GEORGE MASON UNIVERSITY

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

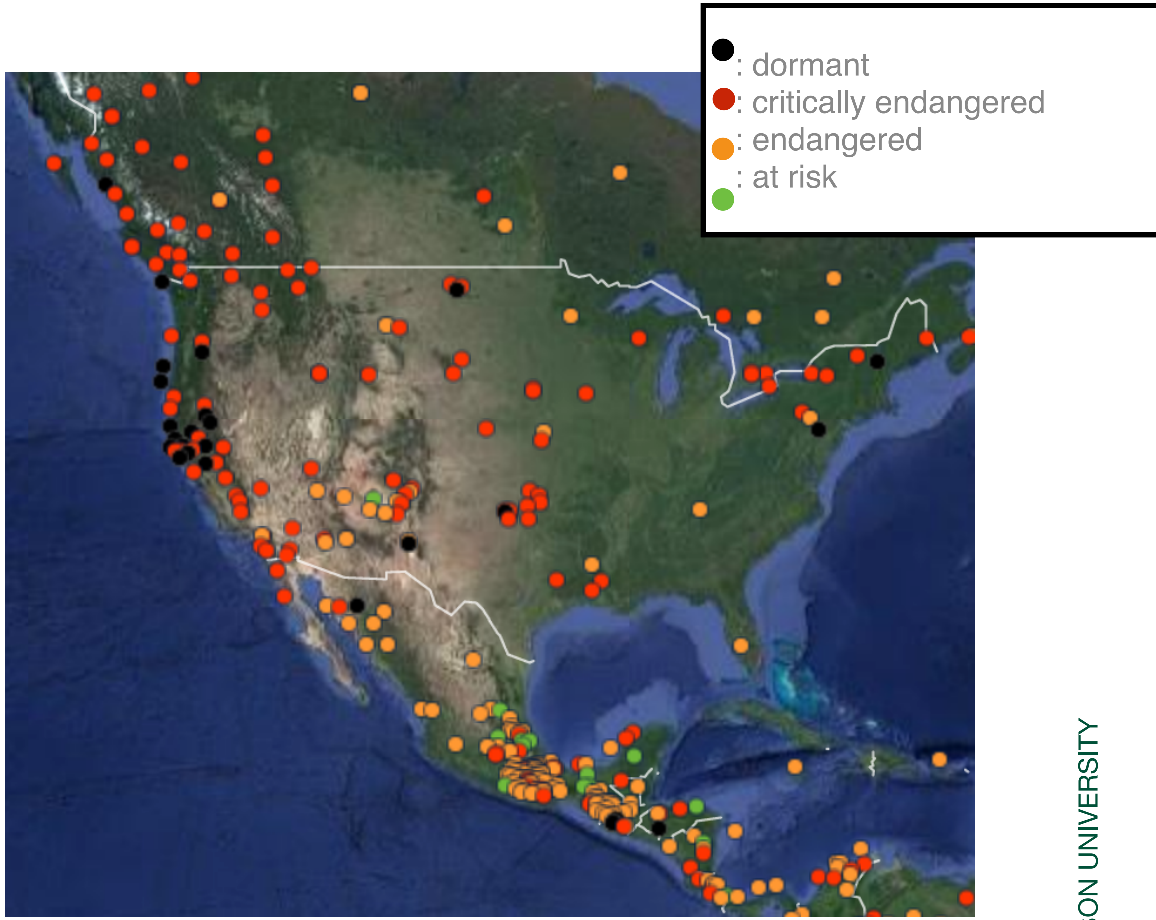
- More than **6000** languages:
  - → 45% oral

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

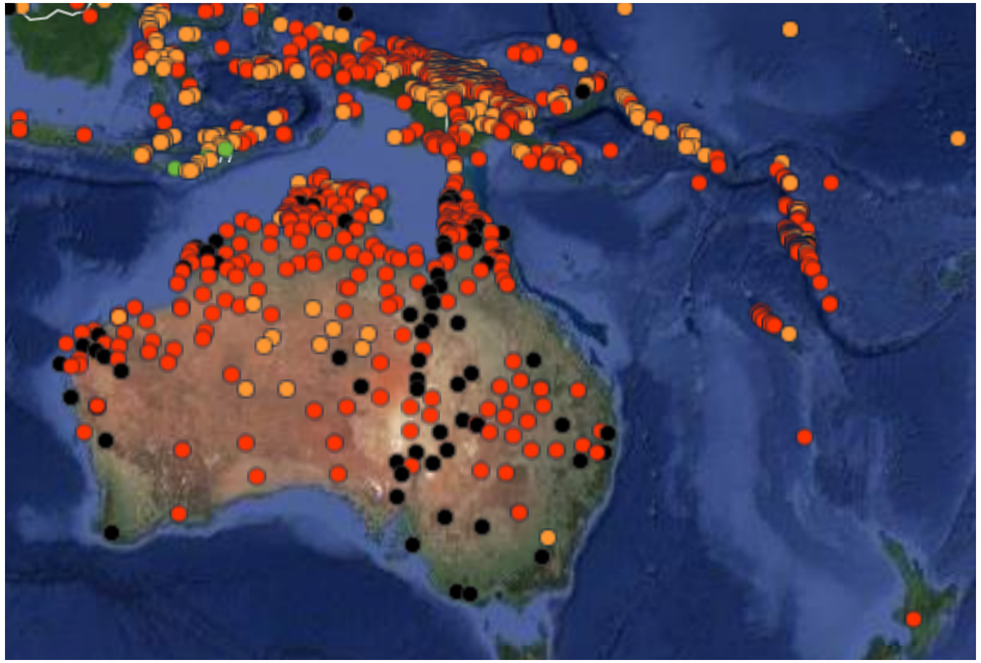


Source: the Endangered Languages Project

The Languages of the World

: dormant
: critically endangered
: endangered
: at risk

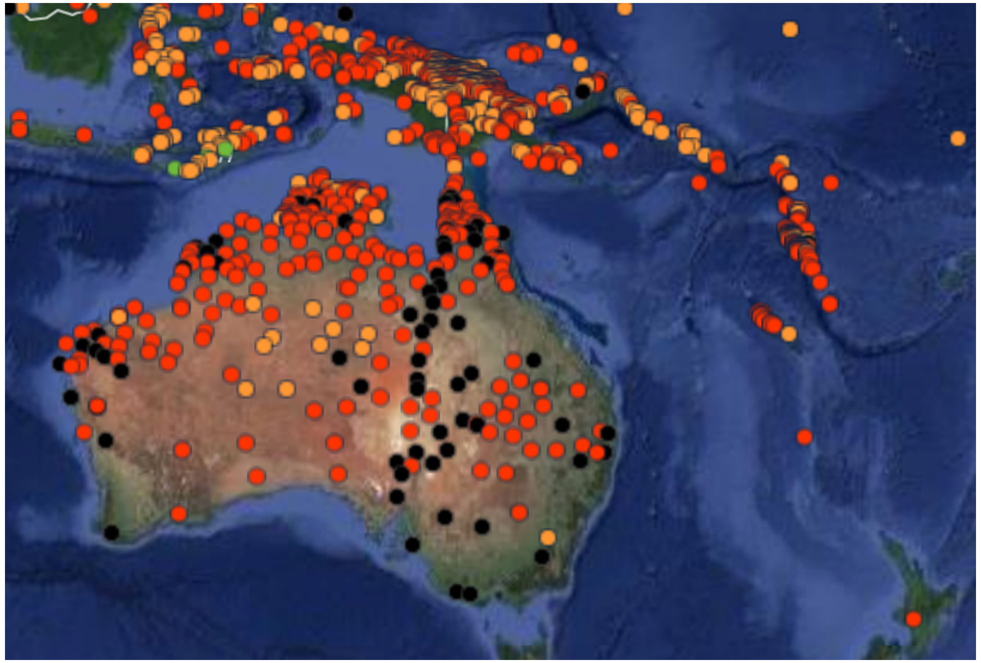
- More than **6000** languages:
  - → 45% oral
  - → 43% endangered or vulnerable
  - → differences in culture, vocabulary



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)

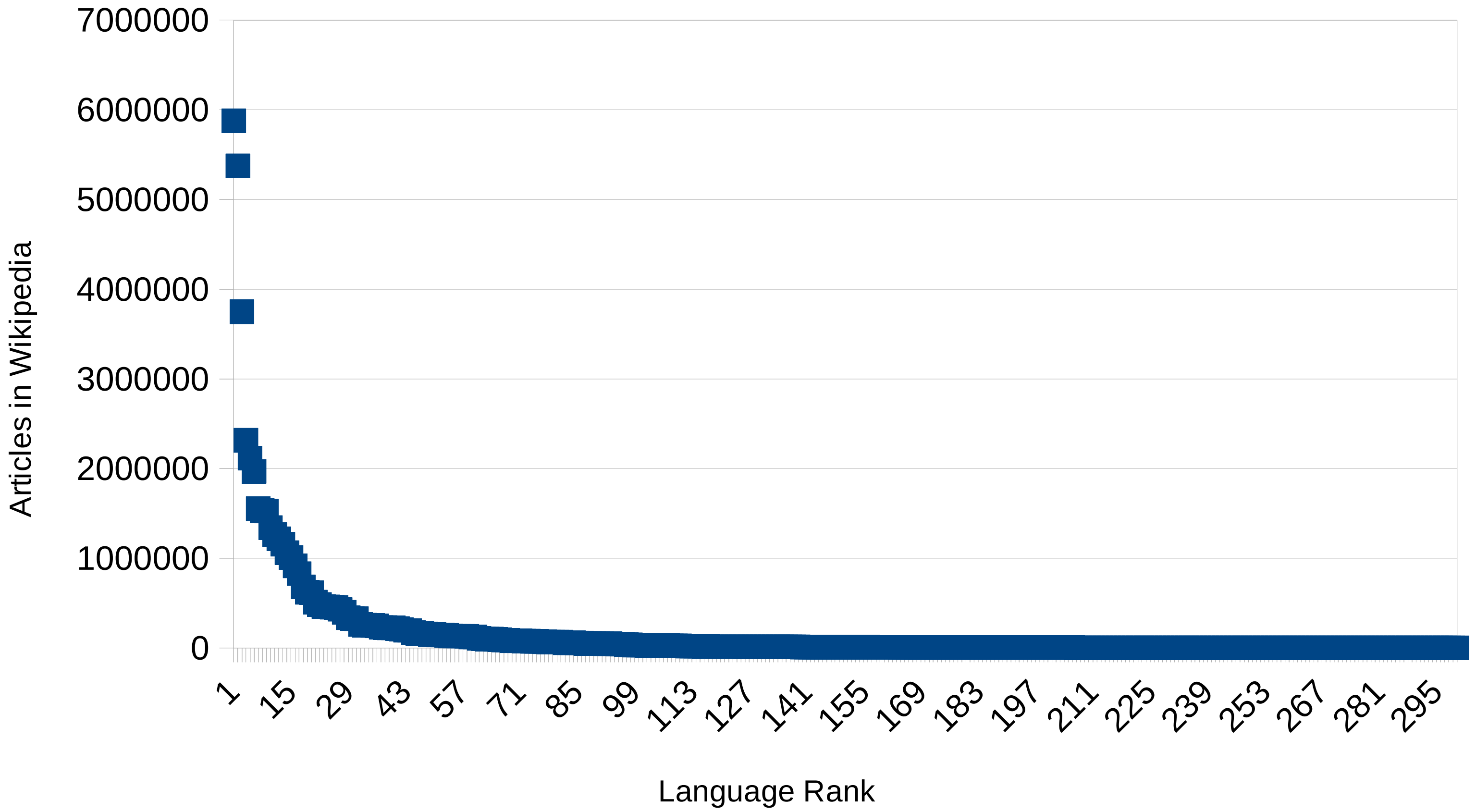
The Languages of the World

- 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

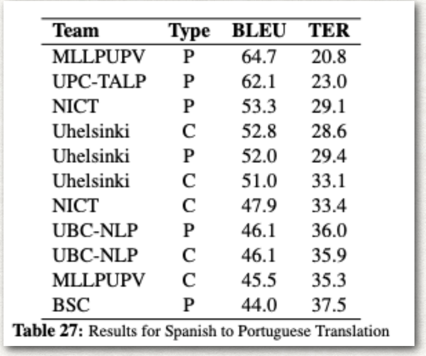
CASE STUDY: TRANSLATION BETWEEN SIMILAR LANGUAGES

Catalan: Què diu aquesta frase?

Spanish: ¿Qué dice esta oración?

Galician: Que di esta frase?

Portuguese: O que esta frase diz?



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?

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? આ વાક્ય શું કહે છે? ಈ ವಾಕ್ಯ ಏನು ಹೇಳುತ್ತದೆ? ਇਹ ਸਜ਼ਾ ਕੀ ਕਹਿੰਦੀ ਹੈ?

ഈ വാചകം എന്താണ് പറയുന്നത്? यह वाक्य क्या कहता है? हे वाक्य काय म्हणते?

ఈ వాక్యం ఏమి చెబుతుంది? यो वाक्यले के भन्छ? මෙම වාක්‍යය පවසන්නේ කුමක්ද?

- 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?
這句话是什么意思?

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?
這句話是什麼意思?

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?
這句话是什么意思?

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	chн
池 pool	水(氵) water	也	chн
施 impose	方 direction	也	sh
弛 loosen	弓 bow	也	chн
地 land	土 soil	也	dм
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 Transl
through ASCII Encoding,
Nikolov et al., 2019.

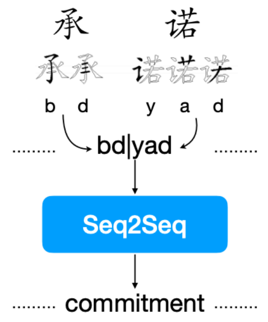


Figure 1: Overview of the **wubi2en** approach to Chinese-to-English translation. A raw Chinese word ('承诺') is encoded into ASCII characters ('bd|yad'), using the Wubi encoding method, before passing it to a Seq2Seq network. The network generates the English translation 'commitment', processing one ASCII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

CASE STUDY: ARABIC

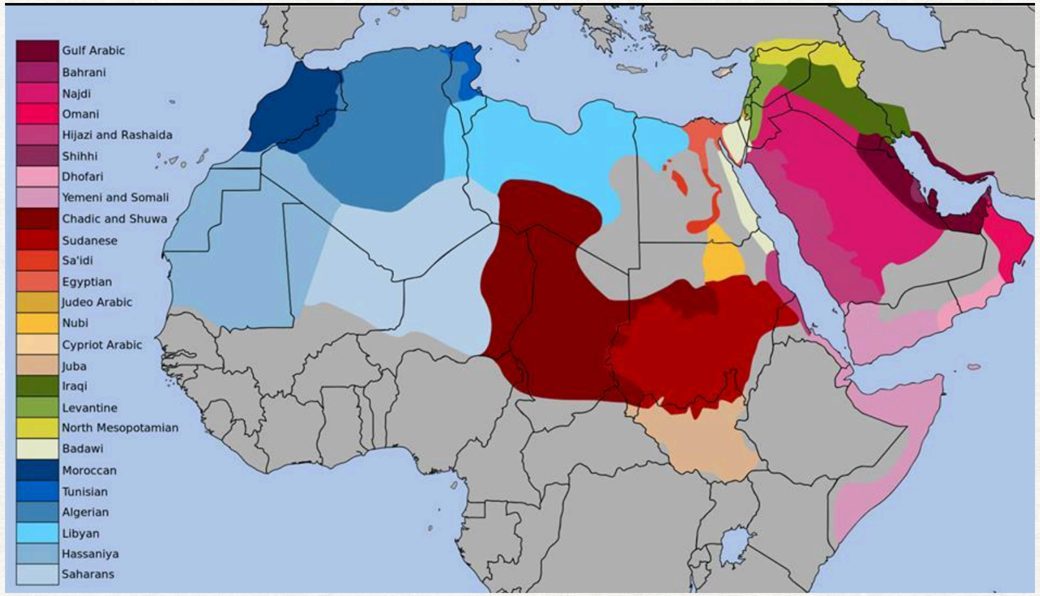
what does this sentence mean?   ماذا تعني هذه الجمله؟

CASE STUDY: ARABIC

what does this sentence mean? ماذا تعني هذه الجمله؟

CASE STUDY: ARABIC

what does this sentence mean? ماذا تعني هذه الجمله؟



CASE STUDY: ARABIC

what does this sentence mean? ماذا تعني هذه الجمله؟

Issue: Root-and-Pattern morphology

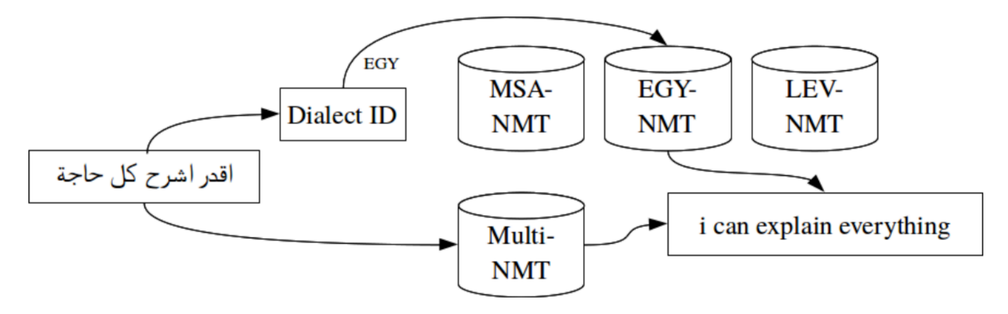
Solution: Morphological Analysis and Disambiguation

Input	wsynhY	Alr}ys	jwlth	bzyArp	AlY	trkyA.	
Gloss	and will finish	the president	tour his	with visit	to	Turkey	.
English	The president will finish his tour with a visit to Turkey.						
ST	wsynhY	Alr}ys	jwlth	bzyArp	AlY	trkyA	.
D1	w+ synhy	Alr}ys	jwlth	bzyArp	<lY	trkyA	.
D2	w+ s+ ynhy	Alr}ys	jwlth	b+ zyArp	<lY	trkyA	.
D3	w+ s+ ynhy	Al+ r}ys	jwlp +P$_{3MS}$	b+ zyArp	<lY	trkyA	.
MR	w+ s+ y+ nhy	Al+ r}ys	jwl +p +h	b+ zyAr +p	<lY	trkyA	.
EN	w+ s+ >nhY$_{VBP}$ +S$_{3MS}$	Al+ r}ys$_{NN}$	jwlp$_{NN}$ +P$_{3MS}$	b+ zyArp$_{NN}$	<lY$_{IN}$	trkyA$_{NNP}$	.

CASE STUDY: ARABIC

what does this sentence mean? ماذا تعني هذه الجمله؟

Handling dialectal data:



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?

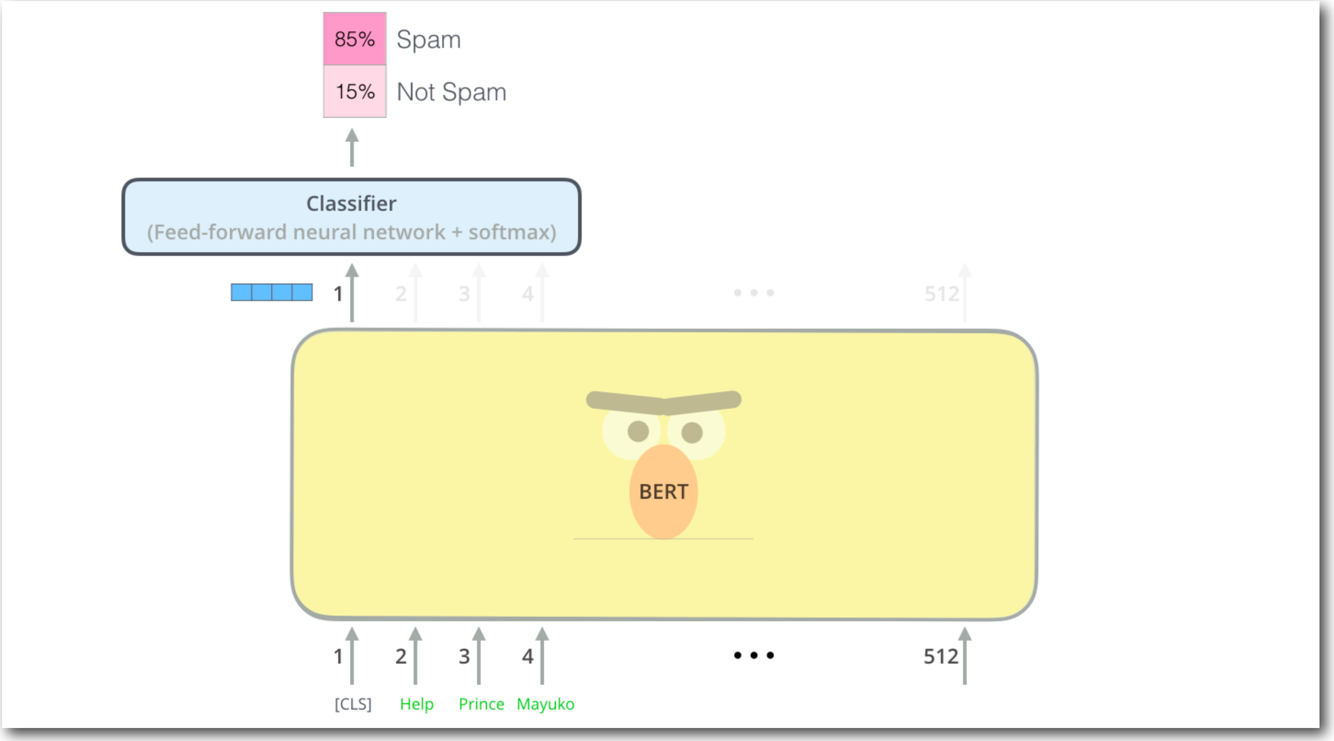
1. Intuition (maaaayyybe ok)
2. Geography (could be misleading)

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



Choose our own adventure

~~Chinchila~~

~~PaLM~~

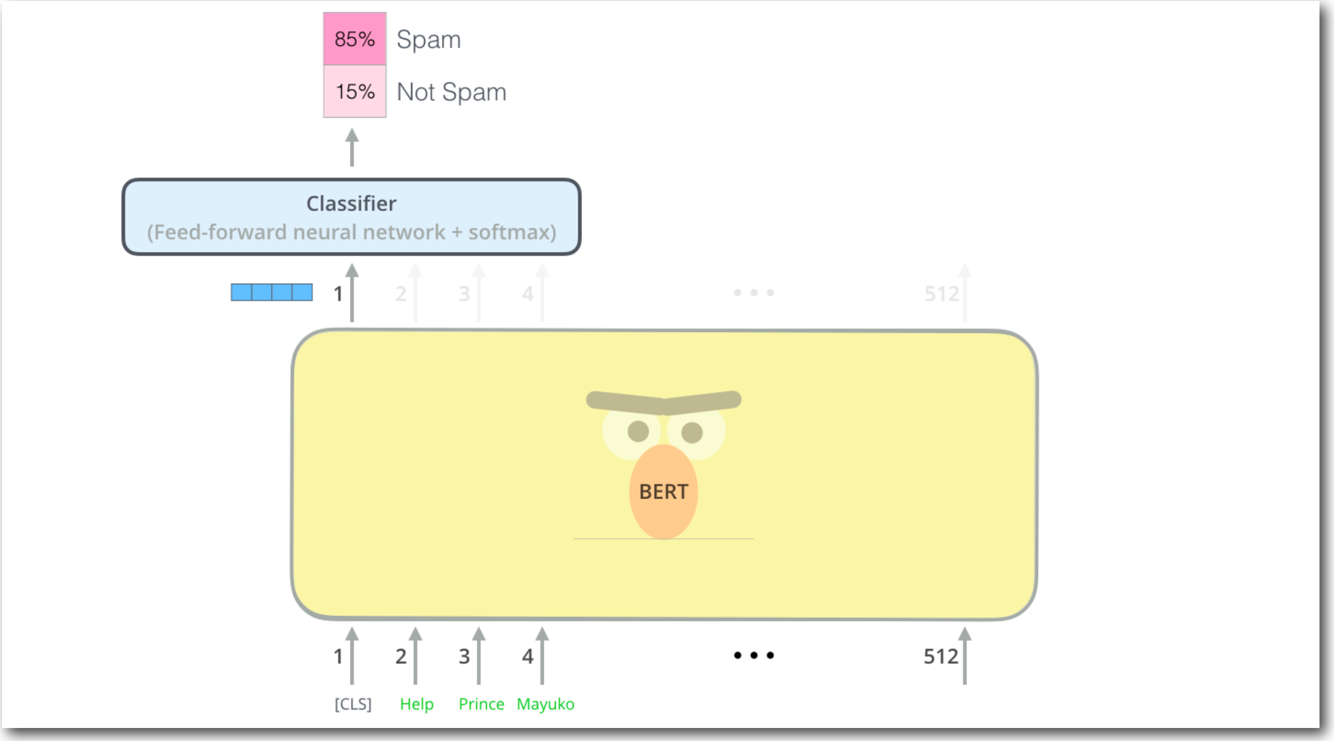
~~GPT-2~~

~~ELECTRA~~

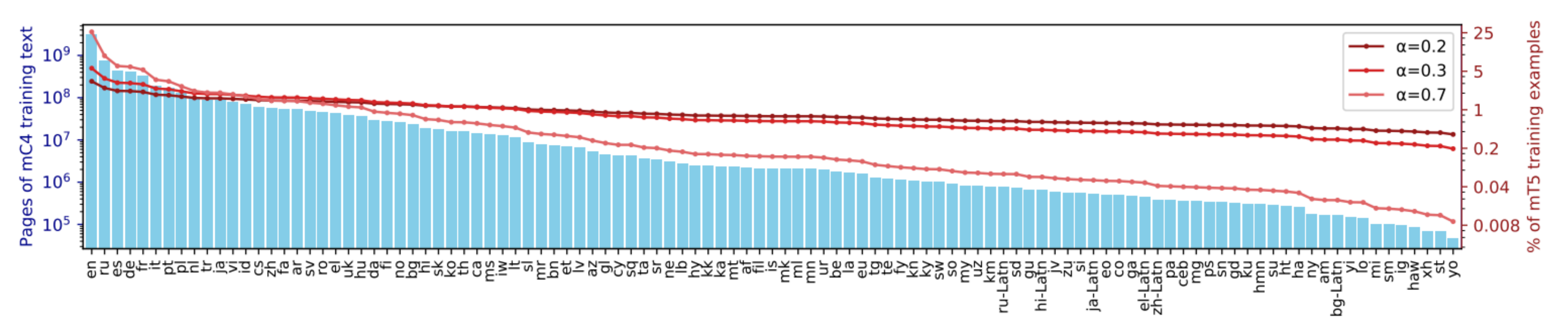
~~XLM-R~~

~~RoBERTa~~

Some recent trends



Make it multilingual!



