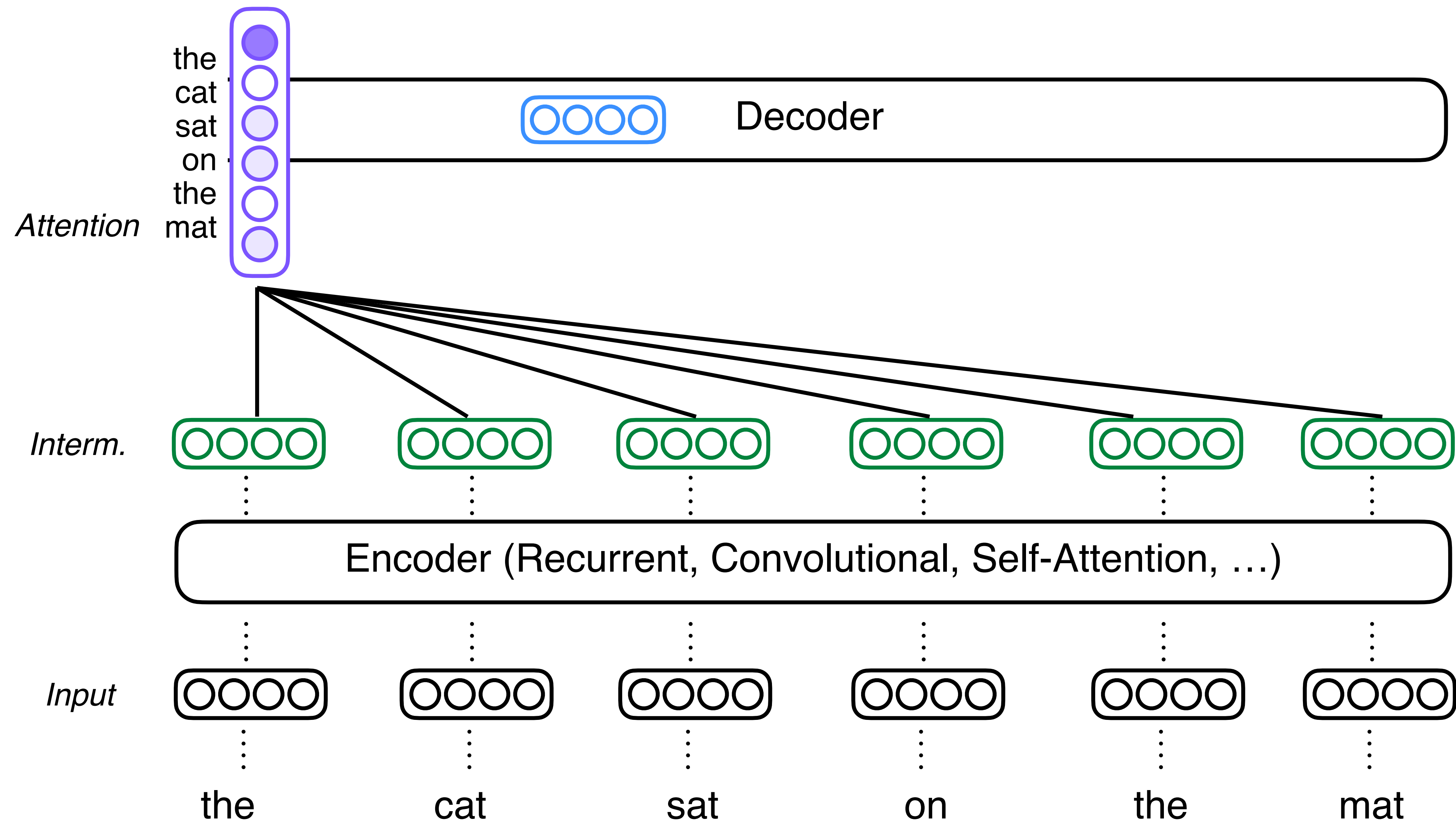
Encoder-Decoder with Attention

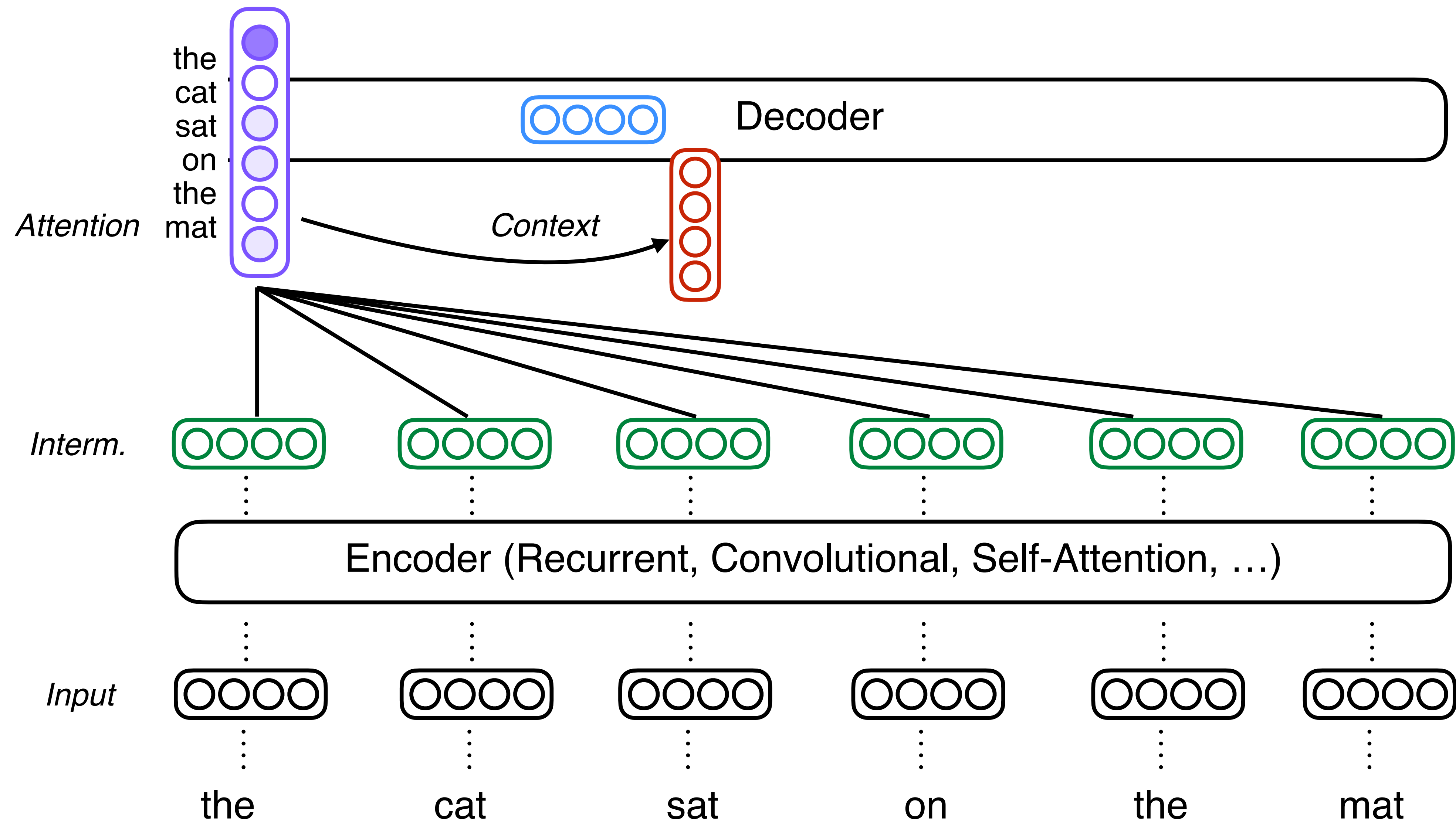


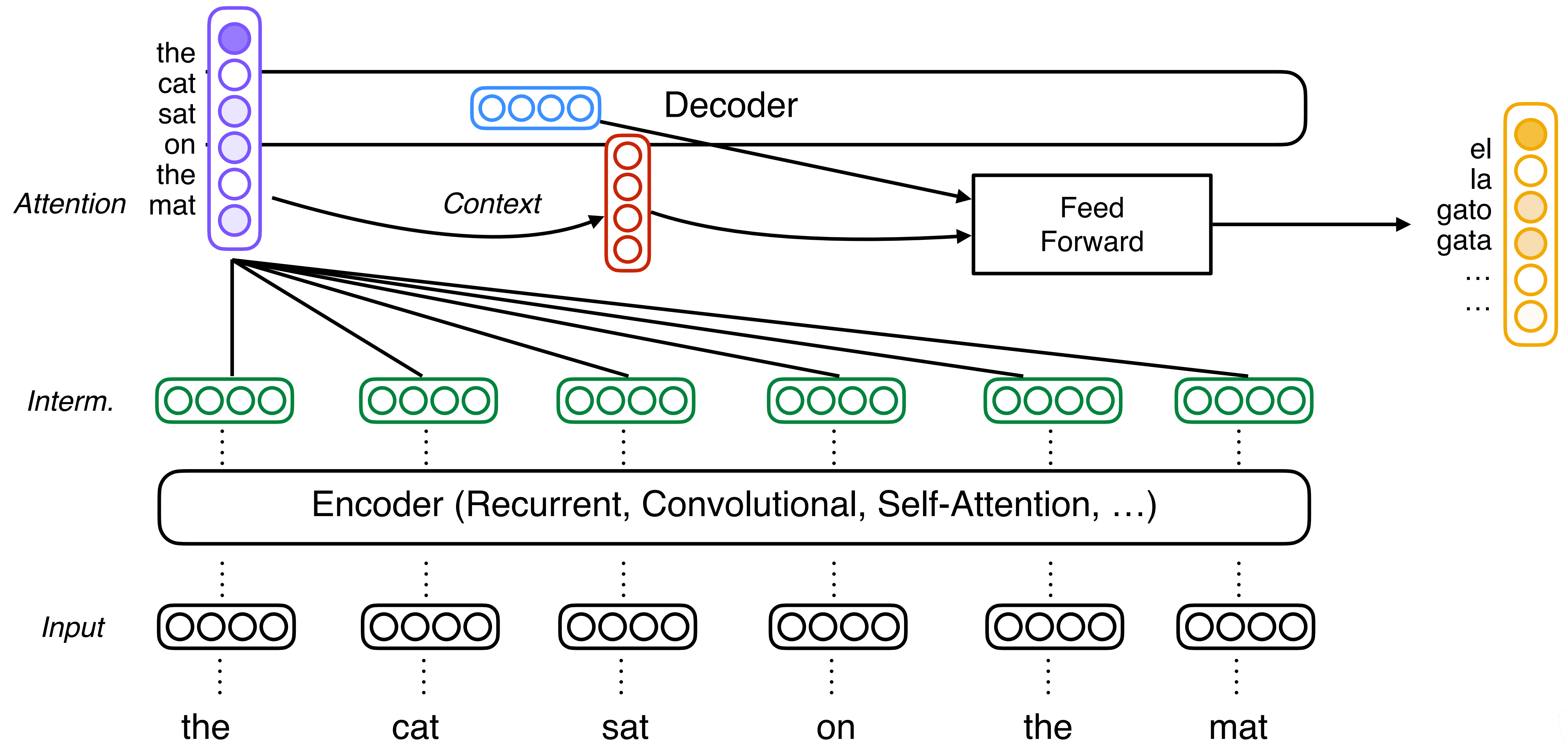












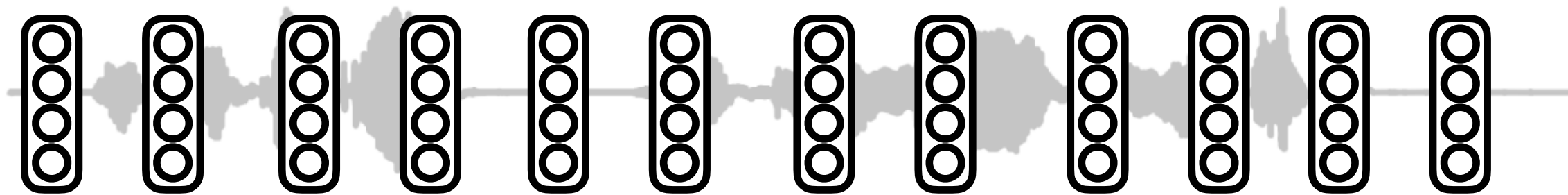
