



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

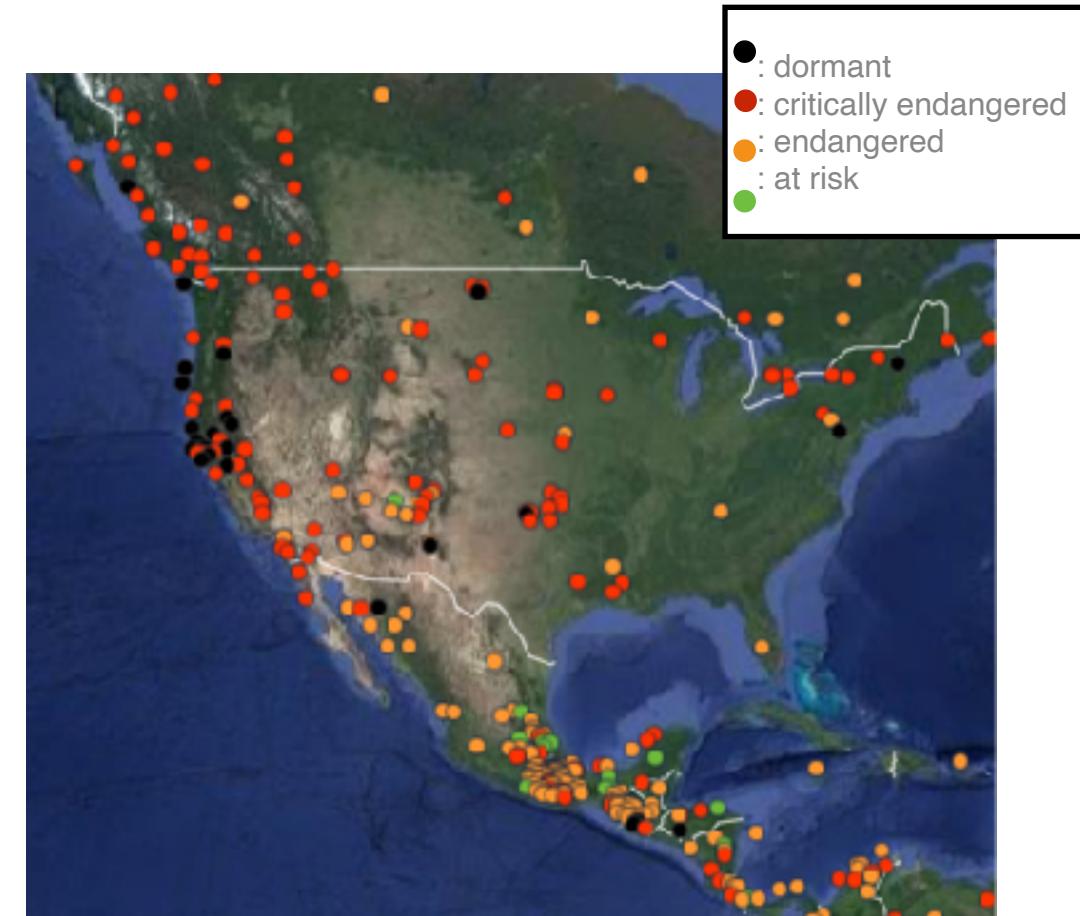


A traditional **Kyrgyz manaschi** performing part of the **Epic of Manas** at a **yurt** camp in **Karakol**

Image Source: Wikipedia

The Languages of the World

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

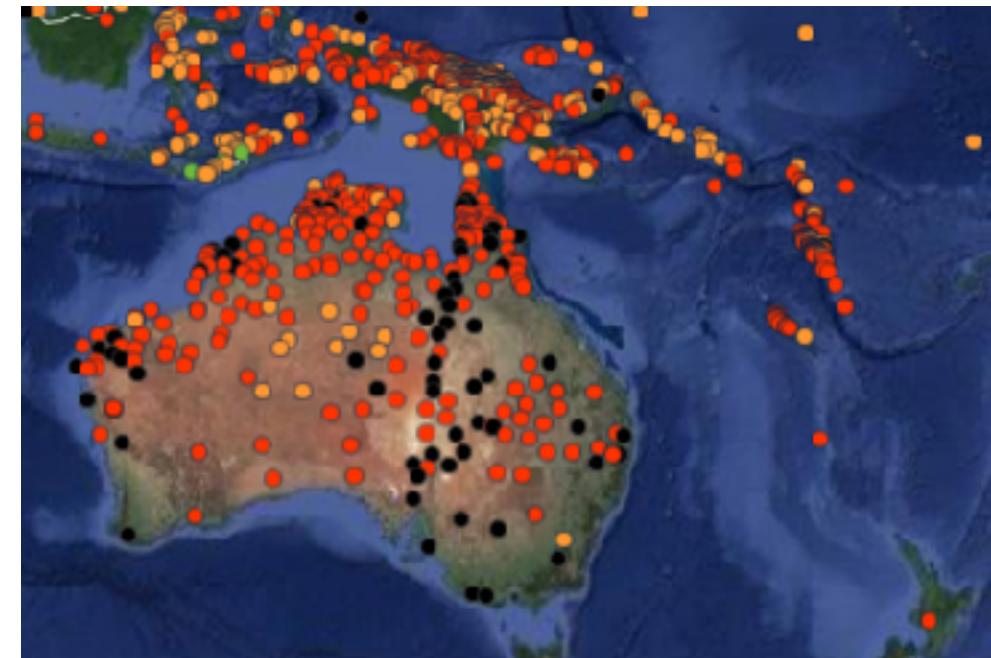


Source: the Endangered Languages Project

The Languages of the World

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

●	: dormant
●	: critically endangered
●	: endangered
●	: at risk

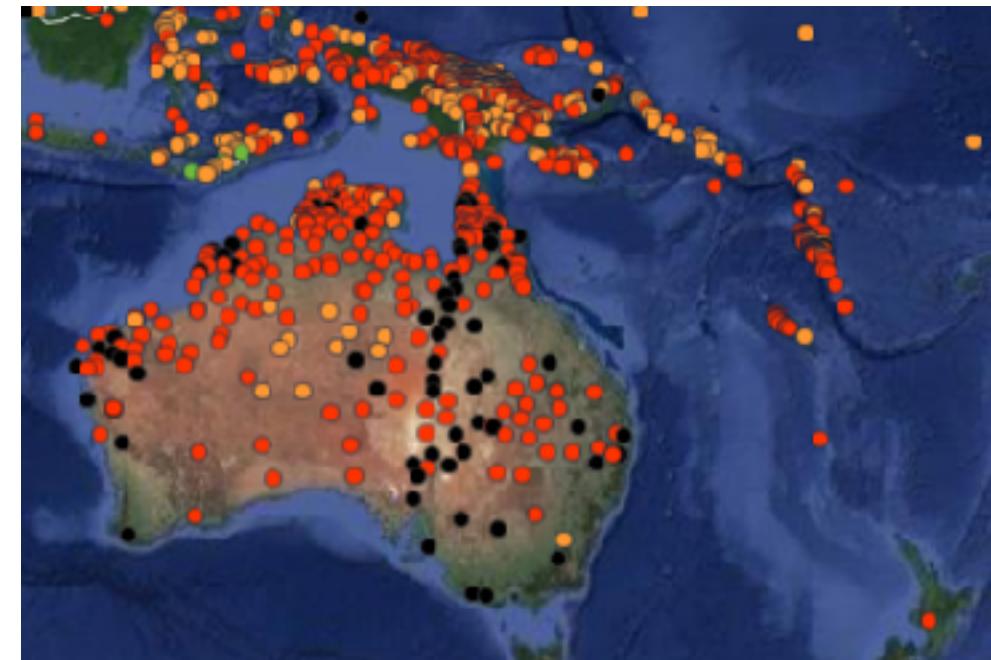


Source: the Endangered Languages Project

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...

- : dormant
- : critically endangered
- : endangered
- : at risk



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

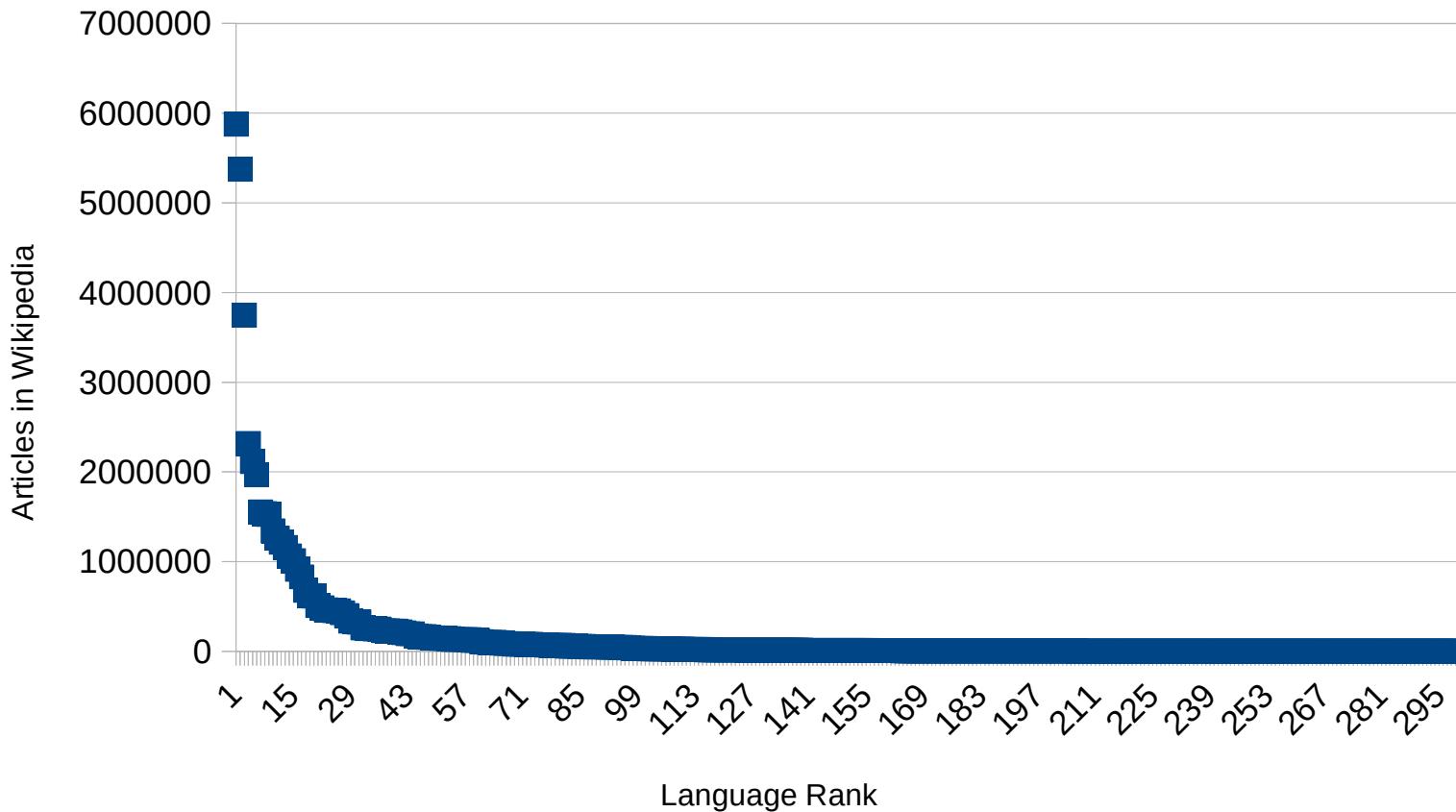
- More than **6000** languages:
 - 45% oral
 - 43% endangered or vulnerable
 - differences in culture, vocabulary
 - differences in morphological complexity, syntax, tonality, word order...

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
MIL-UPPV	P	64.7	20.8
UPC-TALP	P	62.1	23.0
NICT	P	53.3	29.1
Uhelinkki	C	52.8	28.6
Uhelinkki	P	52.0	29.4
Uhelinkki	C	51.0	33.1
NICT	C	47.9	33.4
TBC-NLP	P	46.1	36.0
UBC-NLP	C	46.1	35.9
MIL-UPPV	C	45.5	35.3
BNC	P	44.11	37.5

Table 25: 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	23.1
Uhelinki	C	59.1	24.5
Uhelinki	C	58.6	23.1
Uhelinki	P	58.4	23.3
KYOTOUNIVERSITY	P	56.9	26.9
NICT	C	56.9	28.4
BSC	P	54.8	25.8
UBC-NLP	P	52.3	31.9
UBC-NLP	C	52.2	31.8
MLLPUPV	C	51.9	30.5
MLLPUPV	C	49.7	31.1
BSC	C	48.5	31.1

Table 24: 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
Uhelinki	C	52.8	28.6
Uhelinki	P	52.0	29.4
Uhelinki	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 25: 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
MILIPUPV	P	66.6	19.7
NICT	P	59.9	23.3
Ihelsinki	C	59.1	23.5
Ihelsinki	C	58.6	23.1
Ihelsinki	P	58.4	23.3
KYOTOUNIVERSITY	P	56.9	26.9
NICT	C	56.9	28.4
BSC	P	54.8	25.8
UBC-NLP	P	52.3	31.9
UBC-NLP	C	52.2	31.8
MLLPUPV	C	51.9	30.5
MLLPUPV	C	49.7	31.1
BSC	C	48.5	31.1

Table 24: Results for Portuguese to Spanish Translation

Team	Type	BLEU	TER
MILIPUPV	P	64.7	20.8
UPC-TALP	P	62.1	23.0
NICT	P	53.3	29.1
Ihelsinki	C	52.8	28.6
Ihelsinki	P	52.0	29.4
Ihelsinki	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 25: Results for Spanish to Portuguese Translation

Team	Type	BLEU	TER
NITS-CNLP	C	53.7	36.3
Parligna-KMI	P	11.5	79.1
FMI-VHAN	P	11.1	79.7
UBC-NLP	P	06.2	77.1
UBC-NLP	C	06.2	77.2
NITS-CNLP	P	03.7	-
NITS-CNLP	C	02.6	-
CHLT-HTB	C	03.5	-
Parligna-KMI	C	03.1	-
CHLT-HTB	P	02.8	-
CHLT-HTB	C	02.7	-
Parligna-KMI	C	01.6	-
JUMT	P	01.4	-

Table 26: Results for Hindi to Nepali Translation

Many similarities to utilize

CASE STUDY: INDIAN SUBCONTINENT

ਏਹੋ ਵਾਕਾਤਿ ਕੀ ਵਲ? ਆ ਵਾਕਧ ਥੁੰ ਕਢੇ ਛੇ? ਜਾਂ ਵਾਕ੍ਯ ਦਿੰਨ ਹੈਂਝੁਭੁਦੇ? ਇਹ ਸਜ਼ਾ ਕੀ ਕਹਿੰਦੀ ਹੈ?

ਔਨ ਵਾਚਕਾਂ ਐਗਤਾਣਾਂ ਪਰਿਧੁਨਾਂਵਿੱਖਿ ਵਾਕਧ ਕਿਆ ਕਹਤਾ ਹੈ? ਹੇ ਵਾਕਧ ਕਾਧ ਮਹਣਾਂਤੇ?

ਉਡ ਵਾਕ੍ਯਾਂ ਏਹੀ ਚੇਖਾਤੁਂਦਿ? ਯੋ ਵਾਕਧਾਲੇ ਕੇ ਮਨਤਾਤਮਿ ਵਾਕਧਾਂ ਤਵਿਤਨੀਂਨੇ ਕ੍ਰਮਕੰਢੇ?

- 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?

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
駕 run	馬 horse	也	jiā
池 pool	水(water)	也	chí
施 impose	方 direction	也	shī
弛 loosen	弓 bow	也	chí
地 land	土 soil	也	dì
驅 drive	馬 horse	区	qū

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|ya|d'), 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

ماذا تعني هذه الجملة؟ what does this sentence mean?

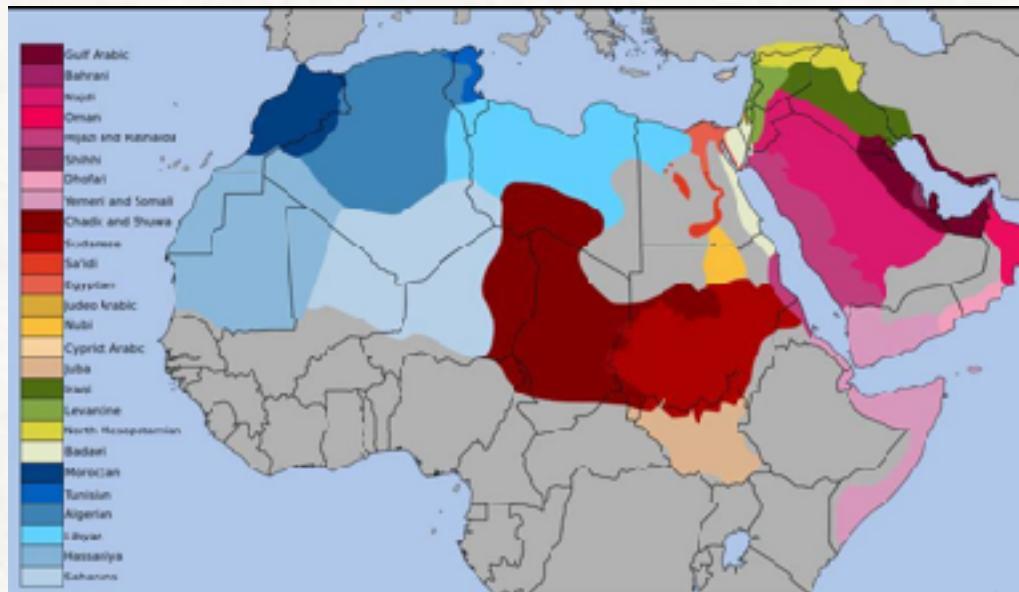
CASE STUDY: ARABIC

ماذا تعني هذه الجملة؟ what does this sentence mean?



CASE STUDY: ARABIC

ماذا تعني هذه الجملة؟ what does this sentence mean?



CASE STUDY: ARABIC

ماذا تعني هذه الجملة؟ what does this sentence mean?

Issue: Root-and-Pattern morphology

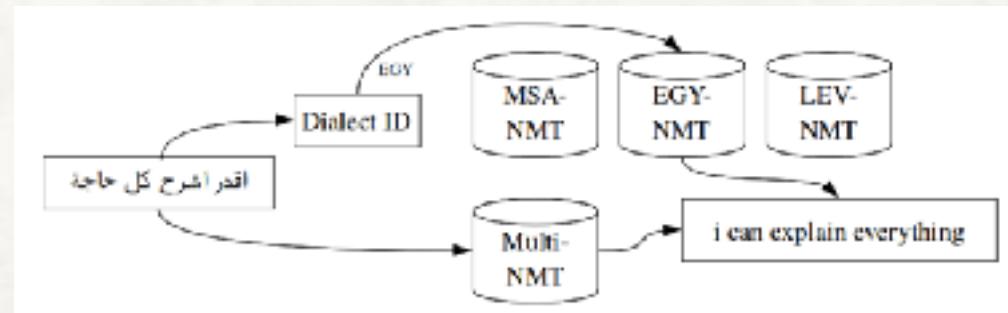
Solution: Morphological Analysis and Disambiguation

<i>Input</i>	wsynhY	Airjys	jwlth	bzyArp	AIY	trkyA.
<i>Gloss</i>	and will finish	the president	tour his	with visit	to	Turkey
<i>English</i>	The president will finish his tour with a visit to Turkey.					
<hr/>						
ST	wsynhY	Airjys	jwlth	bzyArp	AIY	trkyA.
D1	w+ synhy	Airjys	jwlth	bzyArp	<IY	trkyA.
D2	w+ s+ ynhY	Airjys	jwlth	b+ zyArp	<IY	trkyA.
D3	w+ s+ ynhY	Al+ rjys	jwl +P _{ams}	b+ zyArp	<IY	trkyA.
MR	w+ s+ y+ nhY	Al+ rjys	jwl +p +h	b+ zyArp +p	<IY	trkyA.
EN	w+ s+ >nhY< _{VP} > +S _{ams}	Al+ rjys< _{NN}	jwl p< _{NN} > +P _{ams}	b+ zyArp< _{NN}	<IY< _{NN}	trkyA< _{NN} >.

CASE STUDY: ARABIC

ماذا تعني هذه الجملة؟ what does this sentence mean?

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@@ting 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)

