NLP for Low-resource or Endangered Languages and Cross-lingual transfer

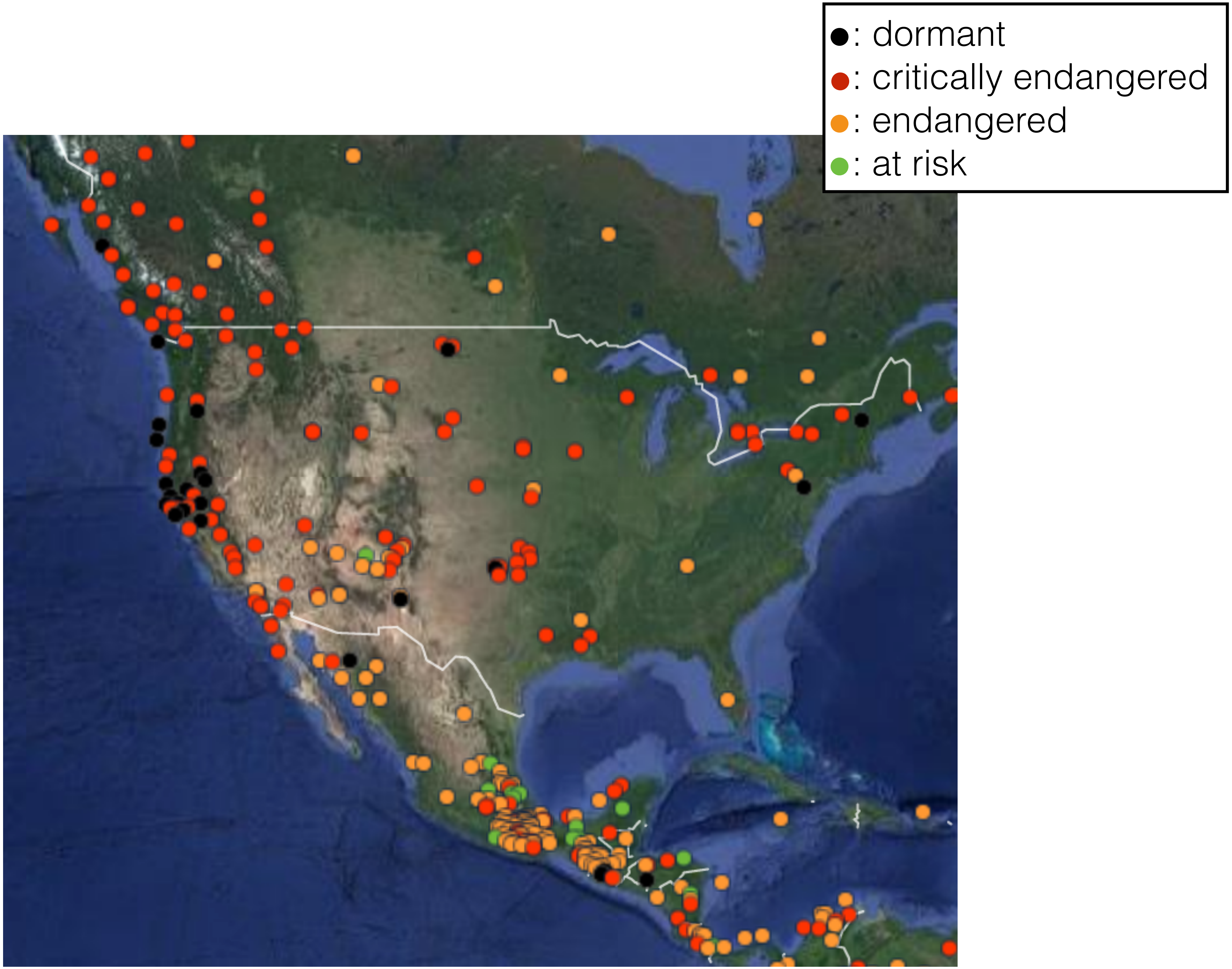
Antonios Anastasopoulos
MTMA

May 31, 2019

There are about ___**7000**___ languages in the world