An Audio-Input model

el gato se sentó en la afobra



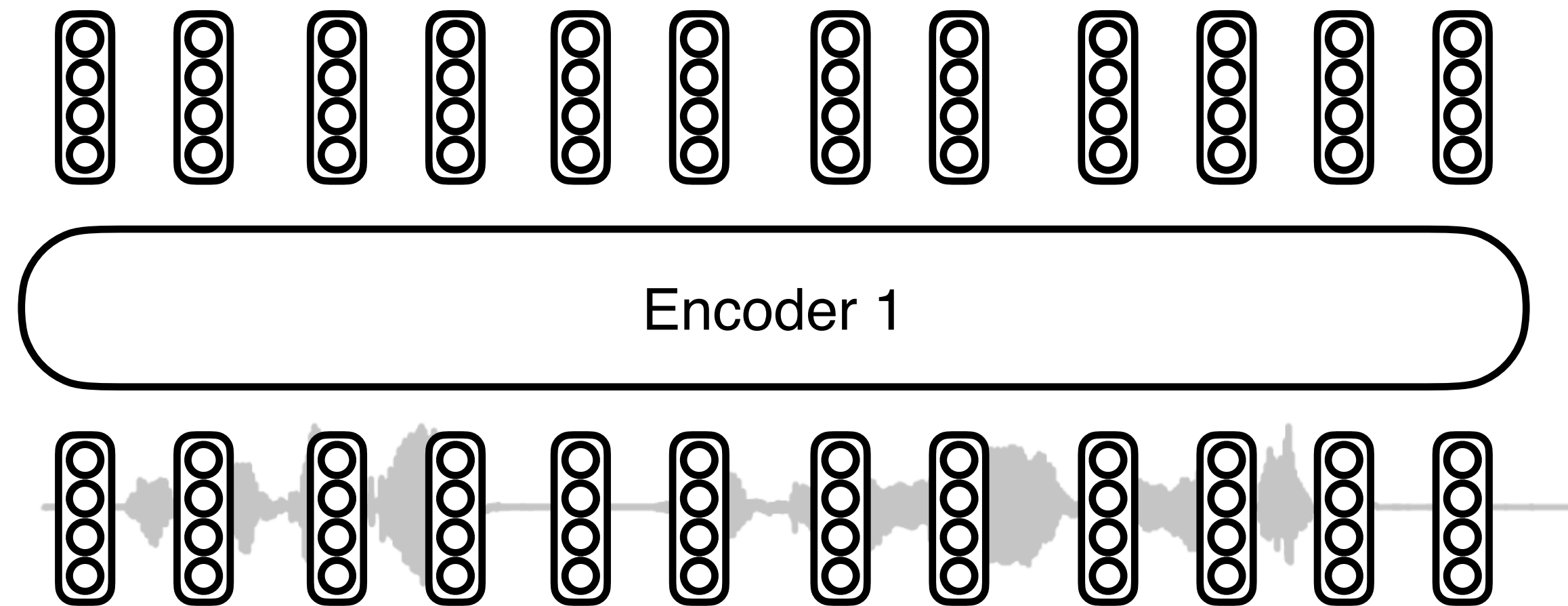
An Audio-Input model

el gato se sentó en la afobra



An Audio-Input model