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

NLP beyond the top-100 languages

Going Beyond the top-100 Languages

GEORGE MASON UNIVERSITY

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***

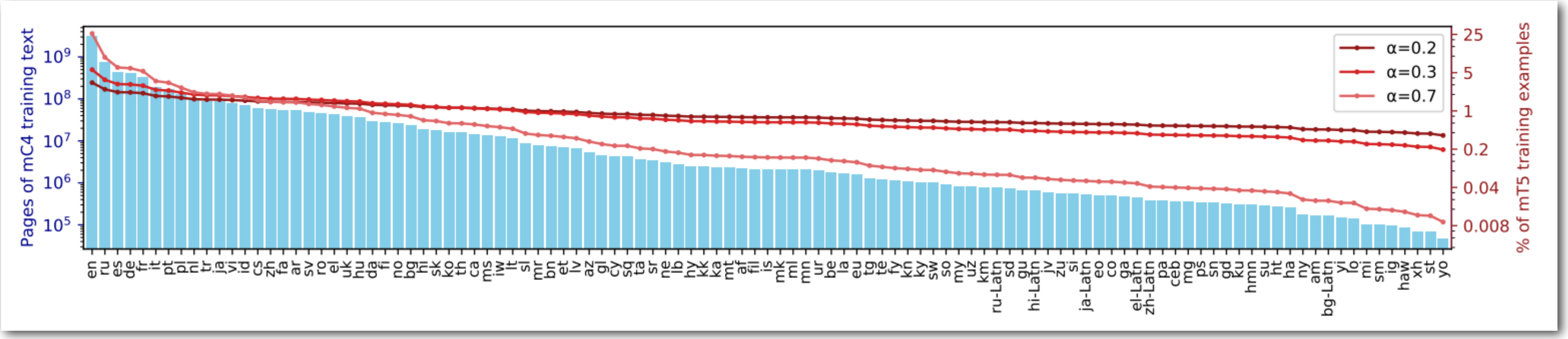
or

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



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

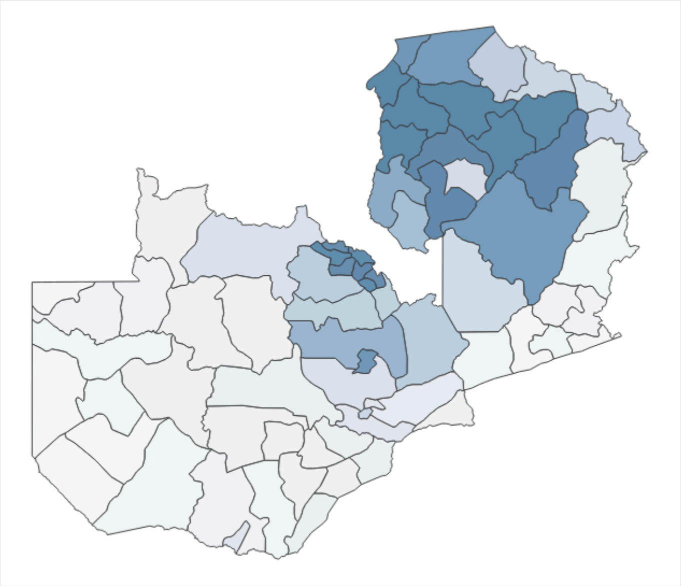
Our Solution: Work with Communities

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%

GEORGE MASON UNIVERSITY

Our Solution: Work with Communities

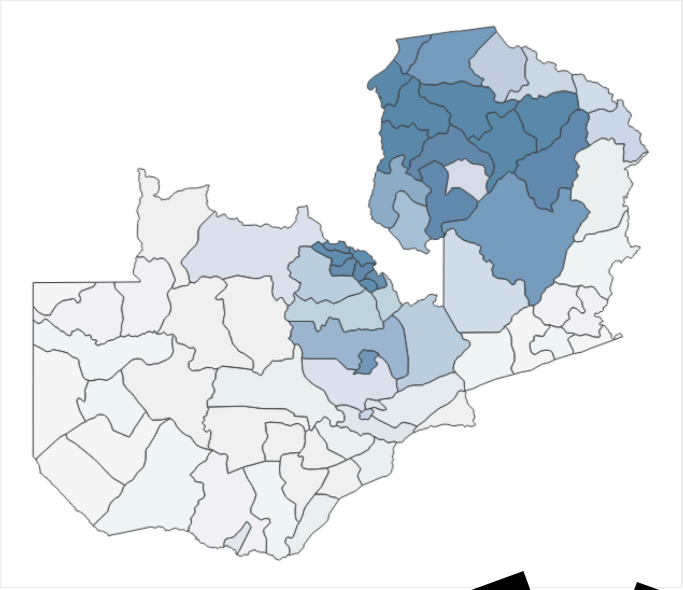
Language map of Zambia

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

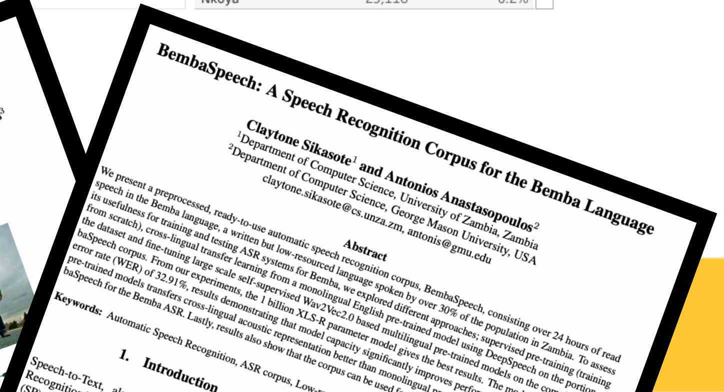


Language	# of speakers	% of population
Bemba	3,727,677	28.9%
Tonga	1,585,877	12.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanj..	1,305,434	10.1%
Lozi	741,755	5.7%
Lunda	520,643	4.0%
Other langua..	464,474	3.6%
Luvale	416,725	3.2%
Lala-Bisa	379,548	2.9%
Nyamwanga	299,337	2.3%
Nsenga	292,814	2.3%
Mambwe-Lun..	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 Sikasote[1], Eunice Mukonde[2], Md Mahfuz Ibn Alam[3], Antonios Anastasopoulos[3]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department of Literature and Languages, University of Zambia, Zambia
[3]Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract



BembaSpeech: A Speech Recognition Corpus for the Bemba Language

Claytone Sikasote[1] and Antonios Anastasopoulos[2]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department 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 complete BembaSpeech corpus. From our experiments, the 1 billion XLS-R parameter model gives the best results. The model achieves a word error rate (WER) of 32.91%, results demonstrating that model capacity significantly improves performance and that multilingual pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained ... BembaSpeech for the Bemba ASR. Lastly, results also show that the corpus can be used ...

Keywords: Automatic Speech Recognition, ASR corpus, Low-

1. Introduction

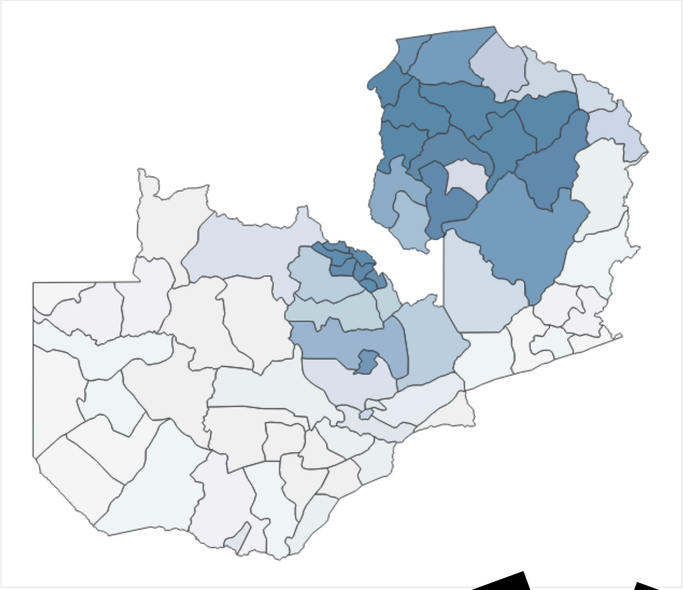
GEORGE MASON UNIVERSITY

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

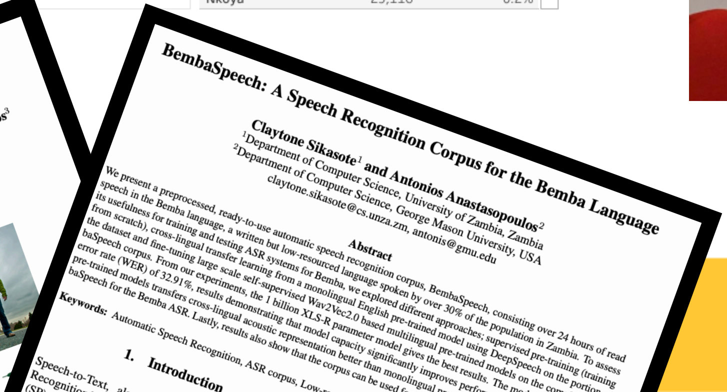


Language	# of speakers	% of population
Bemba	3,727,677	28.9%
Tonga	1,585,877	12.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanj..	1,305,434	10.1%
Lozi	741,755	5.7%
Lunda	520,643	4.0%
Other langua..	464,474	3.6%
Luvale	416,725	3.2%
Lala-Bisa	379,548	2.9%
Nyamwanga	299,337	2.3%
Nsenga	292,814	2.3%
Mambwe-Lun..	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 Sikasote[1], Eunice Mukonde[2], Md Mahfuz Ibn Alam[3], Antonios Anastasopoulos[3]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department of Literature and Languages, University of Zambia, Zambia
[3]Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract



BembaSpeech: A Speech Recognition Corpus for the Bemba Language

Claytone Sikasote[1] and Antonios Anastasopoulos[2]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department 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 complete BembaSpeech corpus. From our experiments, the 1 billion XLS-R parameter model gives the best results. The model achieves a word error rate (WER) of 32.91%, results demonstrating that model capacity significantly improves performance and that multilingual pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained English model on the BembaSpeech for the Bemba ASR. Lastly, results also show that the corpus can be used for

Keywords: Automatic Speech Recognition, ASR corpus, Low-

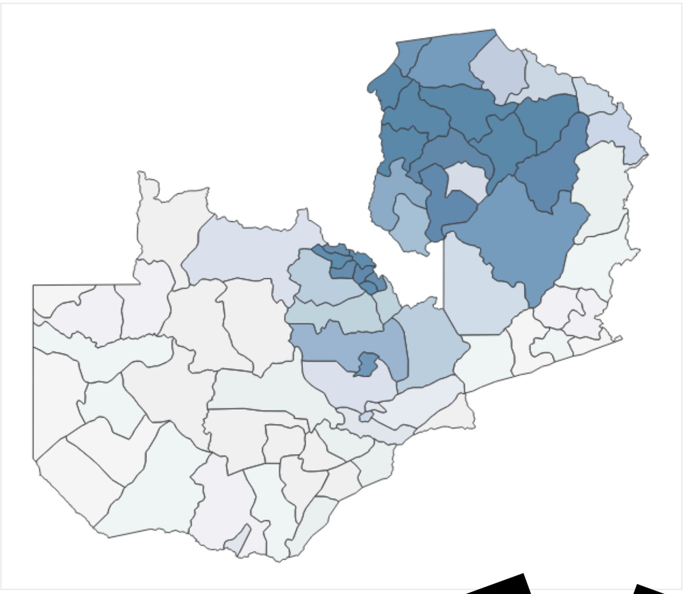
1. Introduction

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

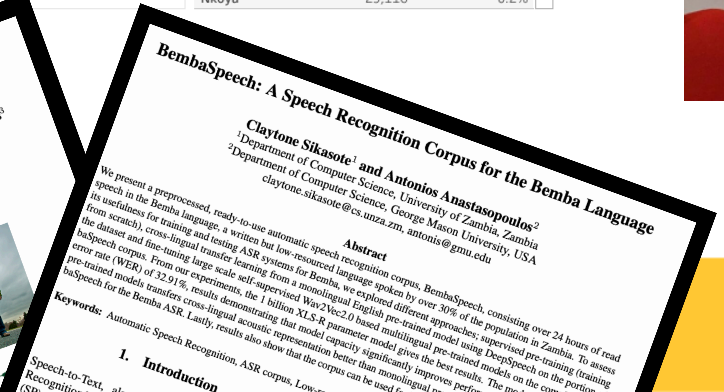


Language	# of speakers	% of population
Bemba	3,727,677	28.9%
Tonga	1,585,877	12.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanj..	1,305,434	10.1%
Lozi	741,755	5.7%
Lunda	520,643	4.0%
Other langua..	464,474	3.6%
Luvale	416,725	3.2%
Lala-Bisa	379,548	2.9%
Nyamwanga	299,337	2.3%
Nsenga	292,814	2.3%
Mambwe-Lun..	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 Sikasote[1], Eunice Mukonde[2], Md Mahfuz Ibn Alam[3], Antonios Anastasopoulos[3]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department of Literature and Languages, University of Zambia, Zambia
[3]Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Abstract


BembaSpeech: A Speech Recognition Corpus for the Bemba Language

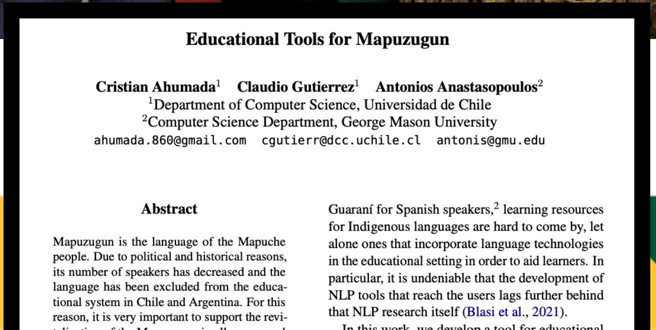
Claytone Sikasote[1] and Antonios Anastasopoulos[2]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu

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 complete BembaSpeech corpus. From our experiments, the 1 billion XLS-R parameter model gives the best results. The model achieves a word error rate (WER) of 32.91%, results demonstrating that model capacity significantly improves performance and that multilingual pre-trained models transfers cross-lingual acoustic representation better than monolingual pre-trained English model on the BembaSpeech for the Bemba ASR. Lastly, results also show that the corpus can be used for ...

Keywords: Automatic Speech Recognition, ASR corpus, Low-...

1. Introduction


Educational Tools for Mapuzugun

Cristian Ahumada[1] Claudio Gutierrez[1] Antonios Anastasopoulos[2]
[1]Department of Computer Science, Universidad de Chile
[2]Computer Science Department, George Mason University
ahumada.860@gmail.com cgutierr@dcc.uchile.cl antonis@gmu.edu

Abstract

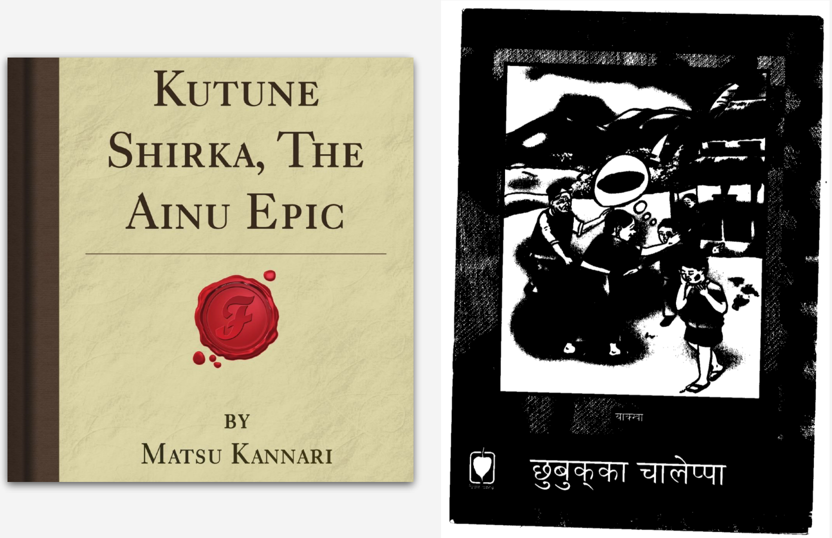
Mapuzugun is the language of the Mapuche people. Due to political and historical reasons, its number of speakers has decreased and the language has been excluded from the educational system in Chile and Argentina. For this reason, it is very important to support the revi-

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).

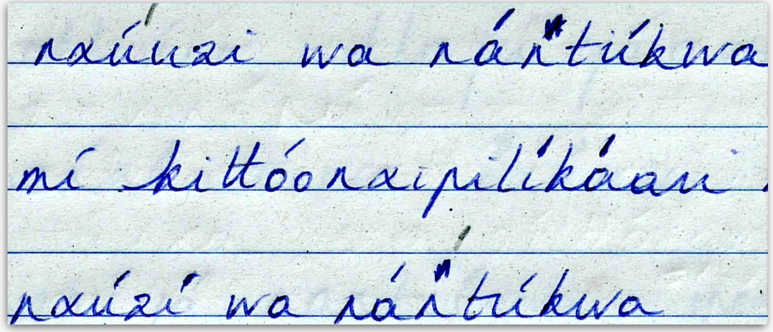
In this work, we develop a tool for educational