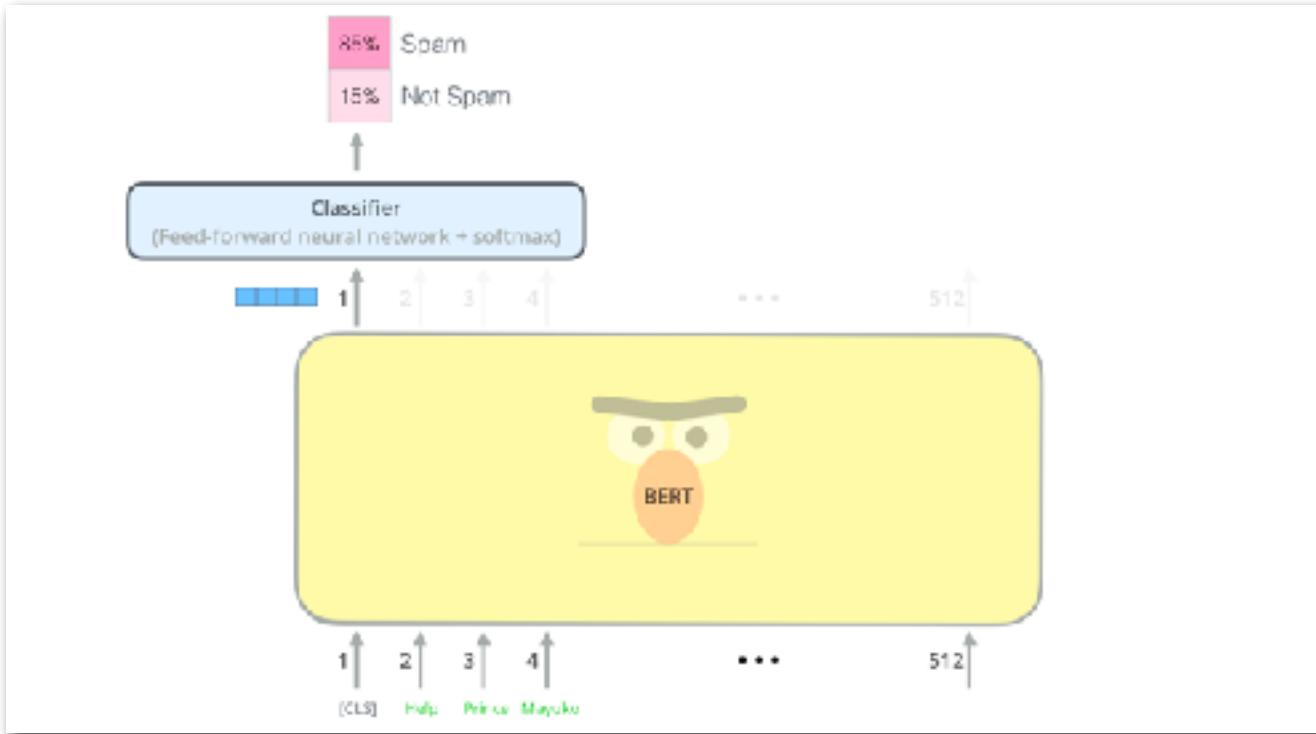


USING RELATED LANGUAGES

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

~~Gemini~~

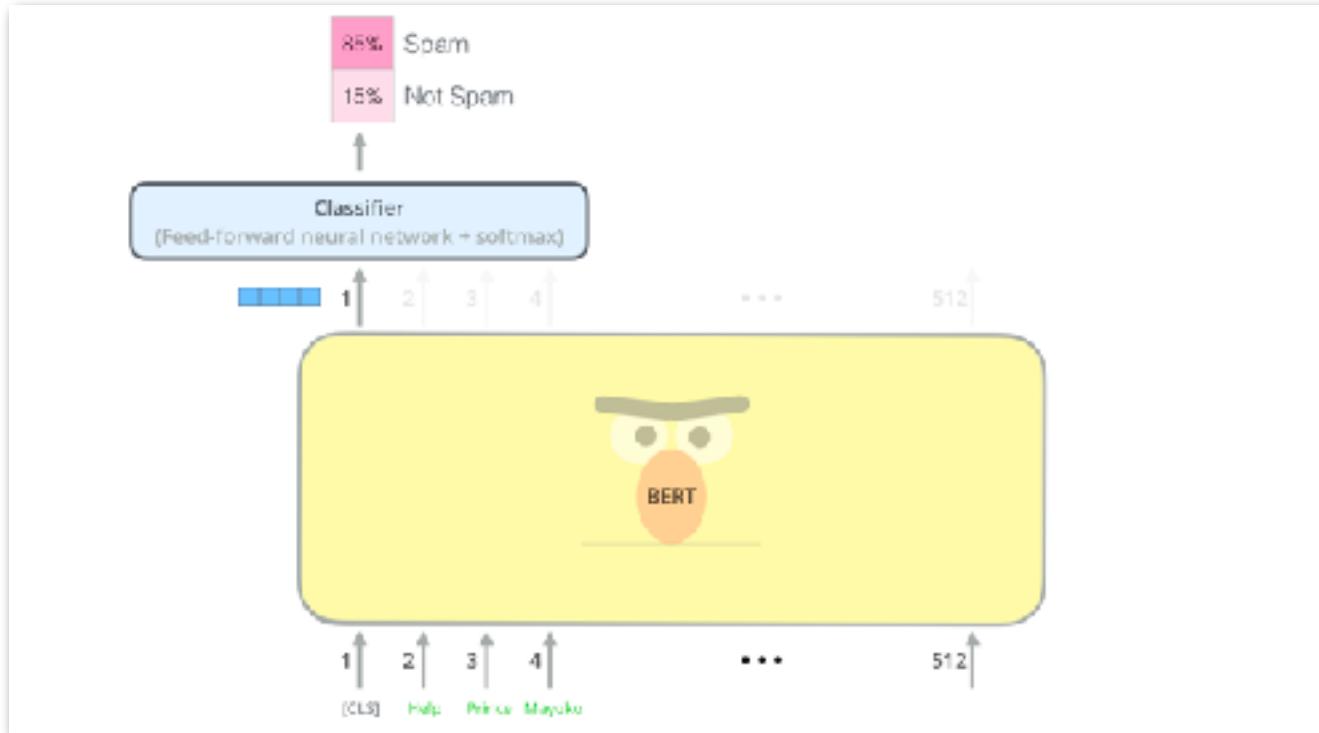
~~GPT-4~~

~~Claude~~

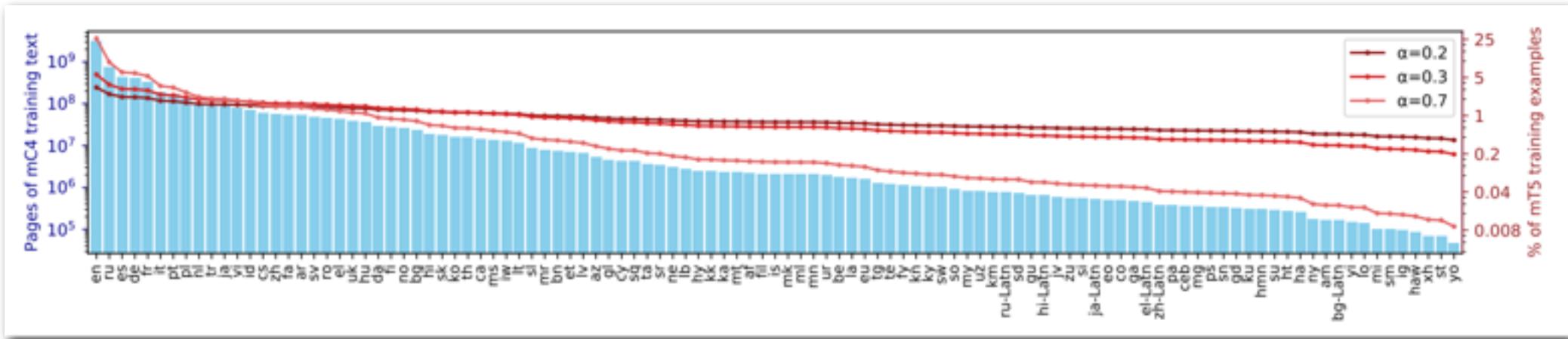
~~PaLM~~

~~GPT-2~~

~~XLM-R~~



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

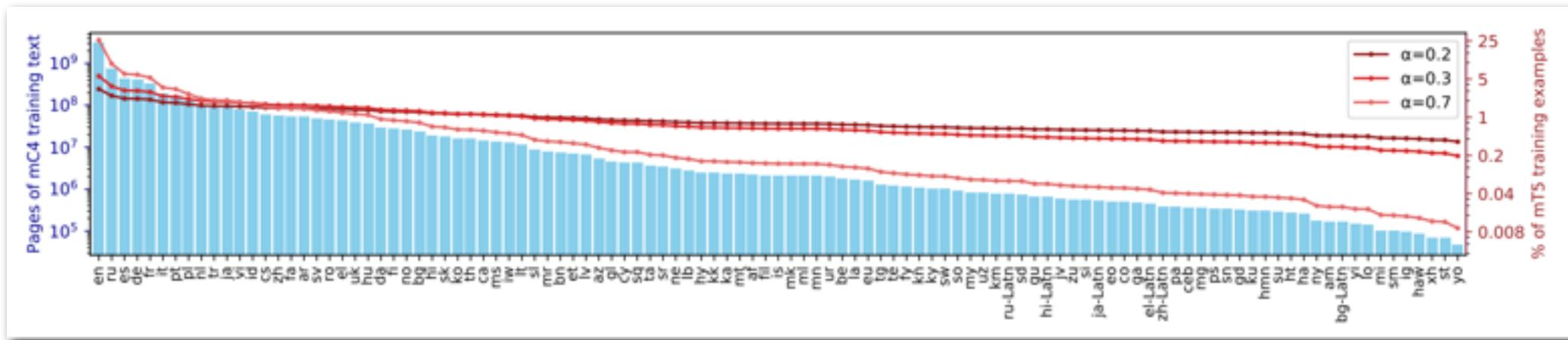
Train on all the internet (GPT-4?) → *incidental multilingualism*
or

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



Getting Data - Internet Crawling

Getting Data - Internet Crawling

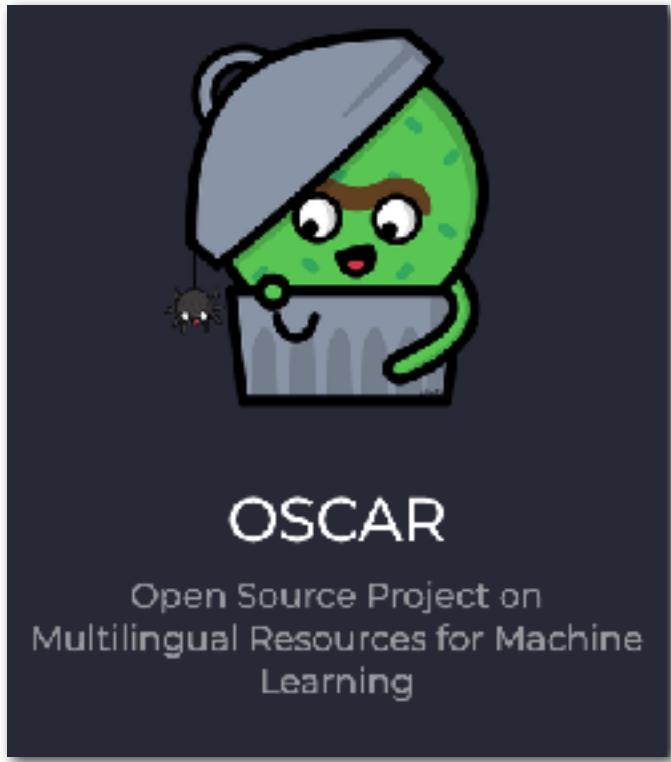
Getting Data - Internet Crawling



OSCAR

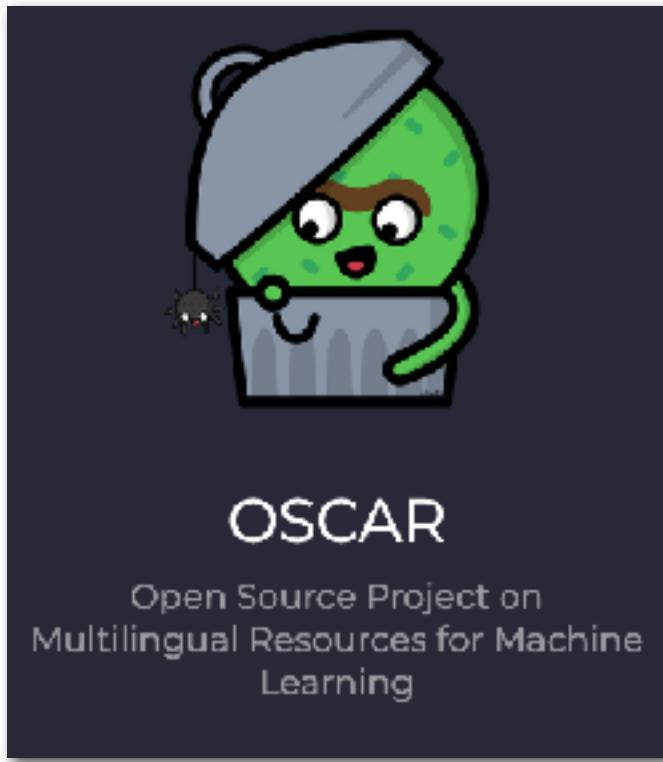
Open Source Project on
Multilingual Resources for Machine
Learning

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



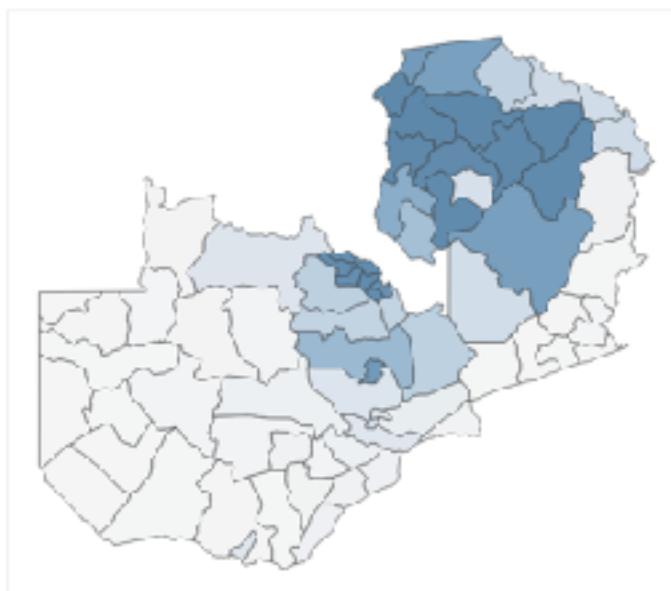
Our Solution: Work with Communities

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

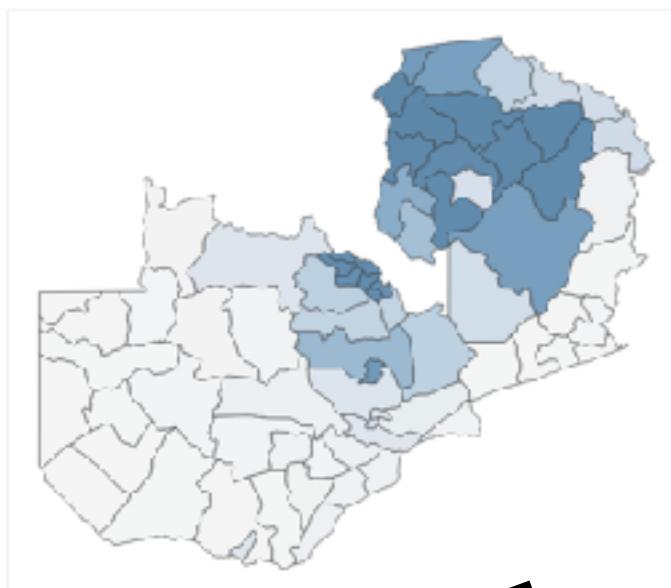
language	# of speakers	% of population
Bemba	3,727,677	29.94
Tonga	1,565,077	13.39
Tumbuka	1,445,111	11.29
Chewa (Nyanja)	1,305,484	10.98
ZL	741,755	6.29
UNSI	560,693	4.69
Other language	464,474	3.69
Swahili	416,763	3.29
Lozi	379,648	3.09
Nyanwanga	269,337	2.29
Wanga	252,814	2.19
Mambwe-Lun	256,006	2.19
Gunda	189,171	1.59
Lamia	189,059	1.59
Gundja	172,380	1.39
Si	150,906	1.29
Zulu (Kwembe)	112,007	1.09
Suli	60,303	0.51
Mwenda	77,004	0.69
English	67,818	0.69
Fulani	42,921	0.39
Anga	46,776	0.41
Swila	20,056	0.21
Wadi	44,625	0.39
Simba	40,983	0.35
Ntuya	29,116	0.25

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	29.9%
Tonga	1,565,077	13.3%
Tumbuka	1,445,111	11.2%
Chewa (Nyanja)	1,305,484	10.3%
Lozi	741,755	5.7%
Lunda	560,693	4.6%
Other languages	464,474	3.6%
Swahili	410,783	3.2%
Lozi	379,648	2.5%
Nyamwanga	260,337	2.2%
Shanga	252,814	2.0%
Mambwe-Luvale	276,006	2.3%
Gunda	189,171	1.5%
Lamia	189,059	1.5%
Gundja	172,360	1.3%
Si	150,906	1.2%
Zulu (Kwembi)	112,001	1.0%
Soli	60,303	0.5%
Mwenda	77,024	0.6%
English	67,818	0.6%
Fulani	42,921	0.4%
Arwa	46,776	0.4%
Swati	20,056	0.2%
Wadi	44,625	0.2%
Siriba	40,985	0.2%
Ntsoya	29,116	0.2%

BIG-C: a Multimodal Multi-Purpose Dataset for Bemba
Cayenne Sikauze¹, Kester Mhembere², Mat Malina (de Almeida)³, Antonio Antunes^{4,5*}
¹Department of Computer Science, University of Zambia, Zambia
²Department of Literature and Languages, George Mason University, USA
³Department of Computer Science, George Mason University, USA
⁴Layton, (lkauzteleka, uta, zt, antao)@gmu.edu

Abstract
BIG-C is a multimodal dataset for Bemba, one of the most spoken languages in Zambia. It contains 10,000 images and 10 hours of audio recordings. The dataset includes images of people speaking Bemba, audio recordings of Bemba speech, and corresponding transcripts. The dataset is designed for tasks such as image captioning, speech recognition, and multilingual NLP. The dataset is available at <http://tiny.cc/meyarw>.

BembaSpeech: A Speech Recognition Corpus for the Bemba Language
Clayton Sikauze¹ and Antonio Antunes^{2,3,4*}
¹Department of Computer Science, University of Zambia, Zambia
²Department of Computer Science, George Mason University, USA
³cayenne.sikauze@zim.ust.zm, antao@gmu.edu

Abstract
We present a large-scale speech recognition corpus named BembaSpeech, containing over 10 hours of oral speech in the Bemba language, written by low-resource language spoken by over 30% of the population in Zambia. To train a speech-to-text system for Bemba, we collected diverse gendered speakers recorded pre-training from various towns in Zambia, using a speaker diarization system. We used a standard English pre-trained gendered speaker diarization model using DeepSpeech on the raw speech data. We then used a speaker diarization system to identify speakers in the dataset and fine-tuned a speaker diarization system. Finally, we used a speaker diarization system to identify speakers in the dataset and fine-tuned a speech-to-text system. The system achieves a word error rate (WER) of 3.49%. Lastly, we also show that the corpus can be used for cross-lingual speech recognition, achieving a WER of 11.6% when trained on Bemba and tested on English.

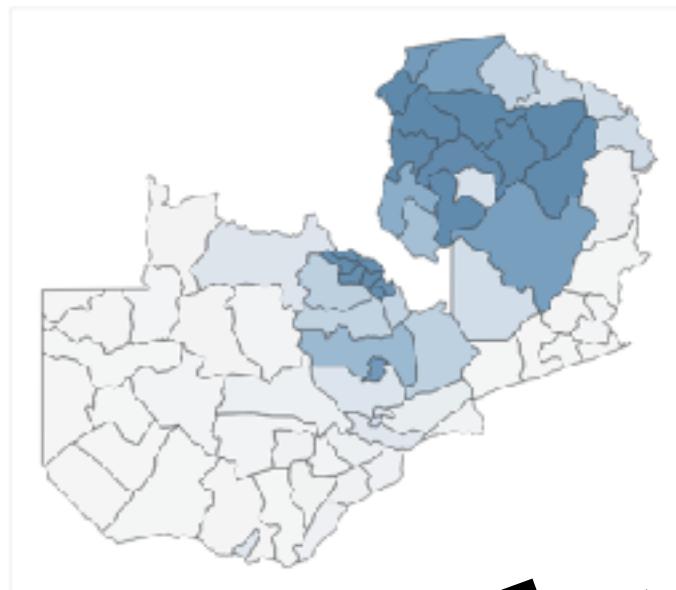
Keywords: Automatic Speech Recognition, ASR corpus, Low-resource languages, Bemba, Zambia

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.

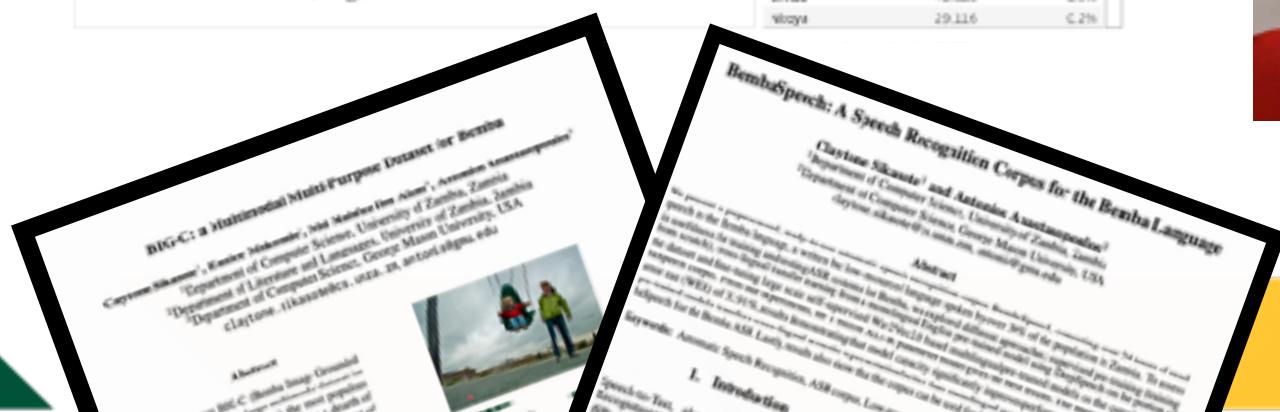
Bemba



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	29.9%
Tonga	1,565,077	13.3%
Tumbuka	1,465,111	11.2%
Chewa (Nyau)	1,305,484	10.3%
ZL	741,755	5.7%
UNSI	560,698	4.6%
Other language	464,474	3.6%
Swahili	416,763	3.2%
Lozi	379,548	3.0%
Nyanwanga	269,337	2.2%
Shona	252,814	2.0%
Mambwe-Lun	256,006	2.0%
Gunda	189,171	1.5%
Lamia	189,059	1.5%
Gundja	172,380	1.3%
Si	150,906	1.2%
Zulu (Kwam)	112,001	1.0%
Soli	60,303	0.5%
Mwenda	77,004	0.6%
English	67,018	0.6%
Fulani	42,921	0.4%
Anga	46,776	0.4%
Swati	20,056	0.2%
Wadi	44,625	0.2%
Sirba	40,985	0.2%
Ntsoya	29,116	0.2%

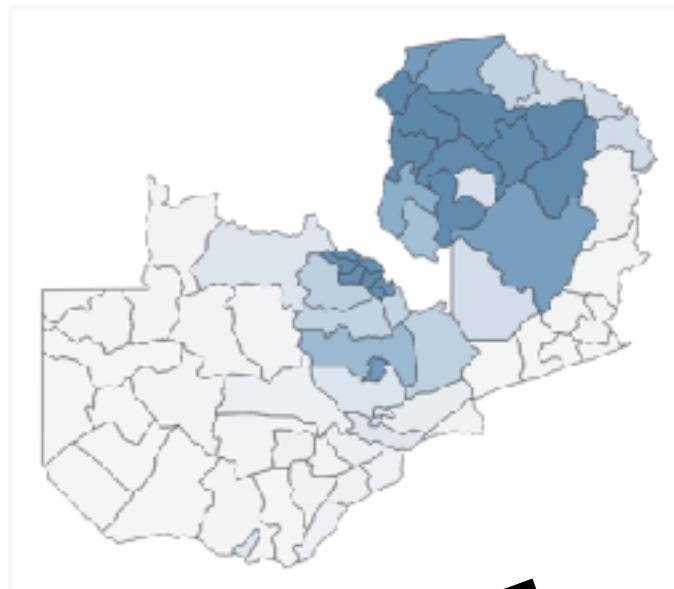


Our Solution: Work with Communities

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	3,727,637	29.9%
Tonga	1,565,077	12.3%
Tumbuka	1,465,111	11.2%
Chewa (Nyau)	1,305,484	10.3%
ZL	741,755	5.7%
UNDI	560,698	4.0%
Other language	464,404	3.6%
Swahili	416,763	3.2%
Lozi	379,548	2.9%
Nyanwanga	269,337	2.2%
Shona	252,814	2.0%
Mambwe-Lun	250,006	2.0%
Gunda	189,171	1.5%
Lamia	189,059	1.5%
Gundja	172,380	1.3%
Si	150,906	1.2%
Zulu (Kwazulu)	112,001	1.0%
Soli	60,303	0.5%
Mwenda	77,004	0.6%
English	67,818	0.6%
Taiwala	42,931	0.4%
Angula	46,776	0.4%
Swati	20,056	0.2%
Wadi	44,625	0.3%
Siriba	40,985	0.3%
Ntsoya	29,116	0.2%



BIG-C: a Multimodal Multi-Purpose Dataset for Bemba
Cayenne Nkamusela¹, Karine Mekhora¹, Matilde Jiménez¹, Antonio Antón-García², Clayton Sikasule¹, Antonia Antón-García²
¹Department of Computer Science, University of Zambia, Zambia
²Department of Literature and Languages, George Mason University, USA
clayton.sikasule@zim.ac.zm, matilde.mj@gsu.edu

BembaSpeech: A Speech Recognition Corpus for the Bemba Language
Clayton Sikasule¹ and Antonia Antón-García²
¹Department of Computer Science, University of Zambia, Zambia
²Department of Computer Science, George Mason University, USA
clayton.sikasule@zim.ac.zm, matilde.mj@gsu.edu

Abstract
We present a large-scale, multi-modal dataset for Bemba speech recognition, containing over 14 hours of raw speech in the Bemba language, written by low-resource language spoken by over 30% of the population in Zambia. To train a speech-to-text system for Bemba, we recorded different gender speakers separated pre-training from the test set. We also recorded a small dataset for training and testing ASR systems in Bemba, as a post-trained English pre-trained model using DeepSpeech on the raw speech. We used a multi-task learning approach to train our system. Our system achieves a word error rate (WER) of 3.40%. Finally, we discuss the challenges that model capacity significantly impacts performance.