el gato se sentó en la afobra

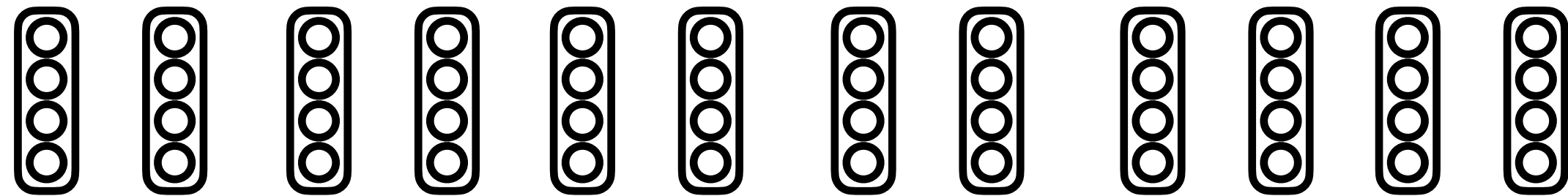


An Audio-Input model

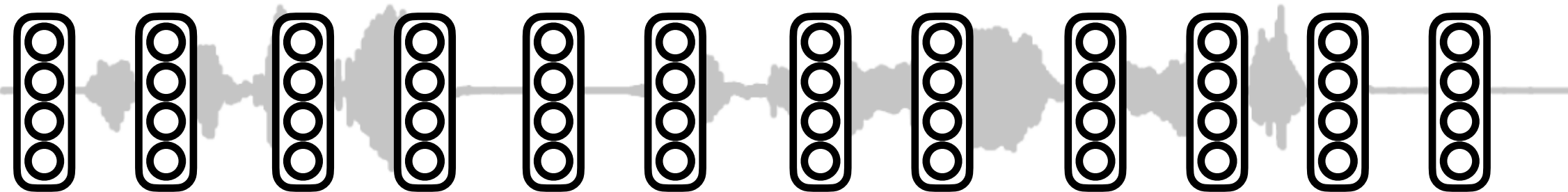
el gato se sentó en la afobra



Decoder