GEORGE MASON UNIVERSITY

Our Solution: Make Existing Data ML-Usable



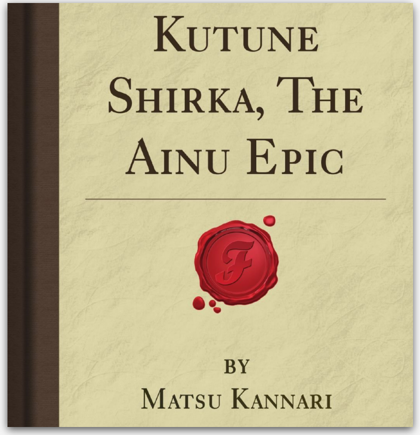
Printed books

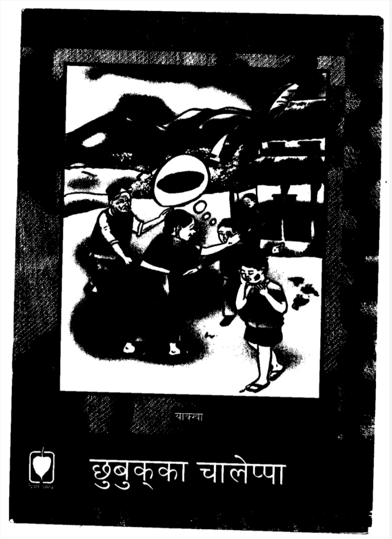


Handwritten notes

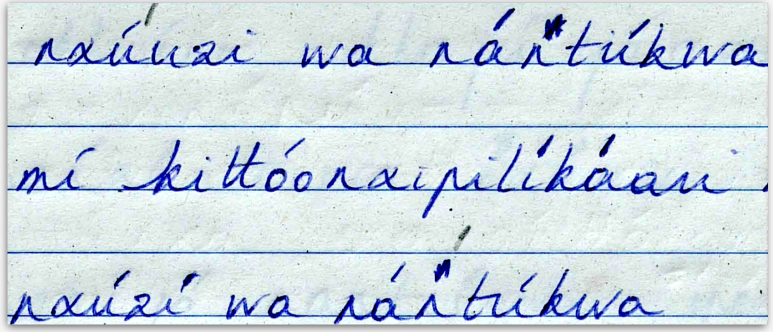
GEORGE MASON UNIVERSITY

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

PRATHAM BOOKS storyweaver

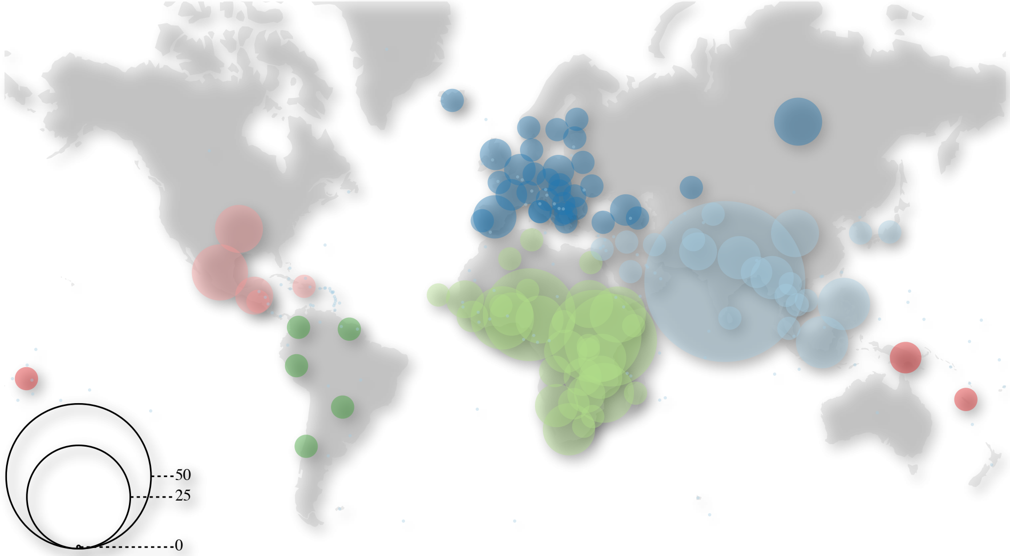
GEORGE MASON UNIVERSITY

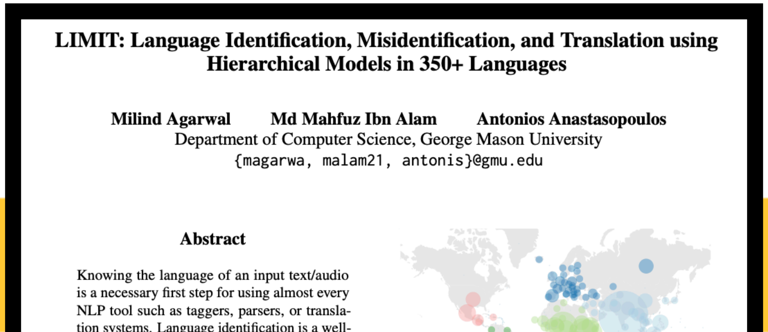
Our Solution: Curation at Scale

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

>350 languages

PRATHAM BOOKS
storyweaver




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-

Language ID at Scale

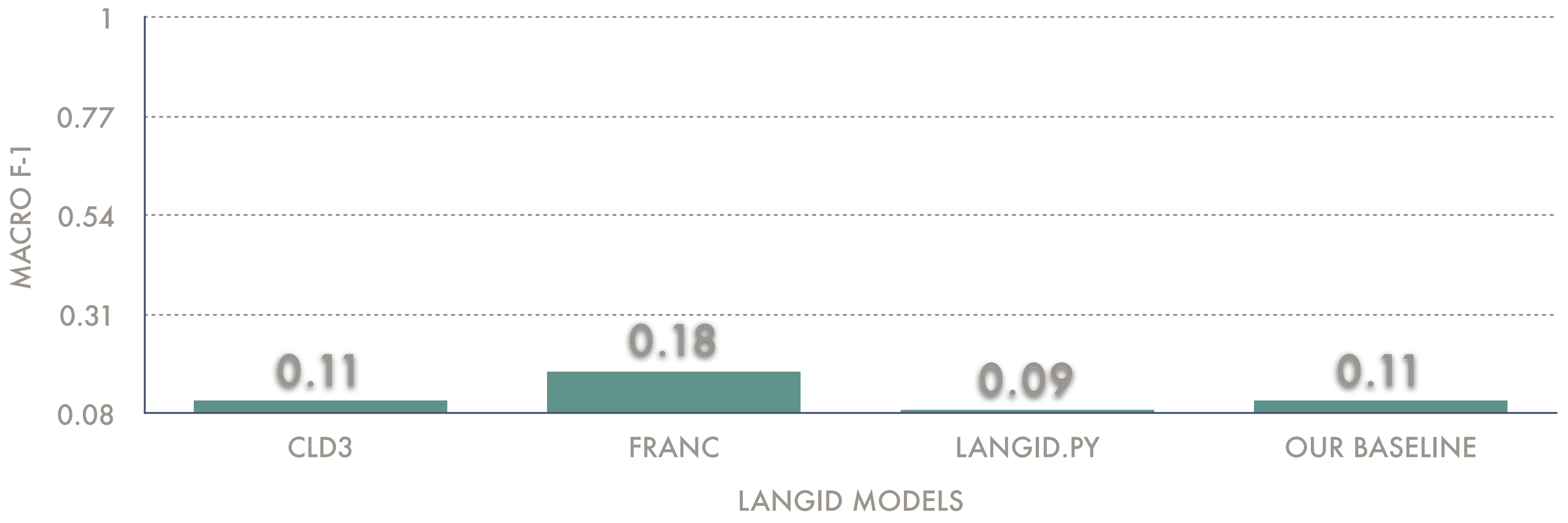
Benchmarking most popular models

Language ID at Scale

Benchmarking most popular models

Language ID at Scale

Benchmarking most popular models



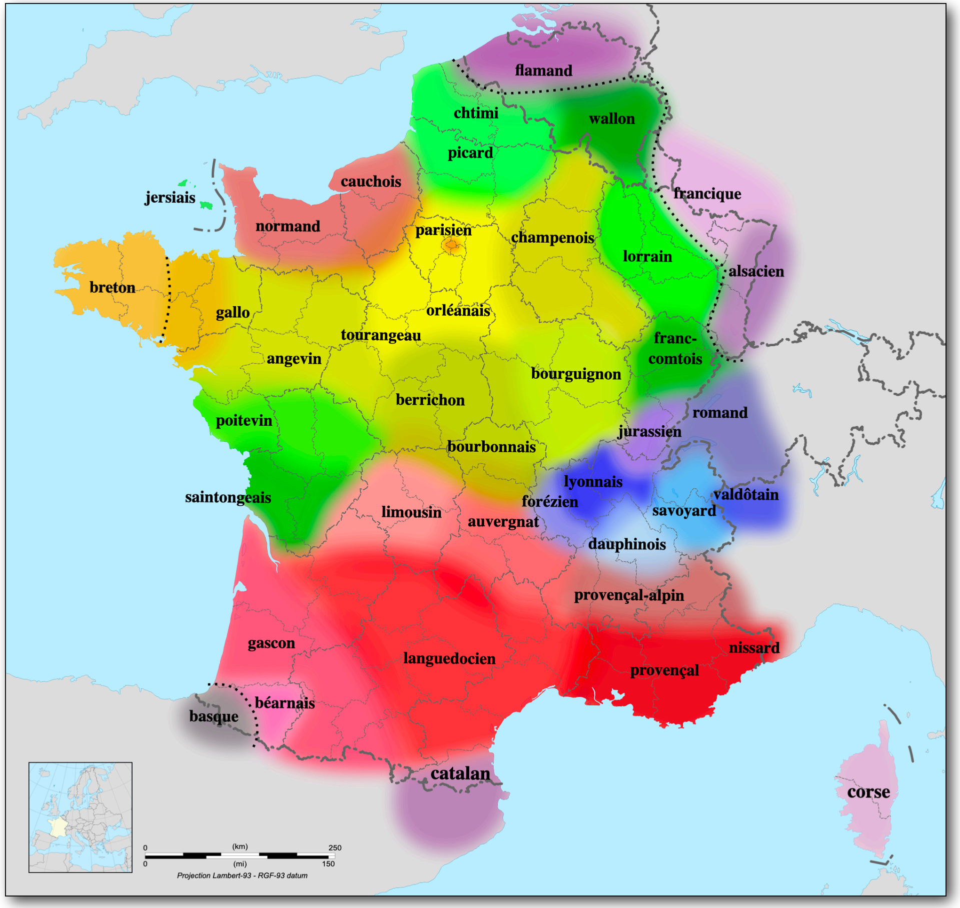
GEORGE MASON UNIVERSITY

Dialects

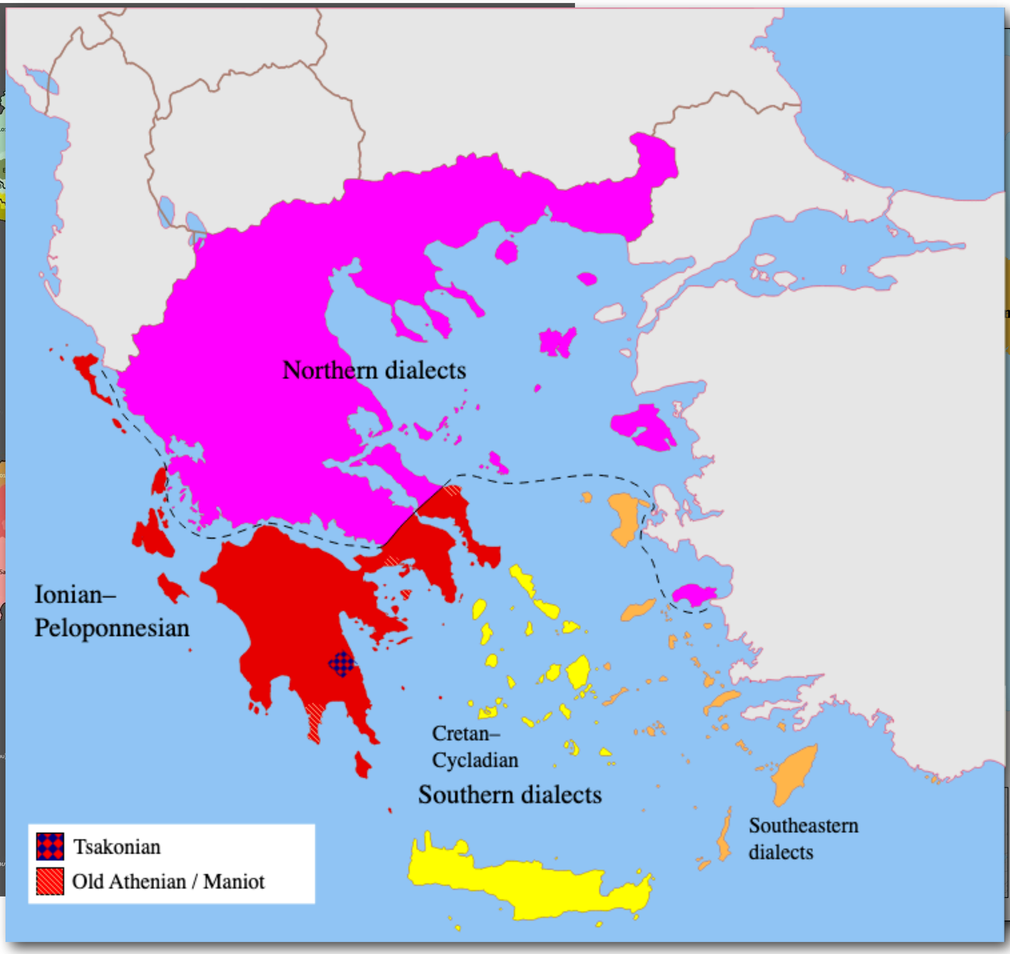
Languages are not Monoliths

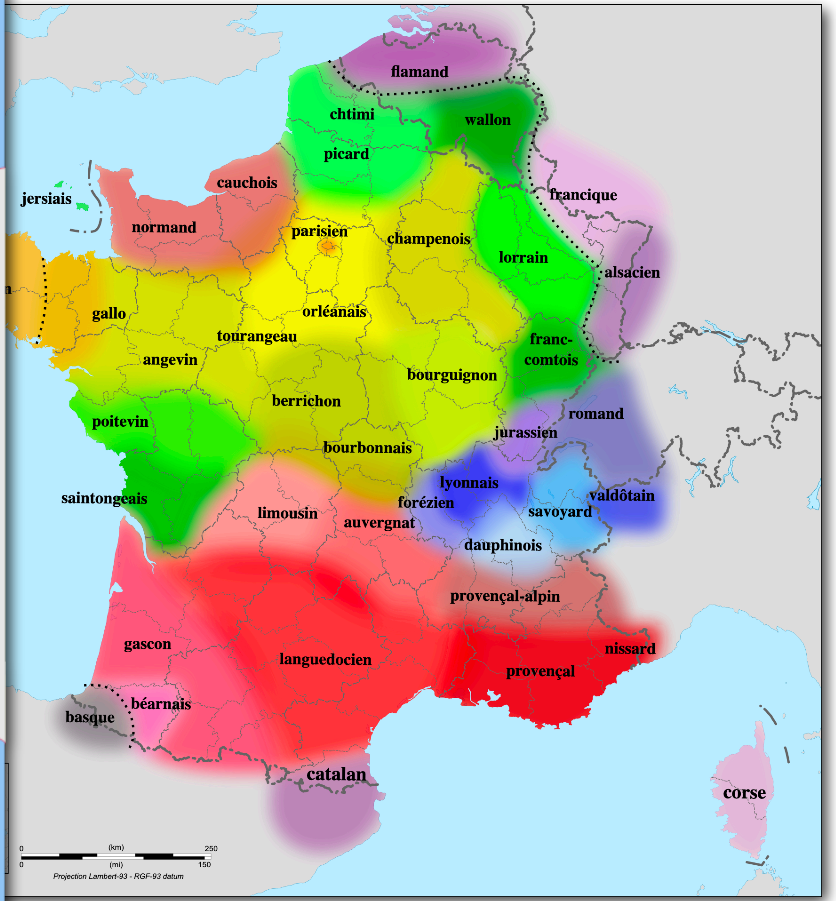
Languages are not Monoliths

Languages are not Monoliths

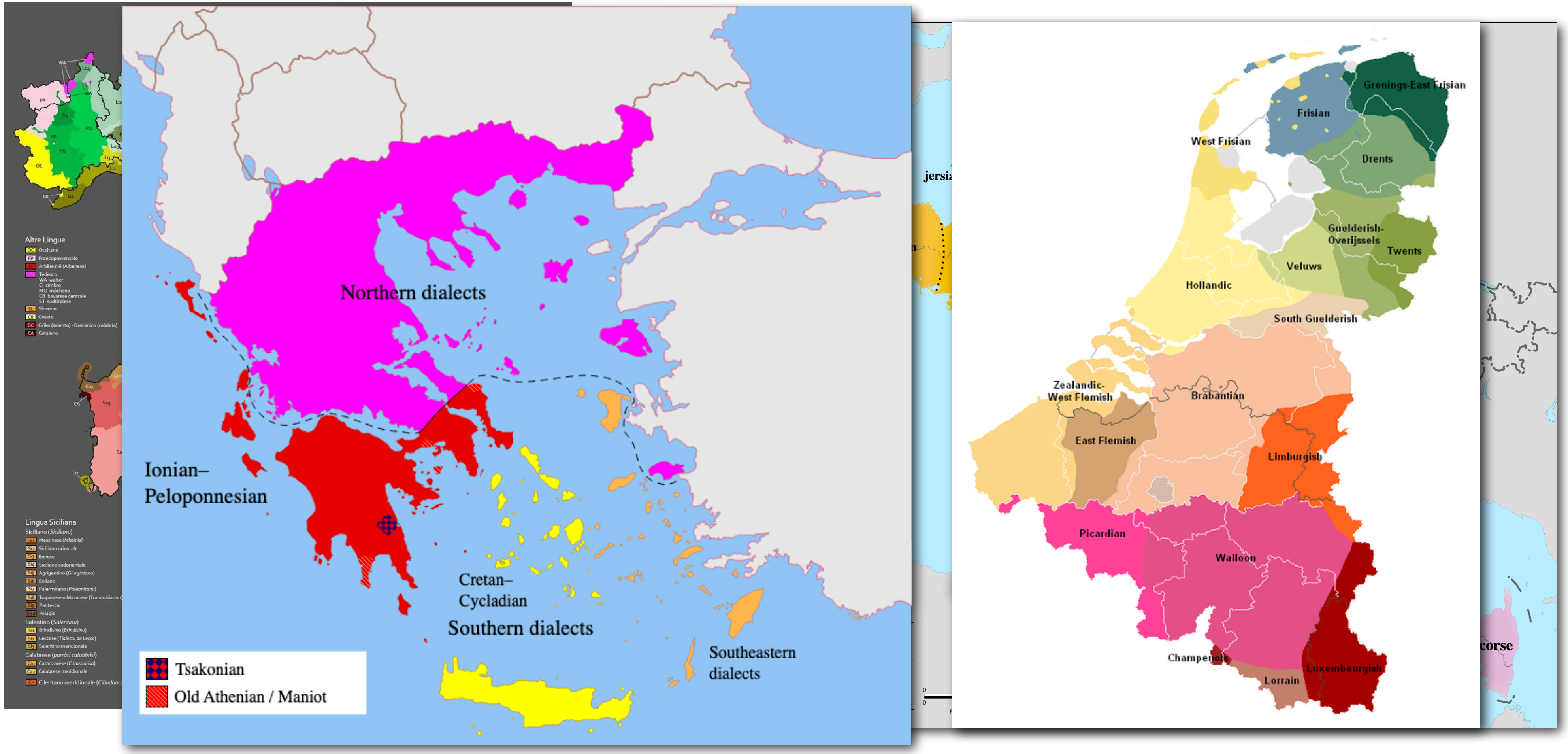


Languages are not Monoliths





Languages are not Monoliths



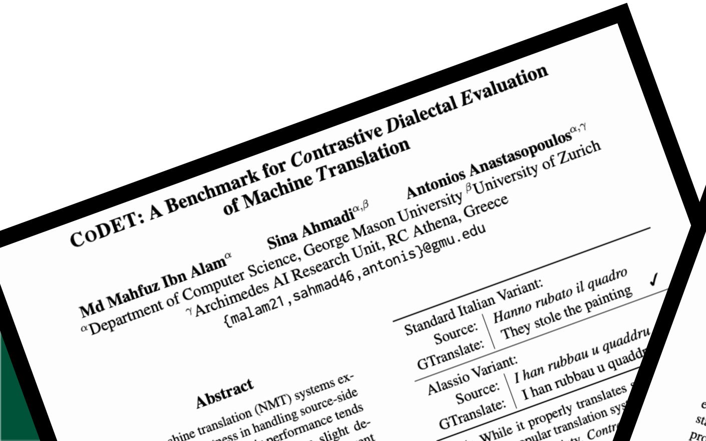
GEORGE MASON UNIVERSITY

DialectBench

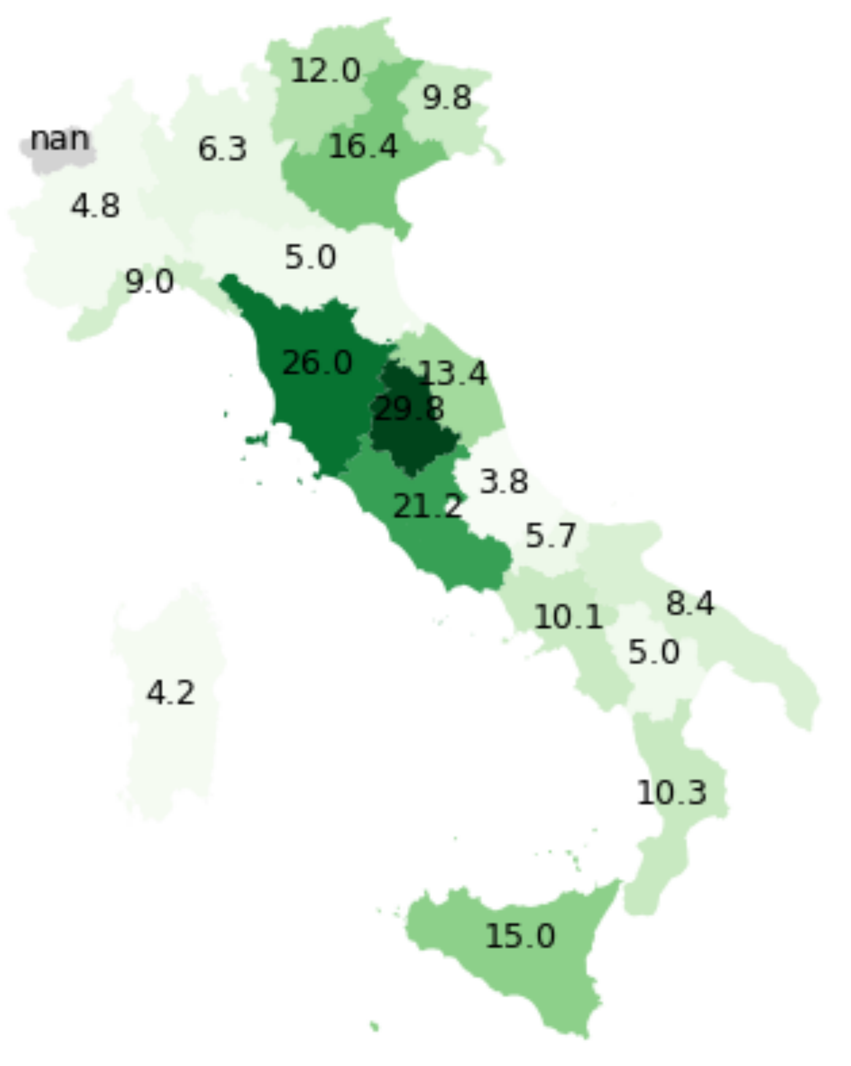
First large-scale benchmark
10 tasks, 40 continua, 281 varieties

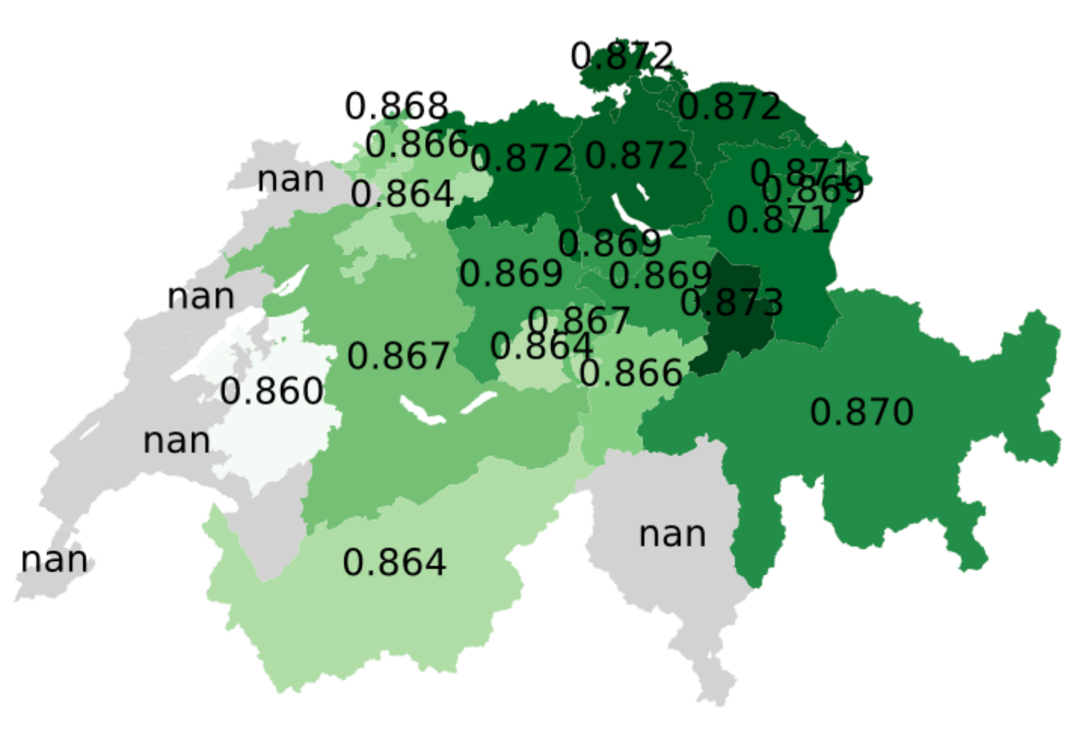
Task	Total	arabic	high german	italian romance	basque	anglic	sinitic	common turkic	sw shift. romance	greek	gallo-rhaetian	norwegian	neva	bengali	gallo-italian	kurdish	komi	serb.-croa.-bosnian	tupi-guarani.	modern dutch	eastern romance	frisian	swahili	Other
DEP.	40	3	2	4		3	3		4		3	3			1		3		3					8
POS.	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									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	32

Language Clusters



DialectBench Results

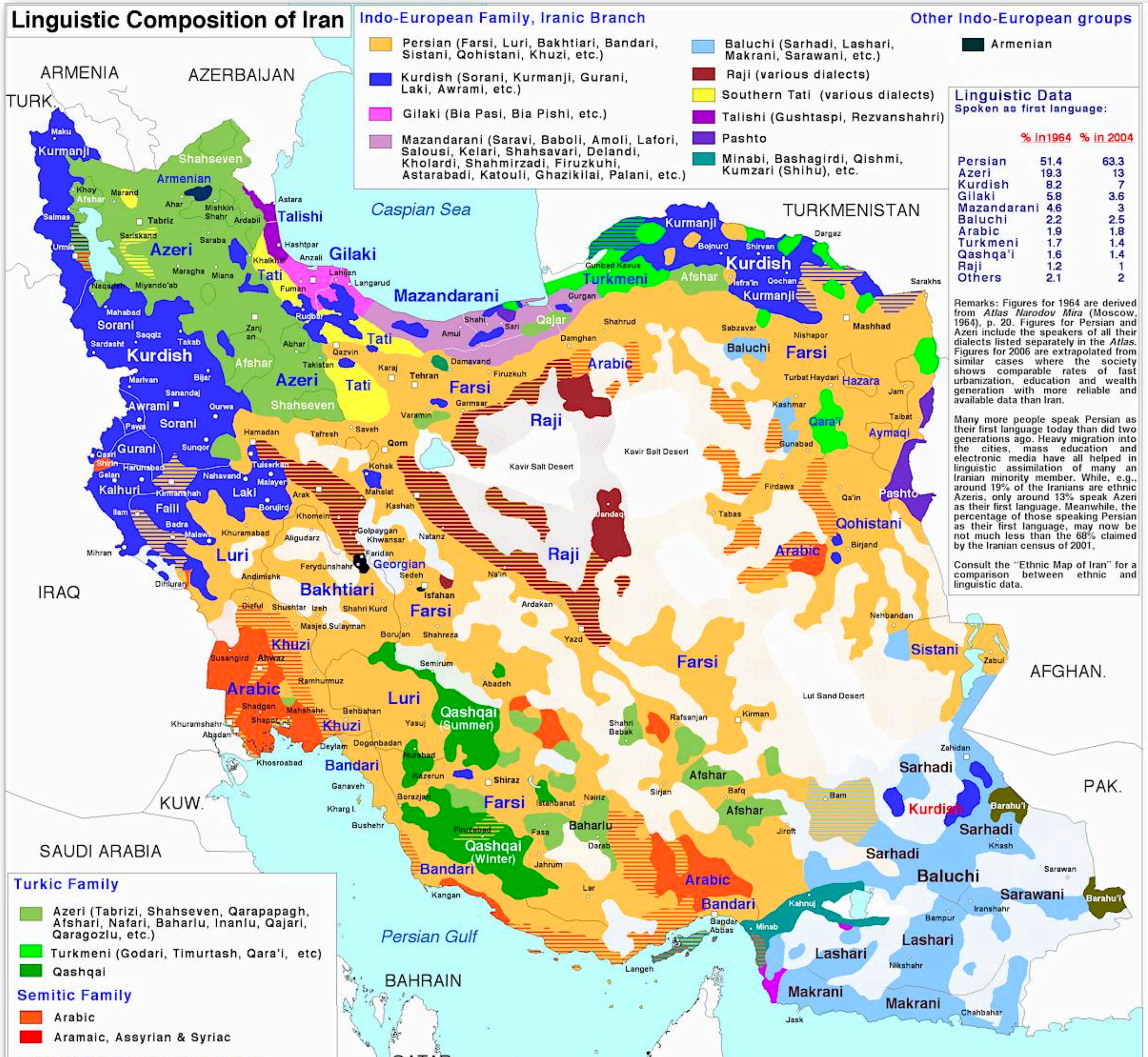




Minority Languages

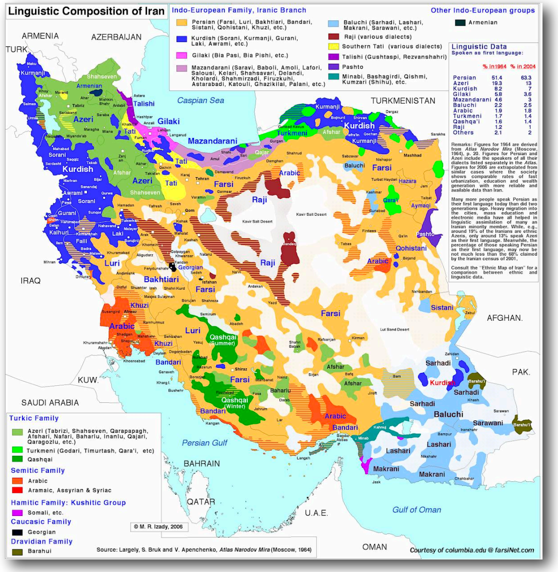
Minority Languages in X-lingual Communities