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

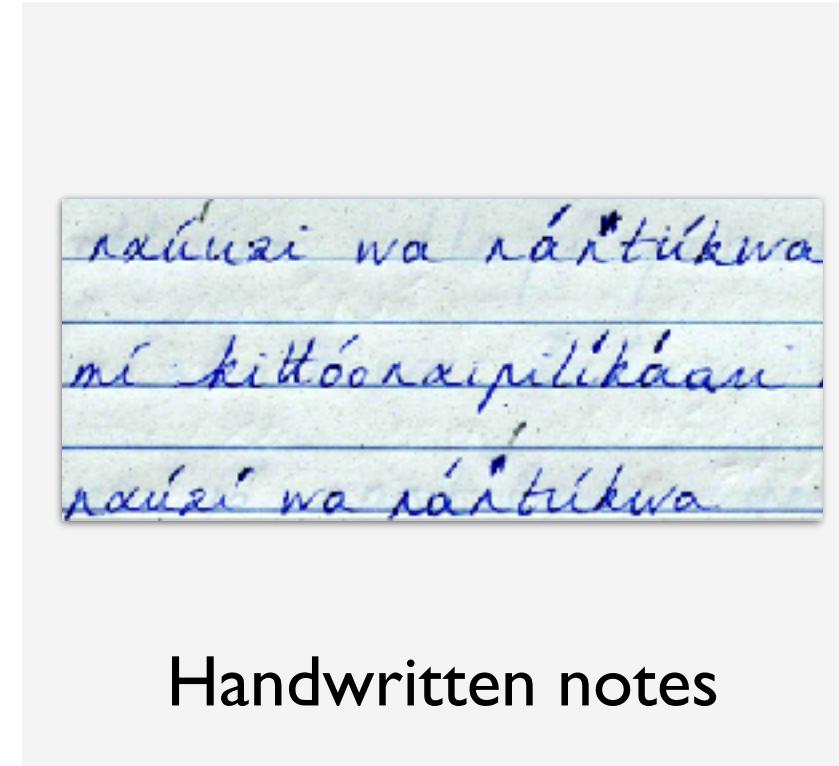
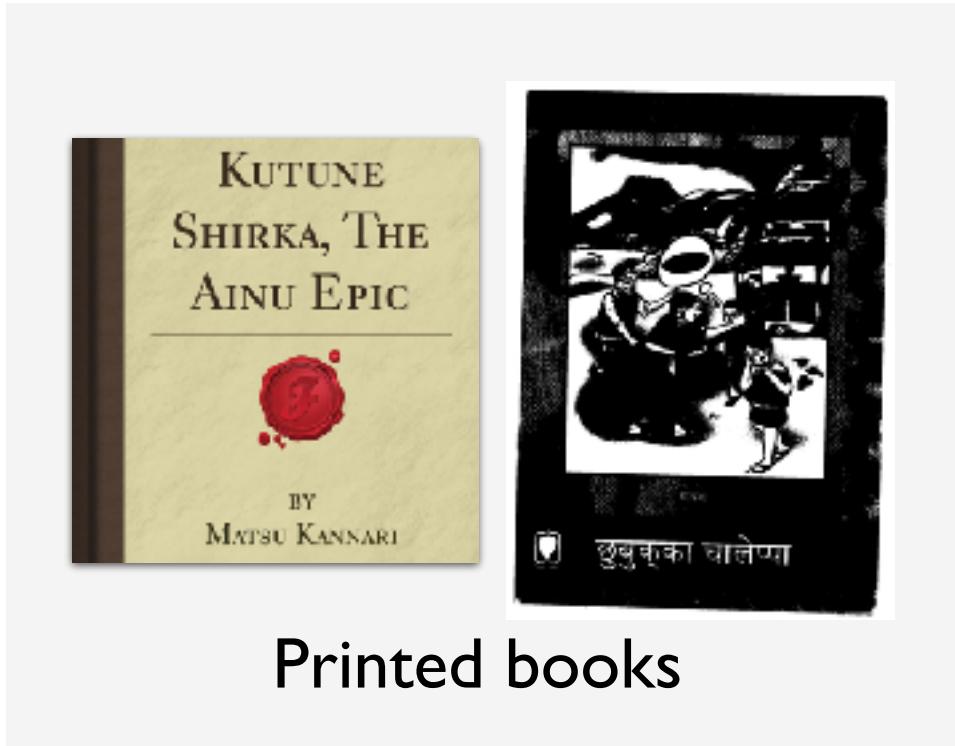
1. Introduction

Educational Tools for Mapuzugan

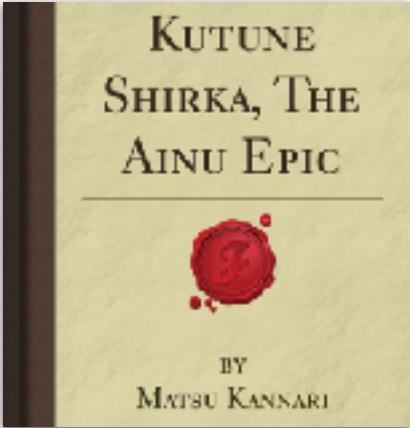
Cristian Almuñoz¹, Claudio Gutiérrez¹, Antonia Antón-García²
¹Department of Computer Science, Universidad de Chile
²Computer Science Department, George Mason University
almudaa.86@gmail.com cgutierrez@dcc.uchile.cl antonti@gsu.edu

Abstract
Mapuzugan 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 setting in order to aid carriers. In particular, it is undesirable that the development of NLP tools that reach the users lag further behind than NLP research itself (Bles et al., 2021).
In this work, we develop a tool for educational

Our Solution: Make Existing Data ML-Usable



Our Solution: Make Existing Data ML-Usable



Printed books

náúsei wa náitíkwa
ní kittóonxipilíkáau
náúsei wa náitíkwa

Handwritten notes

GCK Post Correction for Endangered Language Texts
Shozi Rajbhandari,^{1,2} Aastha Anantwani,^{1,2} Graham Neubig,¹
¹Language Technology Institute, Carnegie Mellon University, ²Microsoft Research, Redmond, WA, USA
arxiv.org/abs/1805.02010

Abstract
There is little to no data available to build accurate language processing models for endangered languages. However, written text is often available in their languages, either from the internet or as a curated dataset. In this work, we propose to use a set of existing text datasets of endangered languages to build a Morpheme-based model of morphology. We create a Morpheme-based dataset of grammatical features for 10 endangered languages. We then use this dataset to fine-tune pre-trained GCK and Transformer models to the downstream NLP tasks of part-of-speech tagging and named entity recognition. Our system can be trained on as few as 1000 words of training data and achieves state-of-the-art performance on the downstream NLP tasks.

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

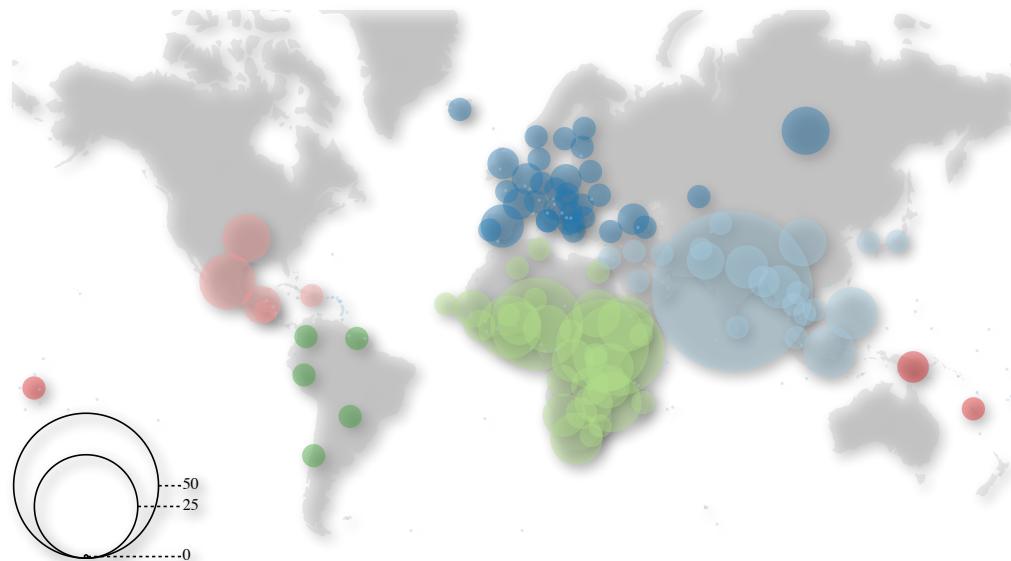
Let's get *small, but high quality* data



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 Mahfuzul Islam Antonios Ananiopoulos
Department of Computer Science, George Mason University
{magarwal, mmlam21, antonios}@gmu.edu

Abstract

Knowing the language of an input sentence is necessary for using almost every NLP tool such as tagger, parser, or translation system. Language identification is a challenging task. Language identification is a challenging task.

Language ID at Scale

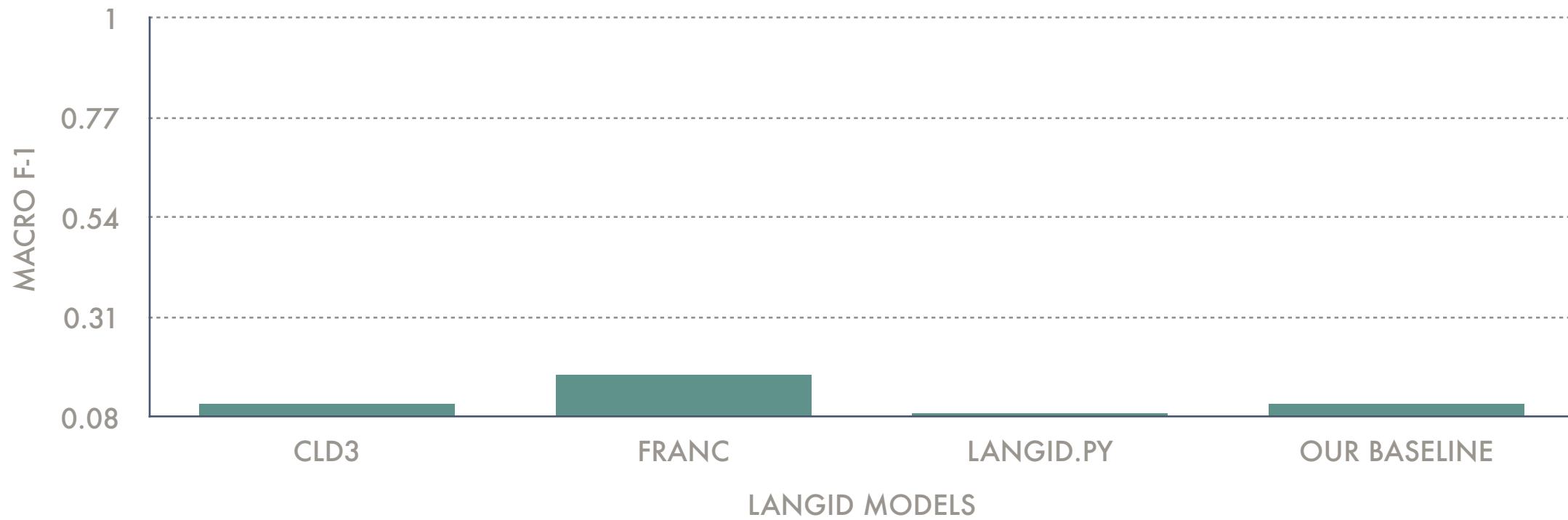
Benchmarking most popular models

Language ID at Scale

Benchmarking most popular models

Language ID at Scale

Benchmarking most popular models





Dialects

Languages are not Monoliths

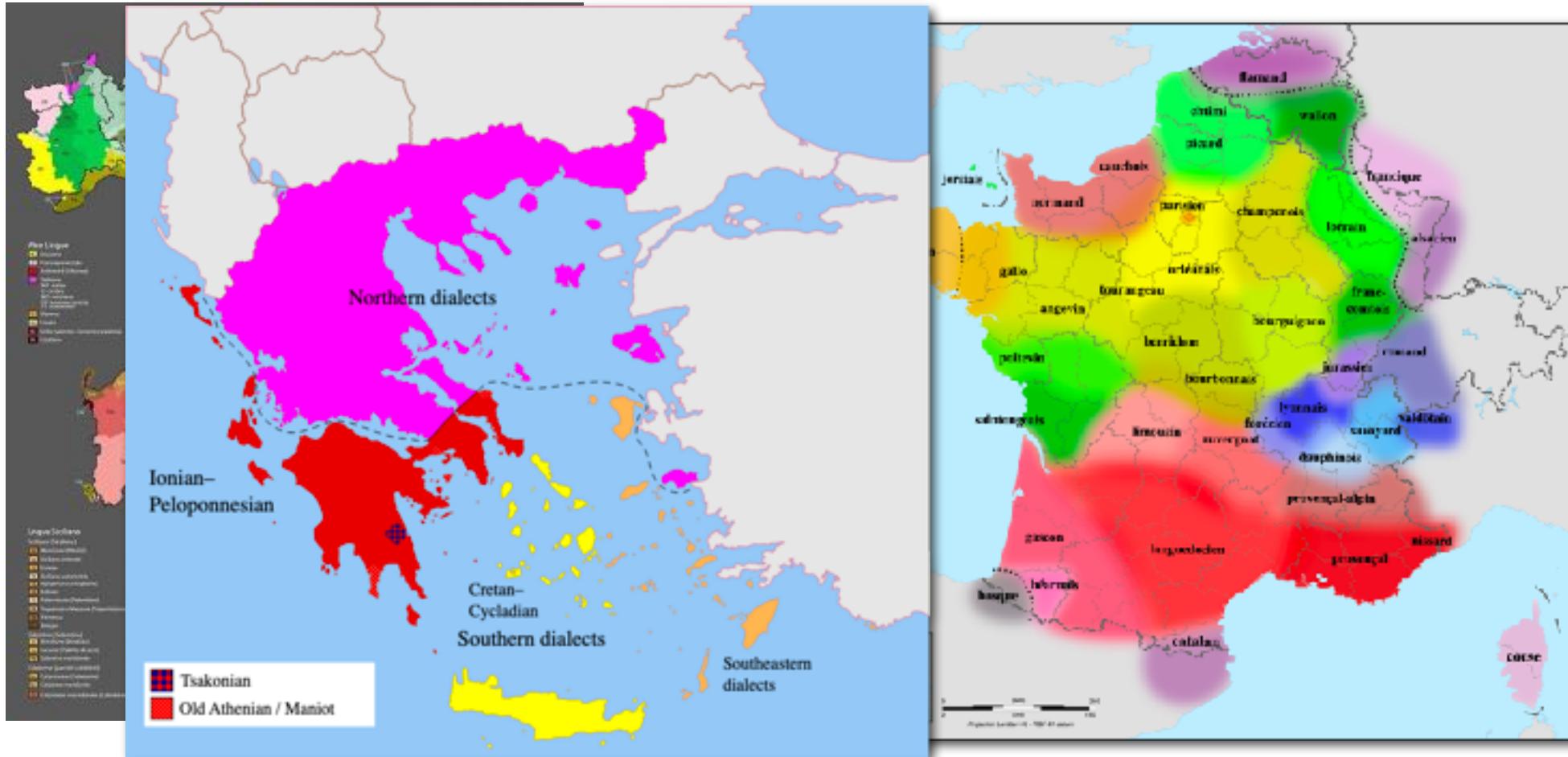
Languages are not Monoliths



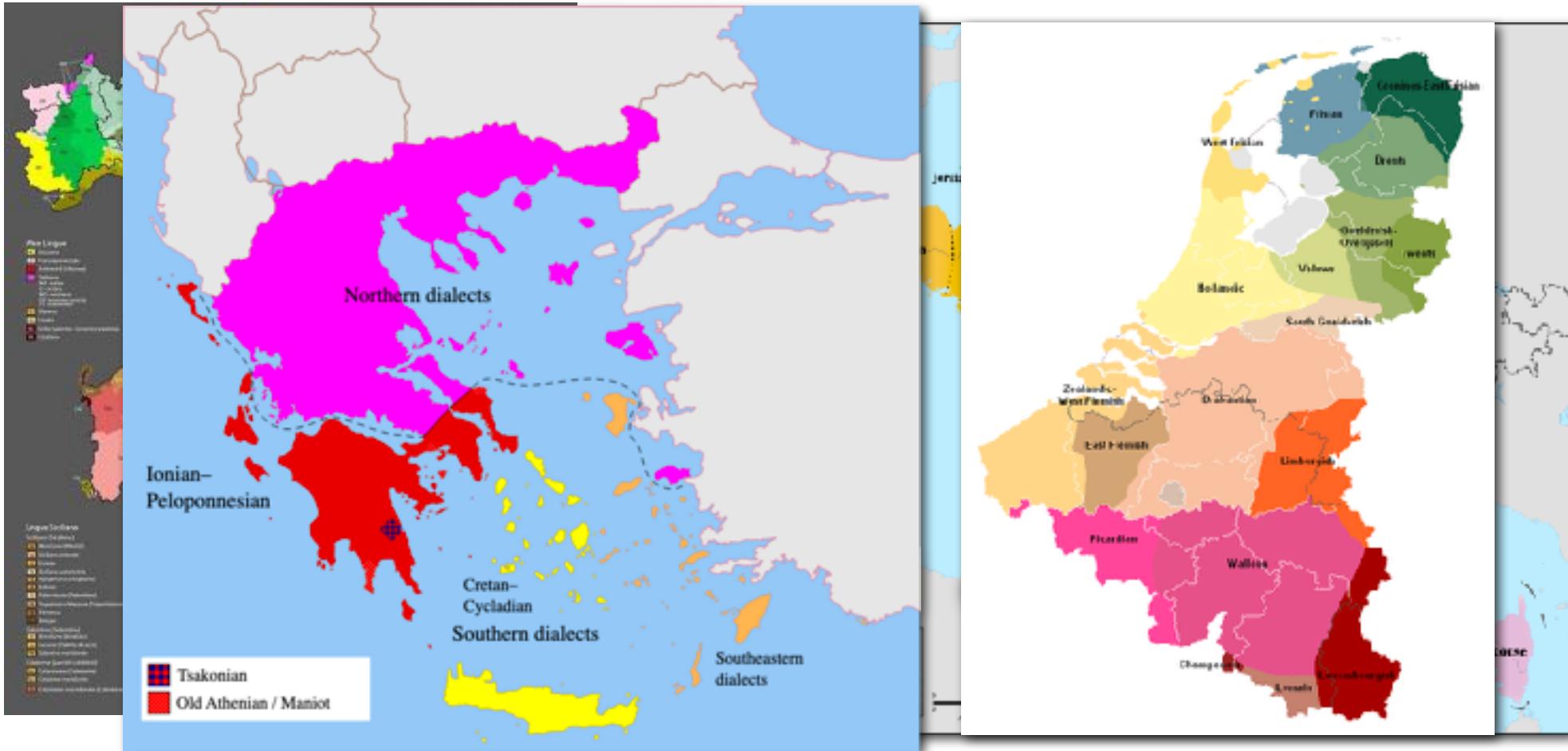
Languages are not Monoliths



Languages are not Monoliths



Languages are not Monoliths

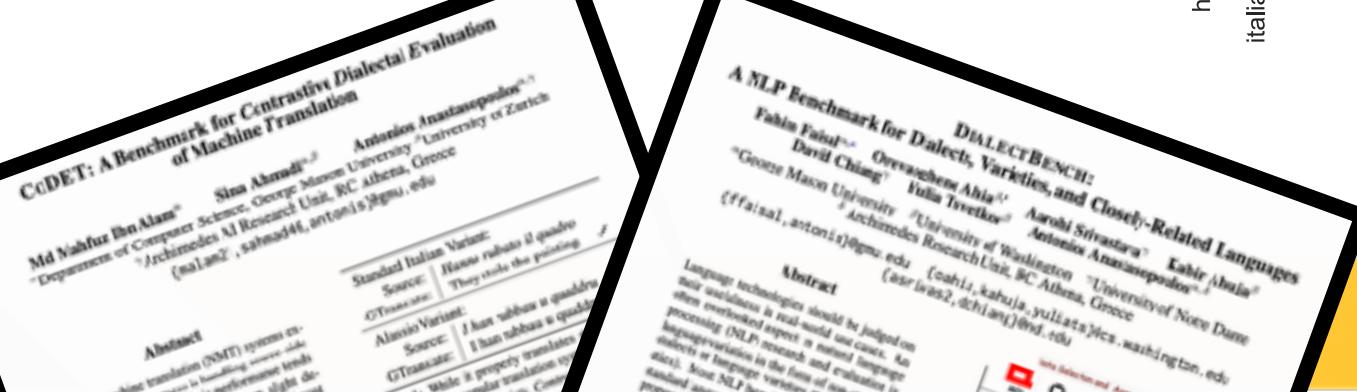


DialectBench

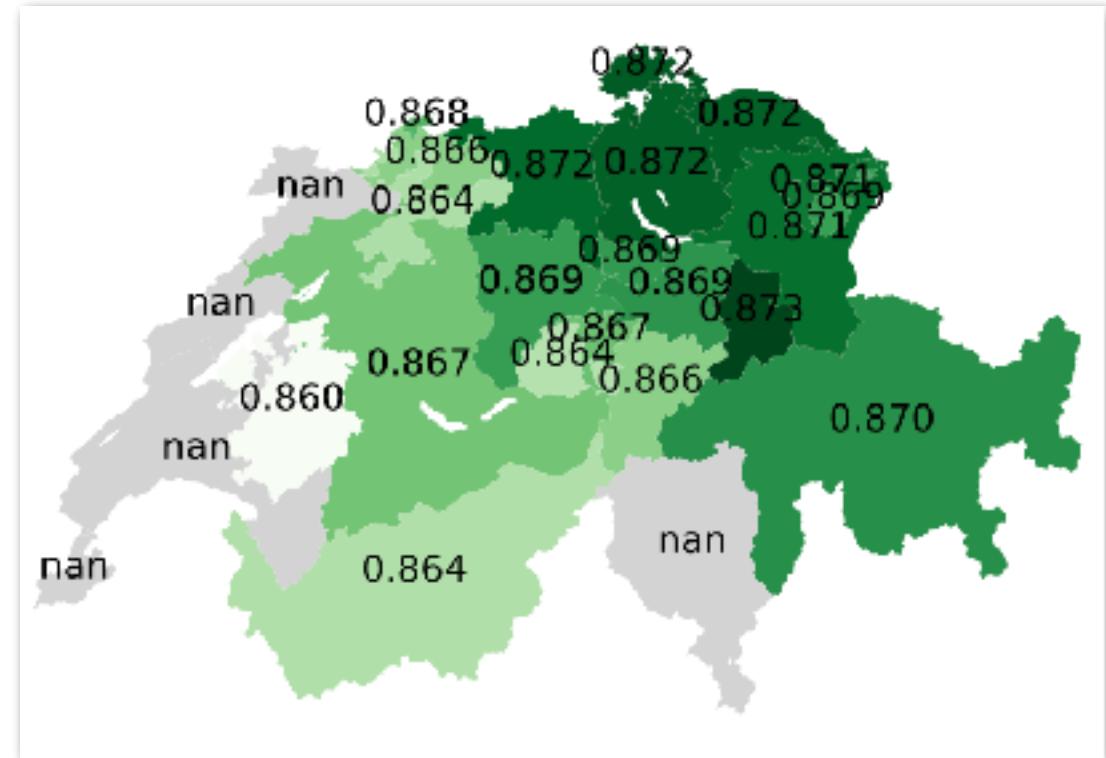
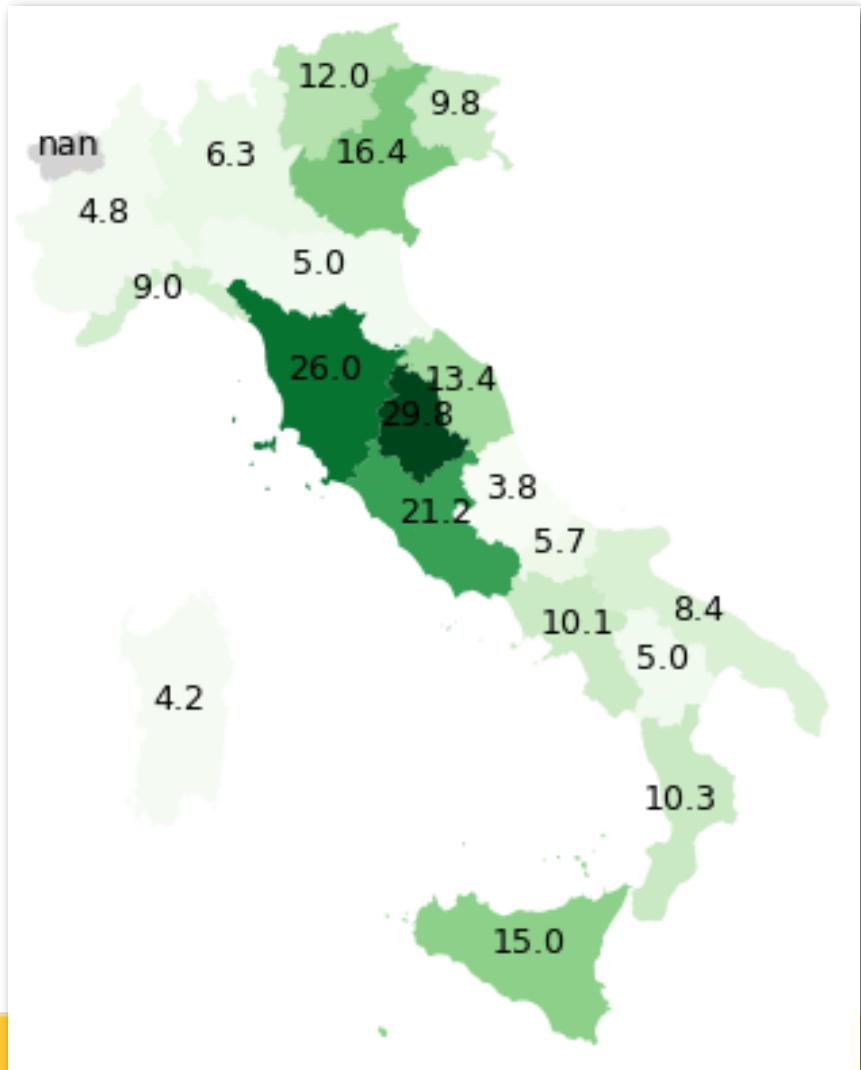
First large-scale benchmark
10 tasks, 40 continua, 281 varieties