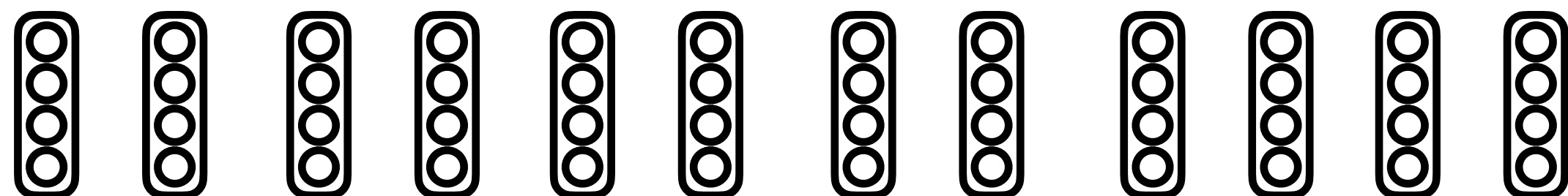


Encoder 1

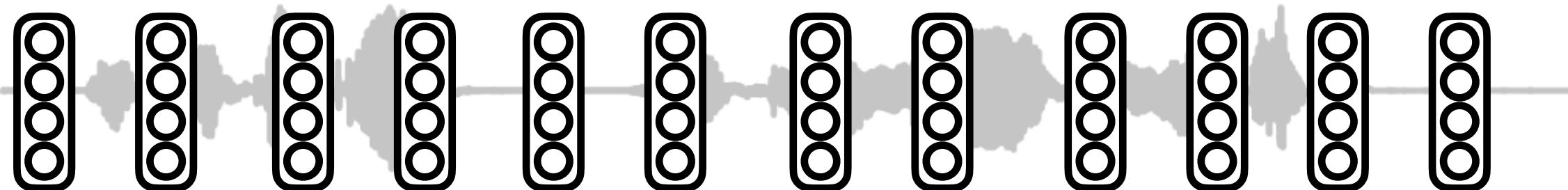


el gato se sentó en la afobra

Decoder



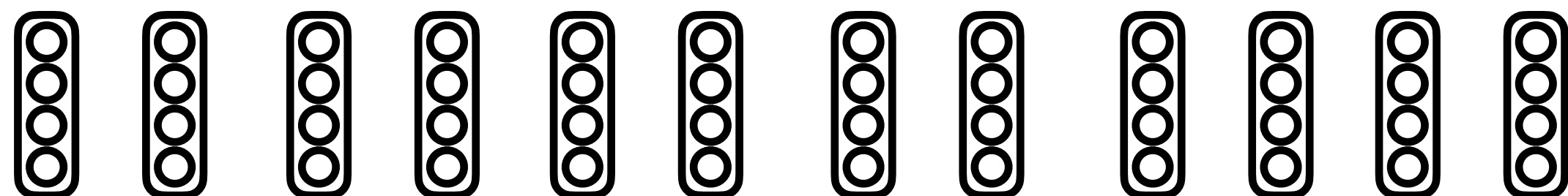
Encoder 1