Minority Languages in X-lingual Communities

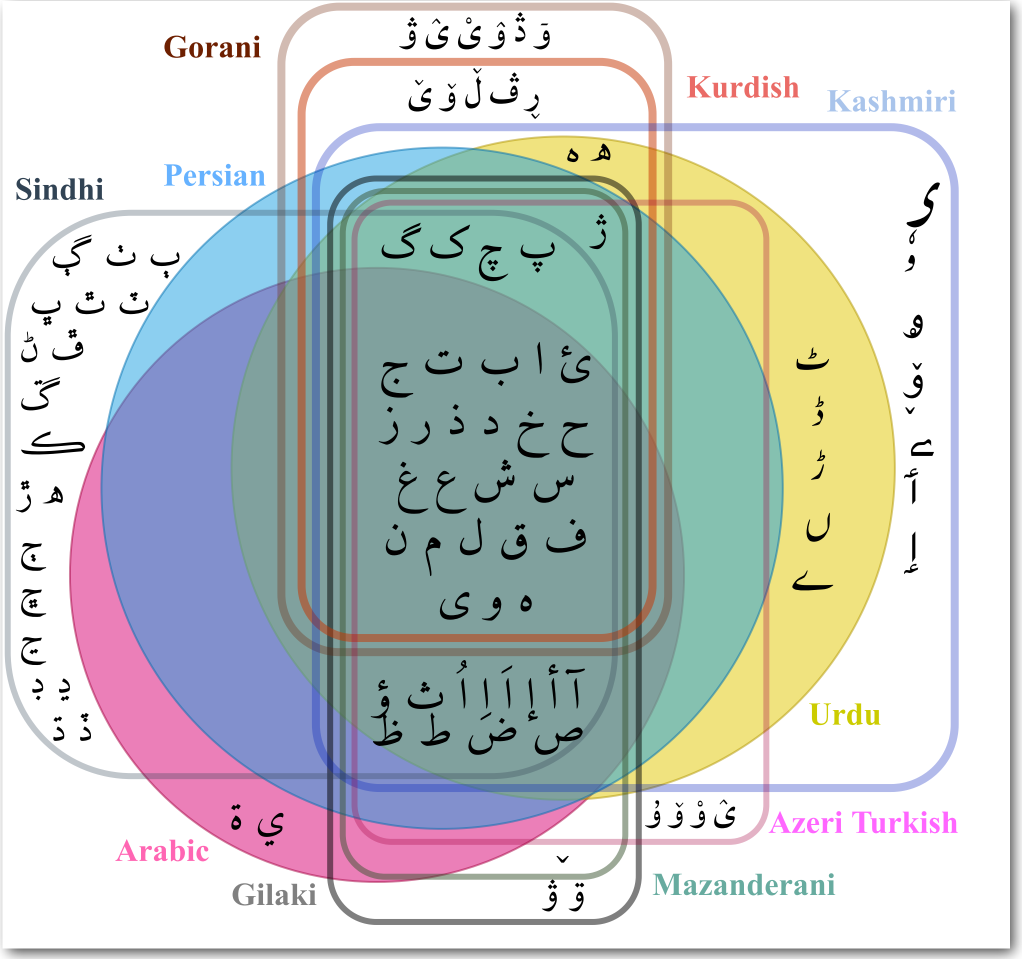


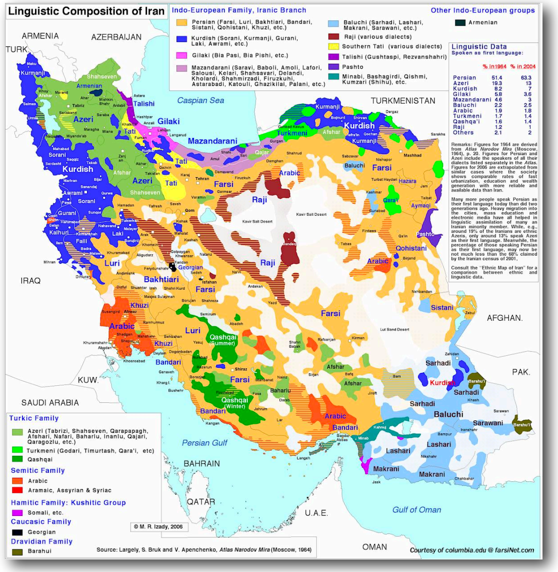
Source:

Minority Languages in X-lingual Communities



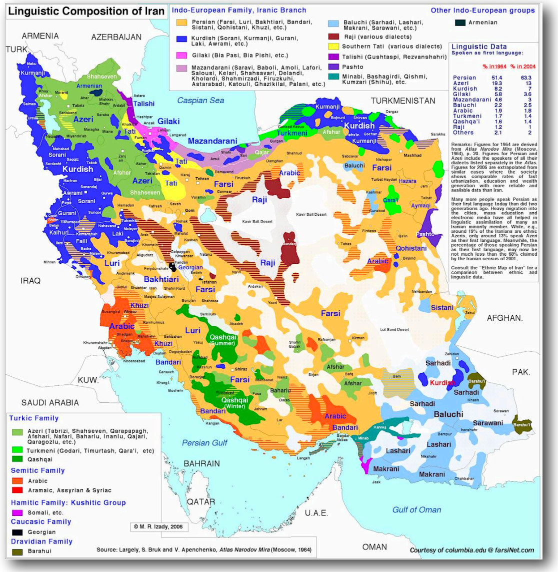
Minority Languages in X-lingual Communities

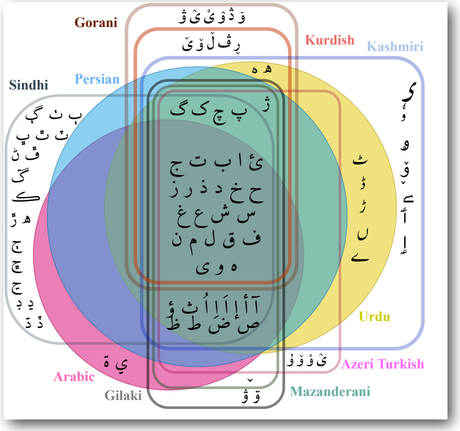




Minority Languages in X-lingual Communities

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

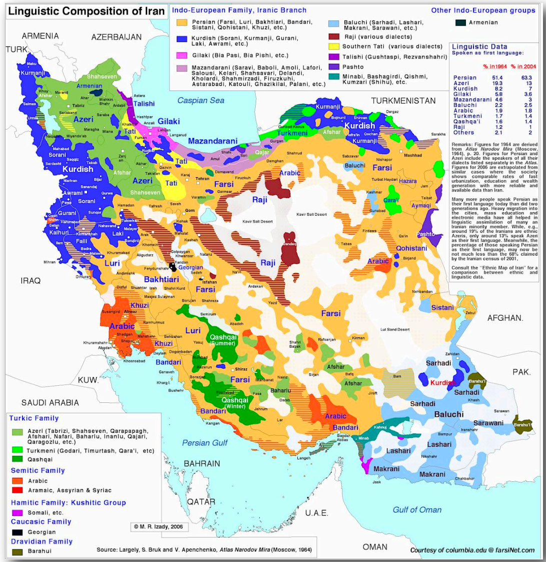


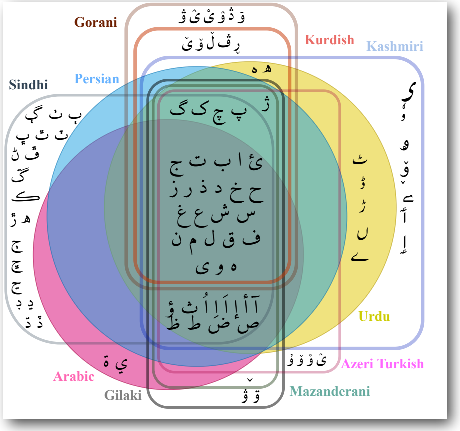


Minority Languages in X-lingual Communities

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

Language	Unconventional script	Unconventional writing	Conventional writing
Gilaki	Persian	یته زون نم هیسه گه گیلکن اون جی گب زنن	یته زوؤن ˇ نؤم هیسه گه گیلکؤن اوّن ˇ جي گب زنن
Kashmiri	Urdu	برور چھ اکھ وراسے جانور۔	برْور چھُ اَکھ وُرأسۍ جانوَر۔
Kurmanji	Arabic	قايمقام امدي بةرثوا پارزكار دهوك دا	قایمقامێ ئامێدیێ بەرسڤا پارێزگارێ دهۆکێ دا
Sorani	Arabic	هةر لة يةكةم شانؤوة ديارة فهديان دةويت	هەر لە یەکەم شانۆوە دیارە فەهەدیان دەوێت
Sindhi	Urdu	مديني ذانهن هجرت وقت فقط هيء گهرواري سانن گڏ هئي	مديني ڏانهن هجرت وقت فقط هيء گهرواري 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.

Lang ID

Lang ID

Terrible performance (F-score < 0.1)
by any existing toolkit

Lang ID

Terrible performance (F-score < 0.1)
by any existing toolkit

→ we trained our own (F-score = 0.88)

Lang ID

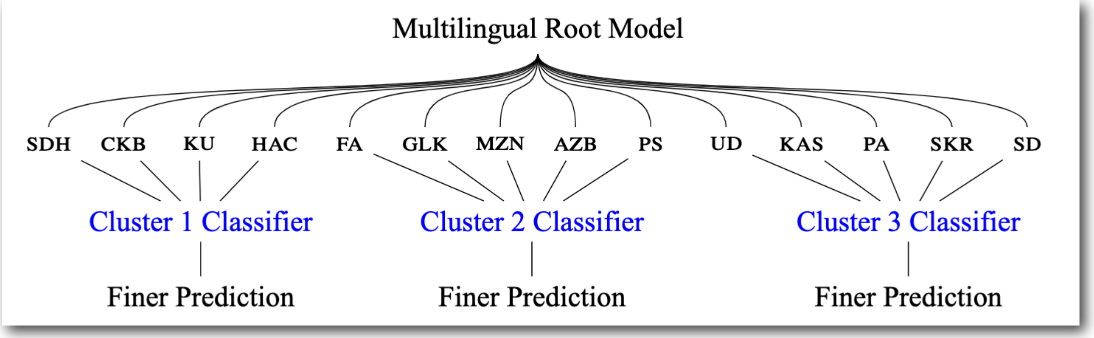
Terrible performance (F-score < 0.1)
by any existing toolkit

→ we trained our own (F-score = 0.88)

	Southern Kurdish	Central Kurdish	Northern Kurdish	Gorani	Persian	Gilaki	Mazanderani	Azeri Turkish	Pashto	Urdu	Kashmiri	Punjabi	Sindhi	Saraiki	Arabic	Balochi	Torwali	Uyghur	Brahui
Southern Kurdish	15643	99	70	113	0	2	0	0	1	0	0	0	0	0	2	0	0	0	0
Central Kurdish	242	15850	94	64	0	1	1	2	0	0	0	1	1	0	0	0	0	1	0
Northern Kurdish	49	29	15800	41	1	0	0	6	2	0	0	1	2	0	1	0	0	5	0
Gorani	59	21	18	15746	0	3	4	3	0	0	0	0	1	1	1	0	0	6	0
Persian	2	0	0	2	15874	50	26	7	8	0	3	1	2	2	7	0	0	0	0
Gilaki	2	0	2	10	63	15778	129	66	1	0	3	1	18	1	3	1	0	1	0
Mazanderani	0	0	0	3	18	92	15709	72	7	0	7	2	3	2	4	0	0	1	0
Azeri Turkish	0	0	2	6	1	44	91	15772	22	4	4	11	4	0	1	0	1	1	0
Pashto	2	1	7	3	21	2	6	34	15916	1	7	14	16	3	1	3	1	3	1
Urdu	0	0	0	0	0	0	0	0	0	15902	4	78	24	32	0	2	14	0	1
Kashmiri	0	0	1	0	0	3	8	7	7	3	15889	28	17	21	2	0	0	2	0
Punjabi	0	0	0	0	2	1	5	8	14	33	33	15782	26	95	0	7	8	1	0
Sindhi	0	0	1	2	1	16	5	1	6	10	5	12	15800	13	17	1	0	0	0
Saraiki	0	0	0	1	8	1	5	11	4	32	37	62	34	15818	0	14	6	2	0
Arabic	1	0	1	1	10	5	7	8	9	0	8	0	43	1	15955	1	0	12	0
Balochi	0	0	0	0	0	1	1	0	1	1	0	0	6	1	1	7464	0	0	0
Torwali	0	0	0	0	0	1	0	0	0	12	0	4	2	8	0	0	3590	0	0
Uyghur	0	0	4	8	0	0	3	3	1	1	0	0	0	0	5	0	0	15965	0
Brahui	0	0	0	0	0	0	0	0	0	1	0	3	1	2	0	0	0	0	286

Lang ID

Terrible performance (F-score < 0.1)
by any existing toolkit
→ we trained our own (F-score = 0.88)

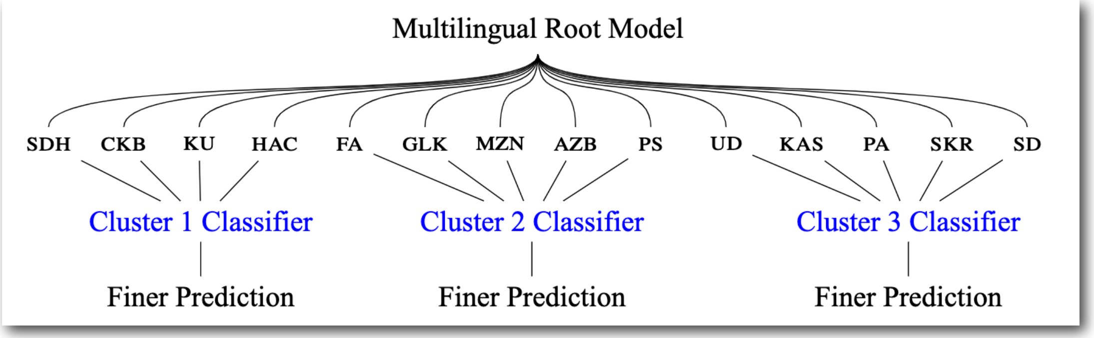


	Southern Kurdish	Central Kurdish	Northern Kurdish	Gorani	Persian	Gilaki	Mazanderani	Azeri Turkish	Pashto	Urdu	Kashmiri	Punjabi	Sindhi	Saraiki	Arabic	Balochi	Torwali	Uyghur	Brahui
Southern Kurdish	15643	99	70	113	0	2	0	0	1	0	0	0	0	0	2	0	0	0	0
Central Kurdish	242	15850	94	64	0	1	1	2	0	0	0	1	1	0	0	0	0	1	0
Northern Kurdish	49	29	15800	41	1	0	0	6	2	0	0	1	2	0	1	0	0	5	0
Gorani	59	21	18	15746	0	3	4	3	0	0	0	0	1	1	1	0	0	6	0
Persian	2	0	0	2	15874	50	26	7	8	0	3	1	2	2	7	0	0	0	0
Gilaki	2	0	2	10	63	15778	129	66	1	0	3	1	18	1	3	1	0	1	0
Mazanderani	0	0	0	3	18	92	15709	72	7	0	7	2	3	2	4	0	0	1	0
Azeri Turkish	0	0	2	6	1	44	91	15772	22	4	4	11	4	0	1	0	1	1	0
Pashto	2	1	7	3	21	2	6	34	15916	1	7	14	16	3	1	3	1	3	1
Urdu	0	0	0	0	0	0	0	0	0	15902	4	78	24	32	0	2	14	0	1
Kashmiri	0	0	1	0	0	3	8	7	7	3	15889	28	17	21	2	0	0	2	0
Punjabi	0	0	0	0	2	1	5	8	14	33	33	15782	26	95	0	7	8	1	0
Sindhi	0	0	1	2	1	16	5	1	6	10	5	12	15800	13	17	1	0	0	0
Saraiki	0	0	0	1	8	1	5	11	4	32	37	62	34	15818	0	14	6	2	0
Arabic	1	0	1	1	10	5	7	8	9	0	8	0	43	1	15955	1	0	12	0
Balochi	0	0	0	0	0	1	1	0	1	1	0	0	6	1	1	7464	0	0	0
Torwali	0	0	0	0	0	1	0	0	0	12	0	4	2	8	0	0	3590	0	0
Uyghur	0	0	4	8	0	0	3	3	1	1	0	0	0	0	5	0	0	15965	0
Brahui	0	0	0	0	0	0	0	0	0	1	0	3	1	2	0	0	0	0	286

Lang ID

Terrible performance (F-score < 0.1)
by any existing toolkit

→ we trained our own (F-score = 0.88)

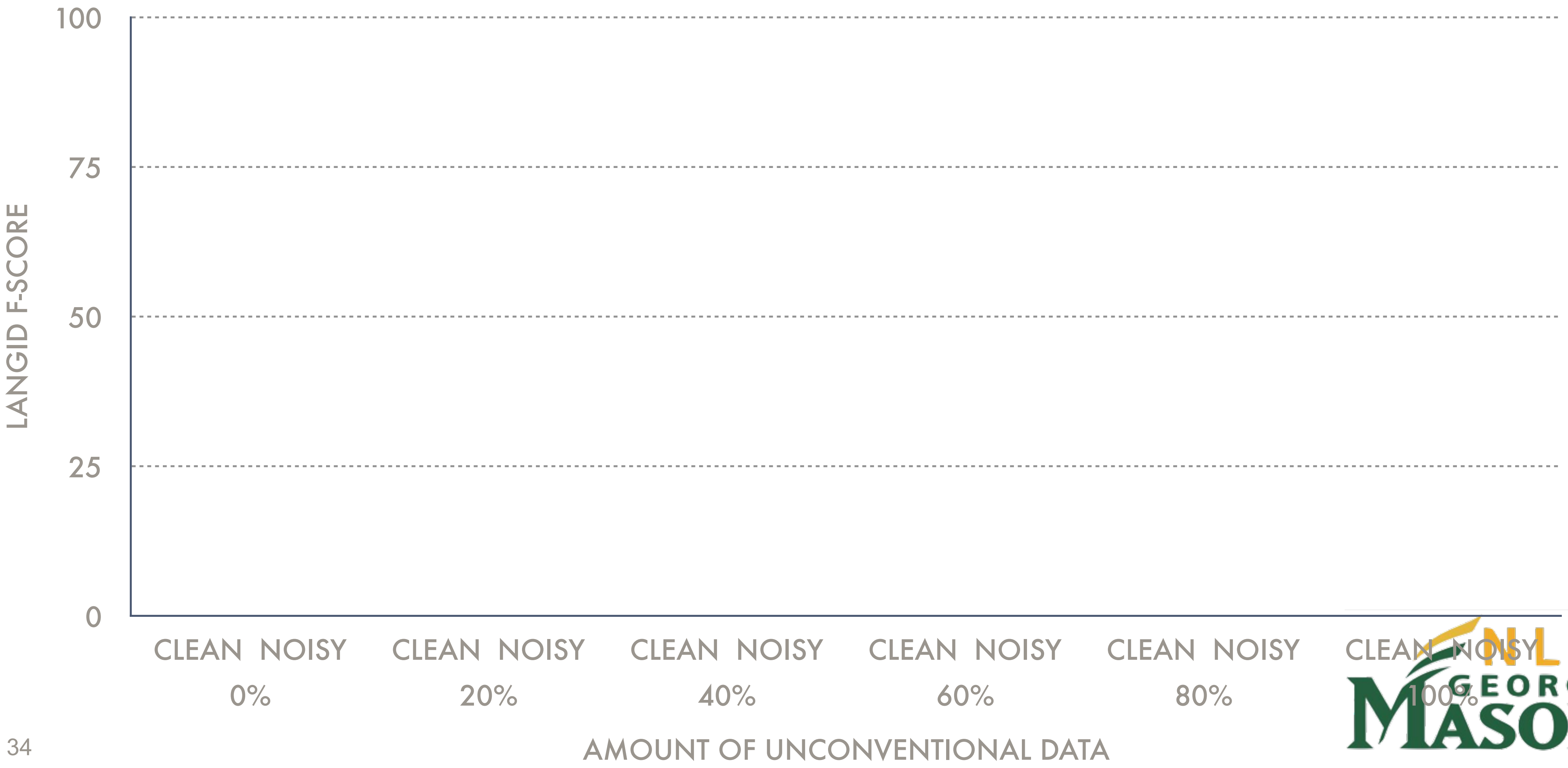


→ hierarchical model (F-score = 0.95)

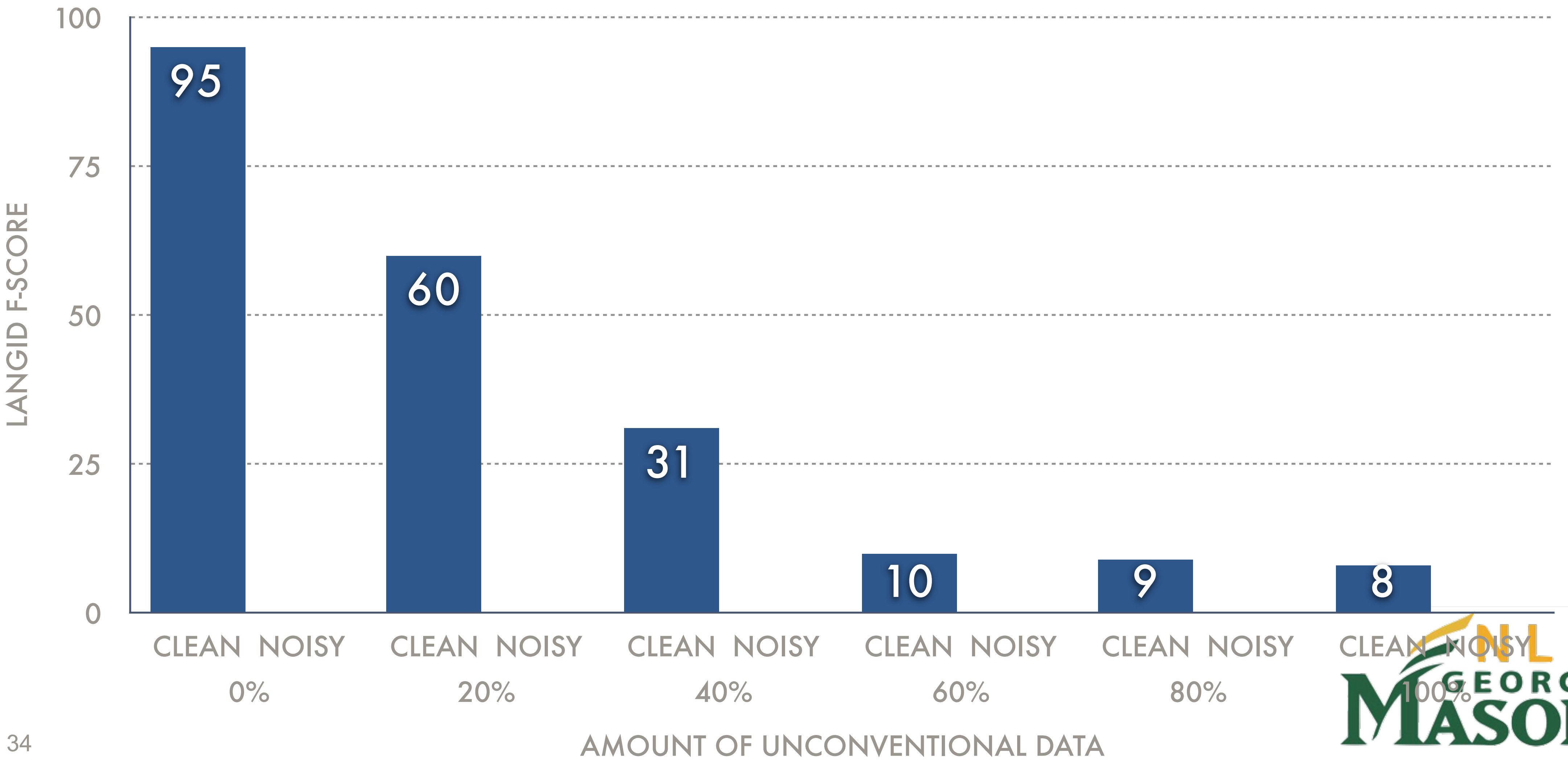
	Southern Kurdish	Central Kurdish	Northern Kurdish	Gorani	Persian	Gilaki	Mazanderani	Azeri Turkish	Pashto	Urdu	Kashmiri	Punjabi	Sindhi	Saraiki	Arabic	Balochi	Torwali	Uyghur	Brahui
Southern Kurdish	15643	99	70	113	0	2	0	0	1	0	0	0	0	0	2	0	0	0	0
Central Kurdish	242	15850	94	64	0	1	1	2	0	0	0	1	1	0	0	0	0	1	0
Northern Kurdish	49	29	15800	41	1	0	0	6	2	0	0	1	2	0	1	0	0	5	0
Gorani	59	21	18	15746	0	3	4	3	0	0	0	0	1	1	1	0	0	6	0
Persian	2	0	0	2	15874	50	26	7	8	0	3	1	2	2	7	0	0	0	0
Gilaki	2	0	2	10	63	15778	129	66	1	0	3	1	18	1	3	1	0	1	0
Mazanderani	0	0	0	3	18	92	15709	72	7	0	7	2	3	2	4	0	0	1	0
Azeri Turkish	0	0	2	6	1	44	91	15772	22	4	4	11	4	0	1	0	1	1	0
Pashto	2	1	7	3	21	2	6	34	15916	1	7	14	16	3	1	3	1	3	1
Urdu	0	0	0	0	0	0	0	0	0	15902	4	78	24	32	0	2	14	0	1
Kashmiri	0	0	1	0	0	3	8	7	7	3	15889	28	17	21	2	0	0	2	0
Punjabi	0	0	0	0	2	1	5	8	14	33	33	15782	26	95	0	7	8	1	0
Sindhi	0	0	1	2	1	16	5	1	6	10	5	12	15800	13	17	1	0	0	0
Saraiki	0	0	0	1	8	1	5	11	4	32	37	62	34	15818	0	14	6	2	0
Arabic	1	0	1	1	10	5	7	8	9	0	8	0	43	1	15955	1	0	12	0
Balochi	0	0	0	0	0	1	1	0	1	1	0	0	6	1	1	7464	0	0	0
Torwali	0	0	0	0	0	1	0	0	0	12	0	4	2	8	0	0	3590	0	0
Uyghur	0	0	4	8	0	0	3	3	1	1	0	0	0	0	5	0	0	15965	0
Brahui	0	0	0	0	0	0	0	0	0	1	0	3	1	2	0	0	0	0	286