Task	DEP.	POS.	NER	EQA	MRC	NLI	TC	SA	DId	MT	Total	arabic	high german	italian romance	basque	anglic	sinitic	common turkic	sw shift. romance	greek	gallo-phaetian	norwegian	neva	bengali	kurkish	komi	serb.-croat.-bosnian	tupi-guarani.	modern dutch	eastern romance	frisian	swahili	Other
DEP.	40	3	2	4		3	3		4		281										1	3	3			3					8		
POS.	51	6	2	4		2	3		5		42										1	3	3			3					8		
NER	85	2	8	4		4	6	4	5	2	85										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		38										3	2				1			5				
TC	38	9	2	2		1	3	3	4		38										3	2				1			5				
SA	9	9																															
DId	49	26	4			3	4		6	6																							
MT	114	25	23	20	21						114										3	5	2							4			
Total	281	42	31	26	21	19	13	12	11	11	281										5	5	2						32				

Language Clusters



DialectBench Results

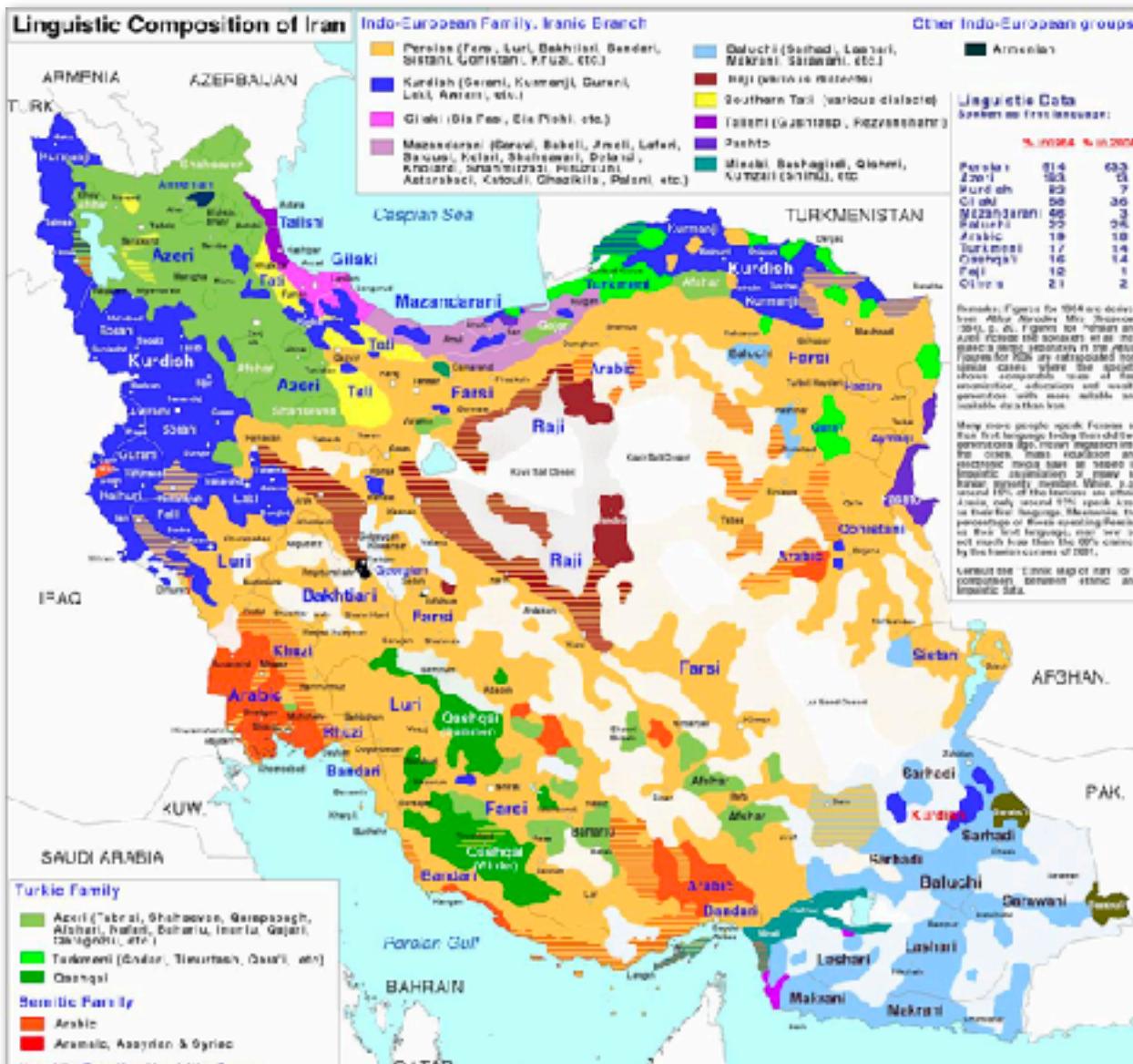




Minority Languages

Minority Languages in X-lingual Communities

Minority Languages in X-lingual Communities

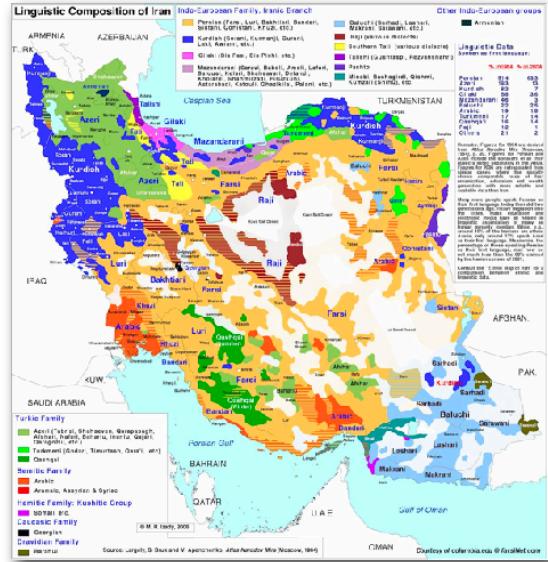
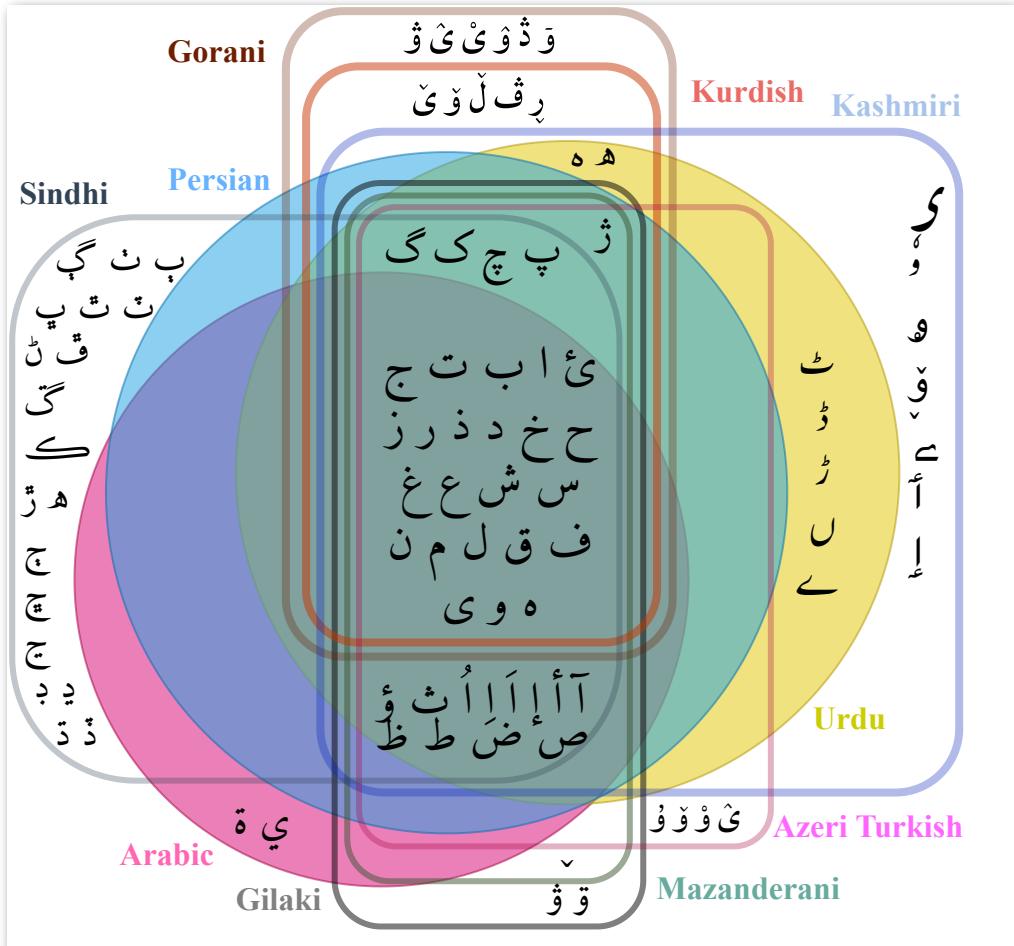


Source:

Minority Languages in X-lingual Communities

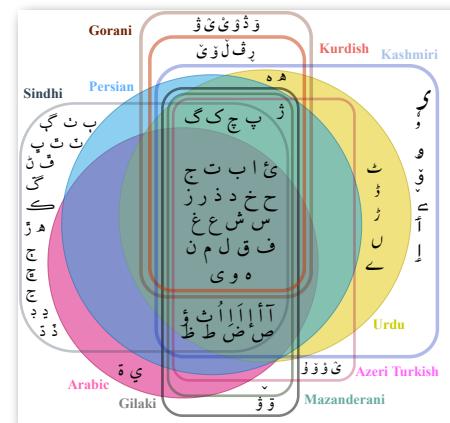
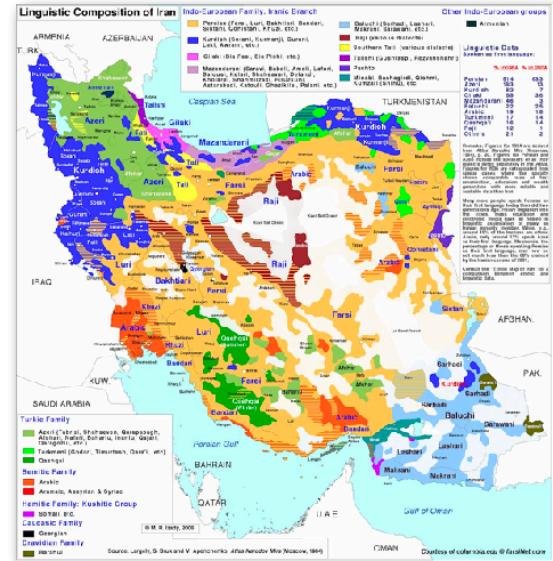


Minority Languages in X-lingual Communities



Minority Languages in X-lingual Communities

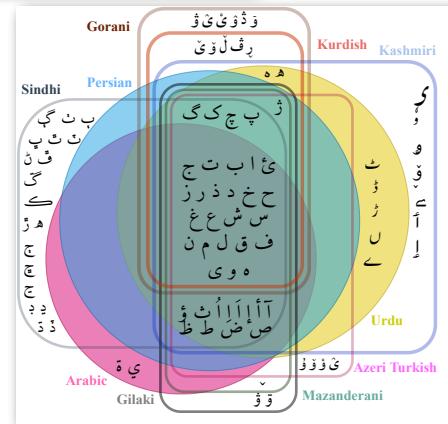
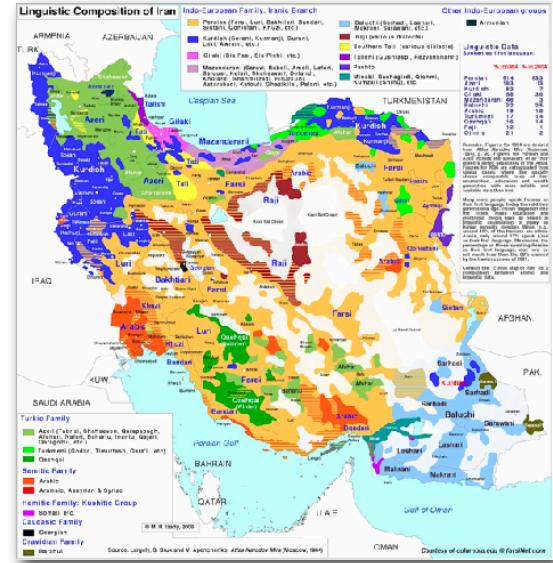
Dominant language (e.g. Farsi) influences the minority one:



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 زن	یته زوئن نؤم هیسه گه گیلکن اون جی گب زن
Kashmiri	Urdu	برور چھ اکھ و راس کے جانور۔	برور چھُ اکھُ وُرَّاَسُ کے جانور۔
Kurmanji	Arabic	قاییقام امدي بقرشو پارزکار دھوک دا	قاییقامی ئامیدی بەرسە پاریزگاری دھوکن دا
Sorani	Arabic	ھەر لە یەکەم شانزوو دیارە فەھەدیان دەویت	ھەر لە یەکەم شانزوو دیارە فەھەدیان دەویت
Sindhi	Urdu	مدینی ڏانهن هجرت وقت فقط هيء گھرواري ساٿن گڏ هيي	مدینی ڏانهن هجرت وقت فقط هيء گھرواري ساٿن گڏ هيي



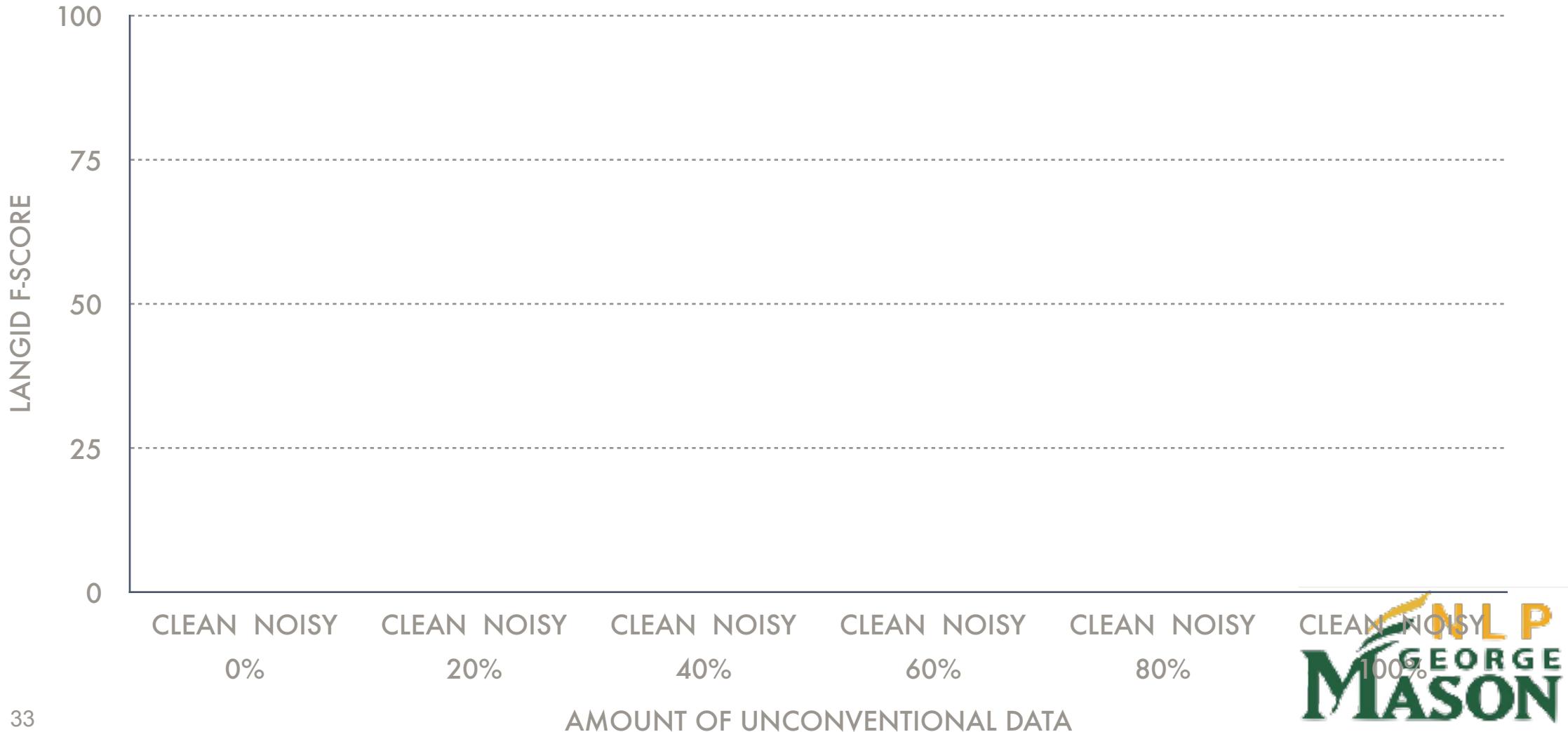
Case Study: Languages using Perso-Arabic Script

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	✗	✗	-

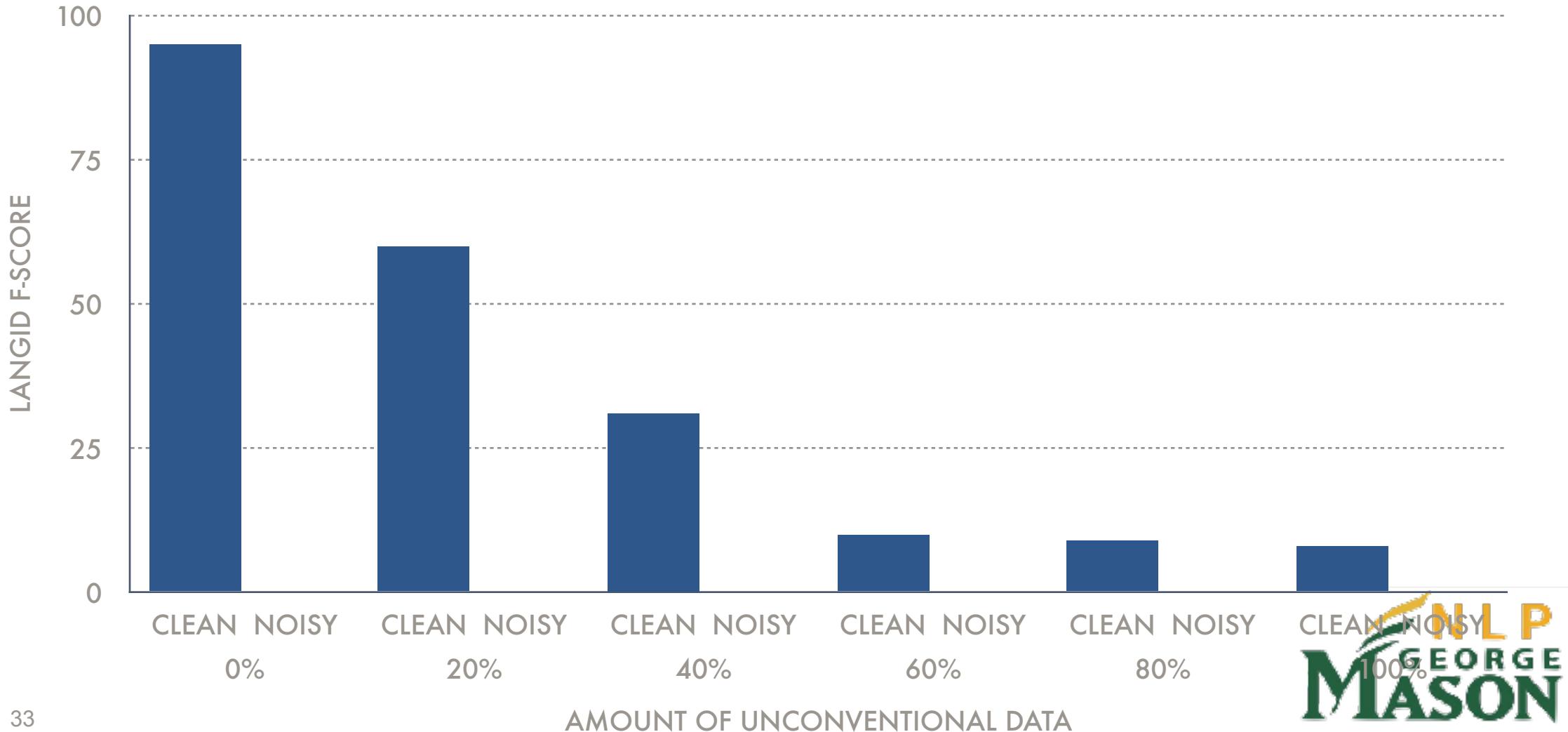
Table 1: Perso-Arabic scripts of the selected languages studied in this paper. Columns 2 and 3 show the codes of the languages in ISO 639-3 and on their specific Wikipedia (WP), if available. The diacritics and zero-width non-joiner (ZWNJ) columns refer to the usage of diacritics (*Harakat*) and ZWNJ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