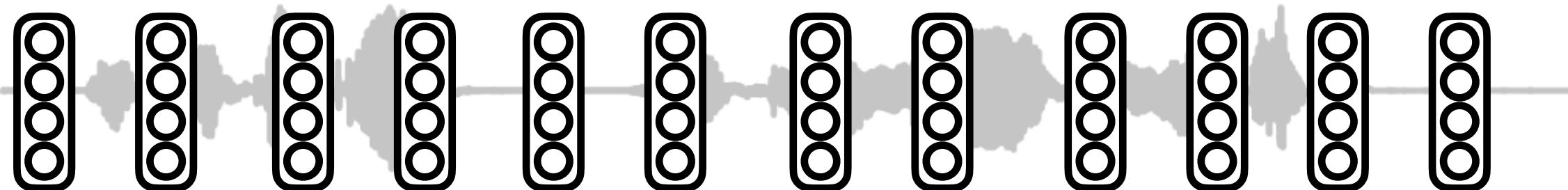
el gato se sentó en la afobra

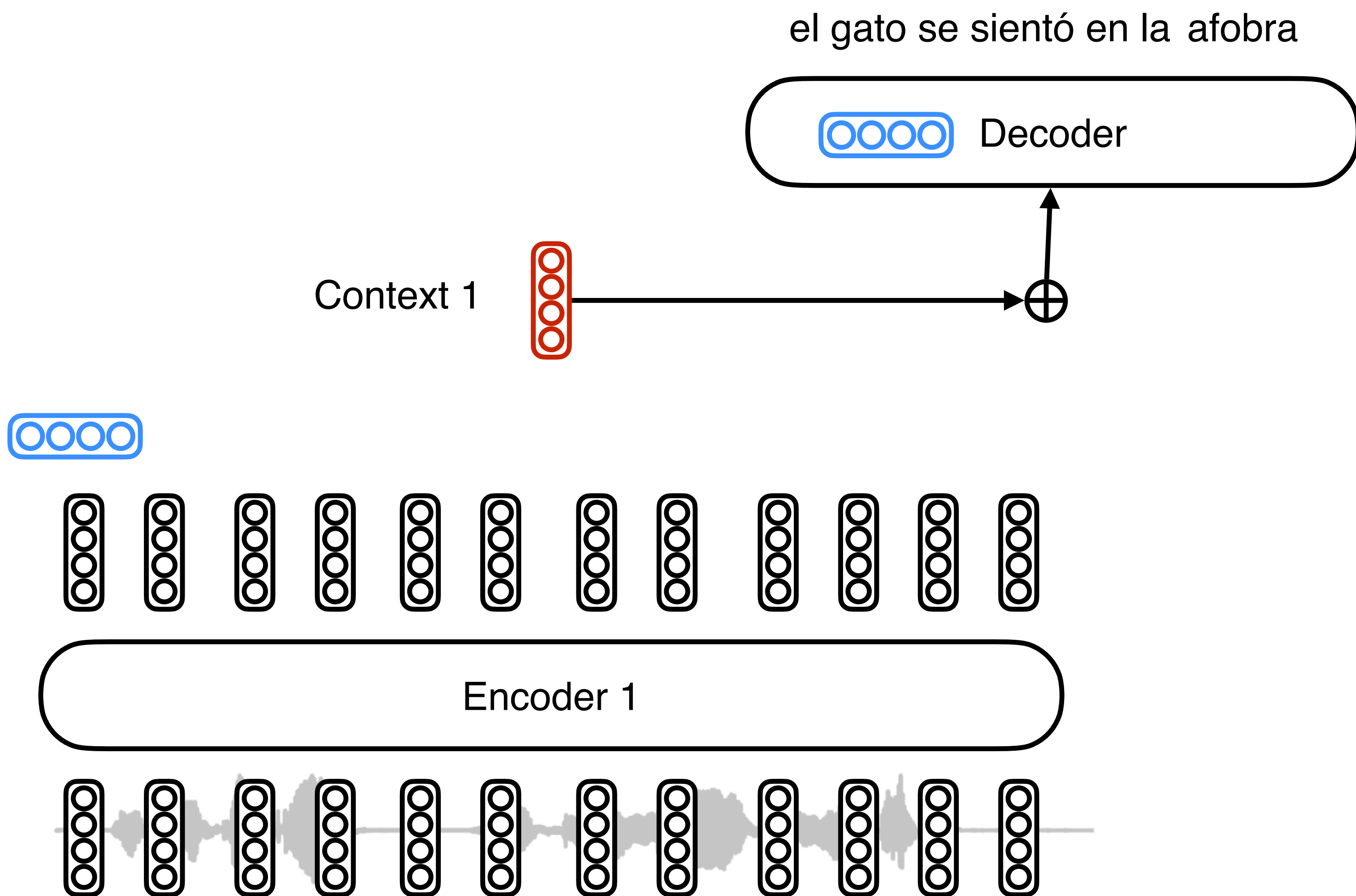
 Decoder





Encoder 1

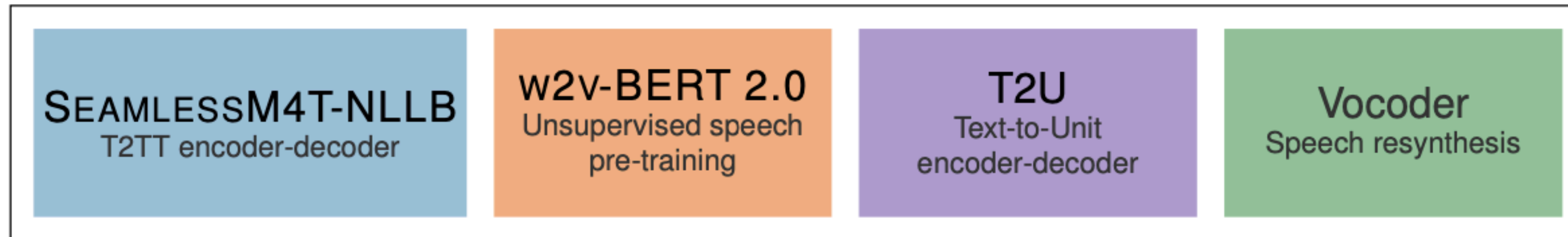