Effect of Unconventional Writing

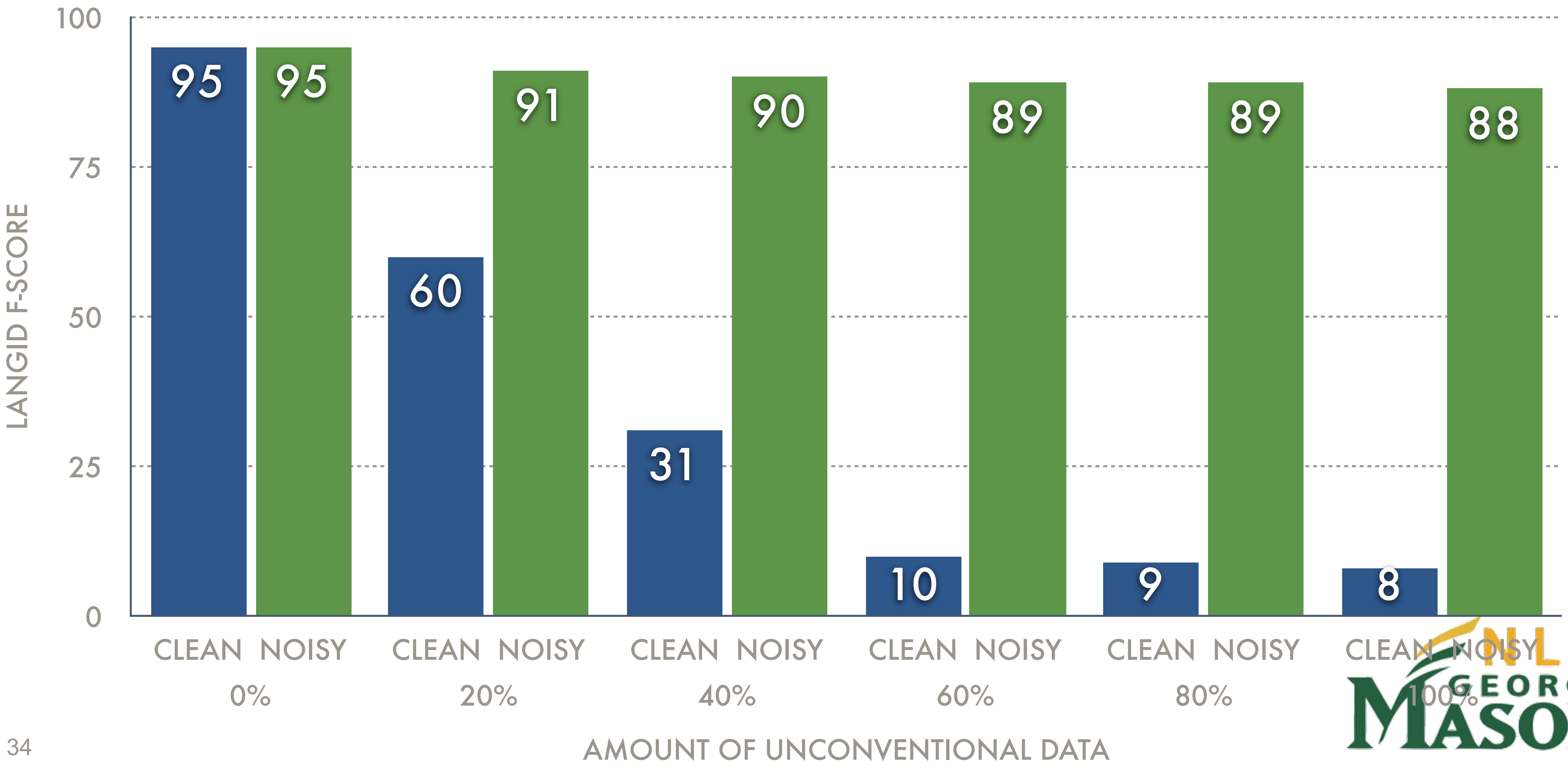
Effect of Unconventional Writing



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

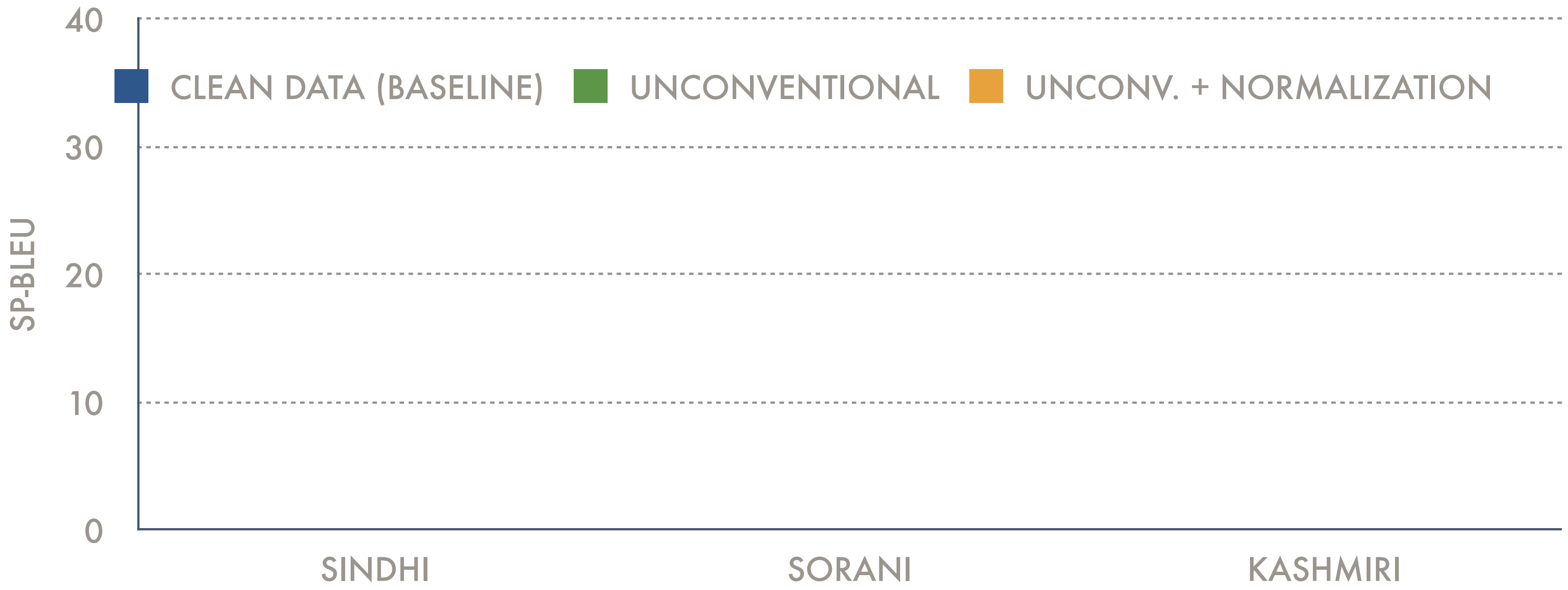
Train a Normalization model

(Encoder-decoder, self-attention based)

Evaluate its effect on Machine Translation

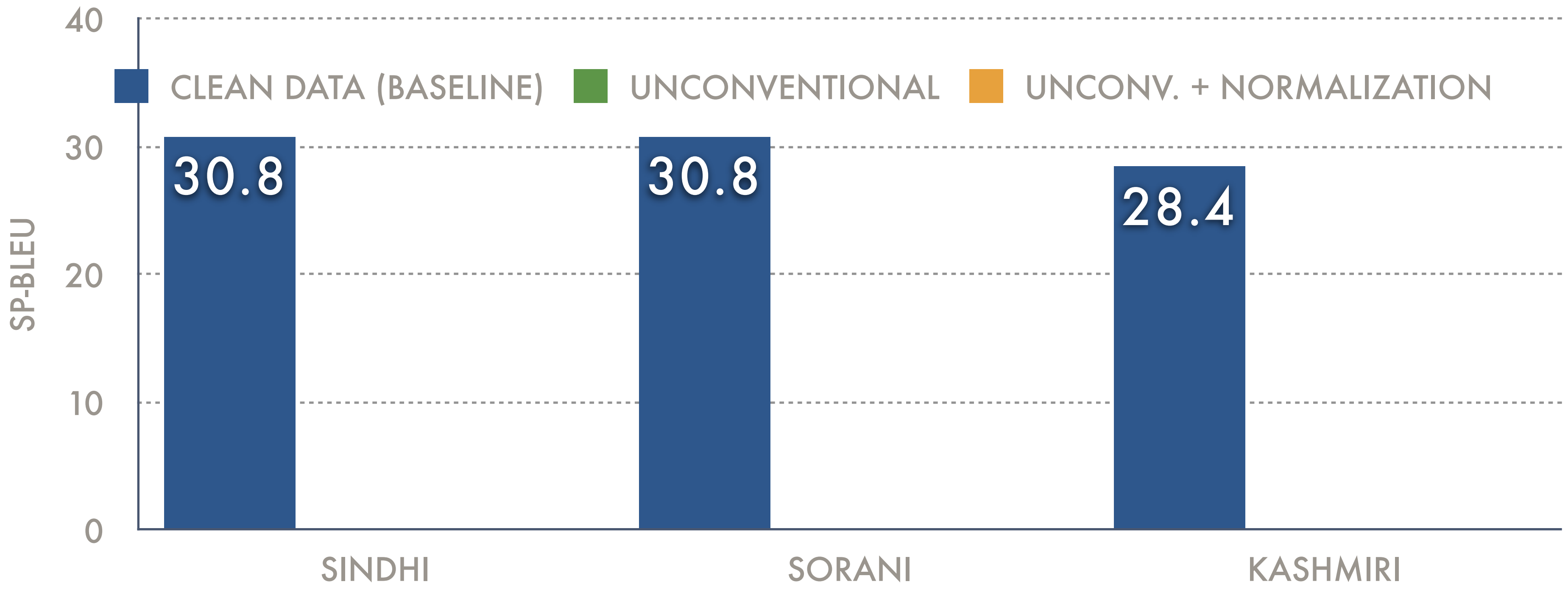
Mitigating the Effect of Unconventional Writing

Train a Normalization model
(Encoder-decoder, self-attention based)
Evaluate its effect on Machine Translation



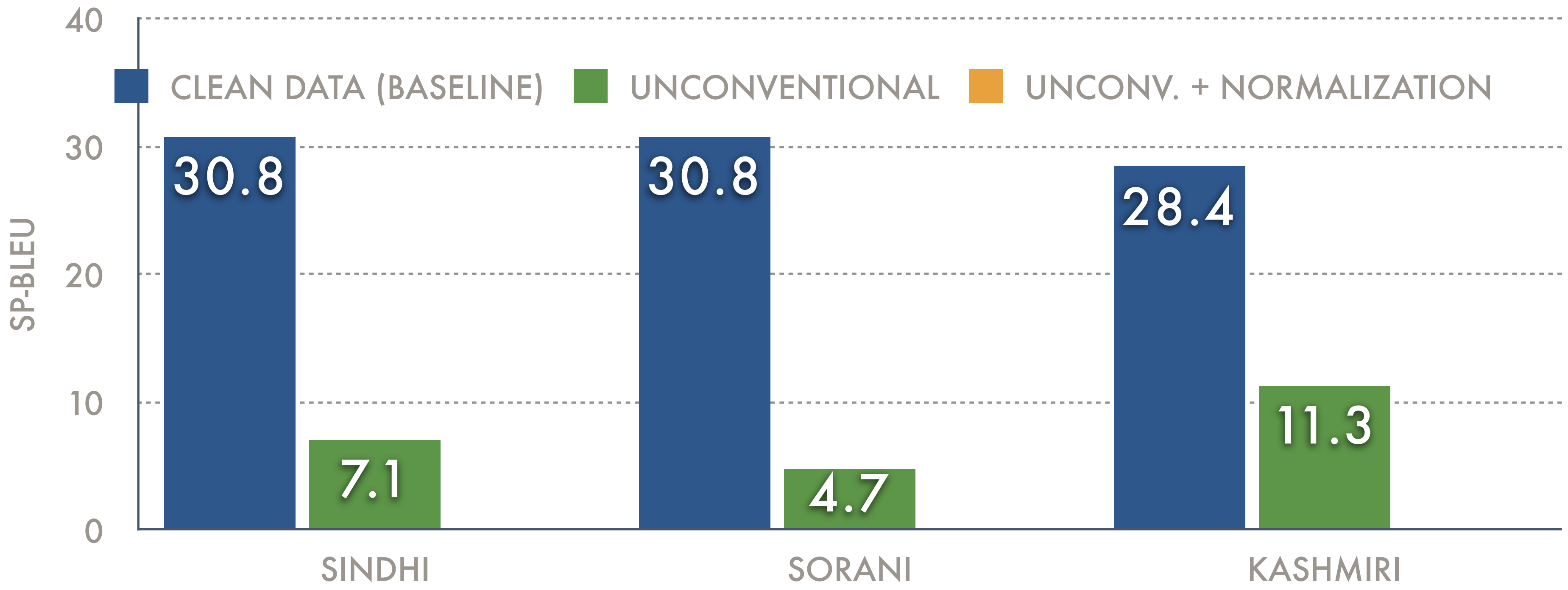
Mitigating the Effect of Unconventional Writing

Train a Normalization model
(Encoder-decoder, self-attention based)
Evaluate its effect on Machine Translation



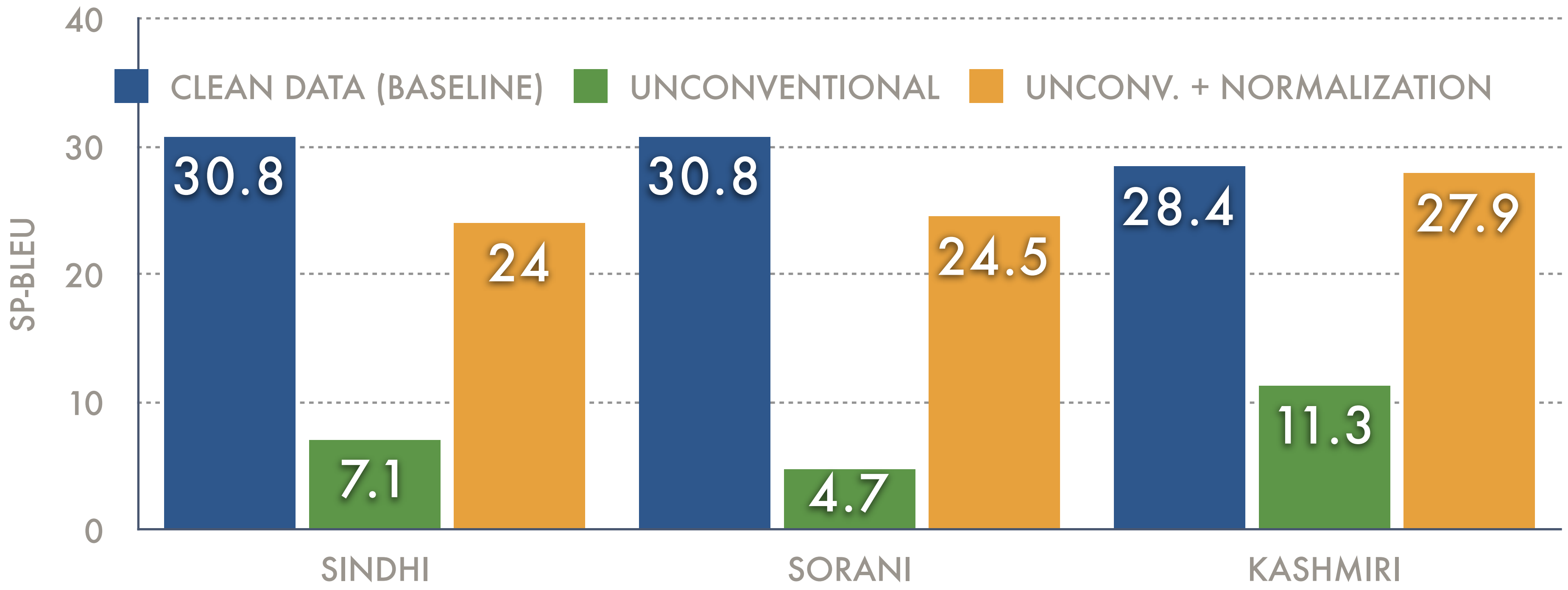
Mitigating the Effect of Unconventional Writing

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)

2005: first ST corpora
Small size/language coverage

2016-2017: first e2e ST models
(Duong et al., 2016, Berard et al., 2016, Weiss et al., 2017, ...)
encoder-decoder architectures based on RNNs

2018: first e2e models at IWSLT
8.7 BLEU points below cascade ST solutions on En-De

2019: significant gap reduction at IWSLT
1.6 BLEU points below cascade ST solutions on En-De

2019: ST adaptation of Transformer
(Di Gangi et al., 2019)

2019-2020: new ST corpora
Larger size/language coverage

2020: the gap almost closed?
+0.24 BLEU on unsegmented En-De test data

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

Reden (engl. speeches)

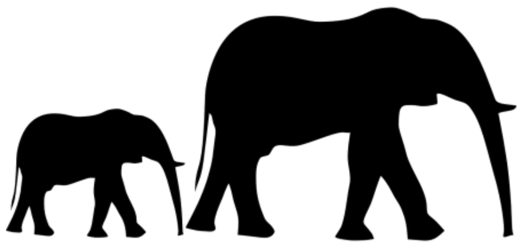
Reben (engl. vines)

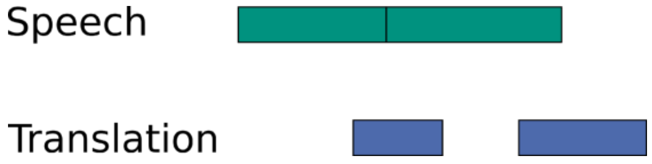
Challenges in translation of speech

- 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

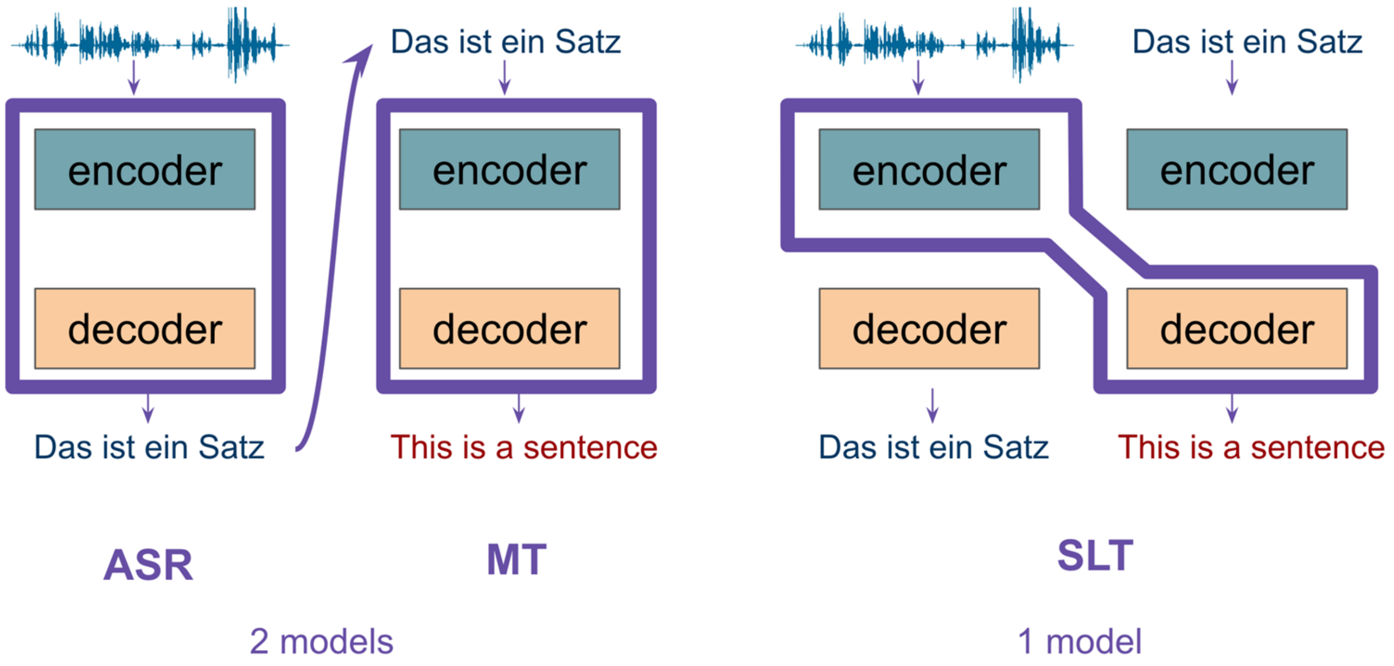
- Audio challenges
- Text-Speech mismatch
- Error propagation
- Data
- Partial information
  - Online: Translate during production of speech
  - Generate translation before full sentence is known