Effect of Unconventional Writing

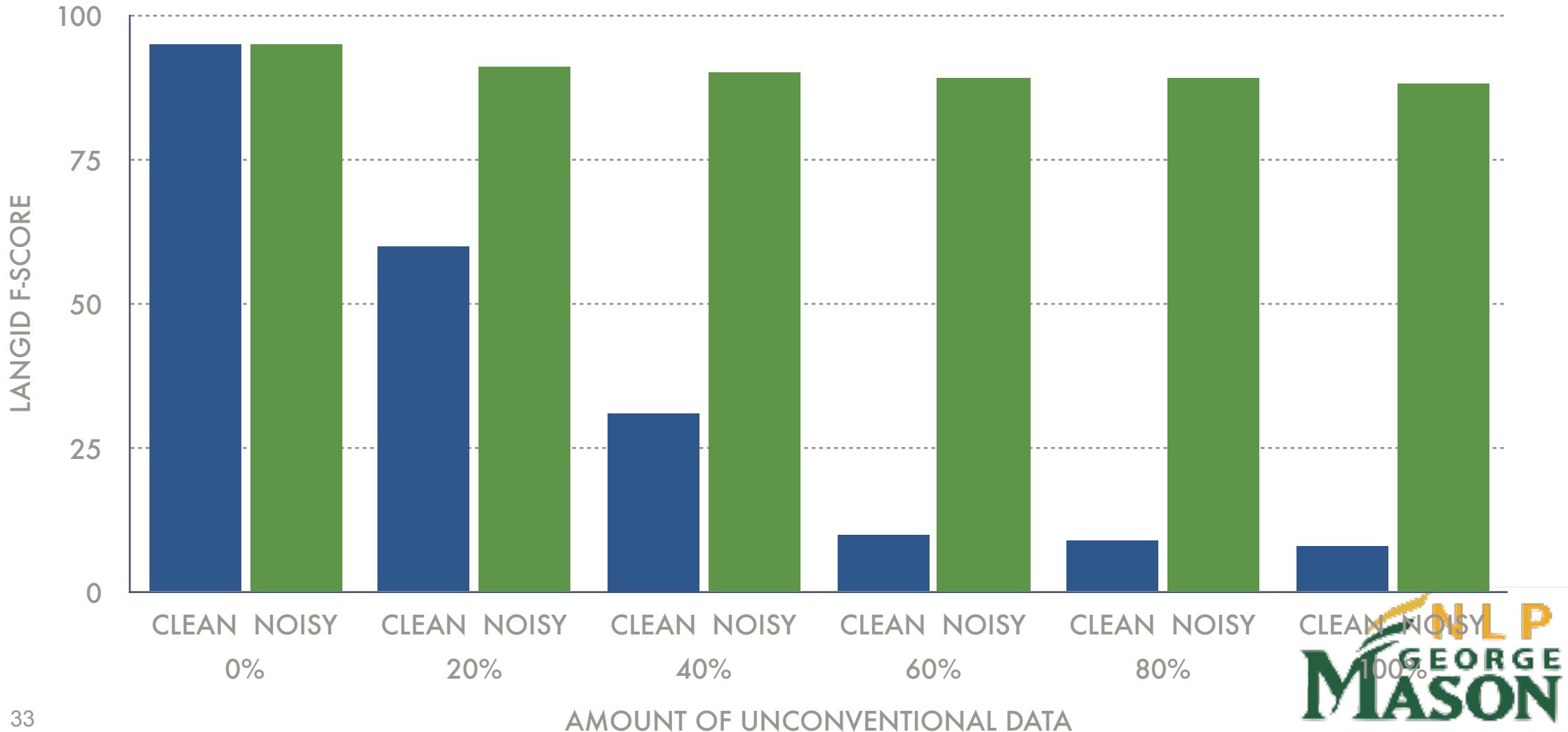
Effect of Unconventional Writing



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

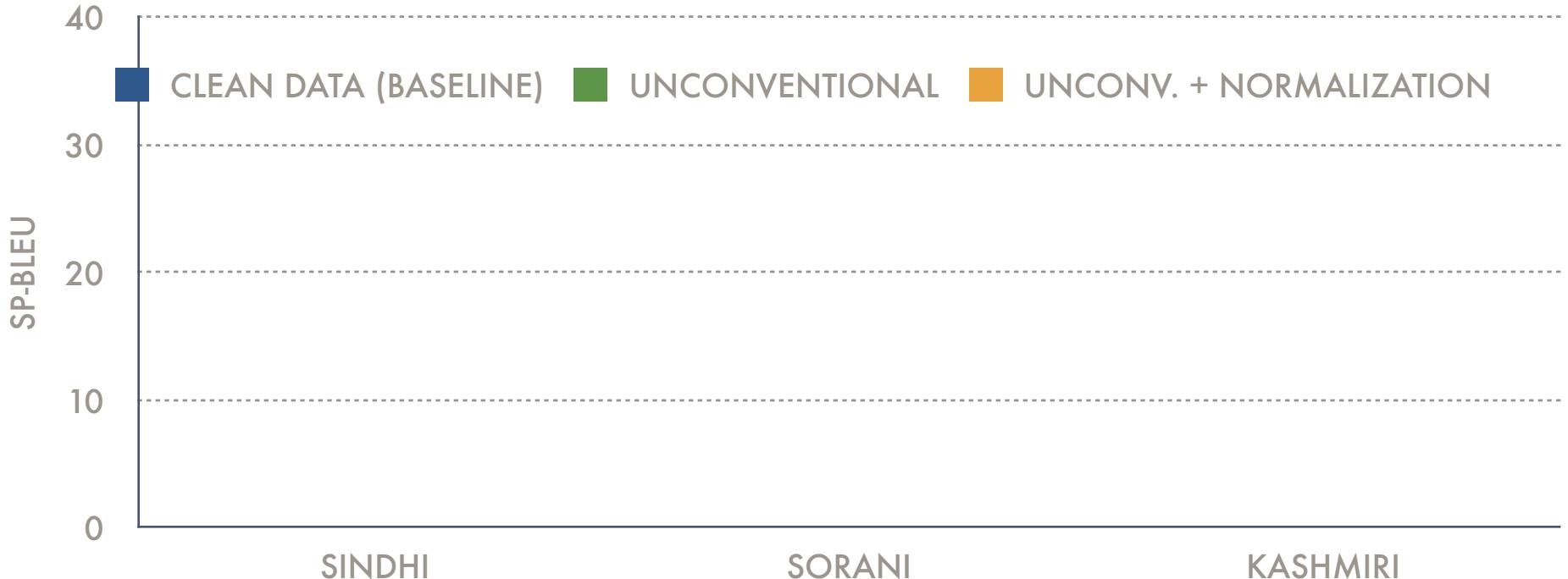
Train a Normalization model

(Encoder-decoder, self-attention based)

Evaluate its effect on Machine Translation

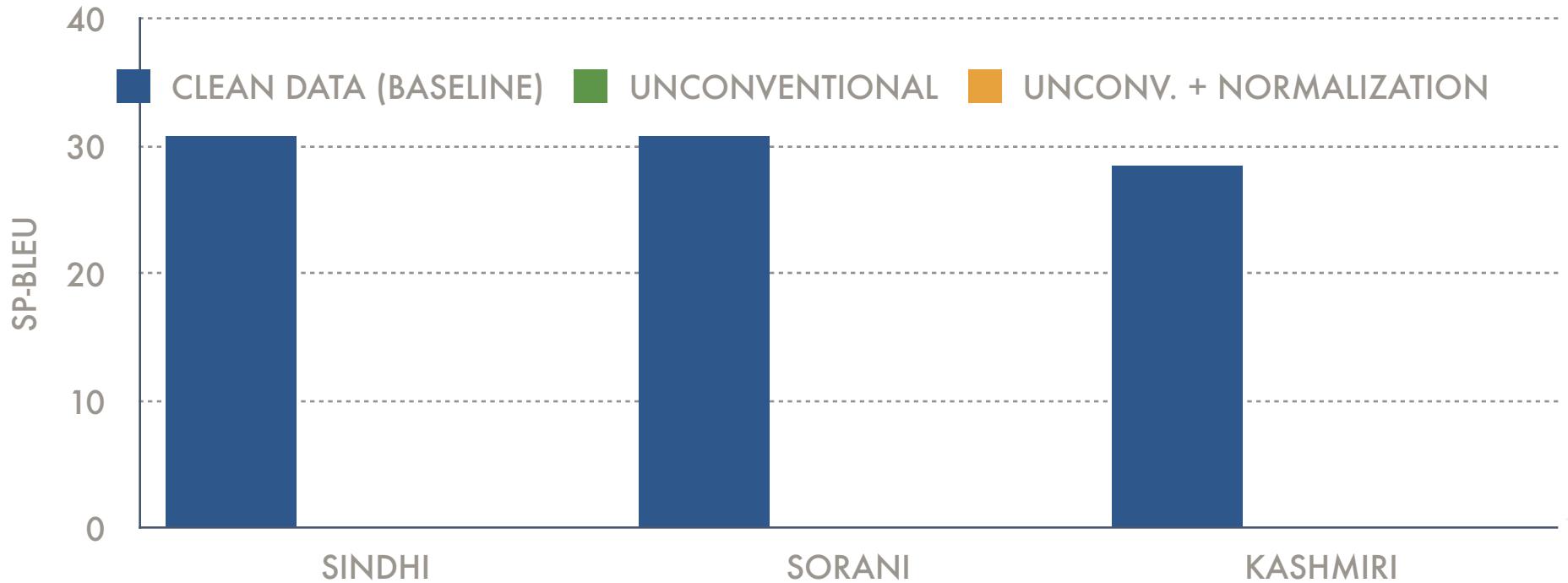
Mitigating the Effect of Unconventional Writing

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



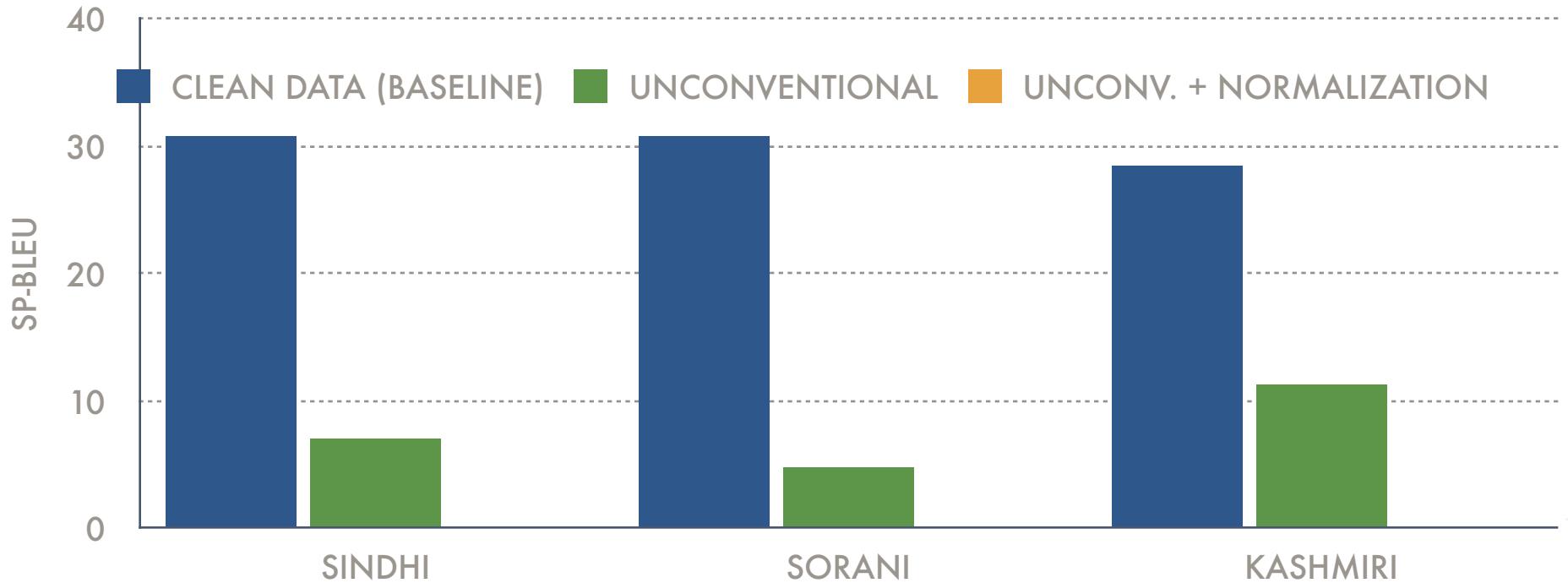
Mitigating the Effect of Unconventional Writing

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



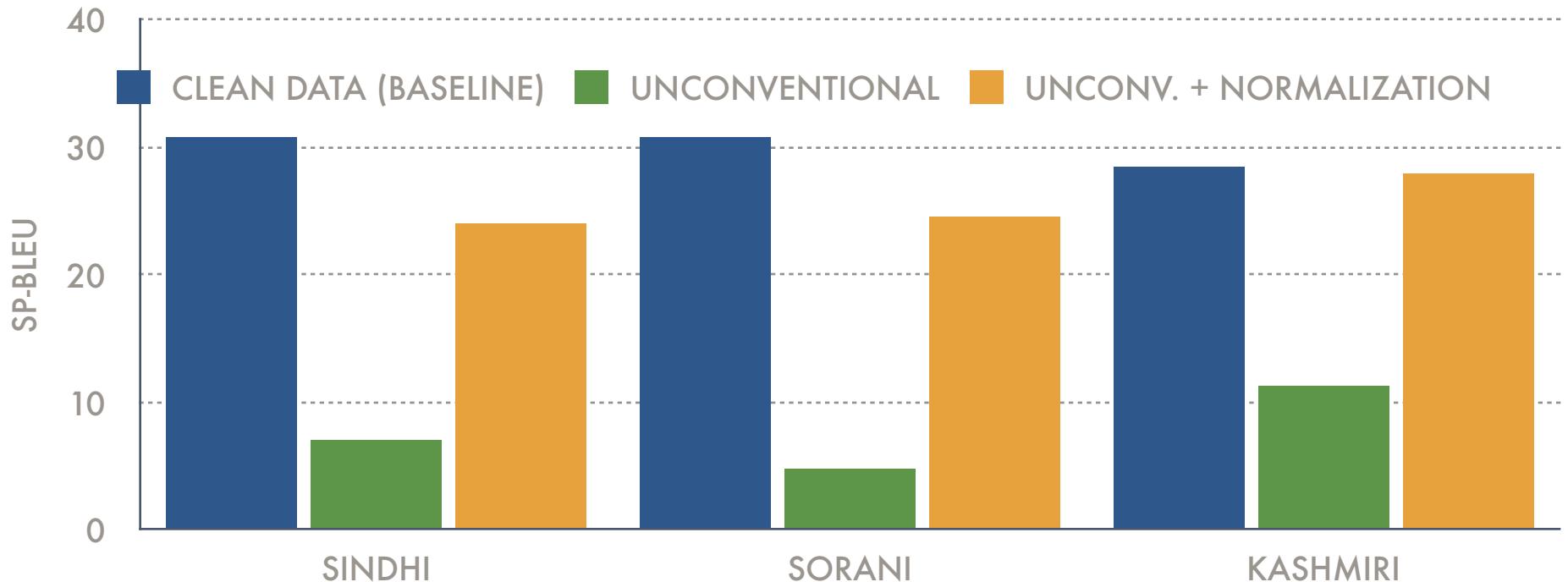
Mitigating the Effect of Unconventional Writing

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

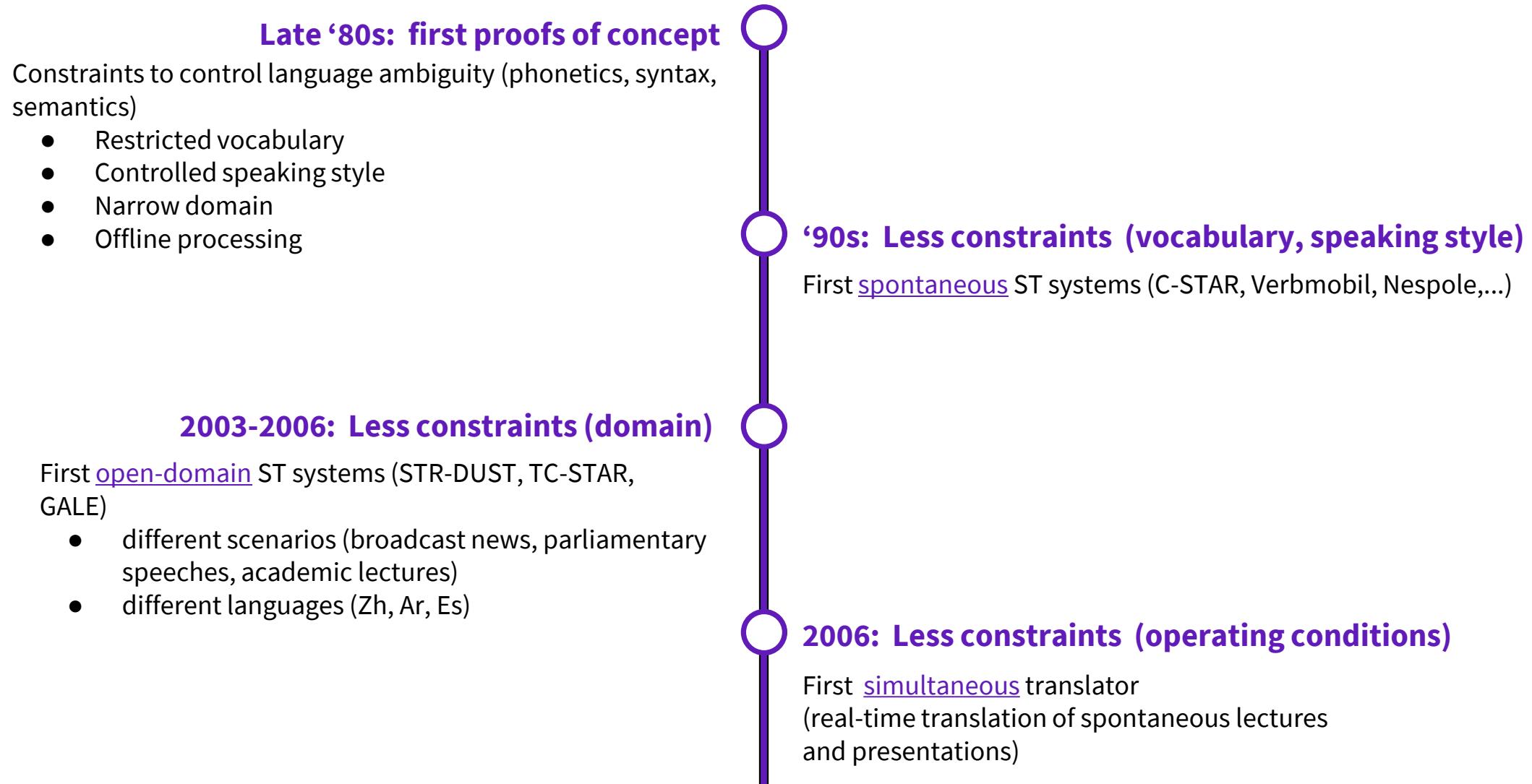




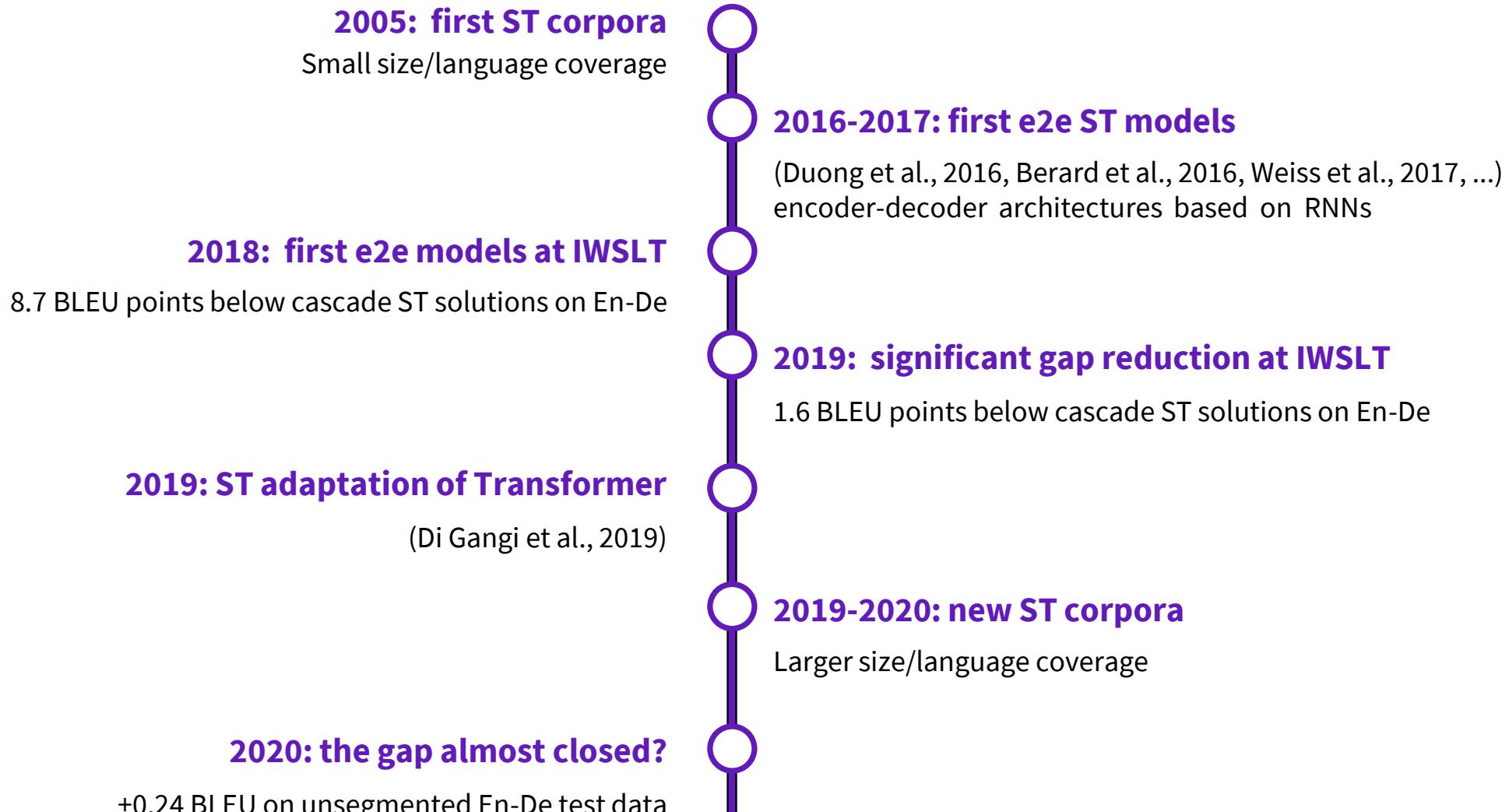
We Need to Handle Speech Input

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)



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



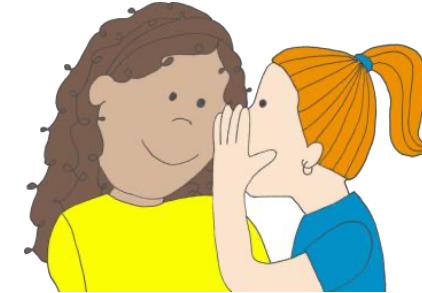
Challenges in translation of speech

- 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



Reden (engl. speeches)



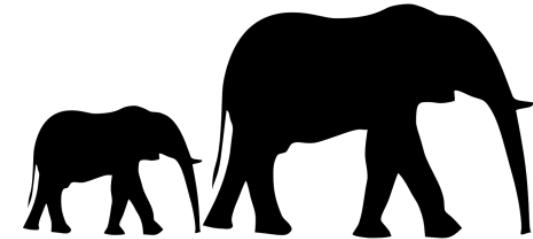
Reben (engl. vines)