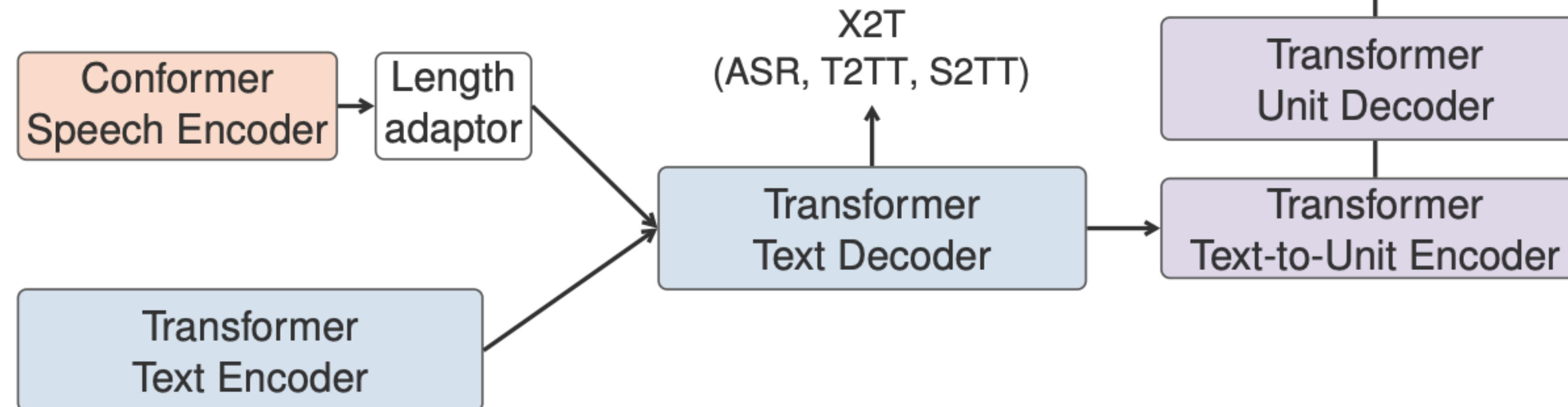
Today: pre-training

The SeamlessM4T model

(1) Pre-trained models

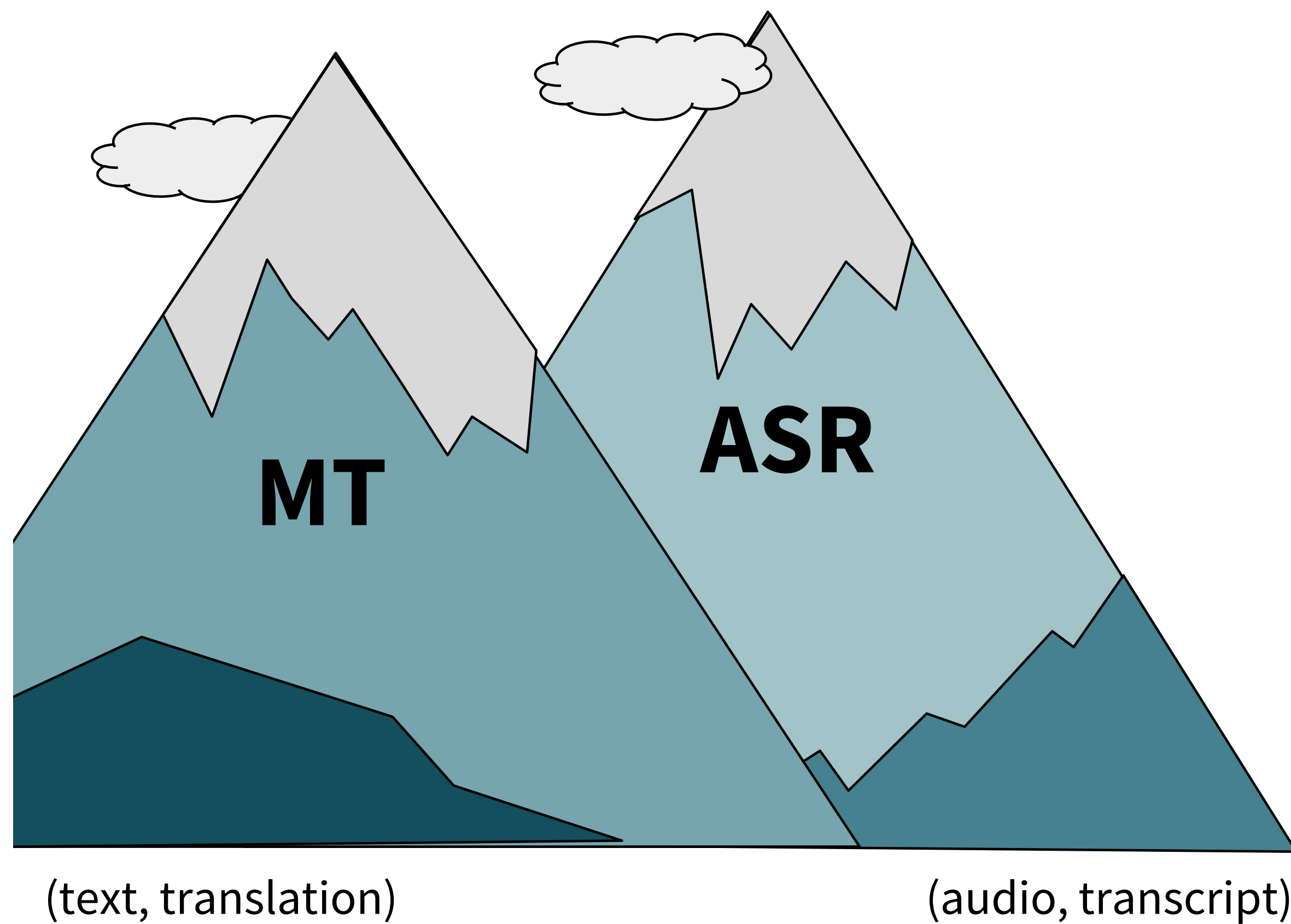


(2) Multitasking UNITY

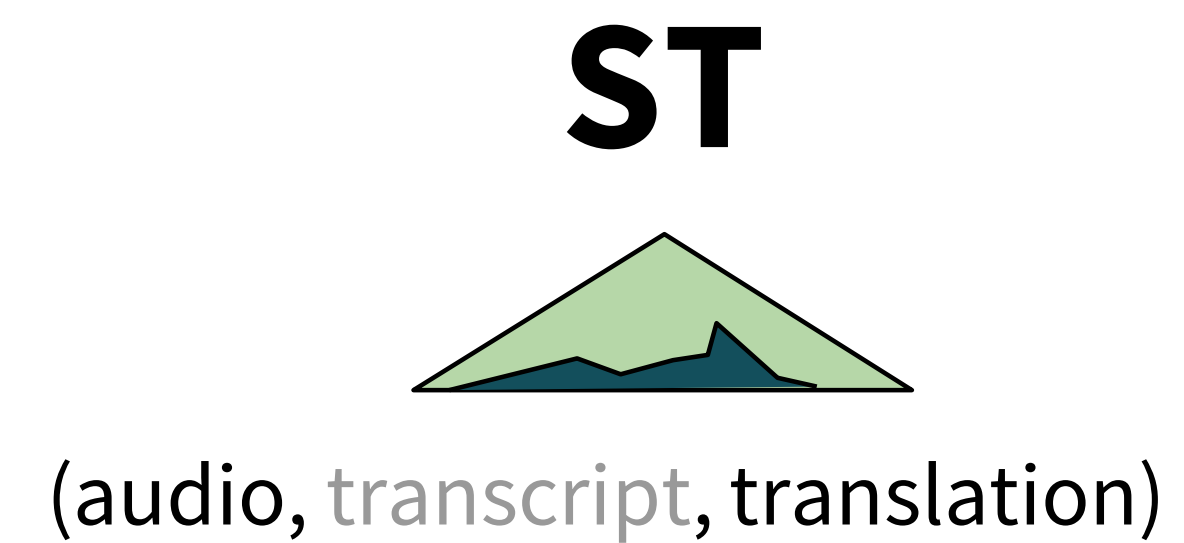