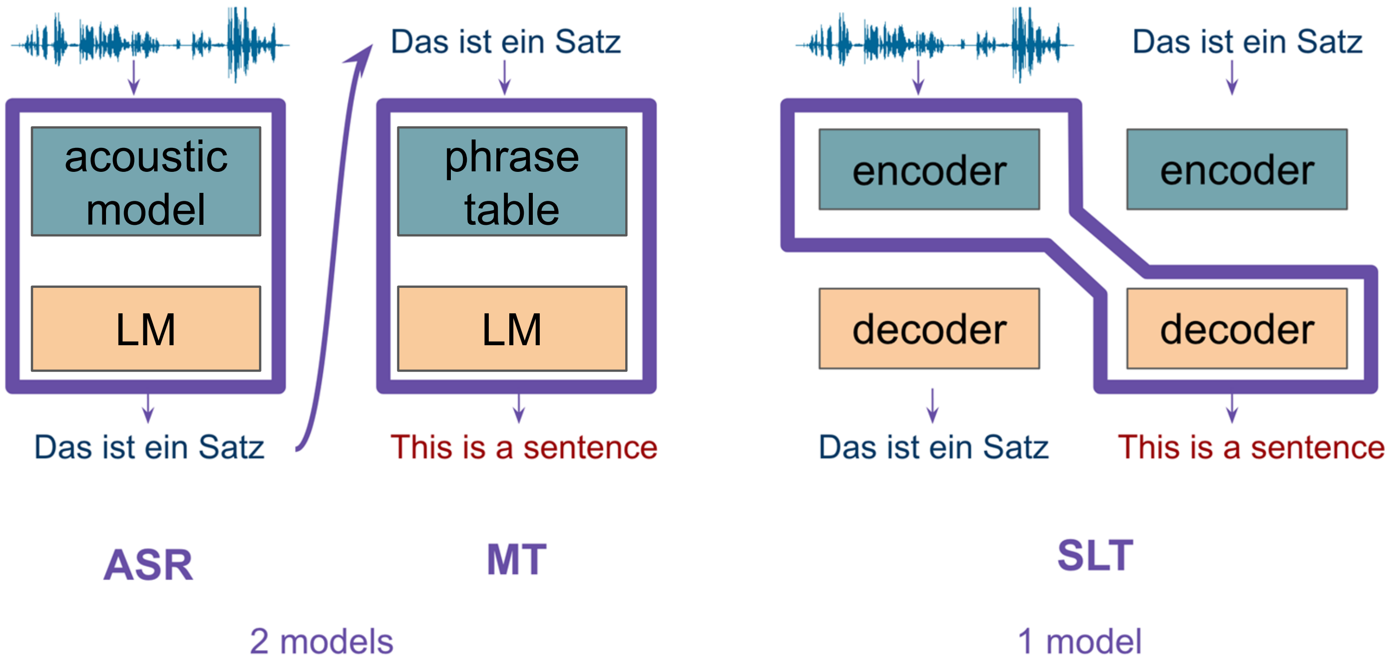


Traditional Cascade Approach

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

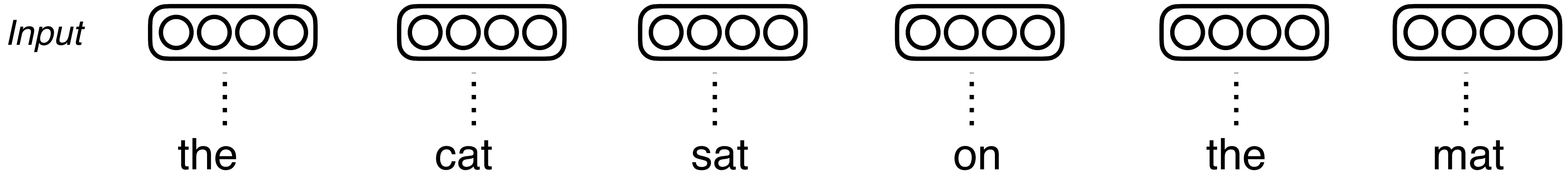
Encoder-Decoder with Attention

the cat sat on the mat

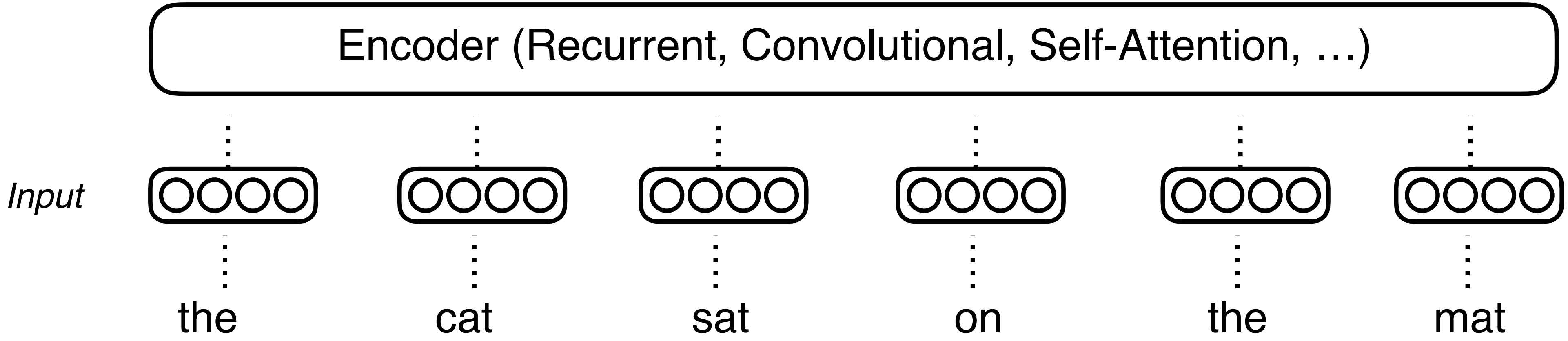
Encoder-Decoder with Attention

the cat sat on the mat

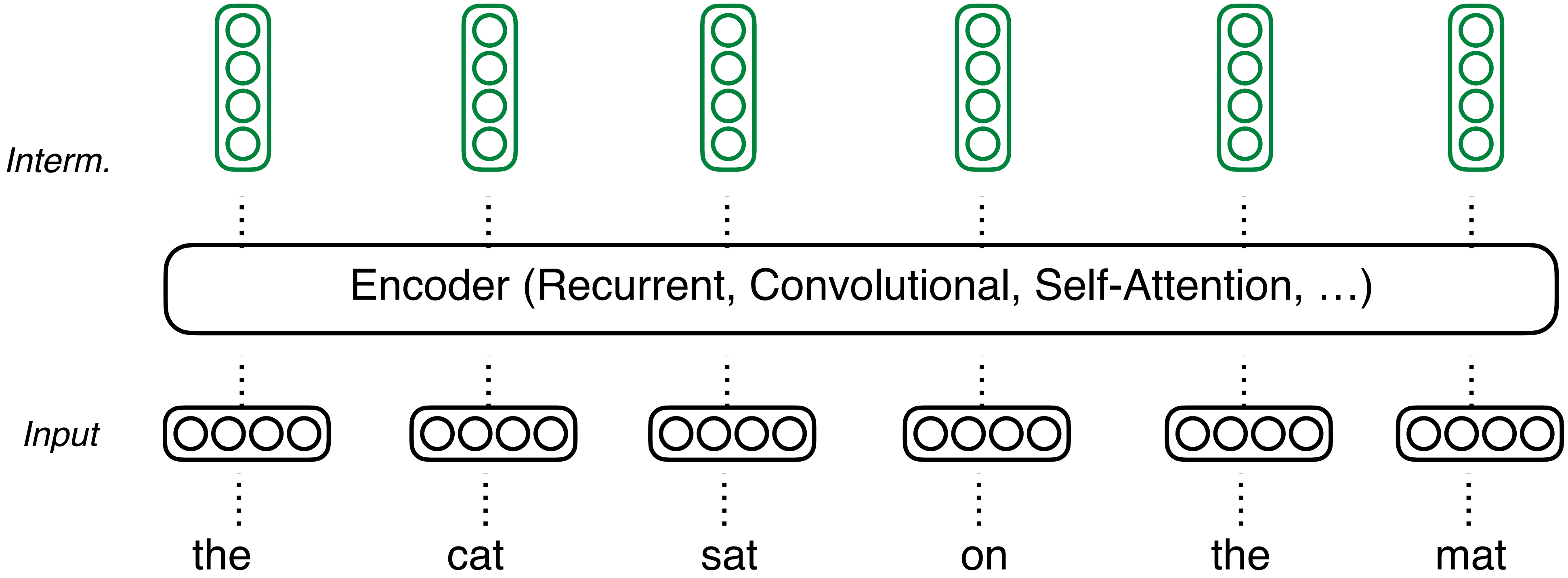
Encoder-Decoder with Attention



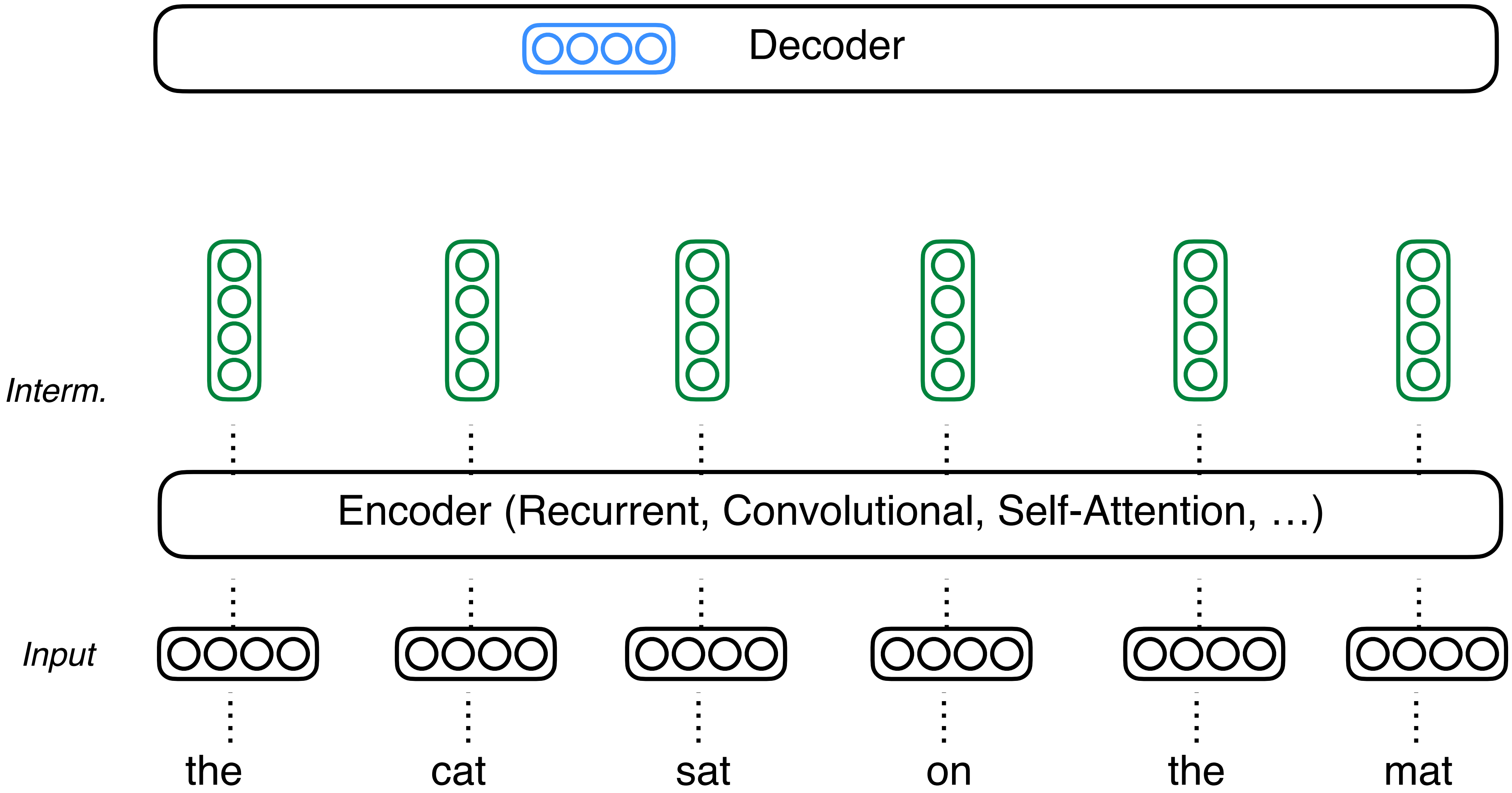
Encoder-Decoder with Attention



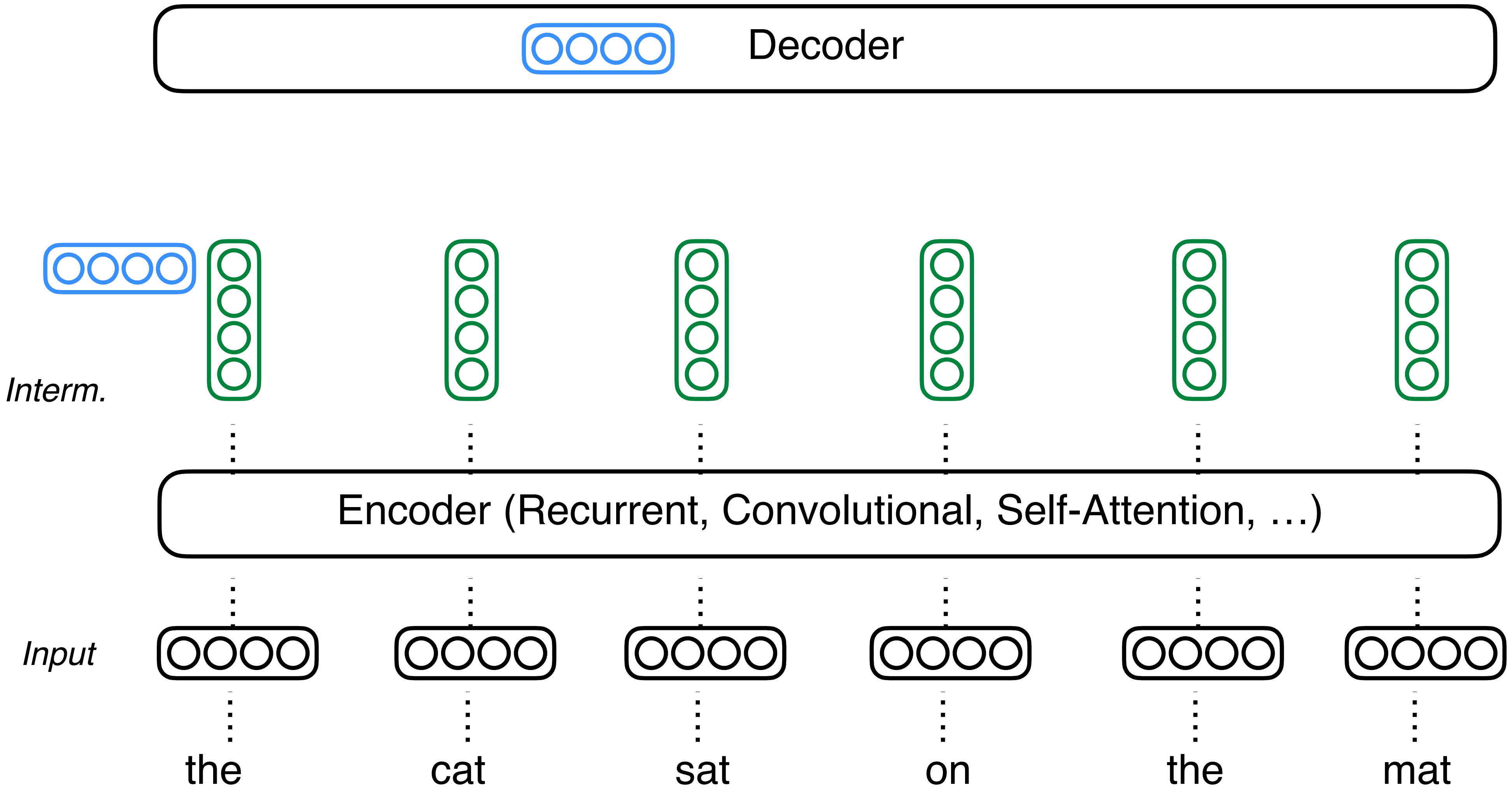
Encoder-Decoder with Attention

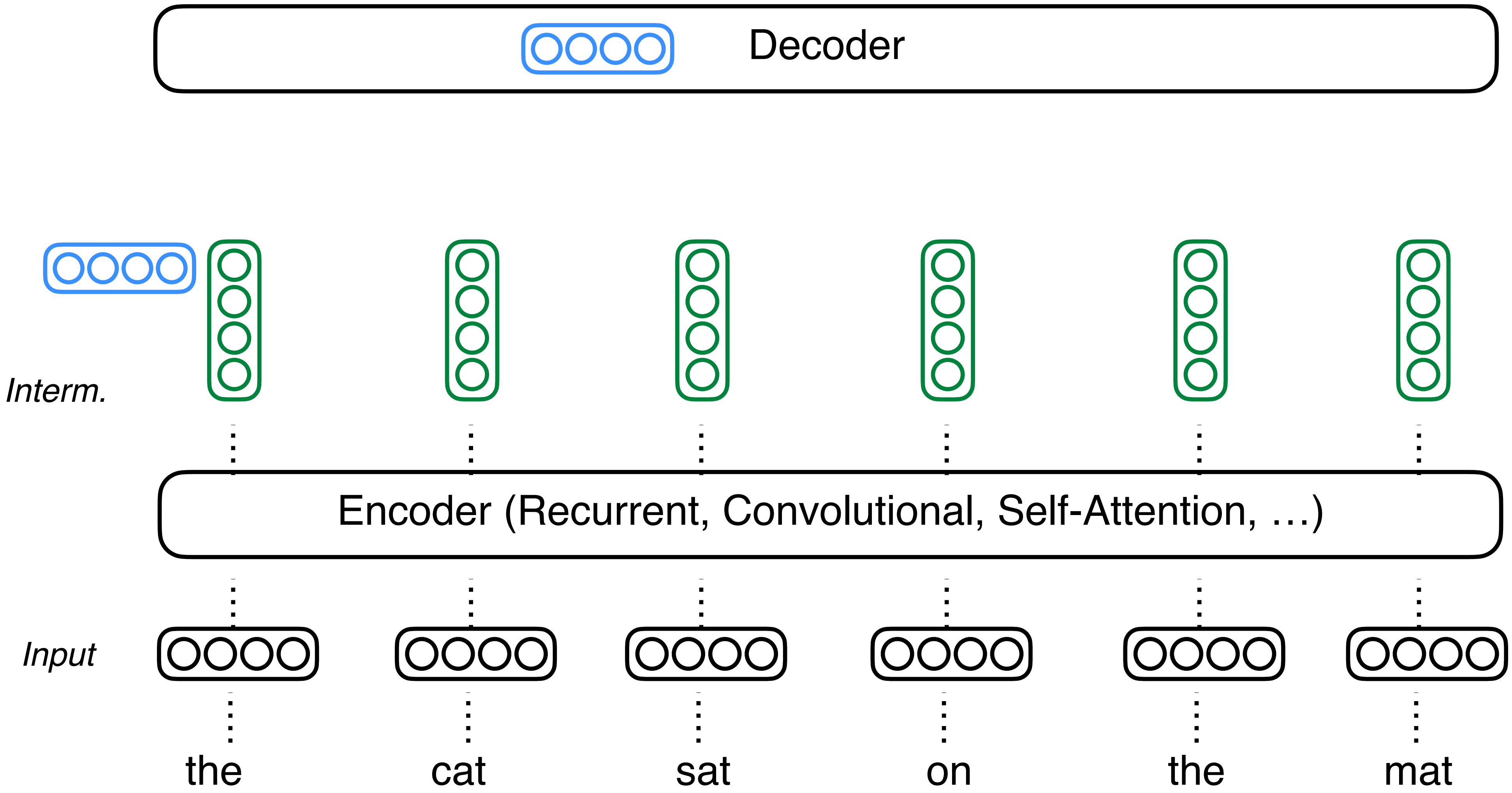


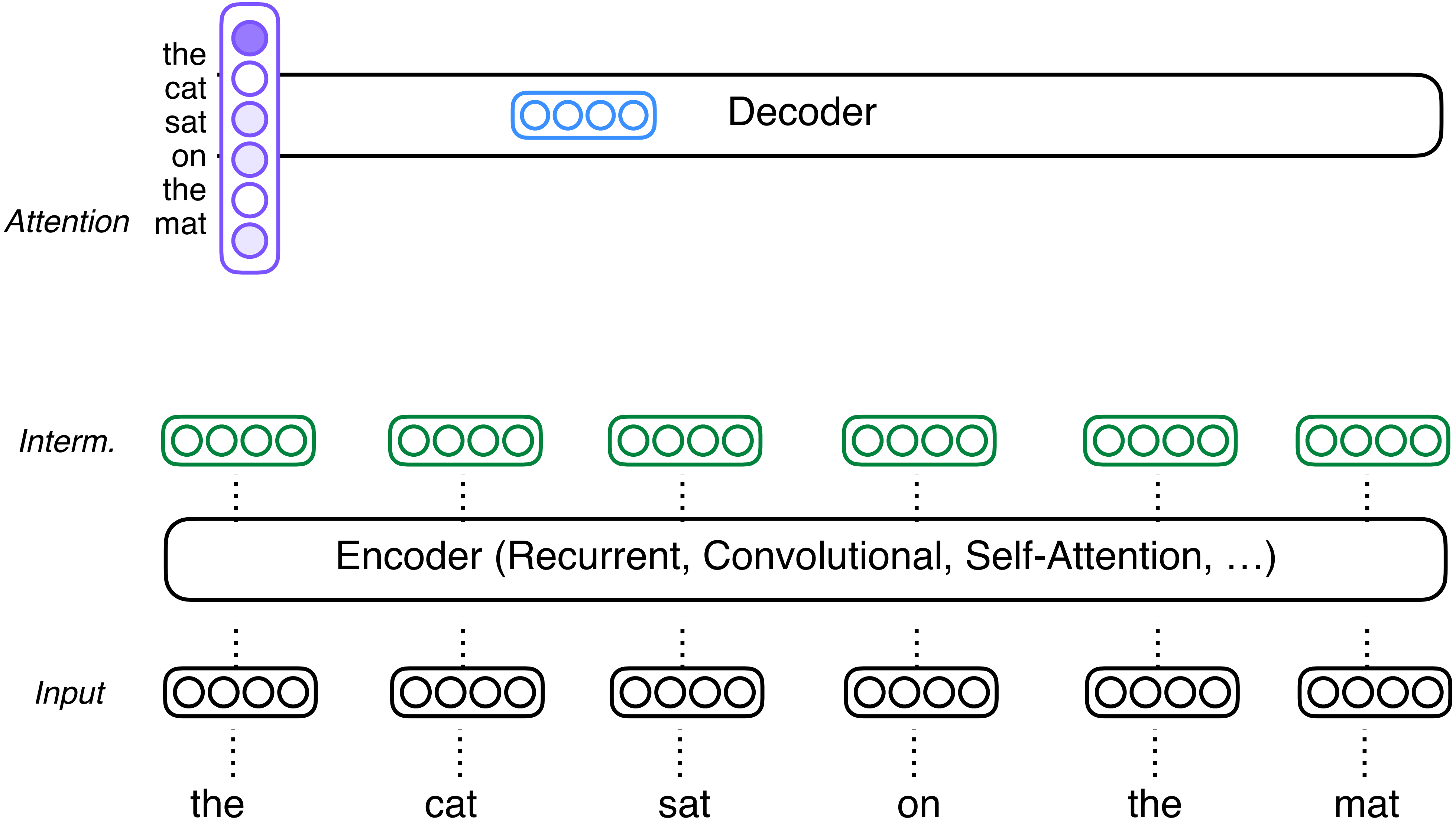
Encoder-Decoder with Attention



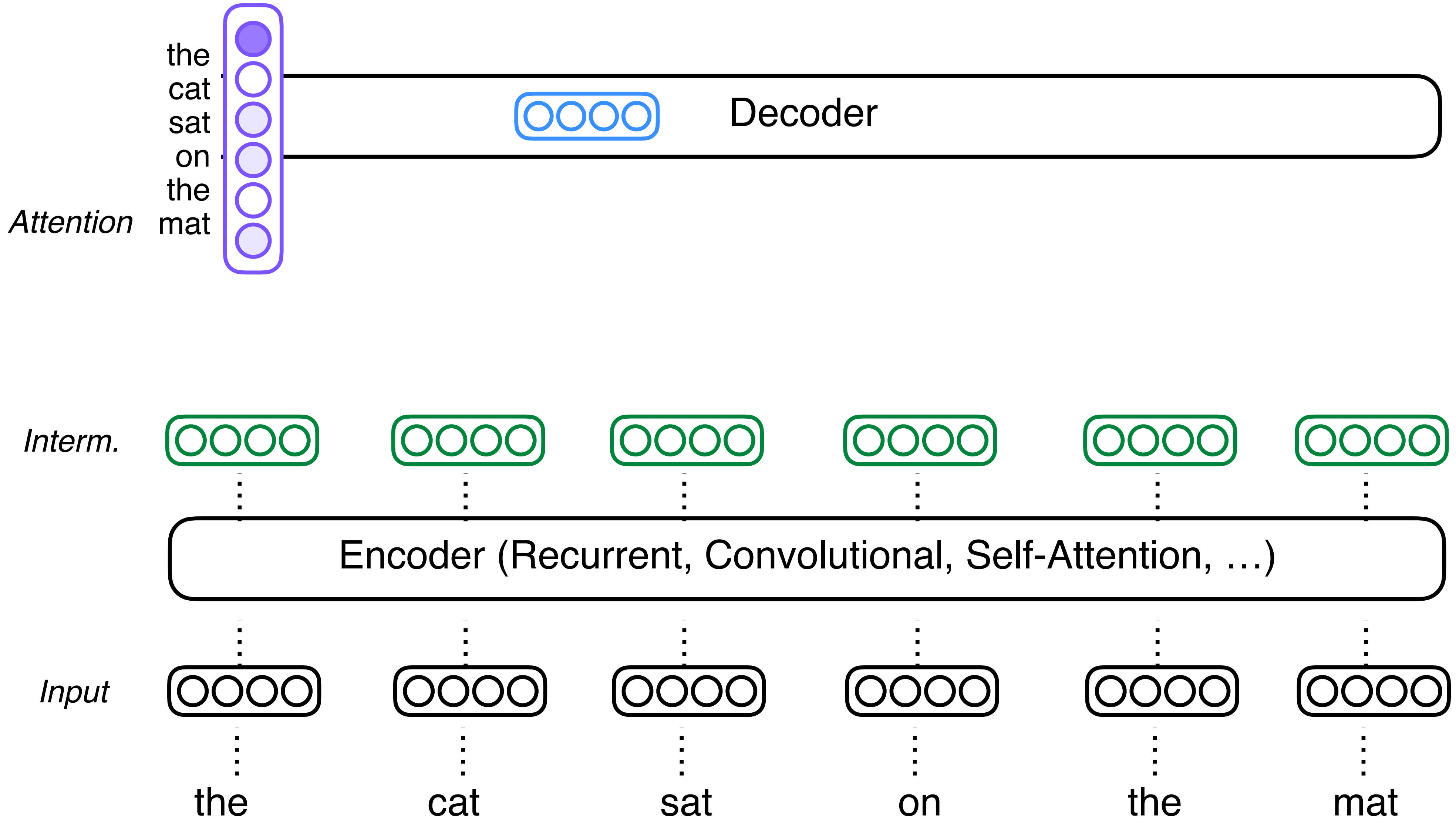
Encoder-Decoder with Attention

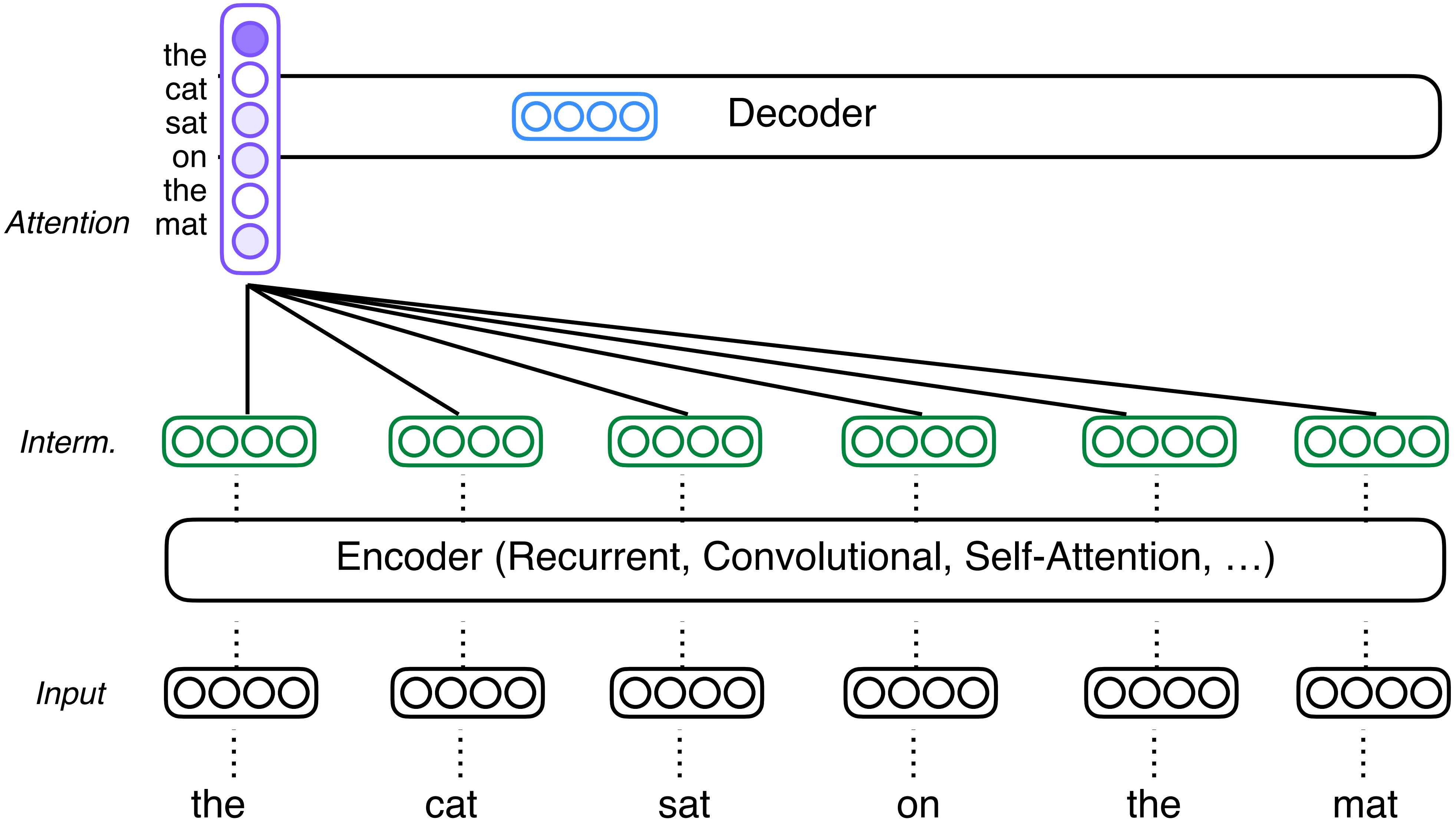


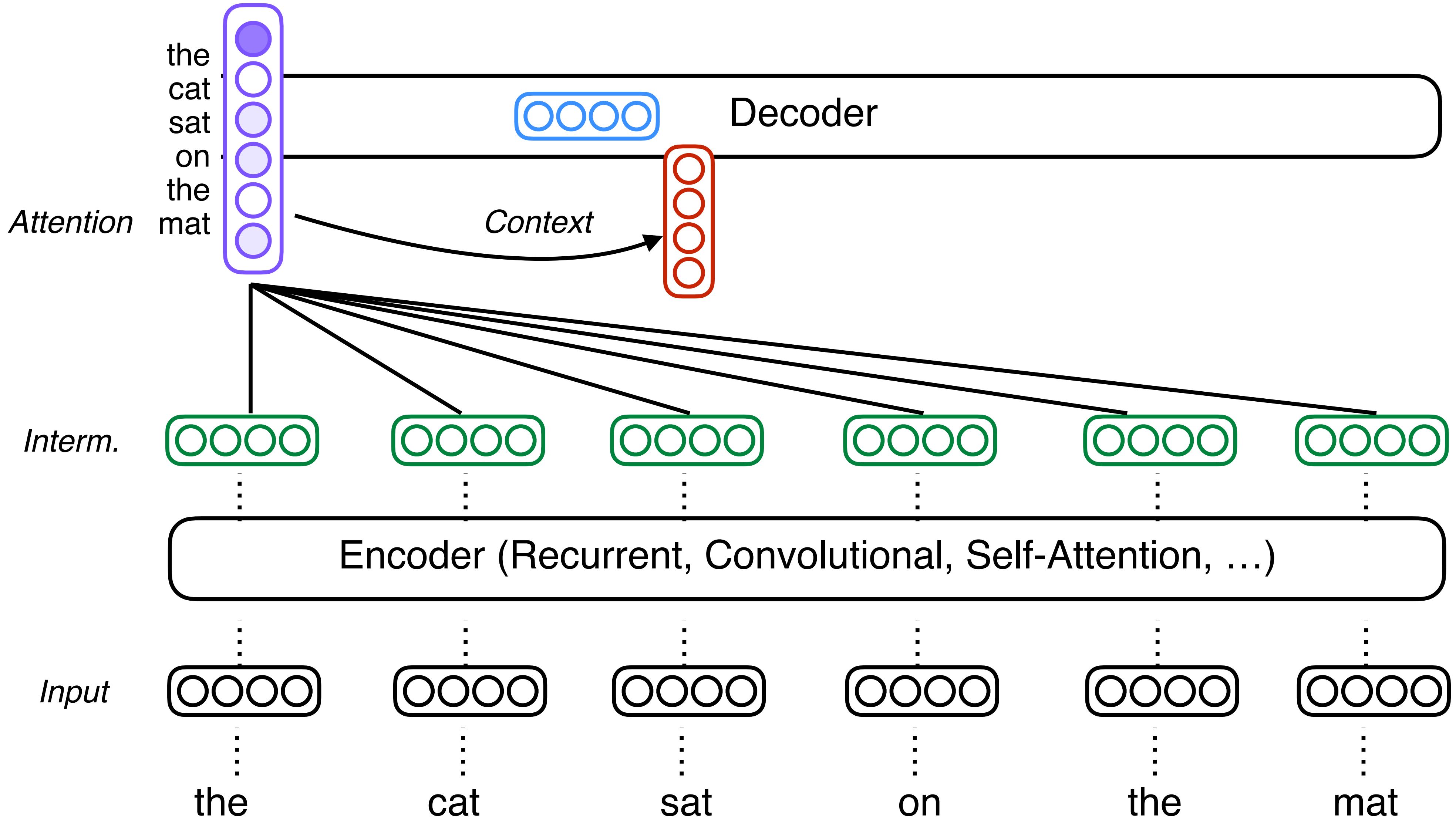

Decoder
Interm.
Encoder (Recurrent, Convolutional, Self-Attention, …)
Input
the
cat
sat
on
the
mat

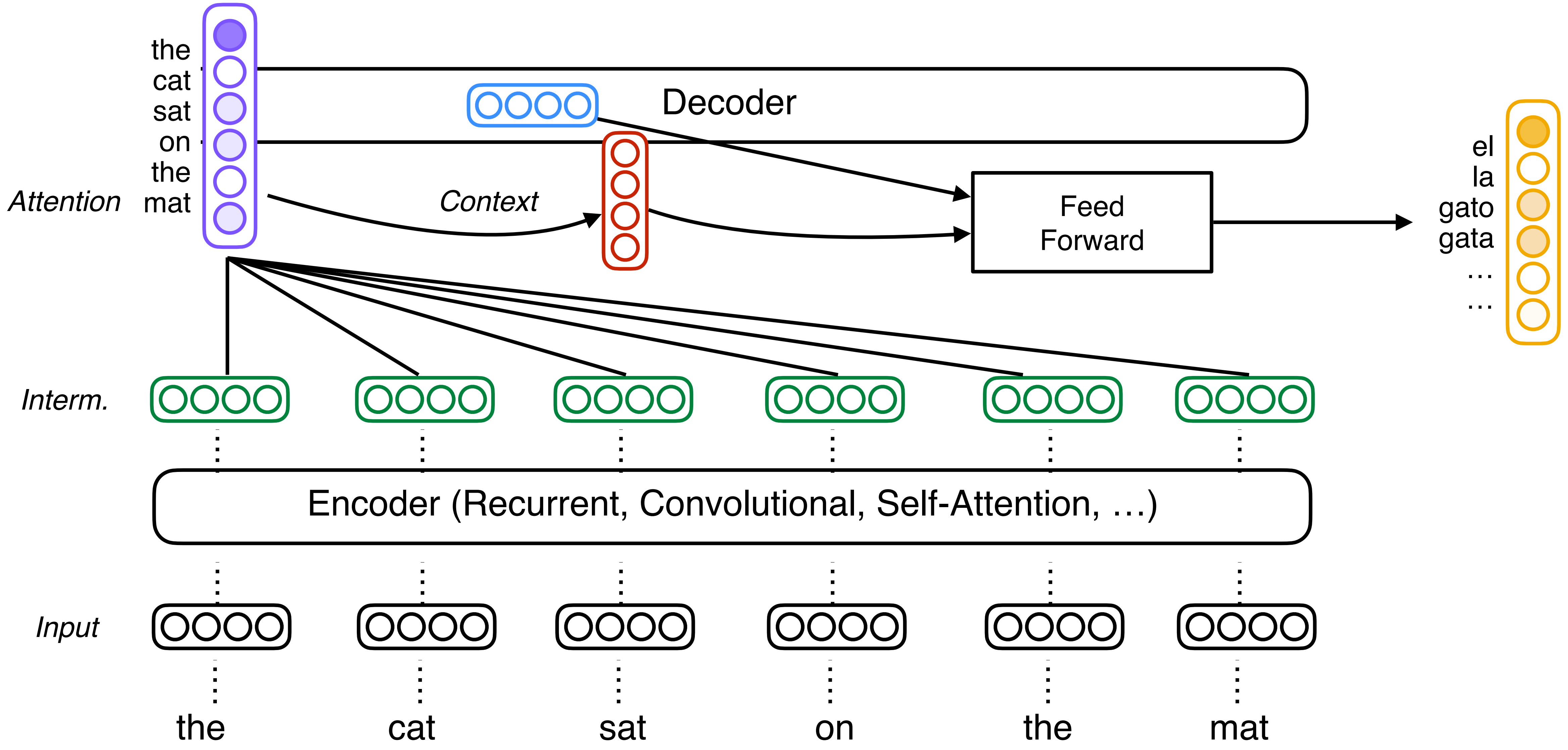


the
cat
sat
on
the
mat
Attention
Decoder
Interm.
Encoder (Recurrent, Convolutional, Self-Attention, …)
Input
the
cat
sat
on
the
mat


the
cat
sat
on
the
mat
Attention
Decoder
Interm.
Encoder (Recurrent, Convolutional, Self-Attention, …)
Input
the
cat
sat
on
the
mat


Decoder
Attention
the
cat
sat
on
the
mat
Interm.
Encoder (Recurrent, Convolutional, Self-Attention, …)
Input
the
cat
sat
on
the
mat


Decoder
Attention
the
cat
sat
on
the
mat
Context
Interm.
Encoder (Recurrent, Convolutional, Self-Attention, …)
Input
the
cat
sat
on
the
mat



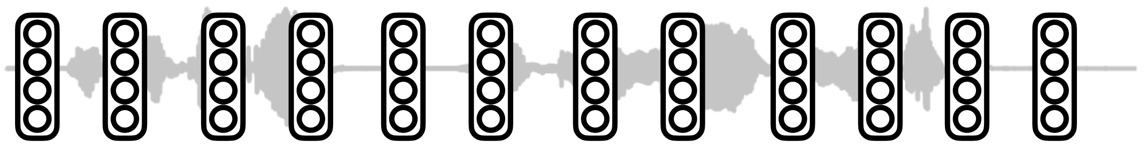
Decoder
the
cat
sat
on
the
mat
Attention
Context
Feed Forward
el
la
gato
gata
…
…
Interm.
Encoder (Recurrent, Convolutional, Self-Attention, …)
Input
the
cat
sat
on
the
mat