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

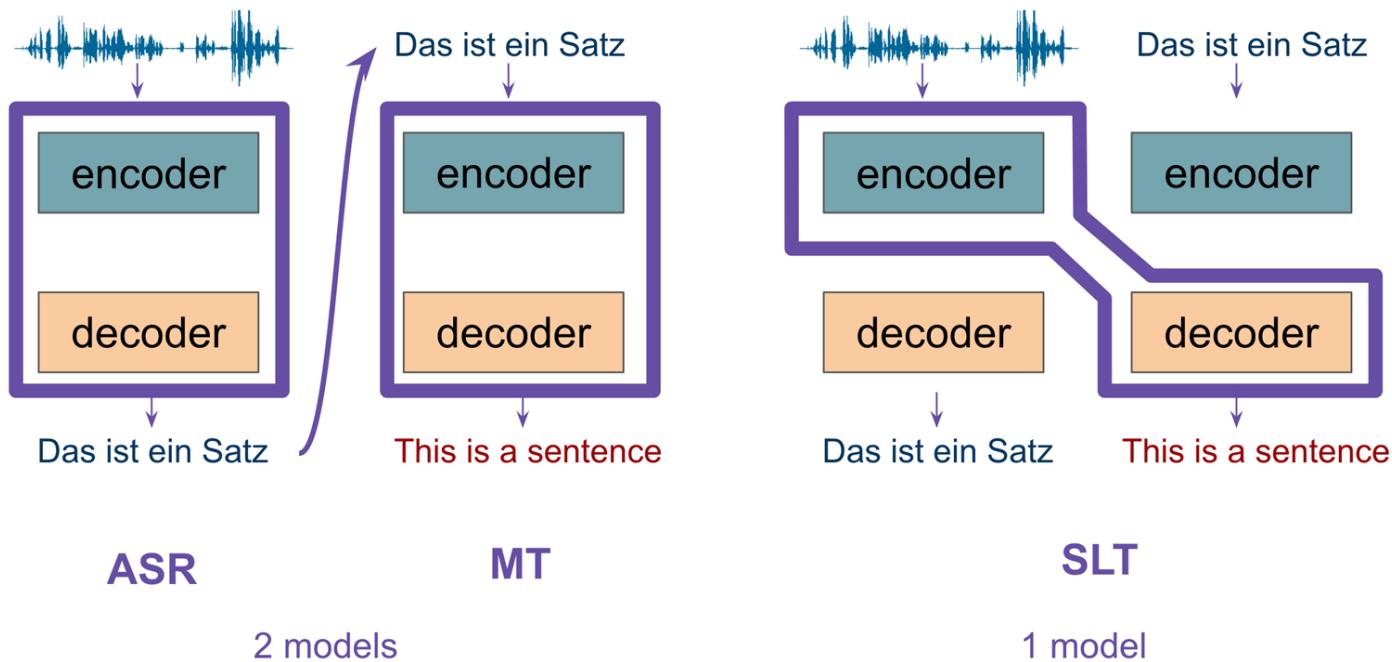
- Audio challenges
- Text-Speech mismatch
- Error propagation
- Data
- Partial information
 - Online: Translate during production of speech
 - Generate translation before full sentence is known



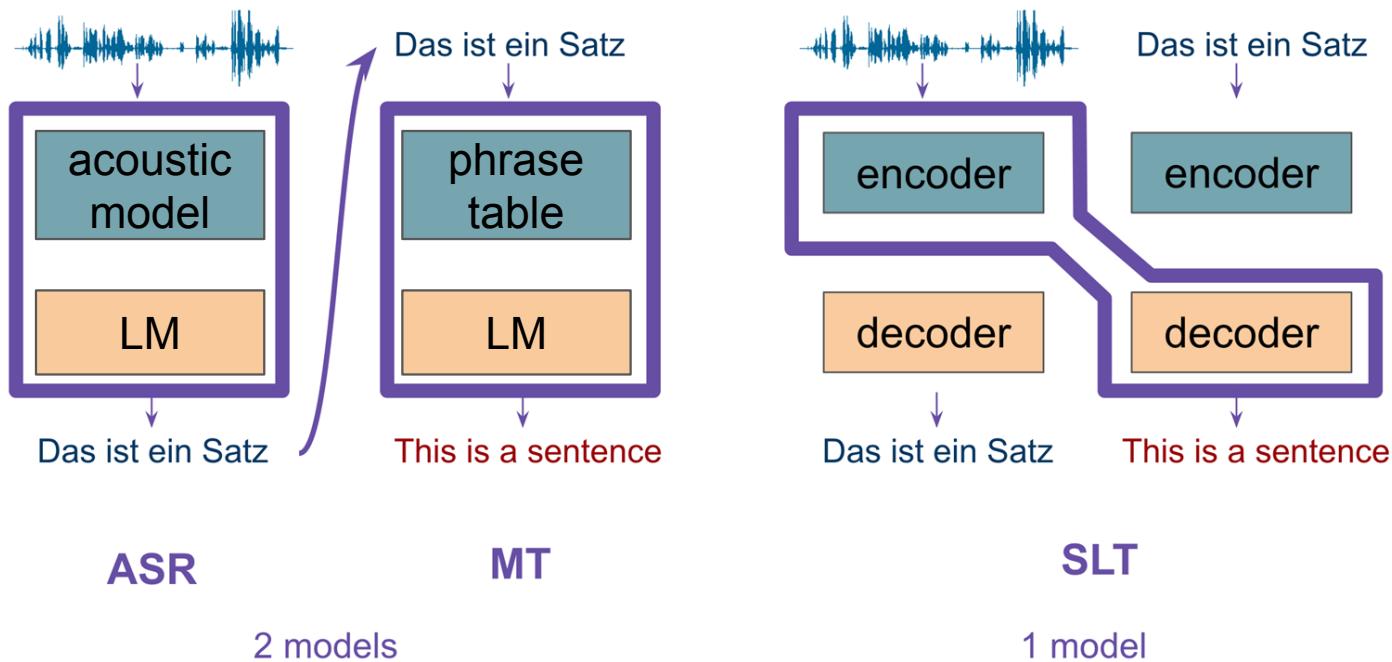


Traditional Cascade Approach

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

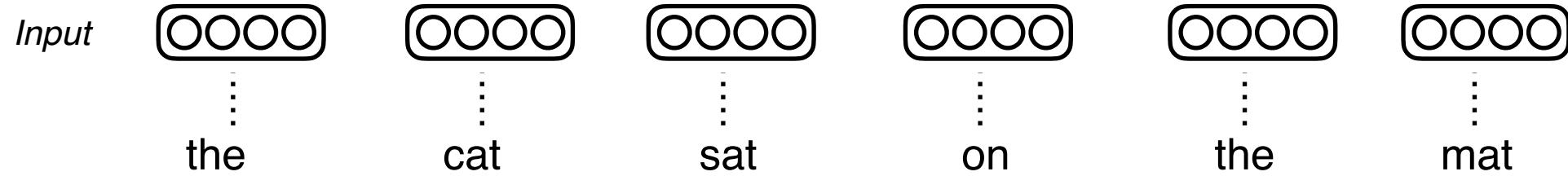
sat

on

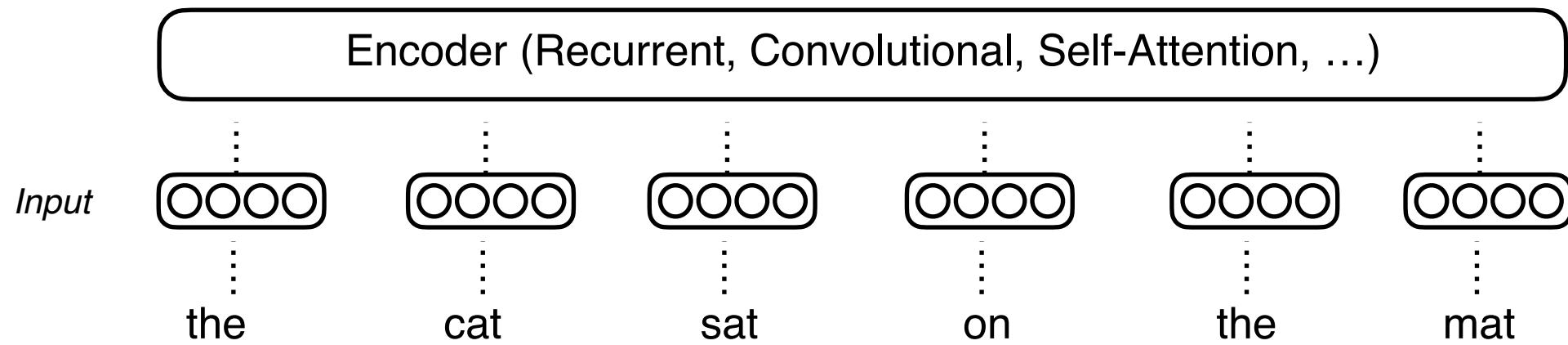
the

mat

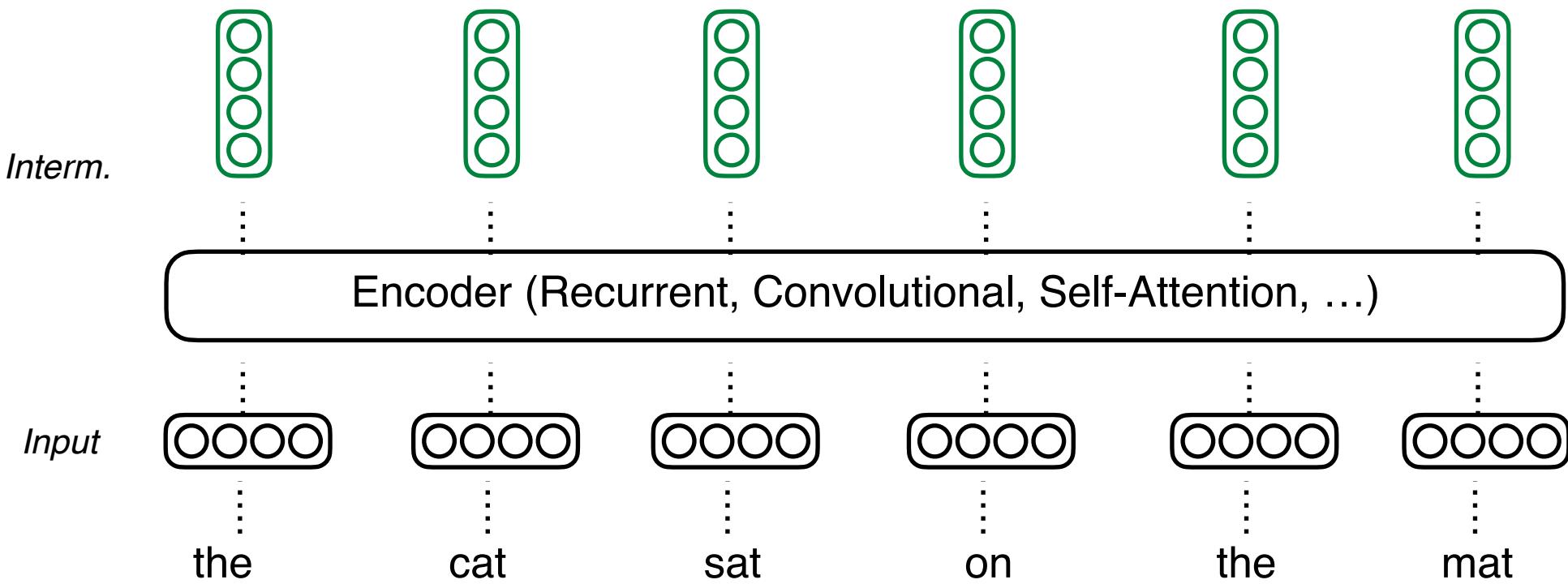
Encoder-Decoder with Attention



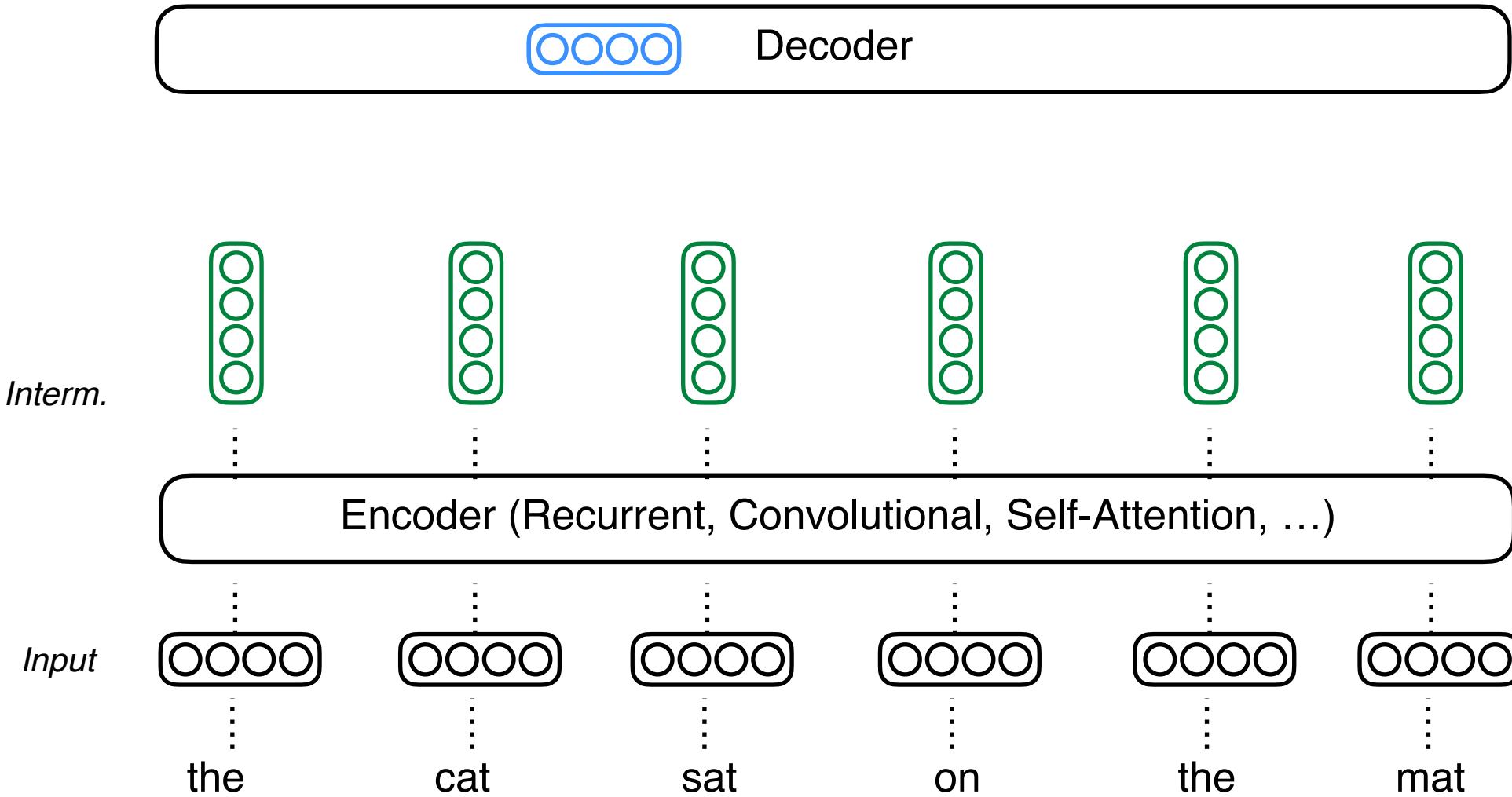
Encoder-Decoder with Attention



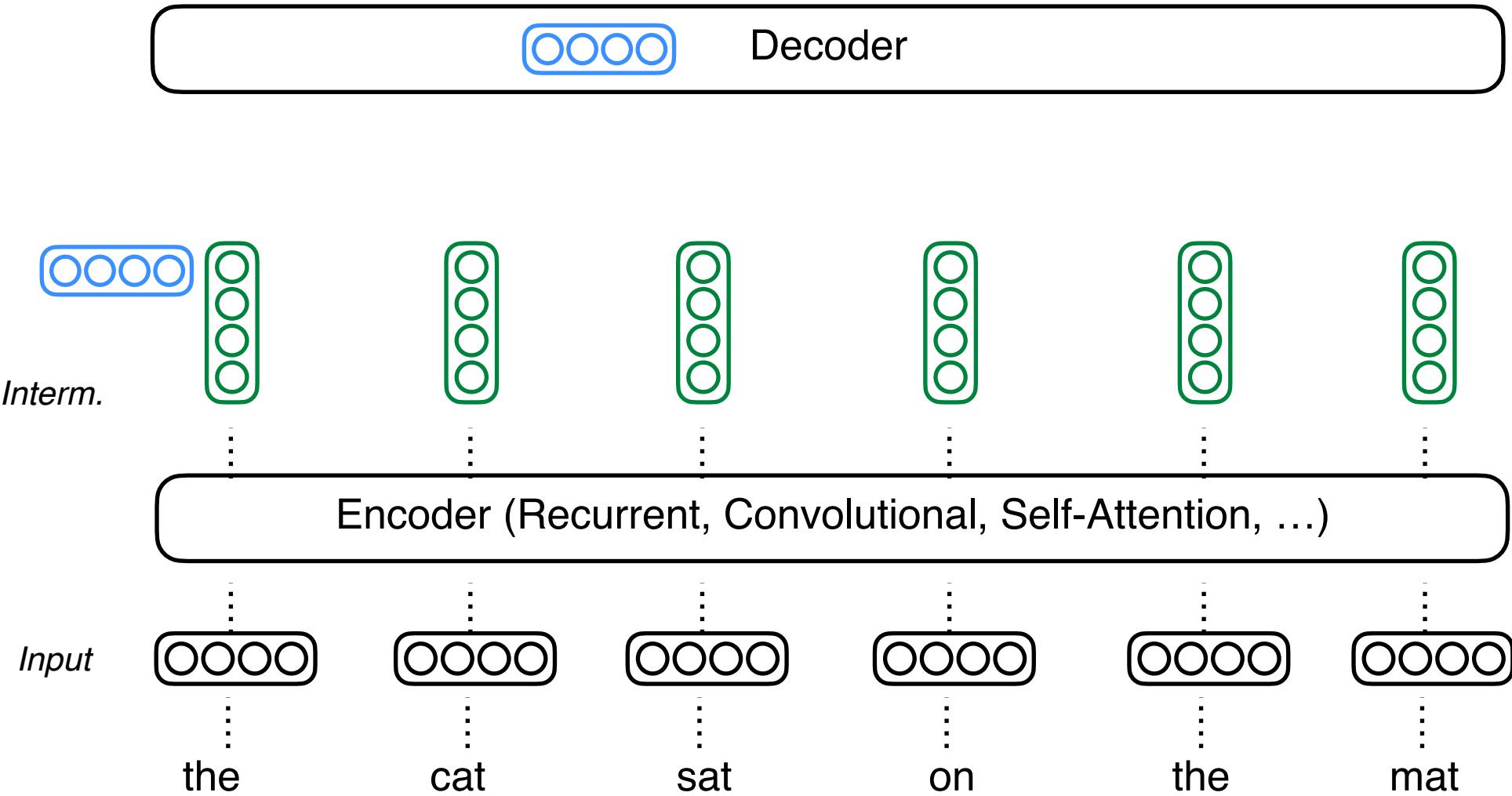
Encoder-Decoder with Attention



Encoder-Decoder with Attention

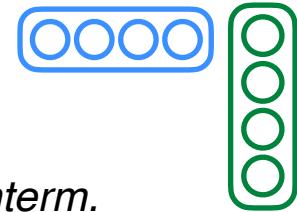


Encoder-Decoder with Attention





Decoder



Interm.



Encoder (Recurrent, Convolutional, Self-Attention, ...)