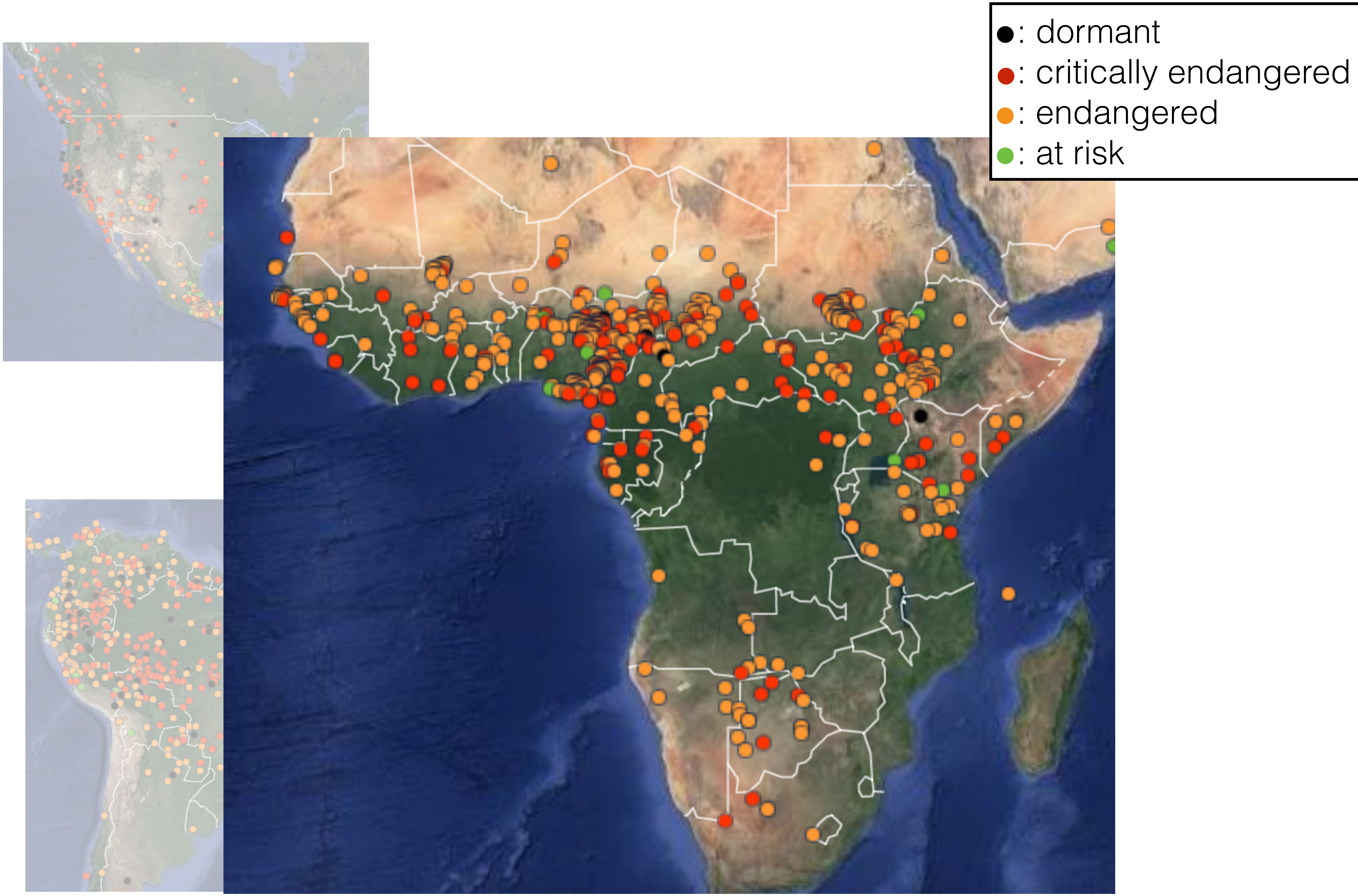
There are about ___**7000**___ languages in the world

According to UNESCO, ___**43**___% of the world's languages are endangered or vulnerable.