An Audio-Input model

el gato se sentó en la afobra

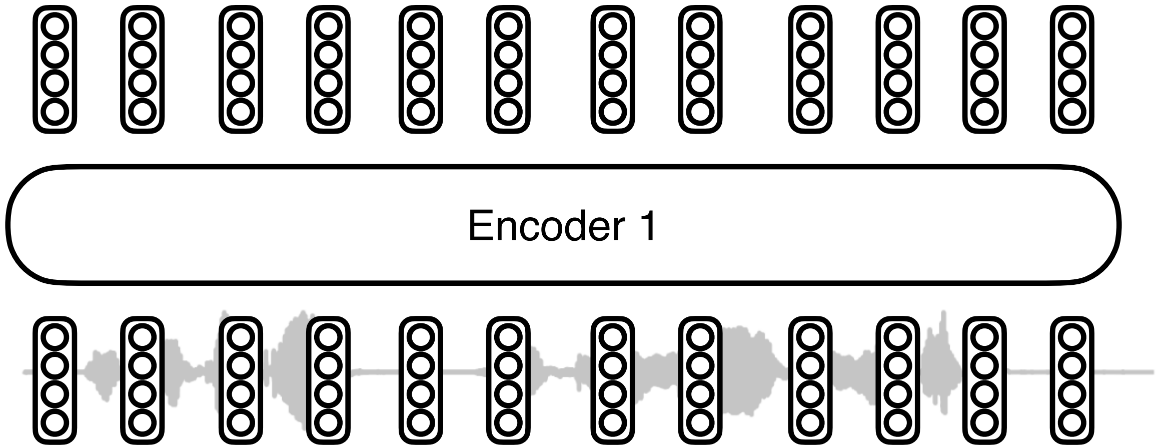
An Audio-Input model

el gato se sentó en la afobra

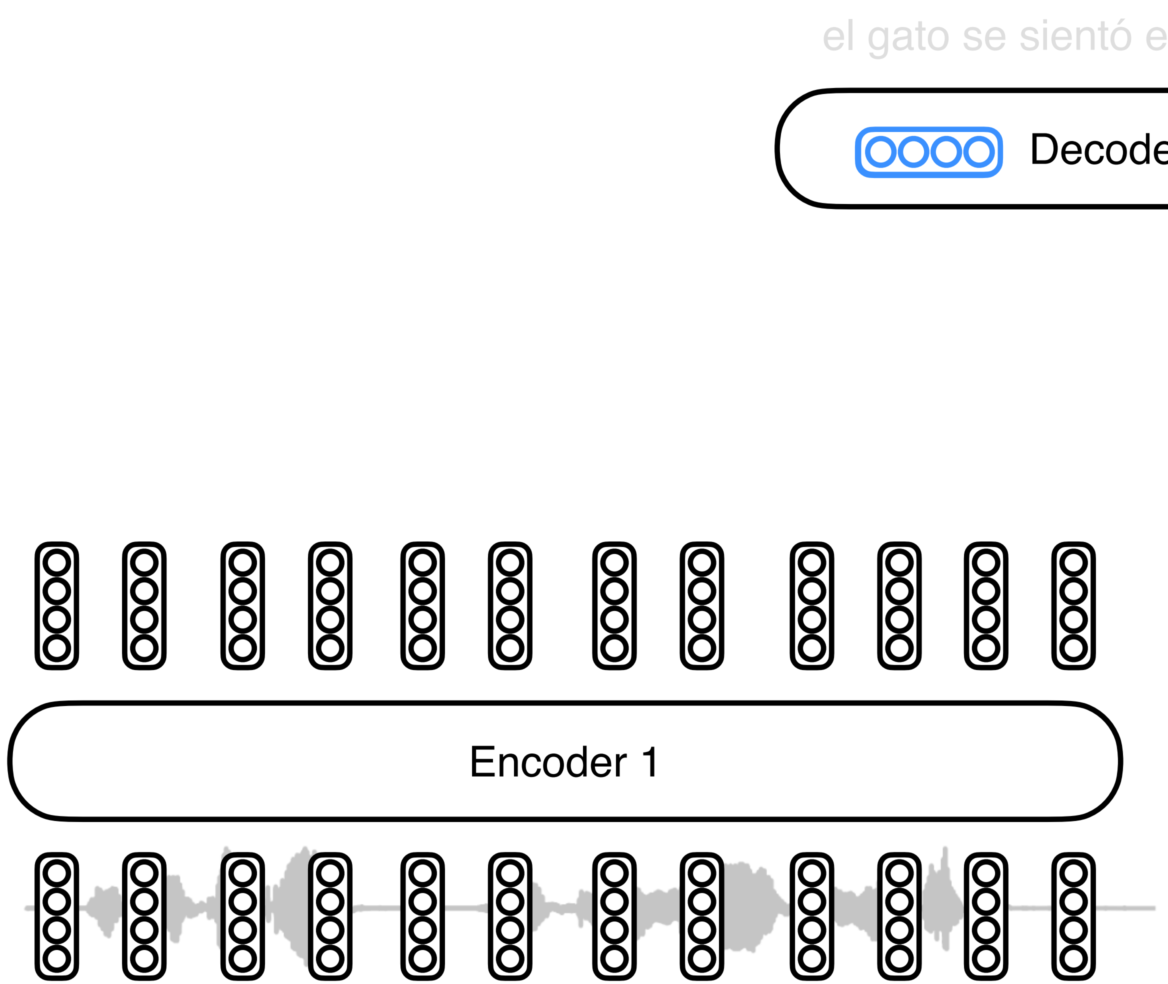


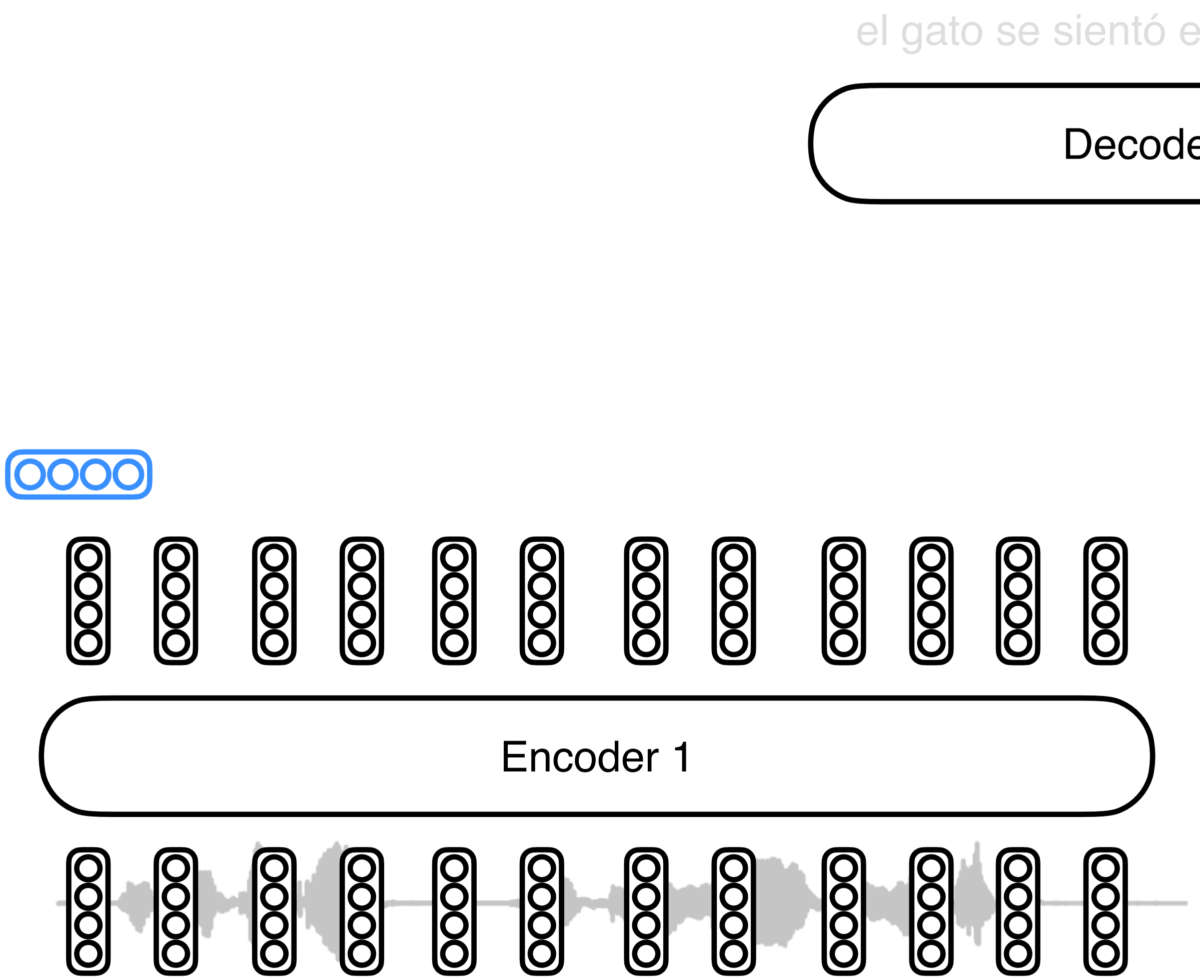
An Audio-Input model

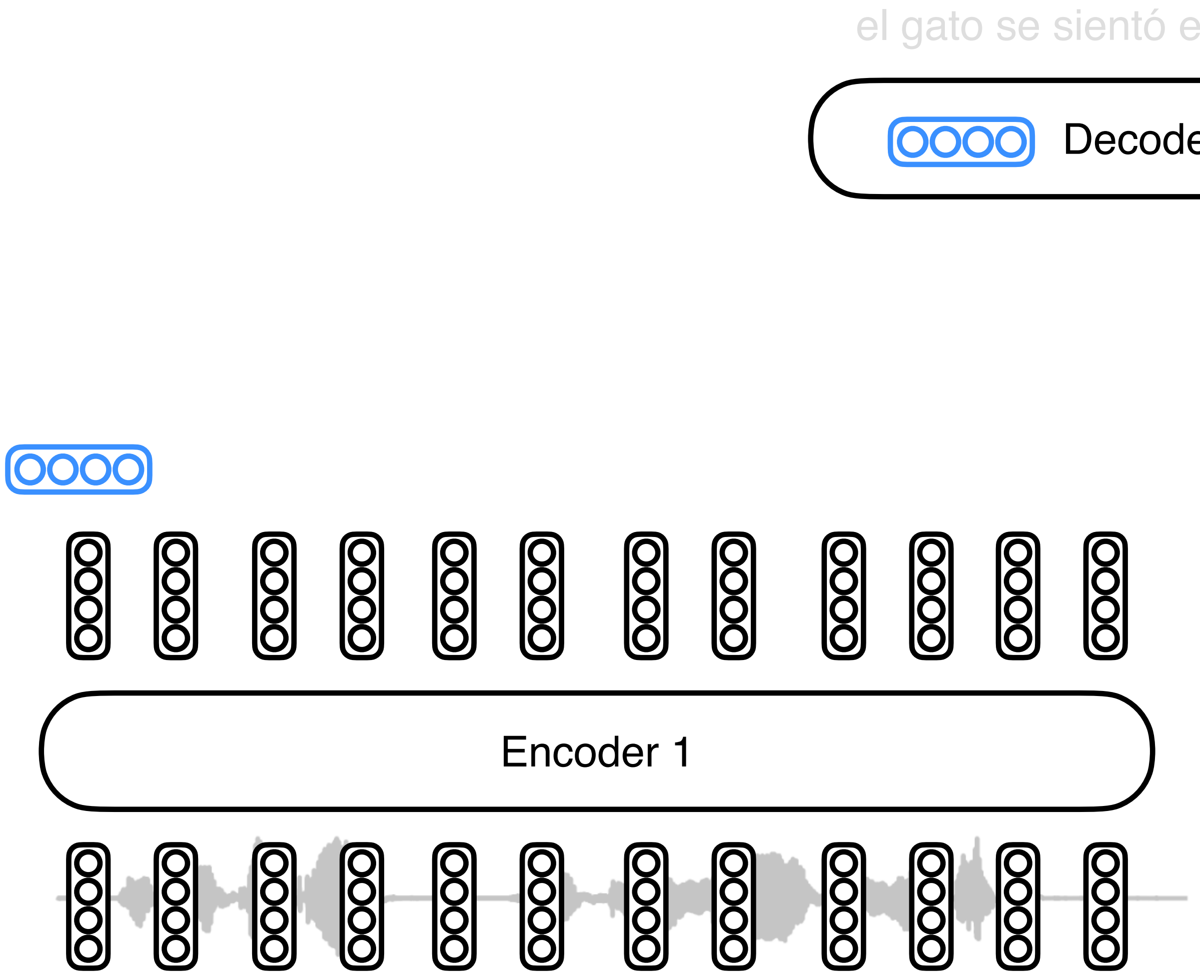
el gato se sentó en la afobra

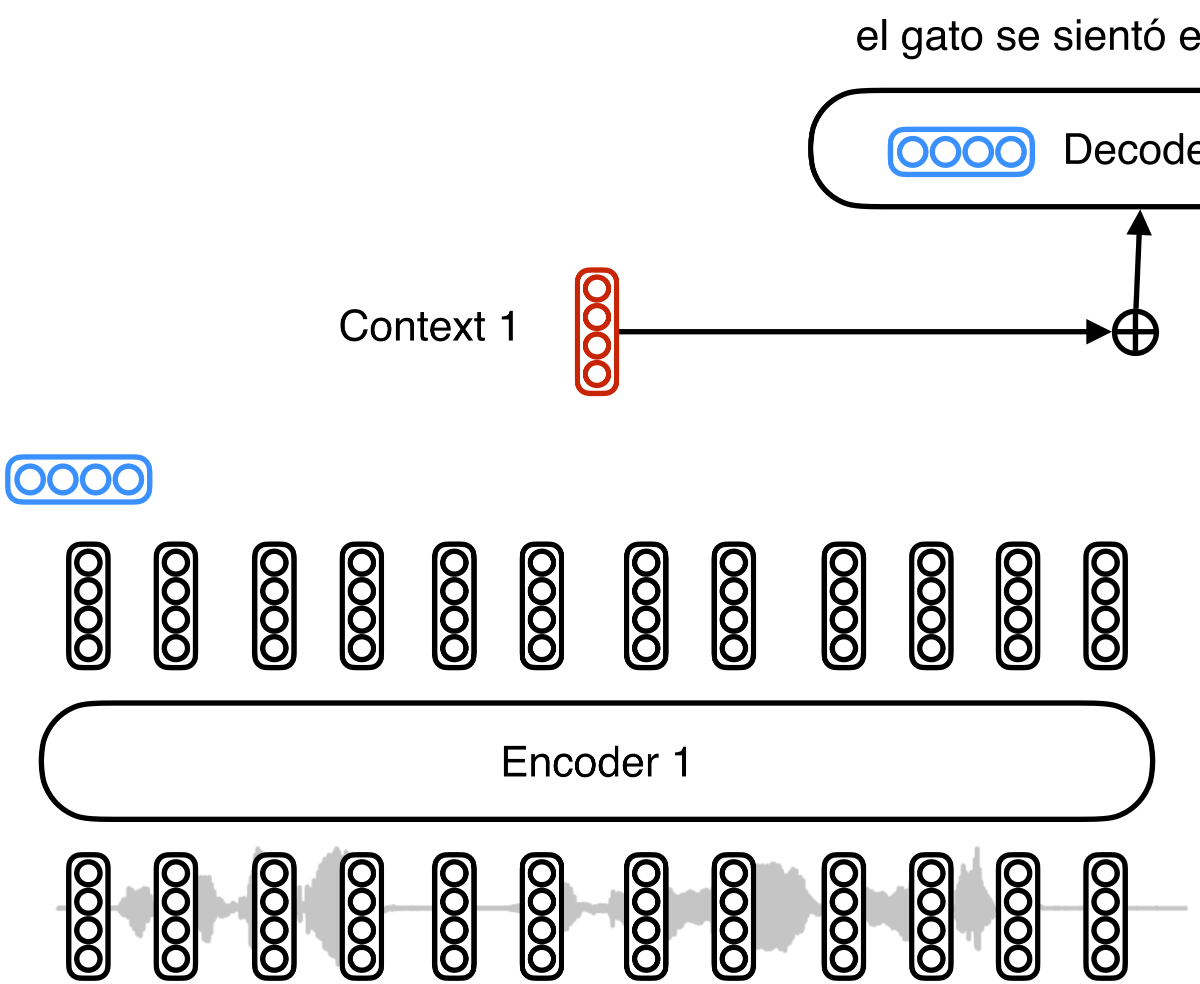


An Audio-Input model



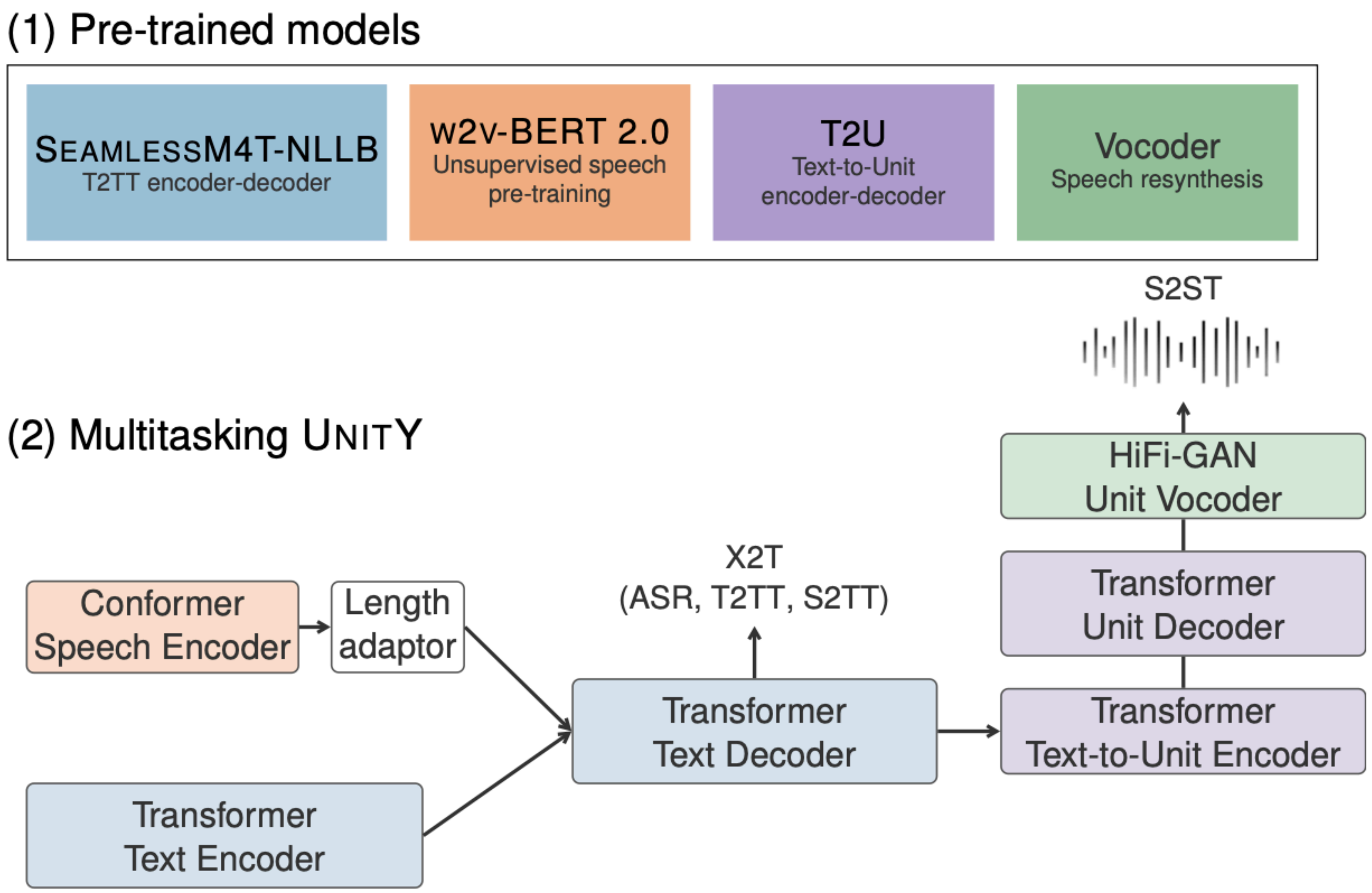

el gato se sentó en la afobra
Decoder
Encoder 1


el gato se sentó en la afobra
Decoder
Encoder 1


el gato se sentó en la afobra
Decoder
Context 1
Encoder 1

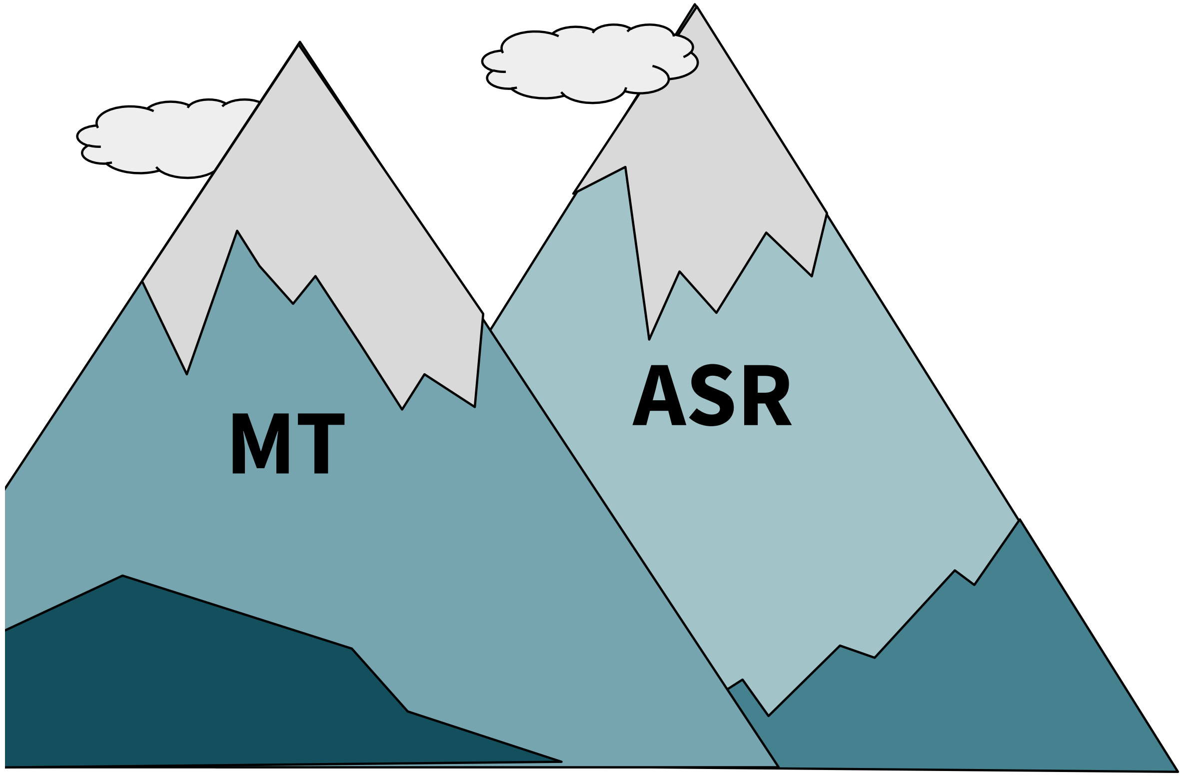
Today: pre-training

The SeamlessM4T model



Today: data mining

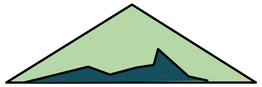
Recap: Available data



(text, translation) (audio, transcript)

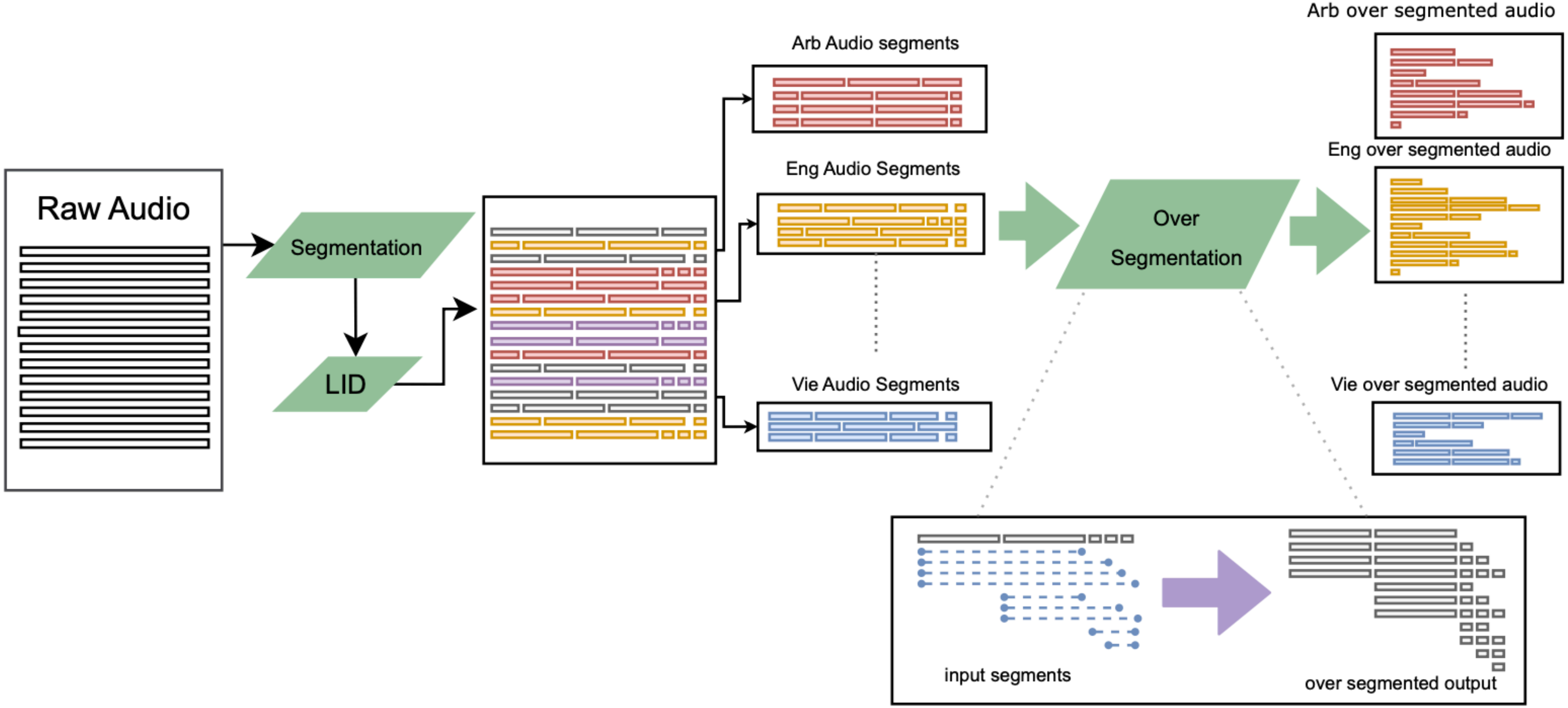
Can we make use of this large amount of data?