Input



the



cat



sat



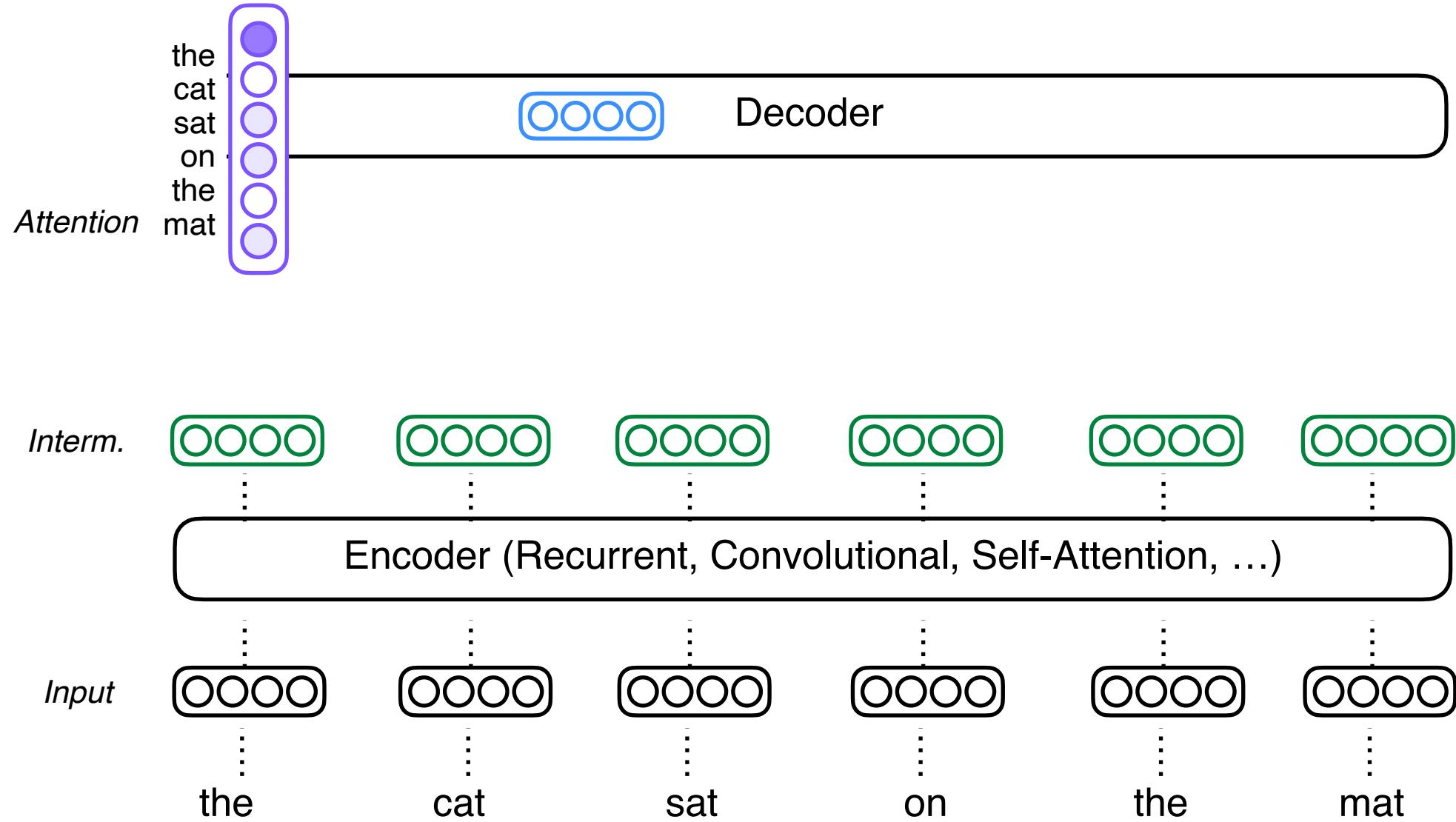
on

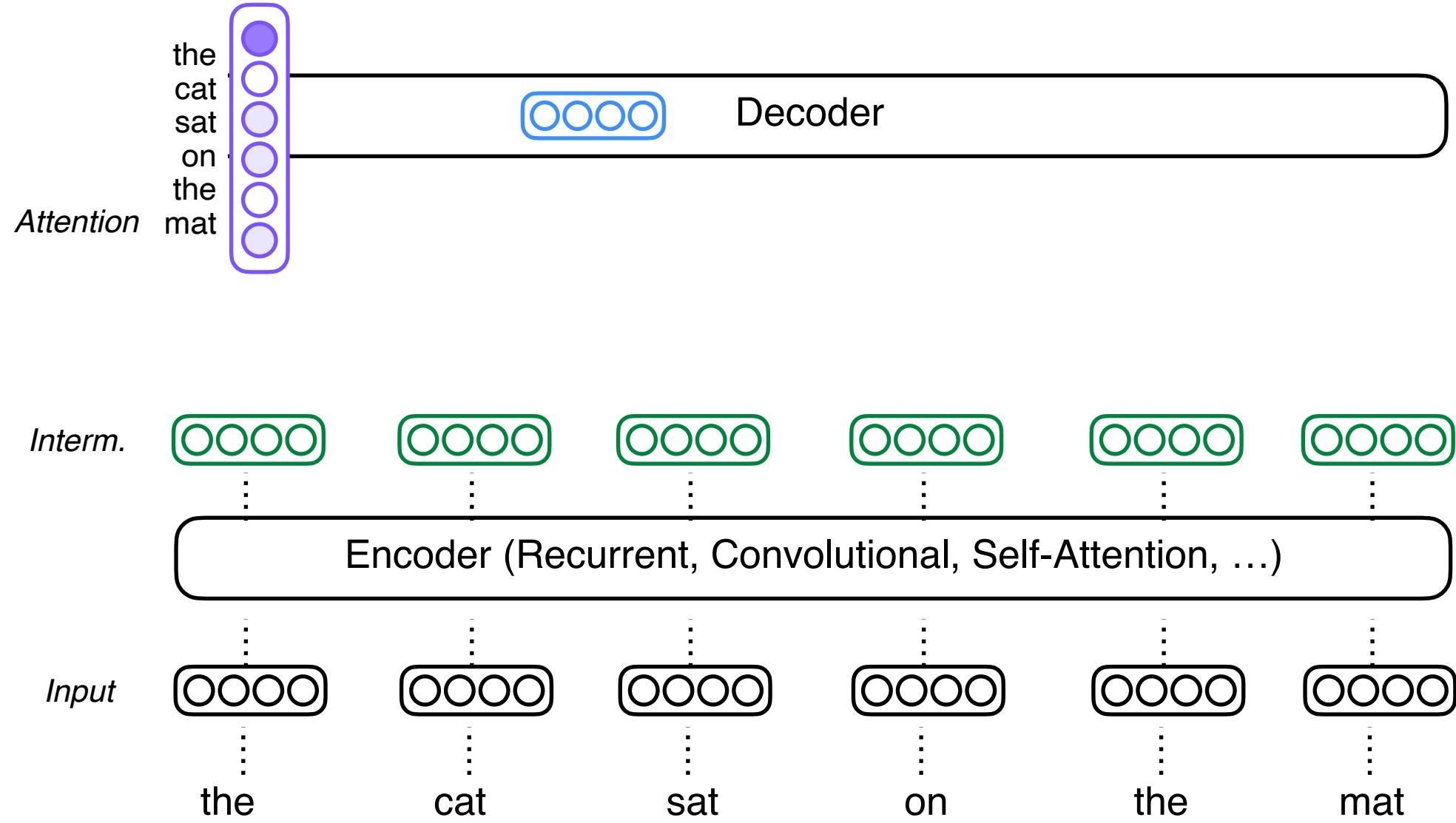


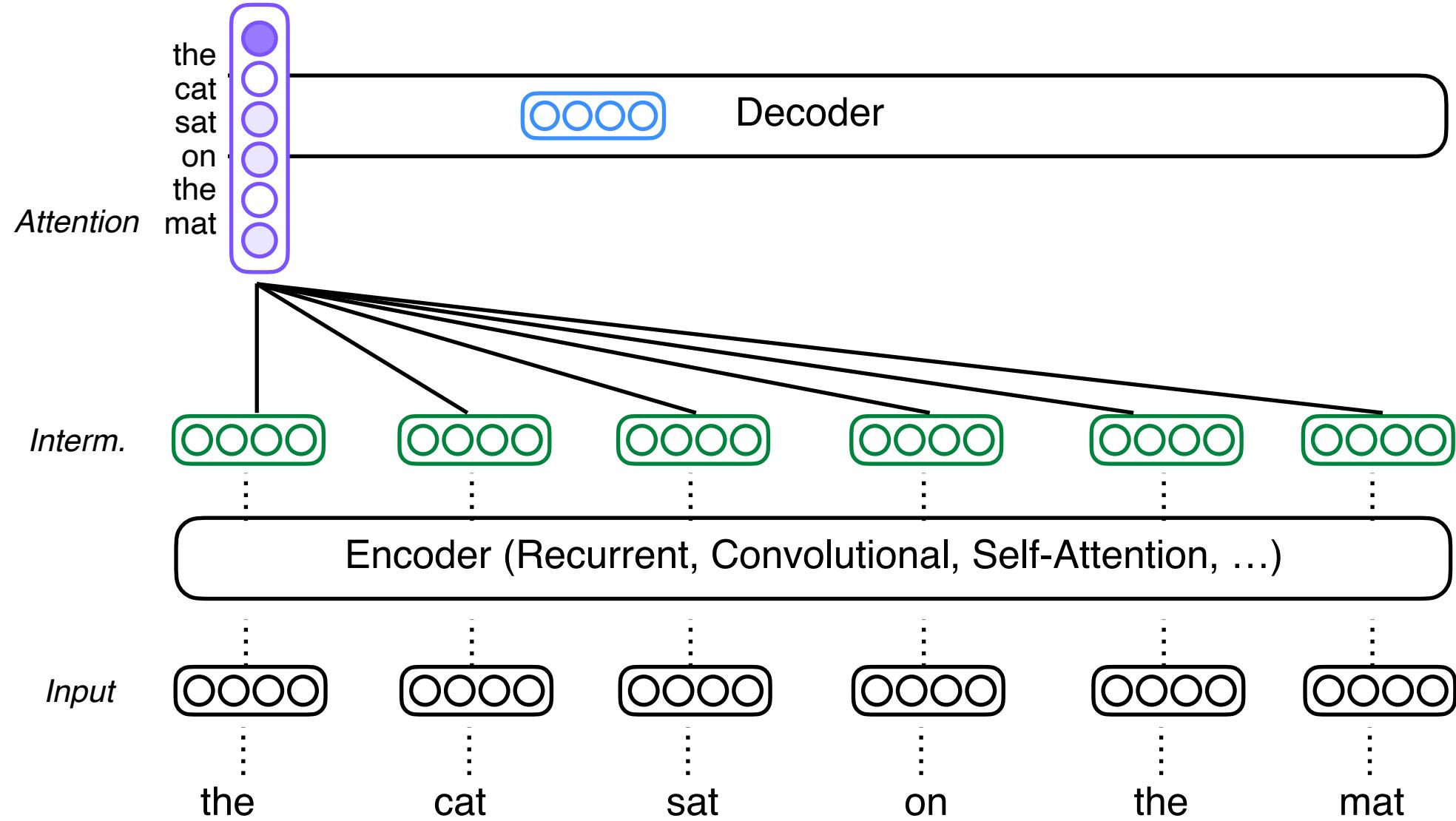
the

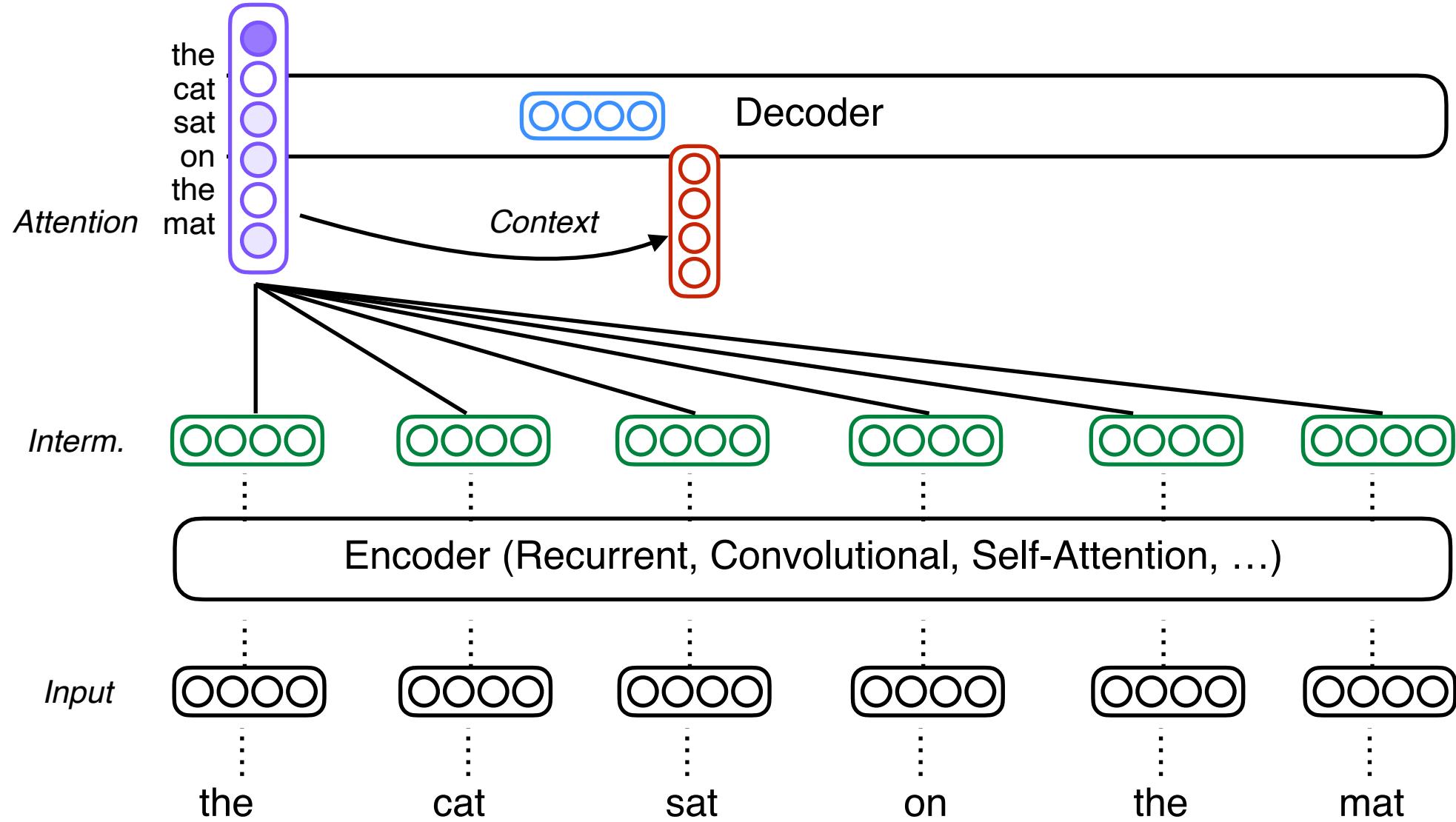


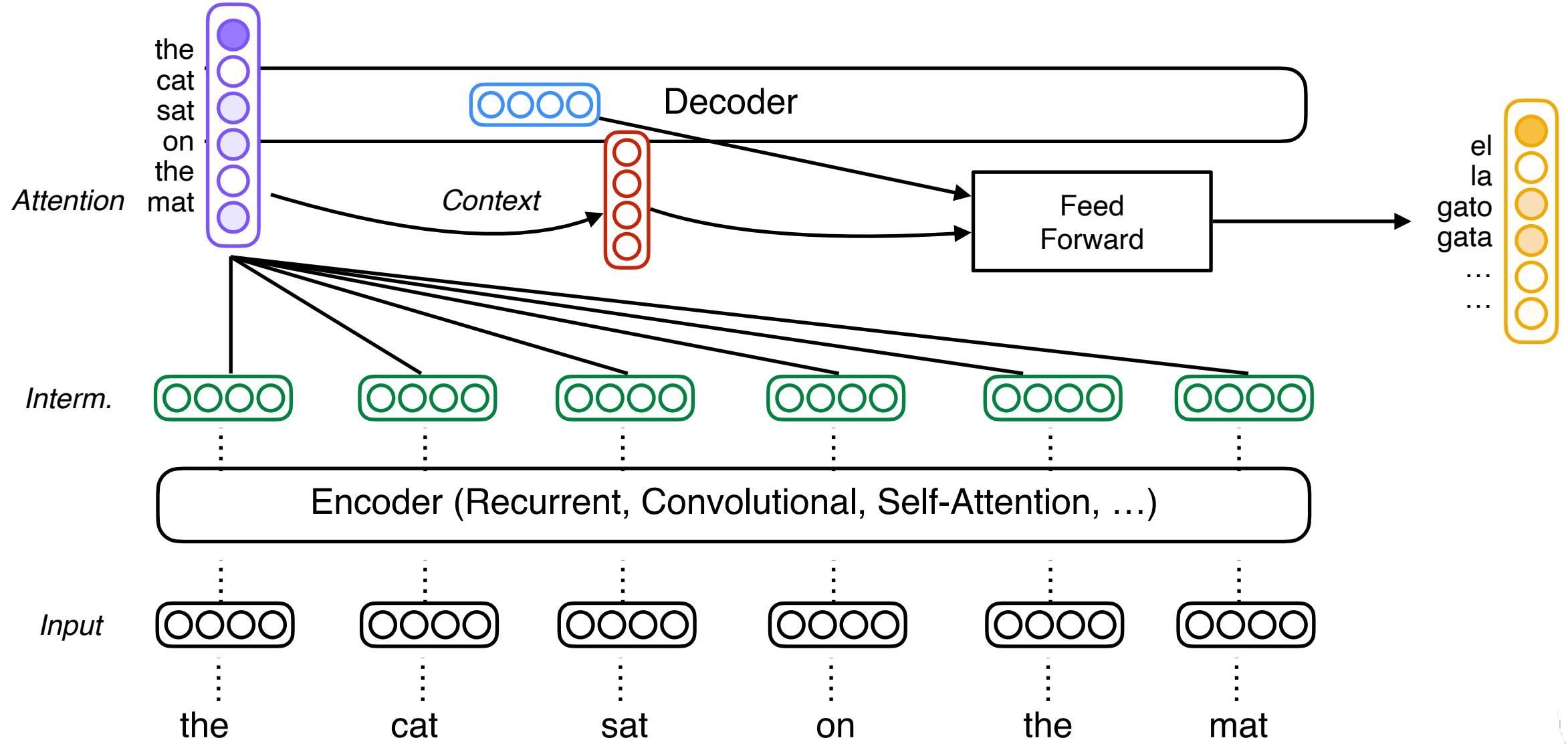
mat







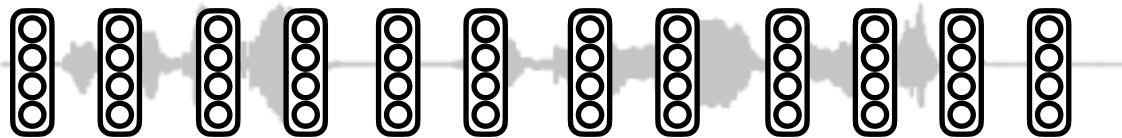




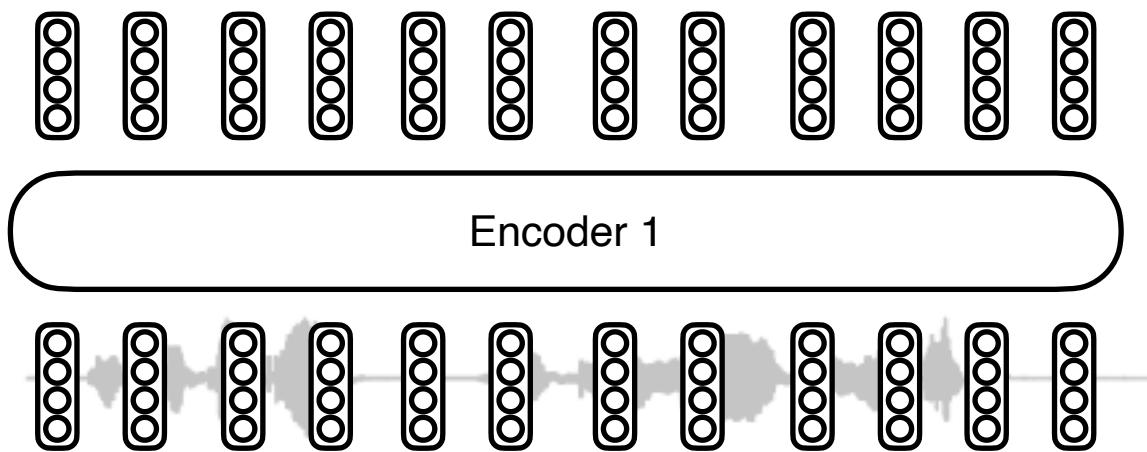
An Audio-Input model



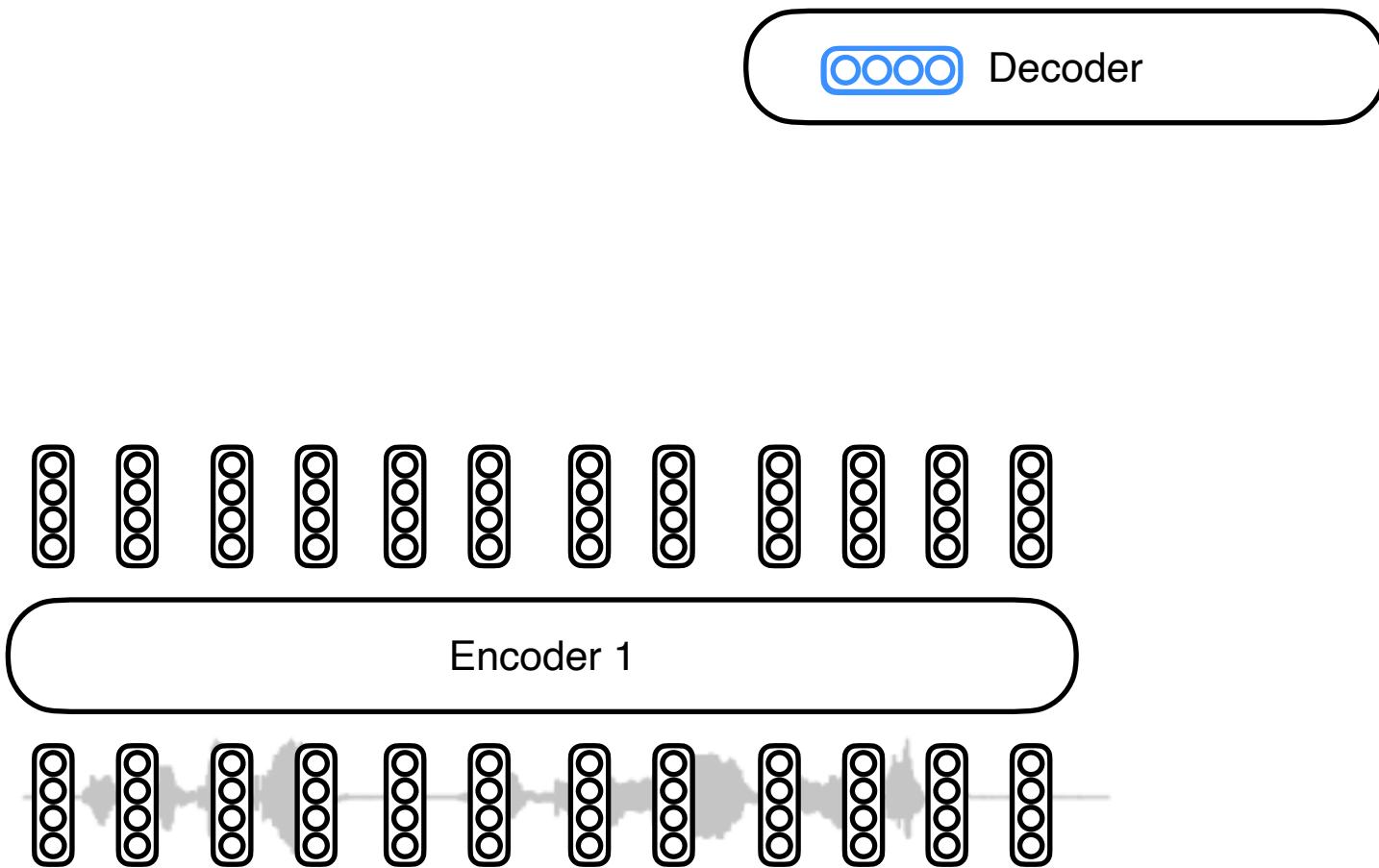
An Audio-Input model

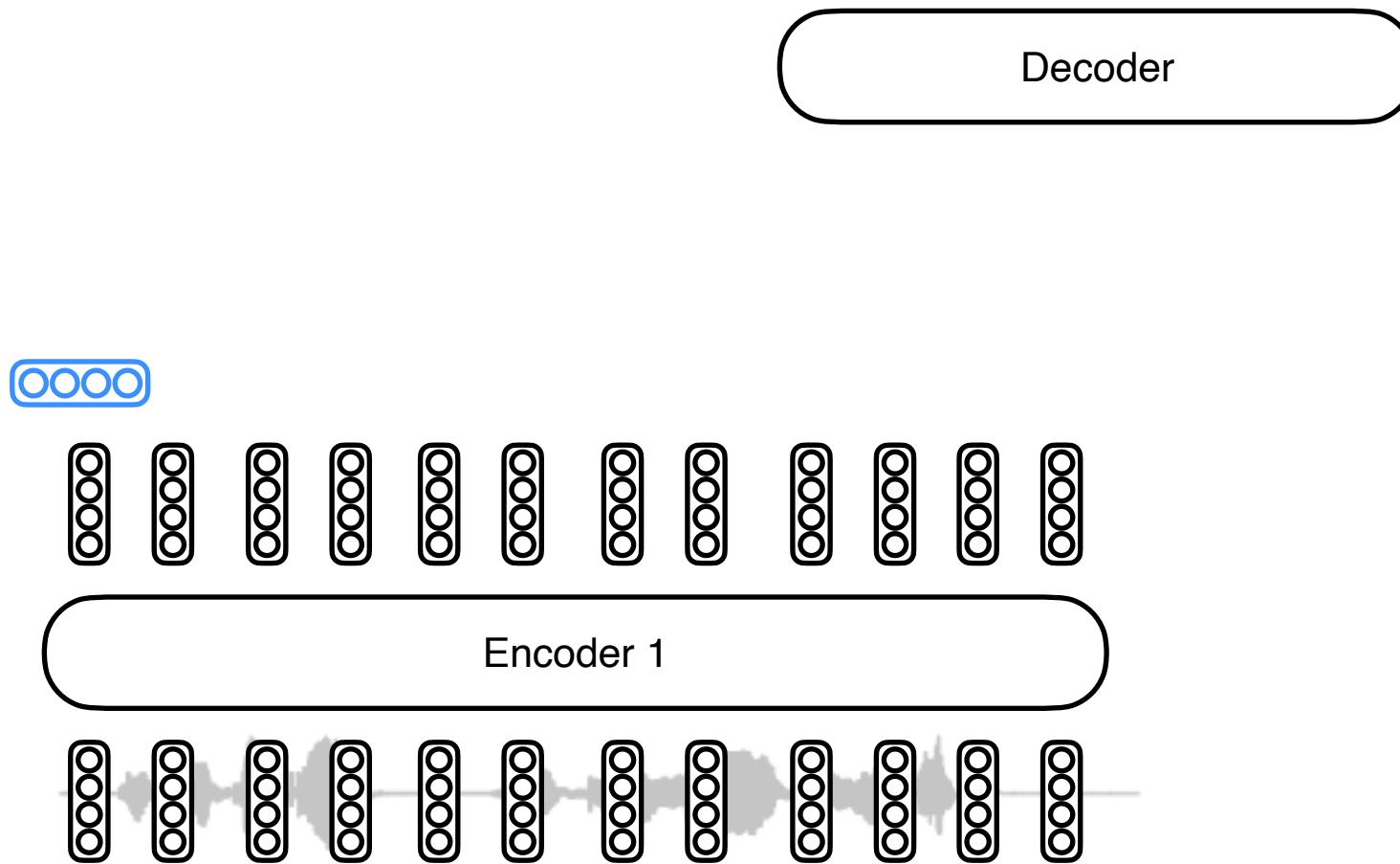


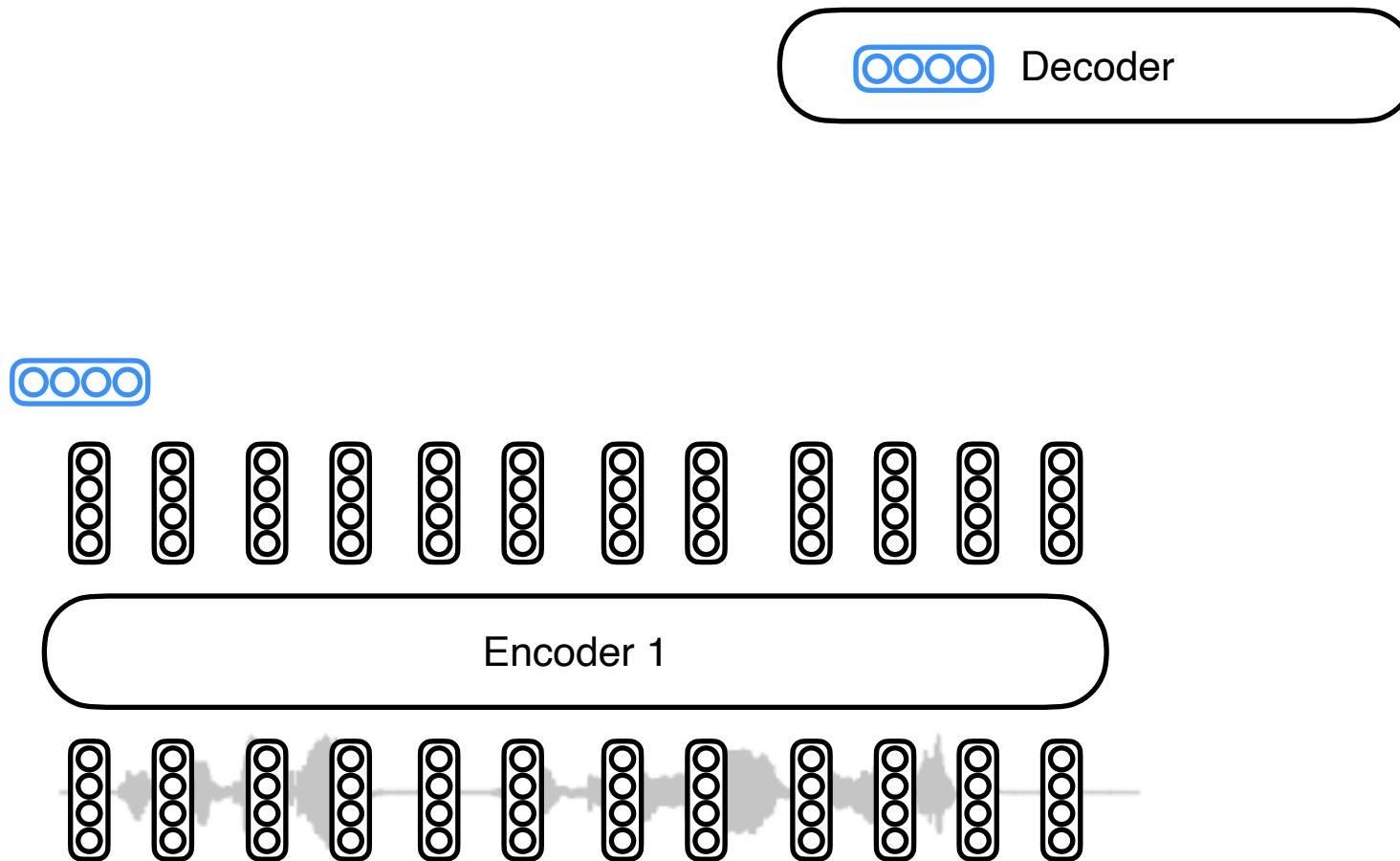
An Audio-Input model

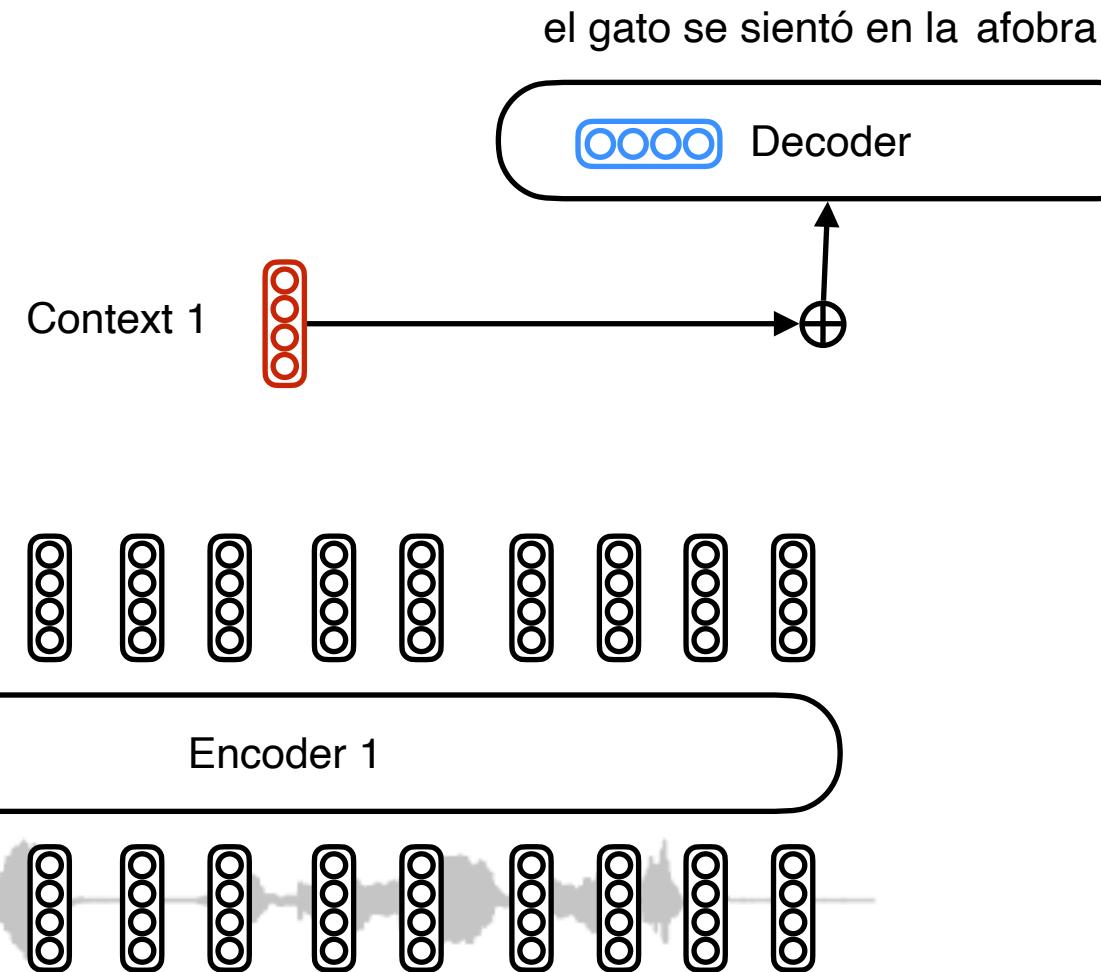