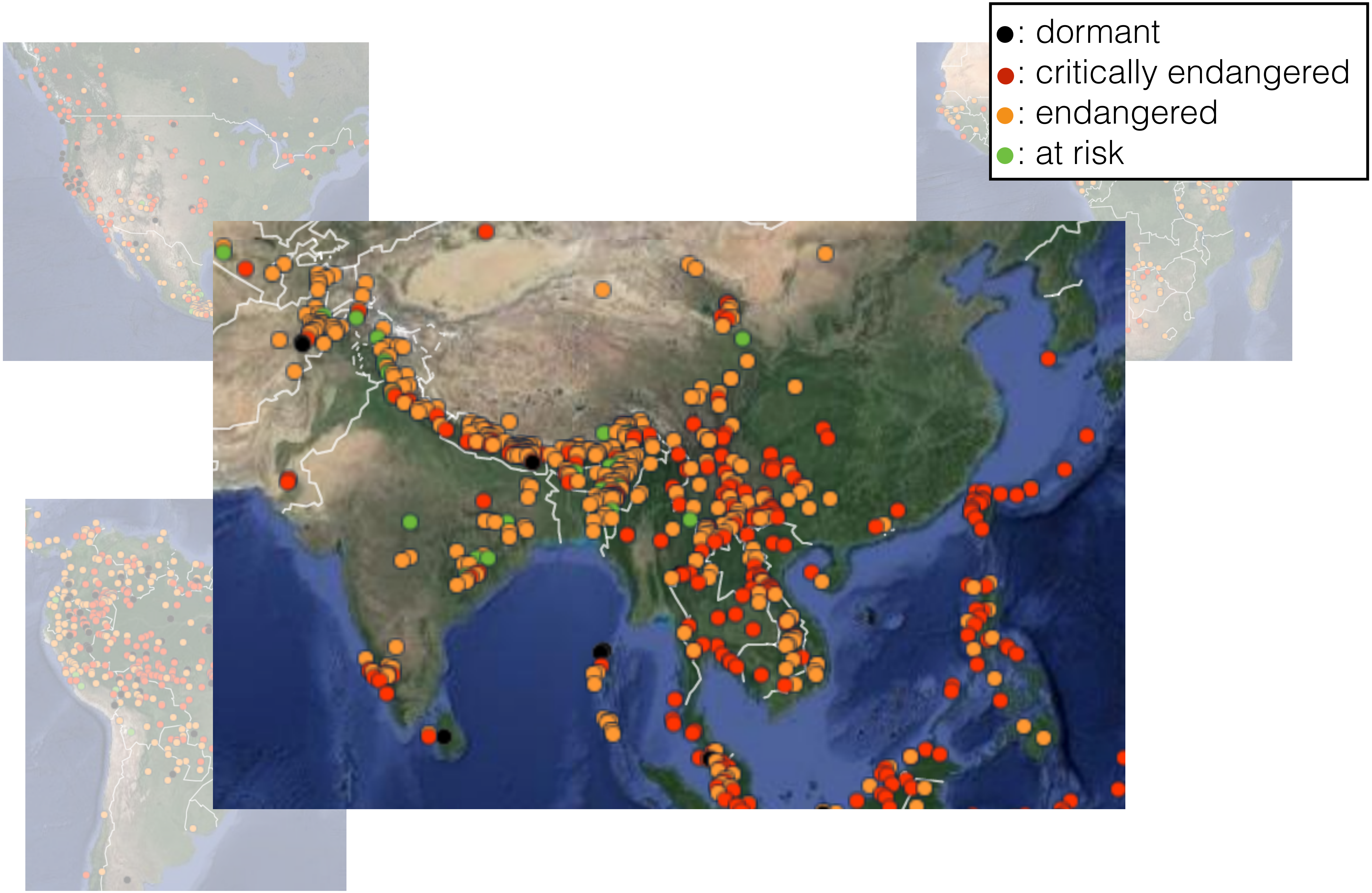


[from The Endangered Languages Project]