ST

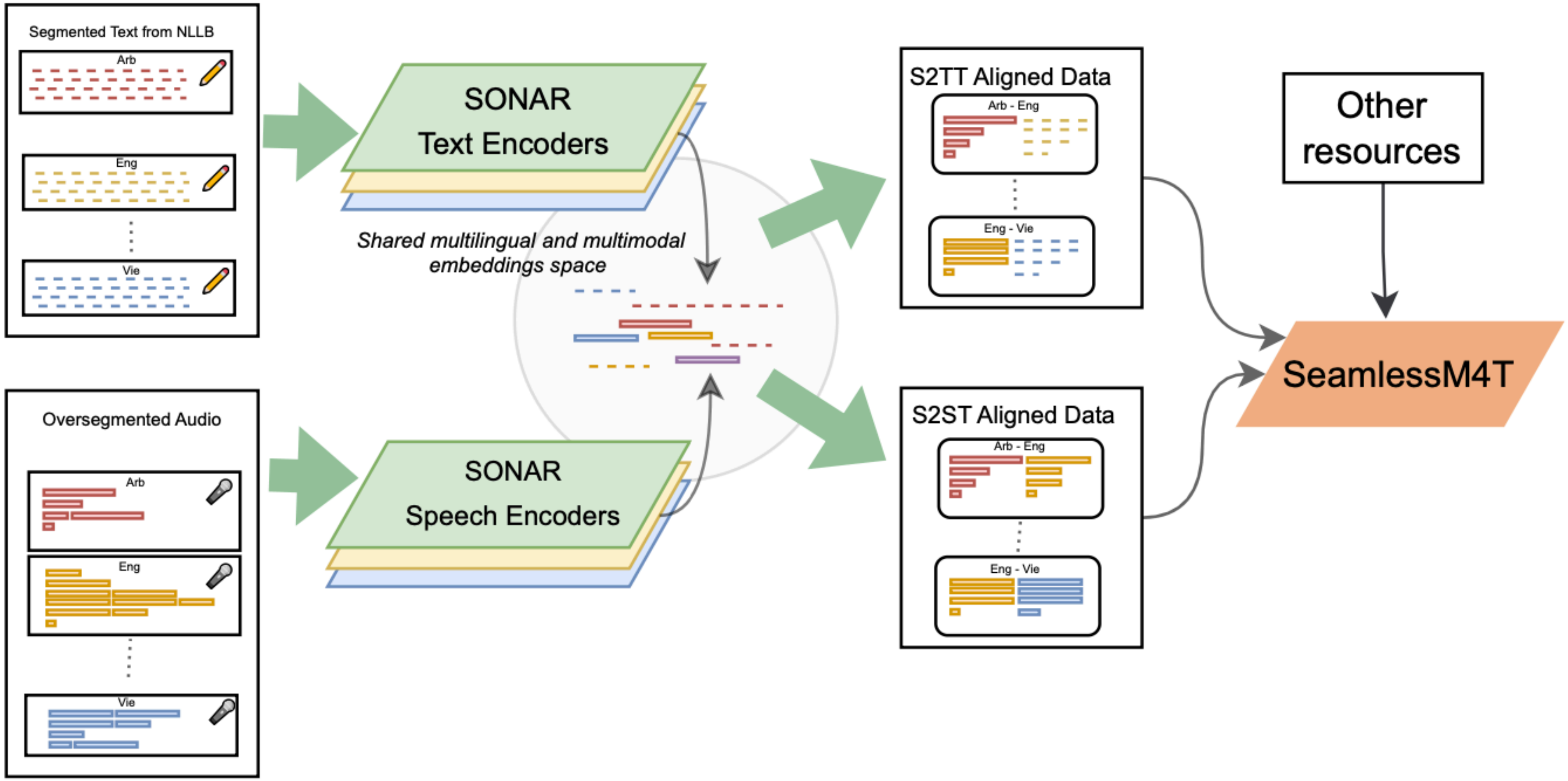


(audio, transcript, translation)

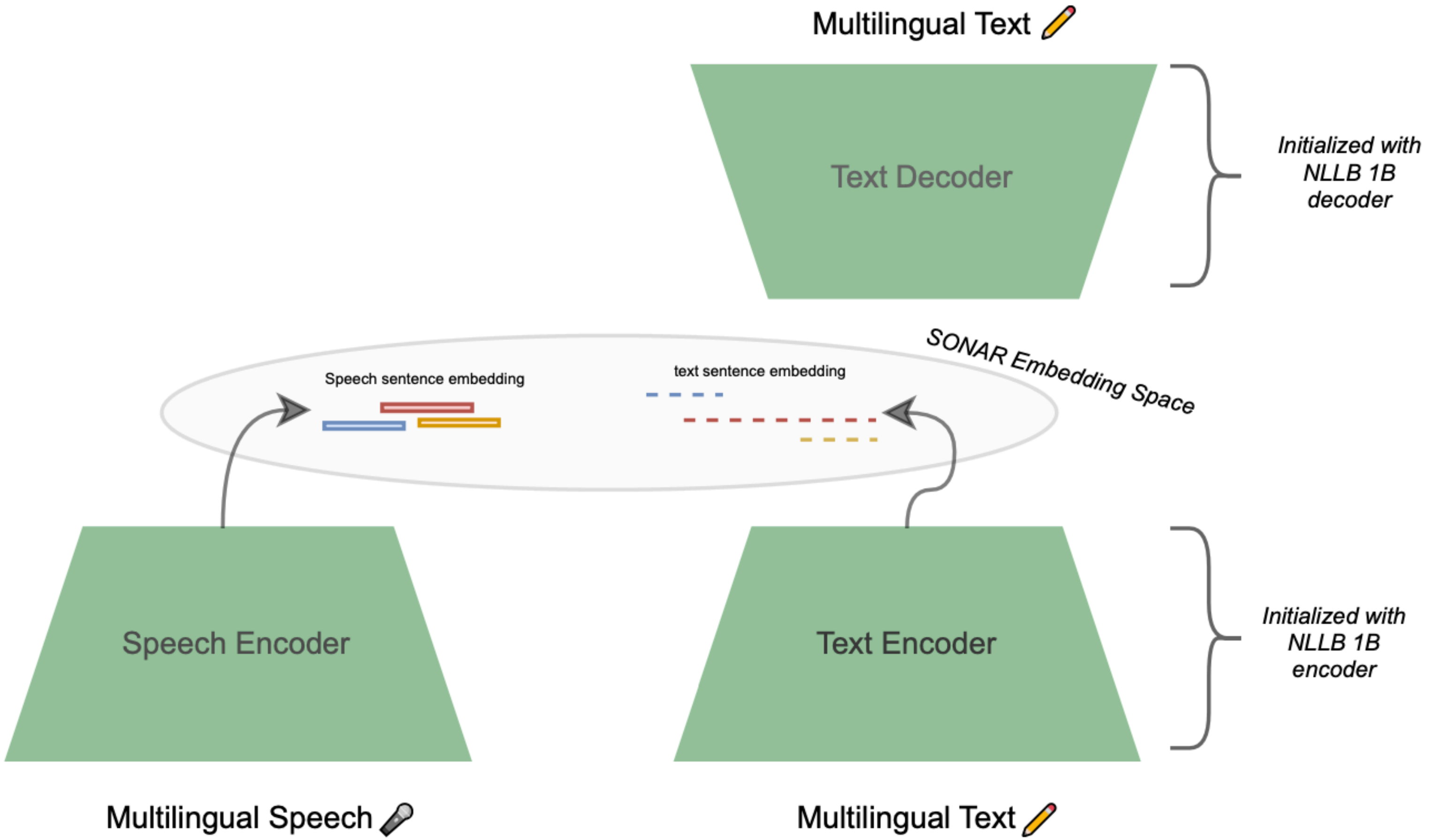
Mining Parallel Speech Data



SONAR Representations



SONAR Representations



SeamlessM4T Results

SeamlessM4T Results

tl;dr: it's great

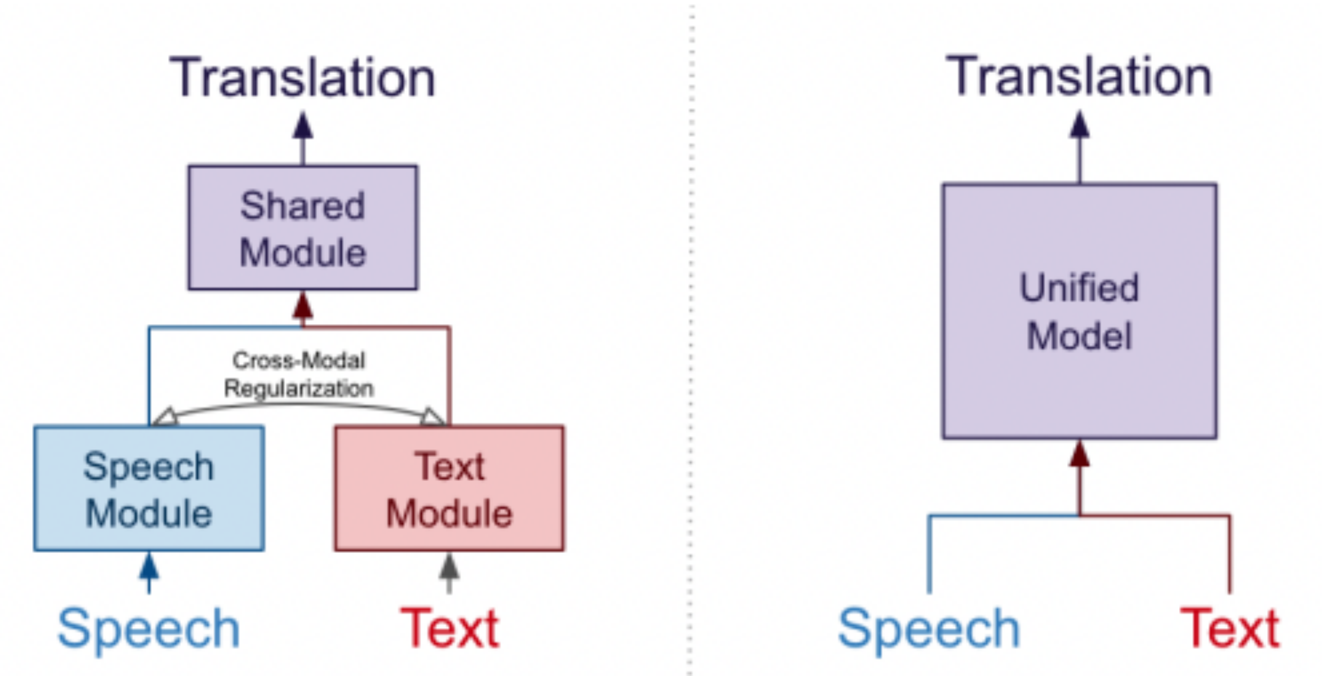
One Model to Rule them All (Yan et al, ICASSP ’24)


CROSS-MODAL MULTI-TASKING FOR SPEECH-TO-TEXT TRANSLATION VIA HARD PARAMETER SHARING

Brian Yan[1], Xuankai Chang[1], Antonios Anastasopoulos[3], Yuya Fujita[4], Shinji Watanabe[1,2]

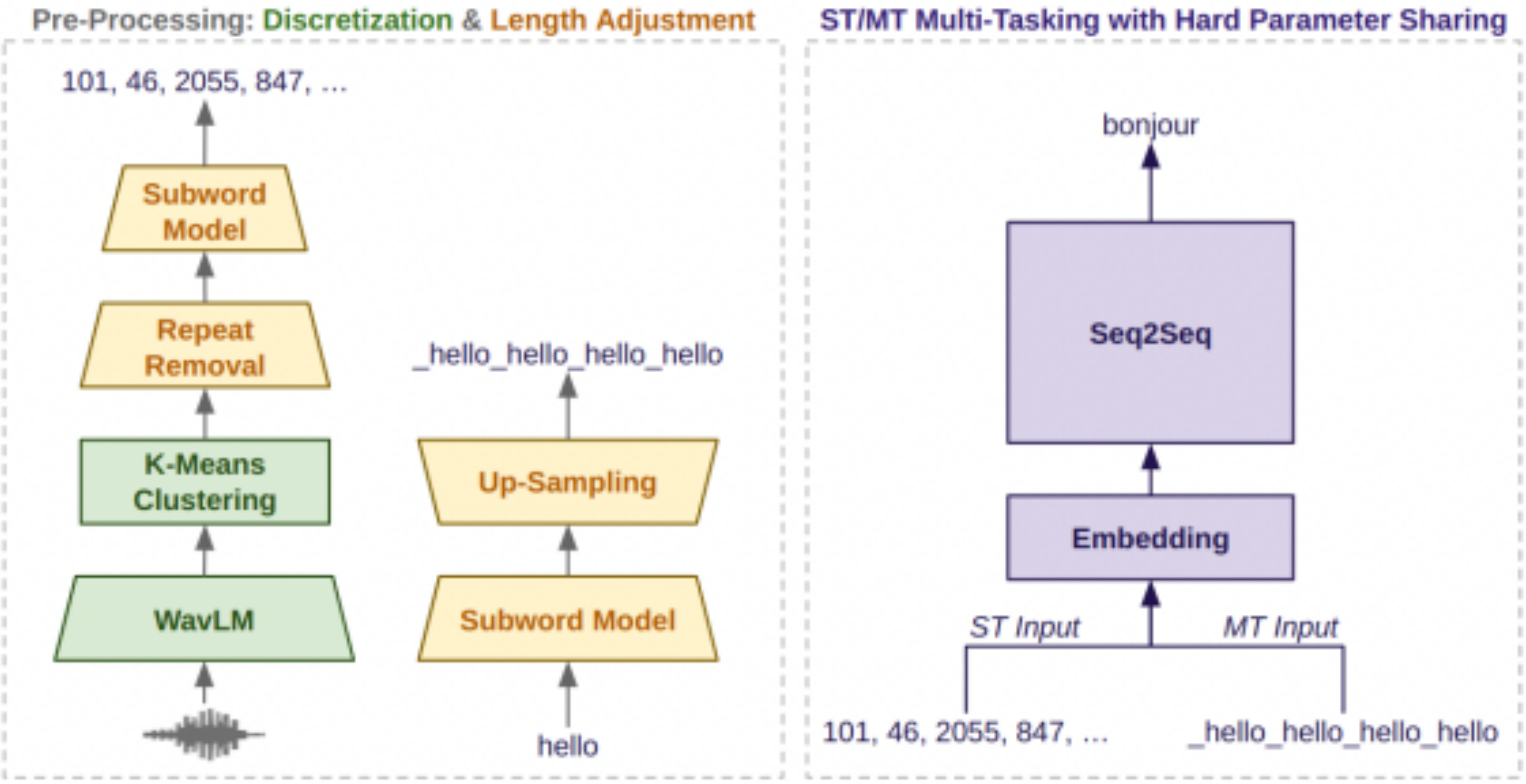
[1]Carnegie Mellon University, US, [2]Johns Hopkins University, US, [3]George Mason University, US, [4]Yahoo Japan Corporation, JP

ABSTRACT



GEORGE MASON UNIVERSITY

Problem: Different Granularities



Results

Results

tl;dr: significantly better than the 2-encoder architecture