An Audio-Input model











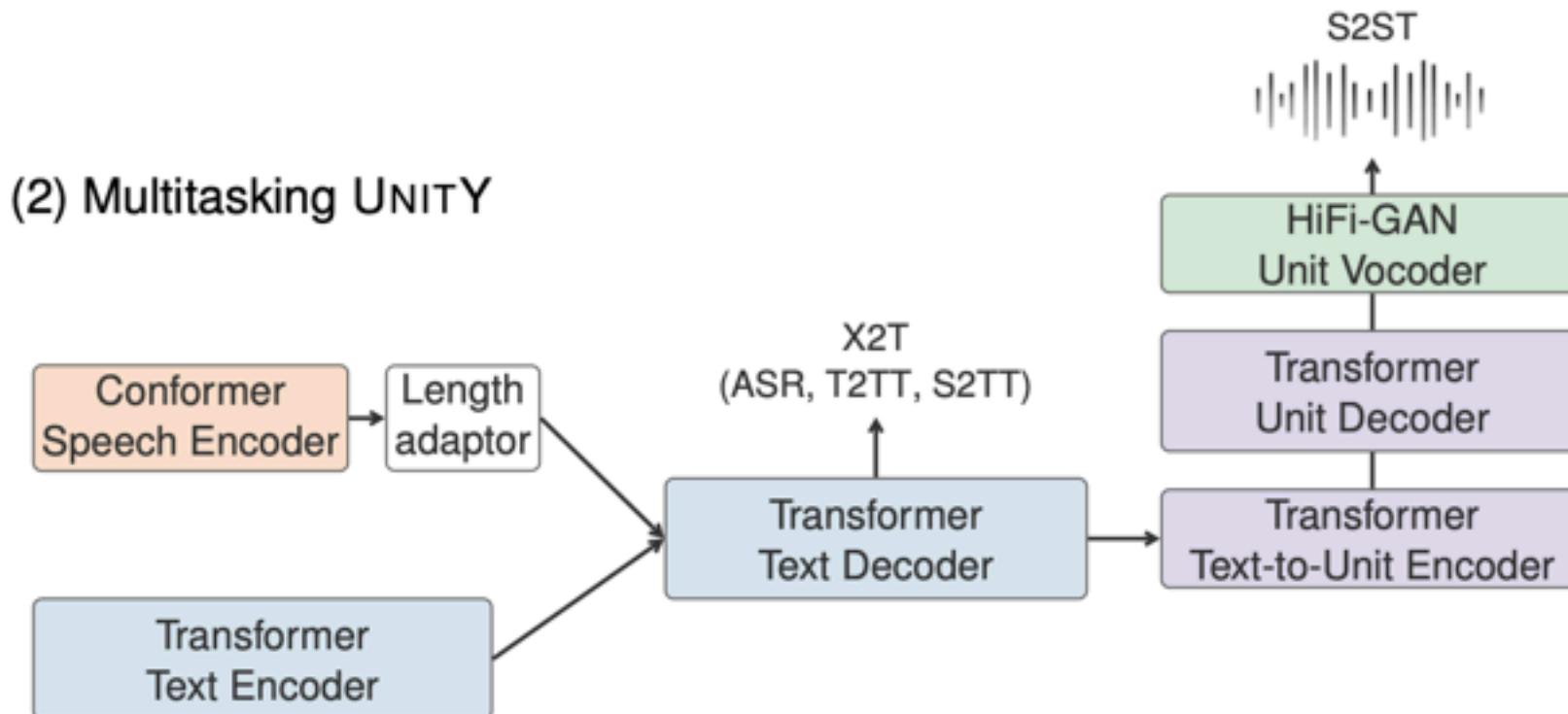
Today: pre-training

The SeamlessM4T model

(1) Pre-trained models



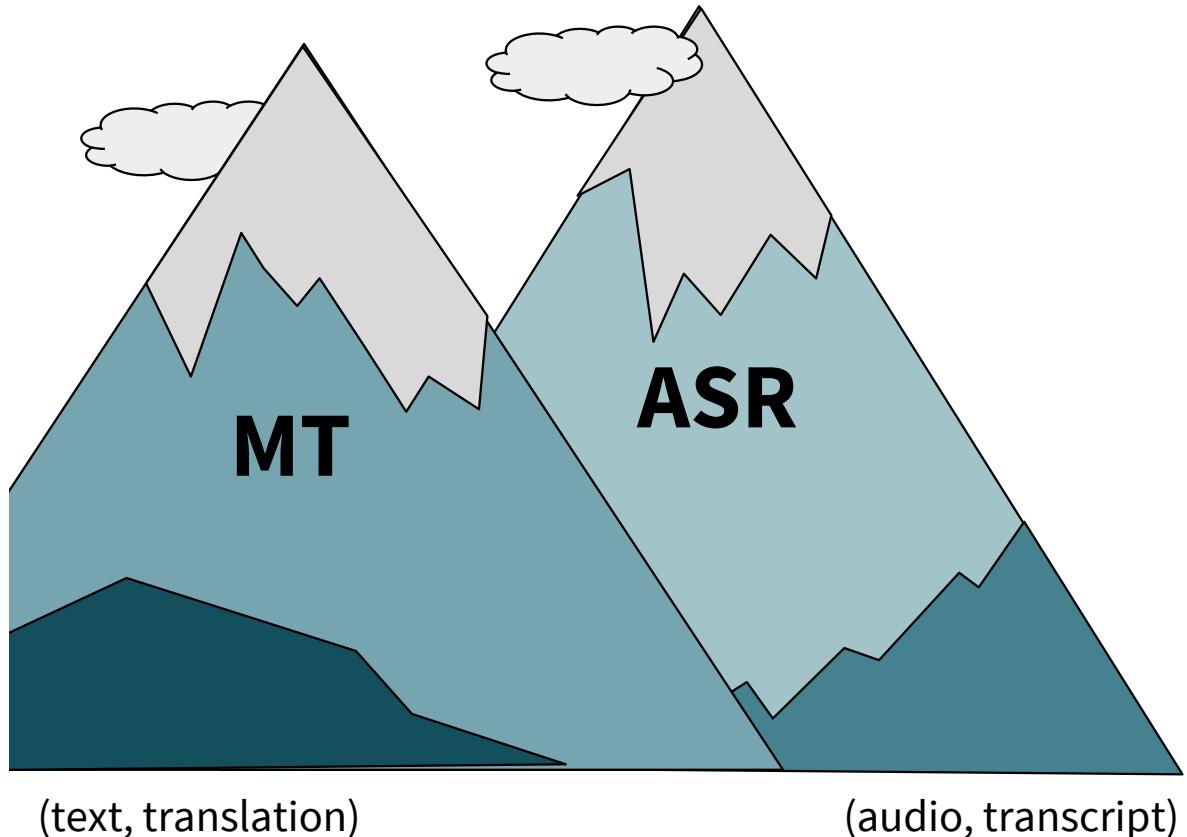
(2) Multitasking UNITY





Today: data mining

Recap: Available data

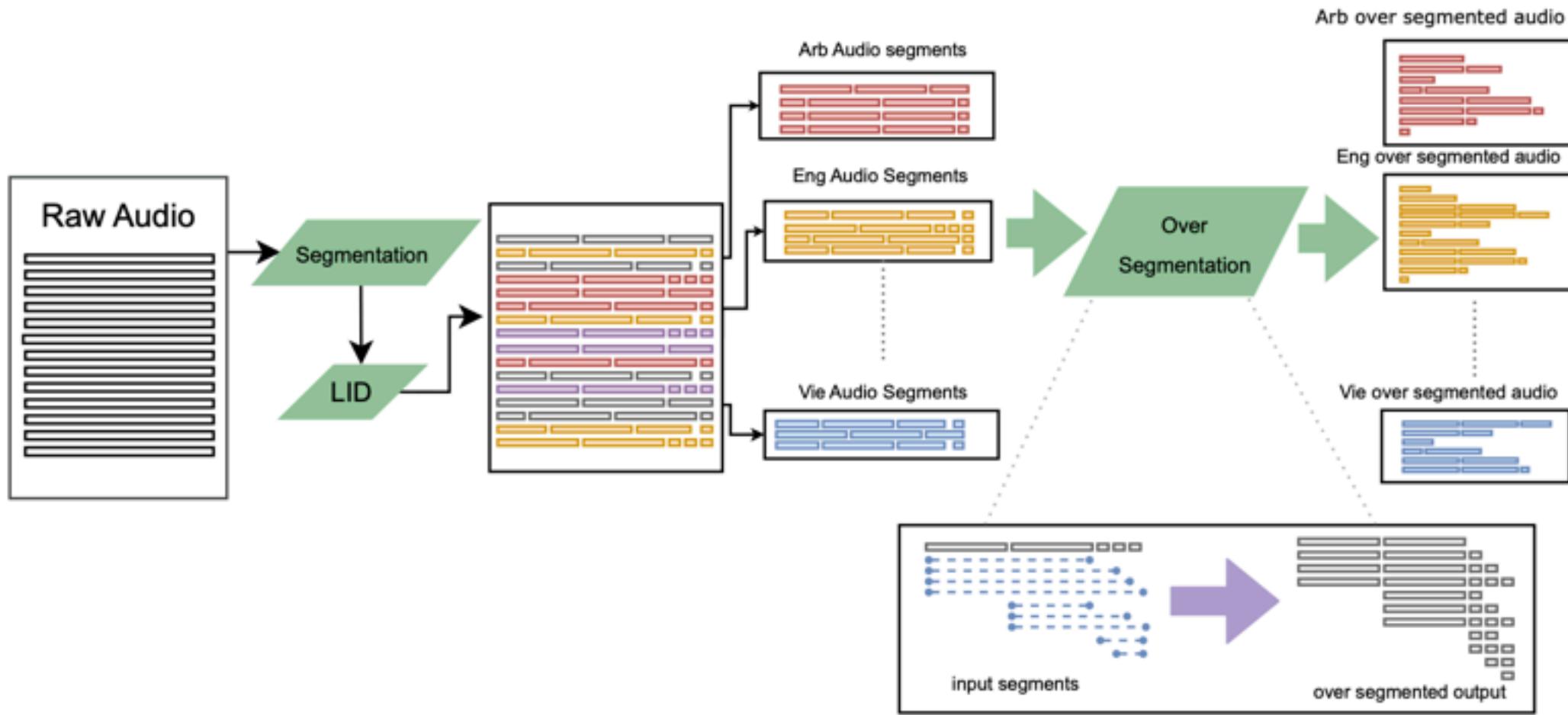


Can we make use of this large amount of data?

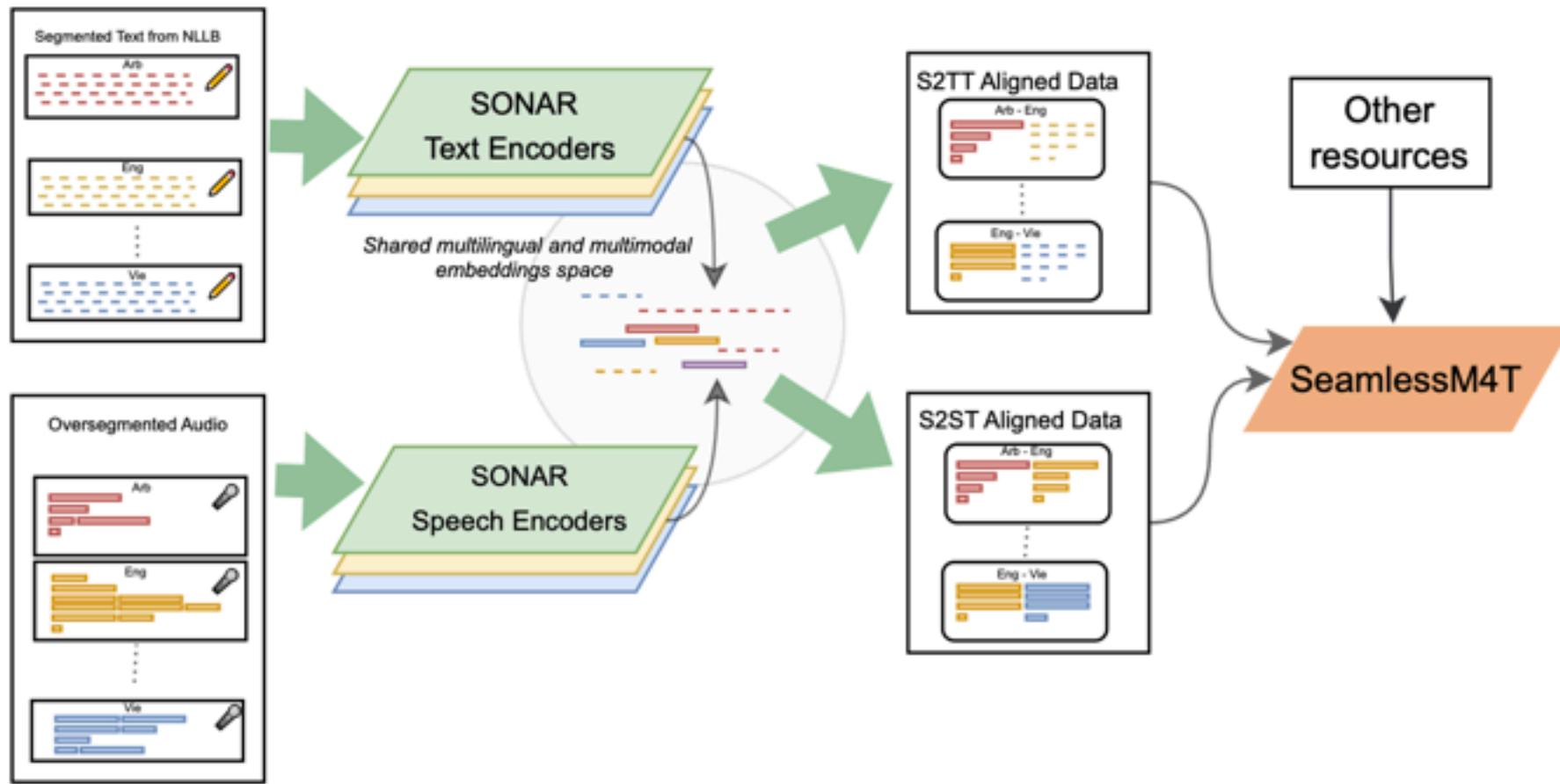


(audio, transcript, translation)

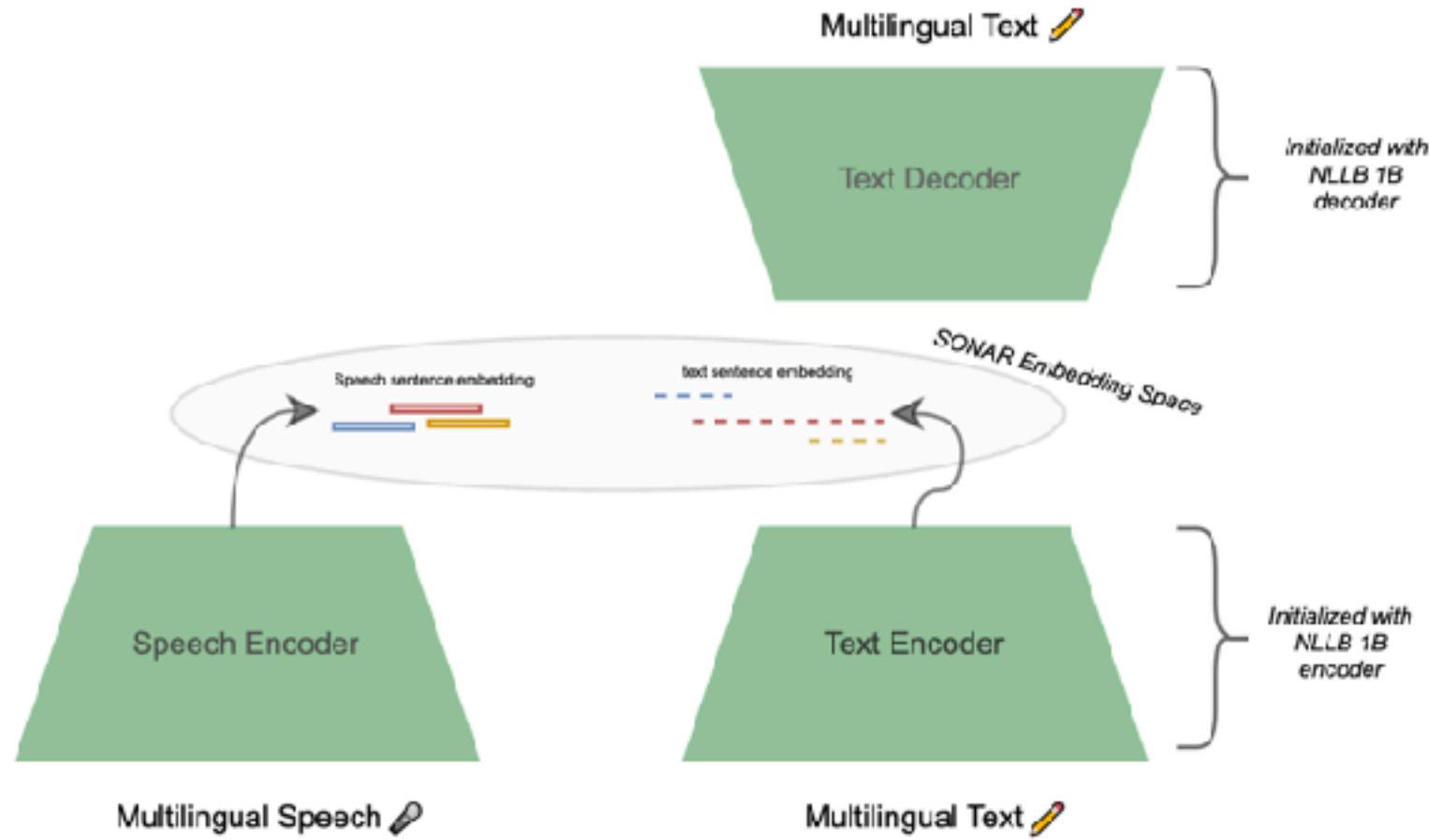
Mining Parallel Speech Data



SONAR Representations



SONAR Representations



SeamlessM4T Results

SeamlessM4T Results

tl;dr: it's great