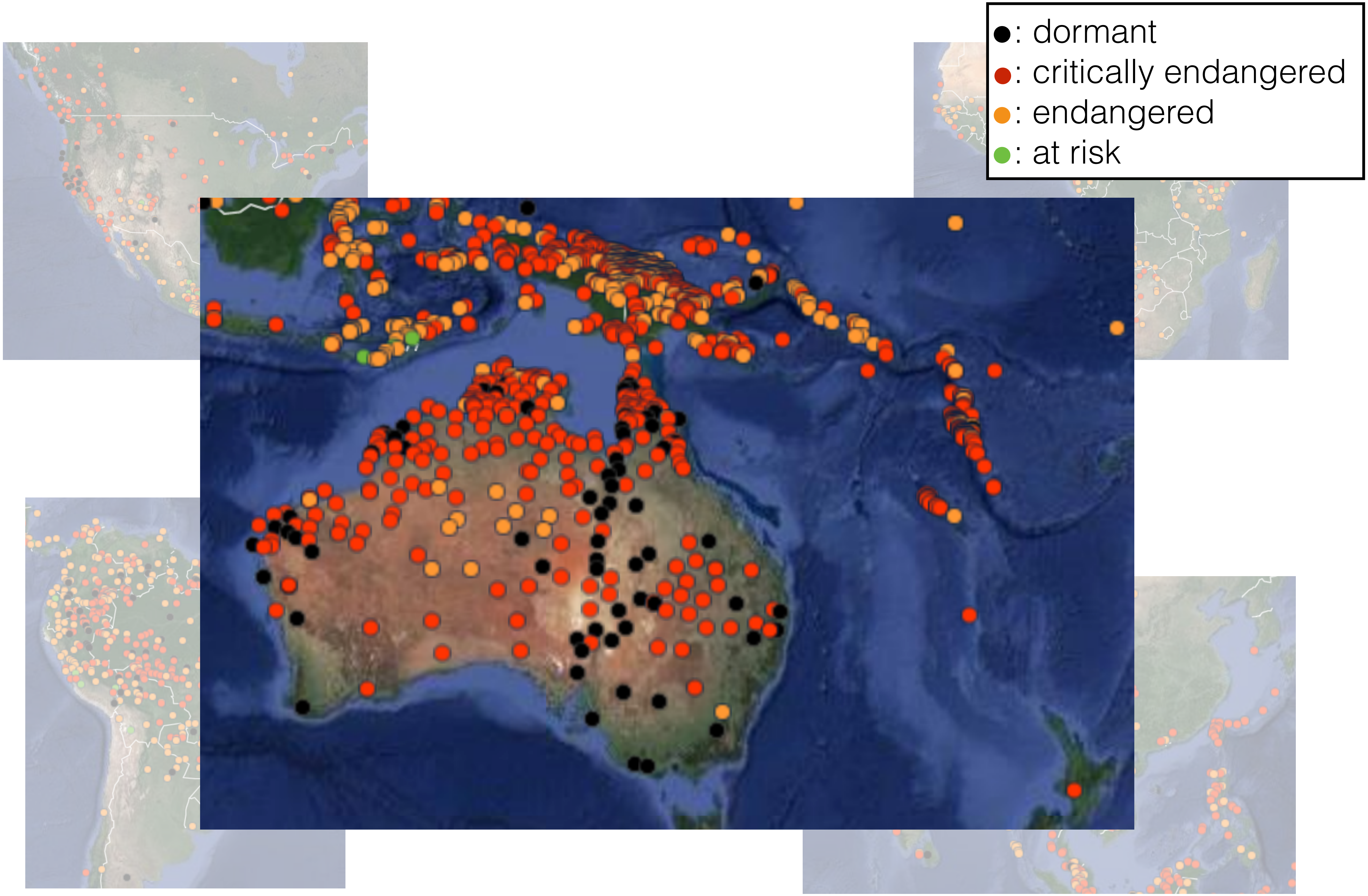
[from The Endangered Languages Project]



[from The Endangered Languages Project]



[from The Endangered Languages Project]



[from The Endangered Languages Project]