Today: data mining

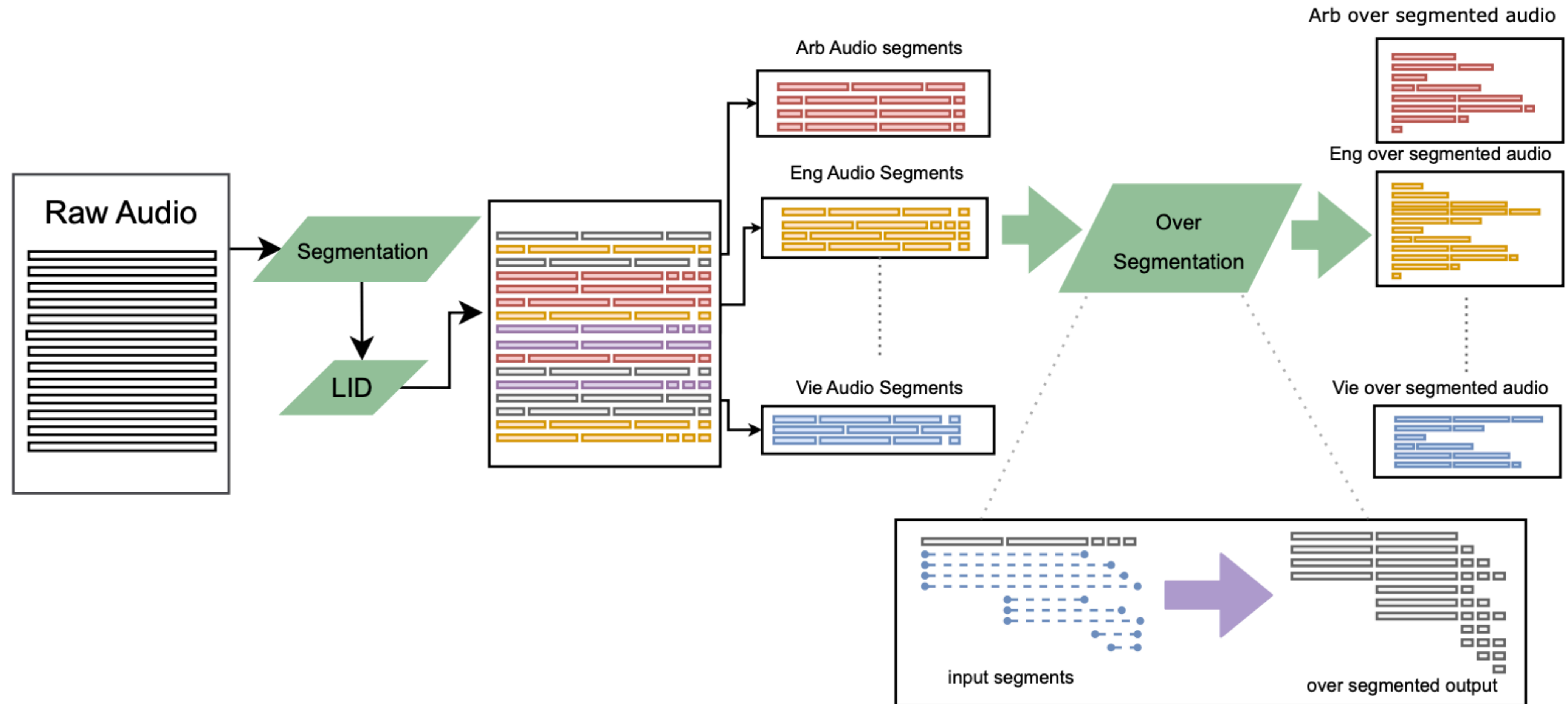
Recap: Available data



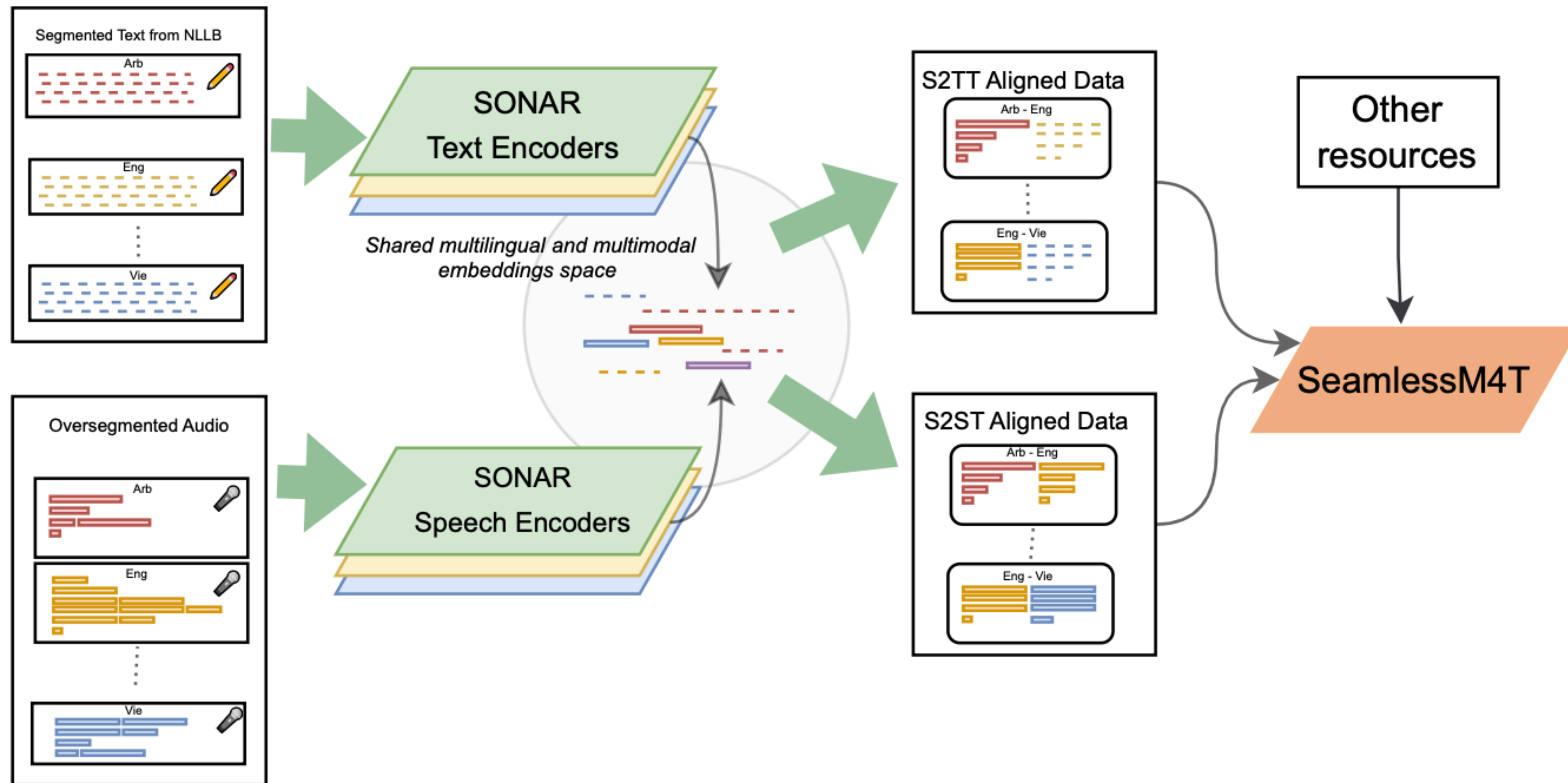
Can we make use of this large amount of data?