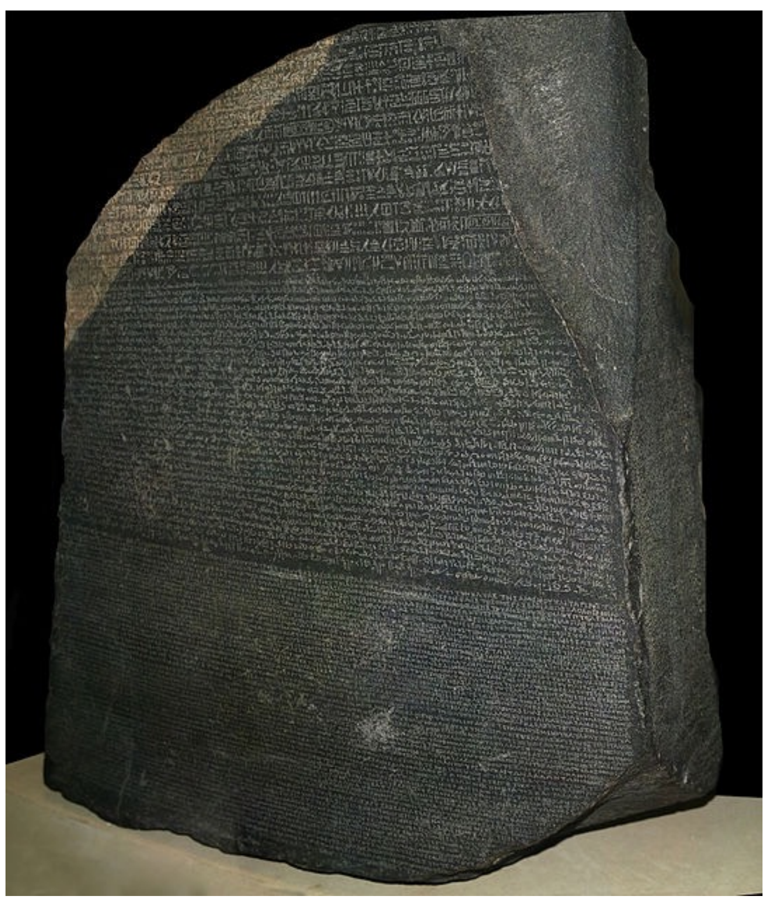
Language Documentation

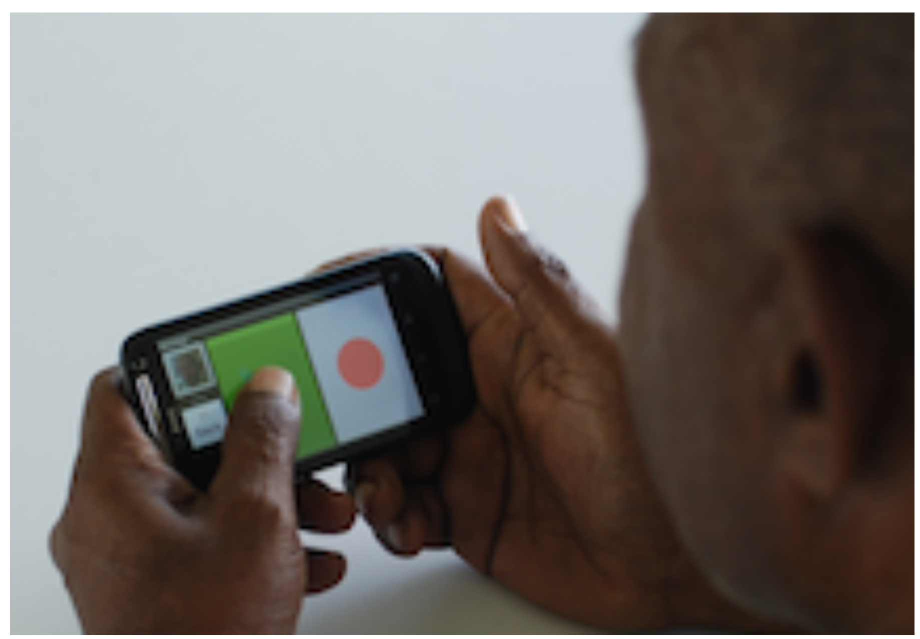
1. Collect (record) data
2. Transcribe and translate
3. Perform analysis
4. Elicit further paradigms
5. Prepare a grammar

Making an
Audio Rosetta Stone

"We collect and archive language recordings now while the speakers are still alive. That's all. We have the whole of the future to transcribe and process the recordings...