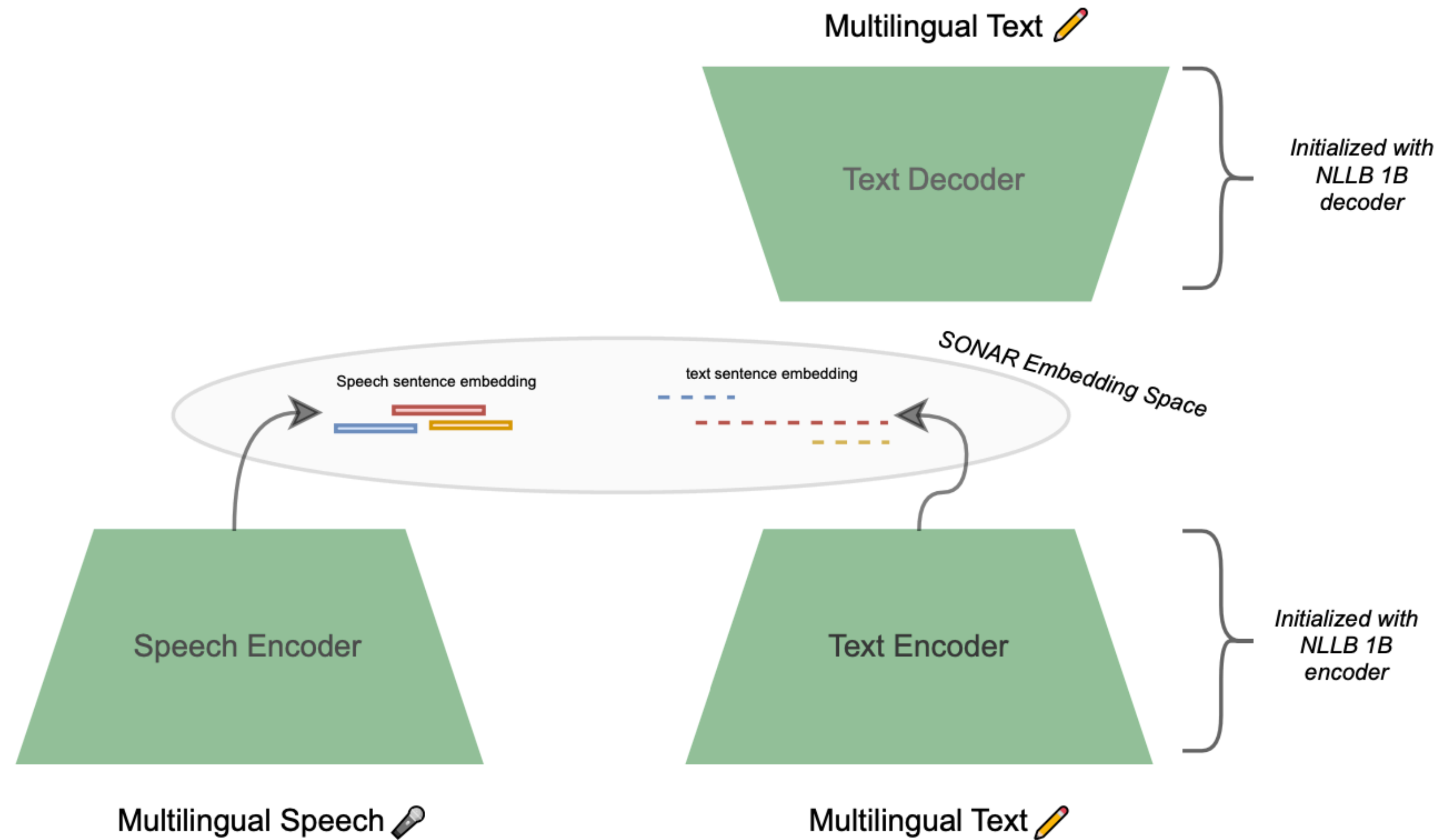
Mining Parallel Speech Data



SONAR Representations



SONAR Representations



SeamlessM4T Results



SeamlessM4T Results

tl;dr: it's great