Steven Bird





AIKUMA

My work

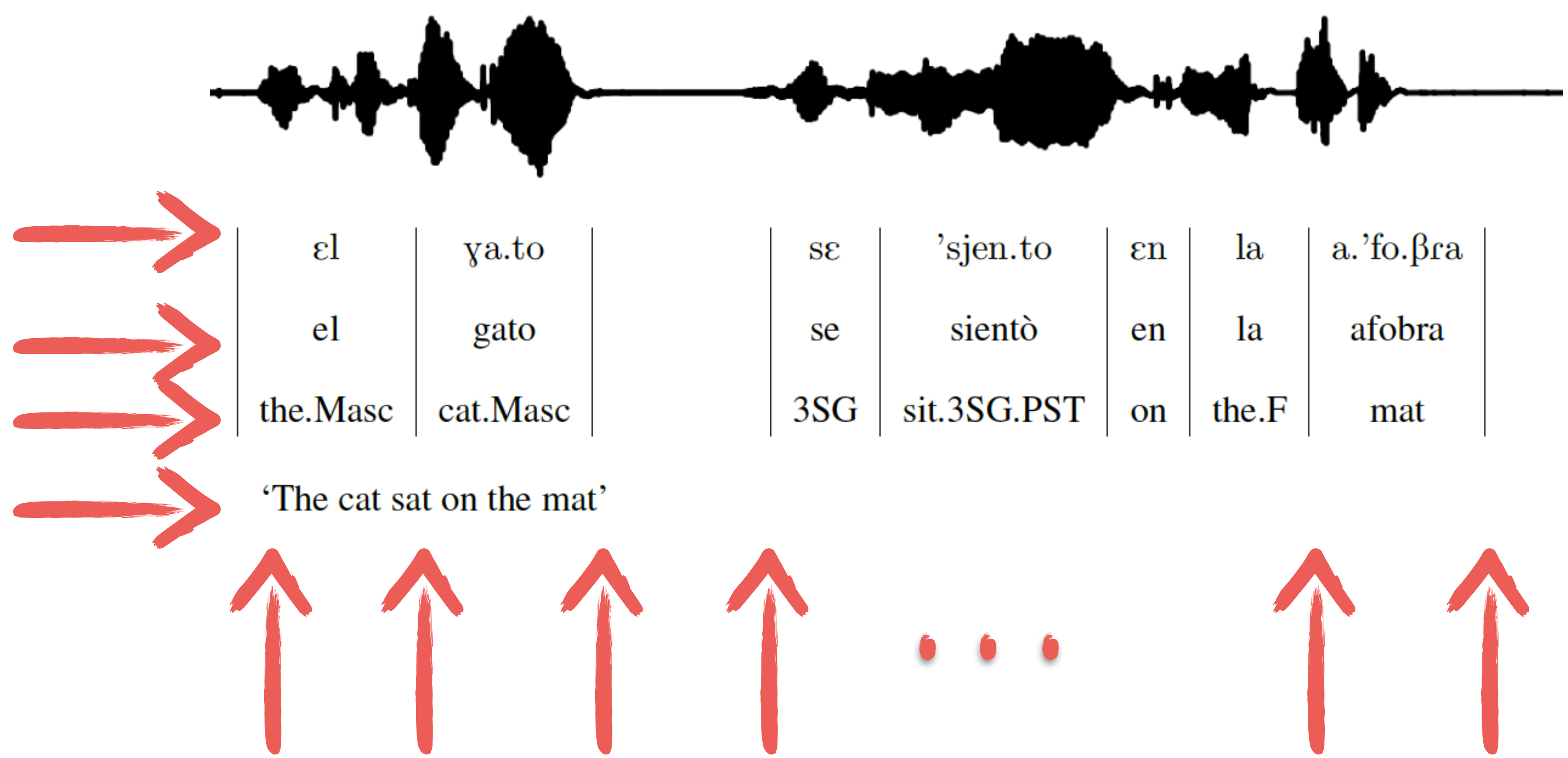
Develop methods that will automate and speed up the language documentation process:

Alignment (segmentation)

Transcription

Translation

Analysis

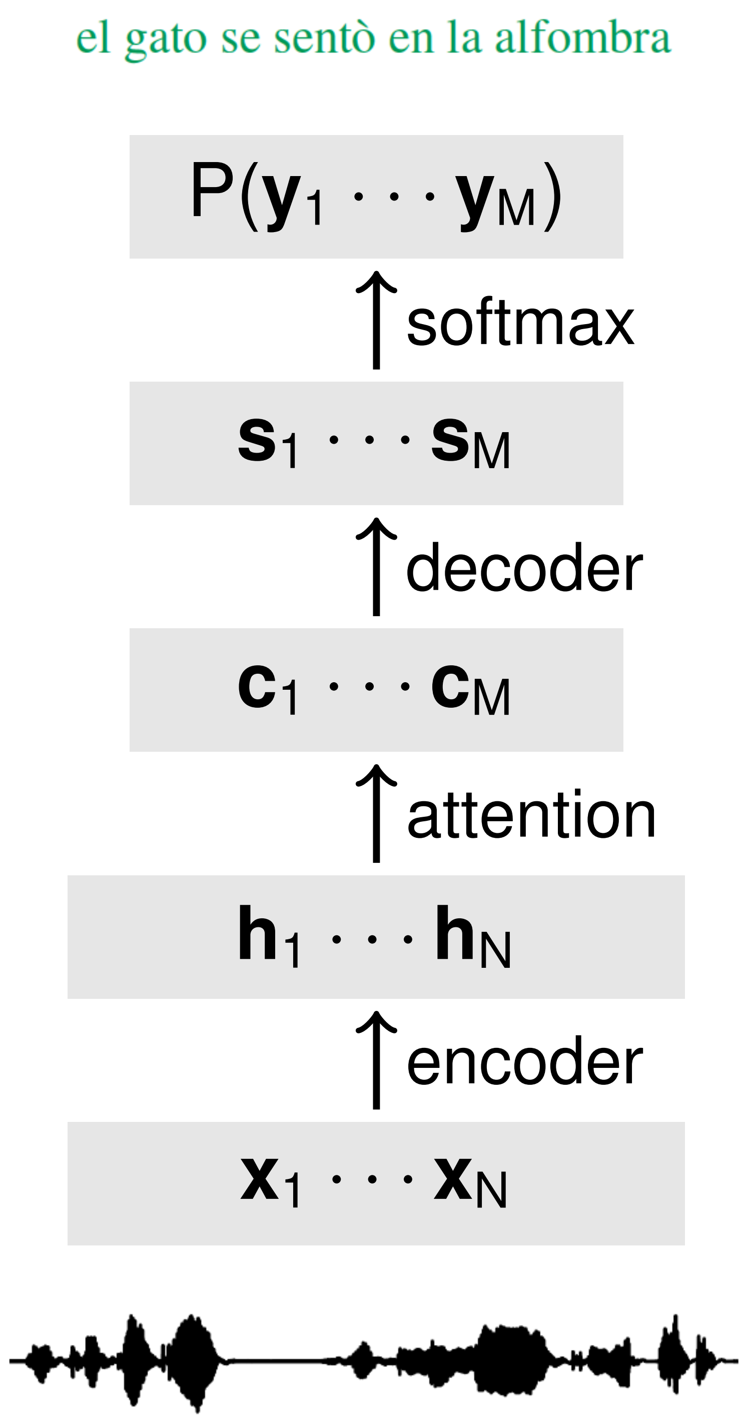


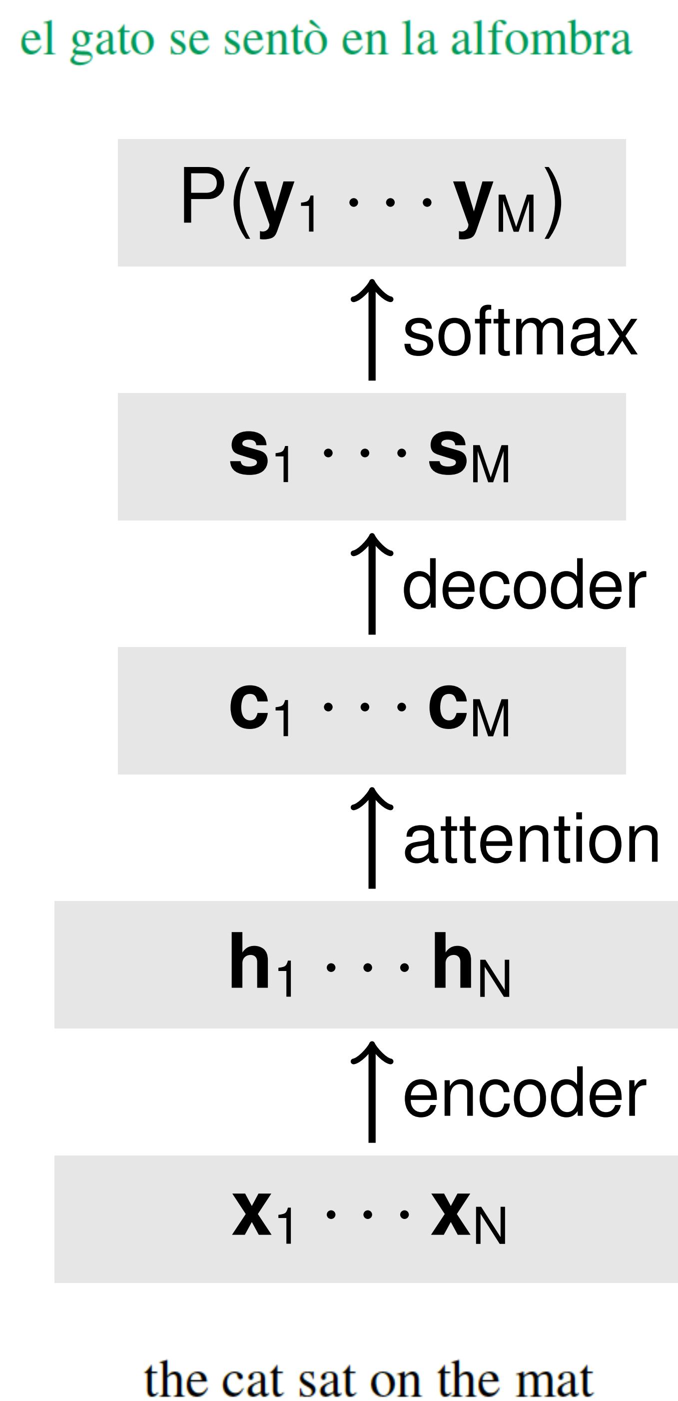
Speech Transcription using Translations

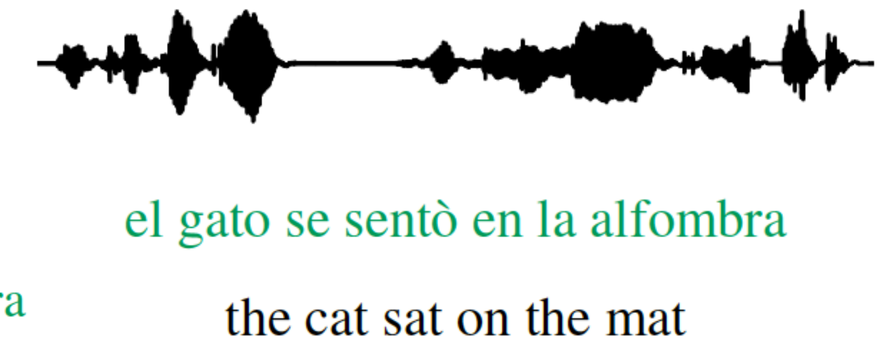
el gato se sentò en la alfombra

the cat sat on the mat

Leveraging Translations for Speech Transcription in Low-Resource Settings
Antonios Anastasopoulos and David Chiang.
Interspeech 2018

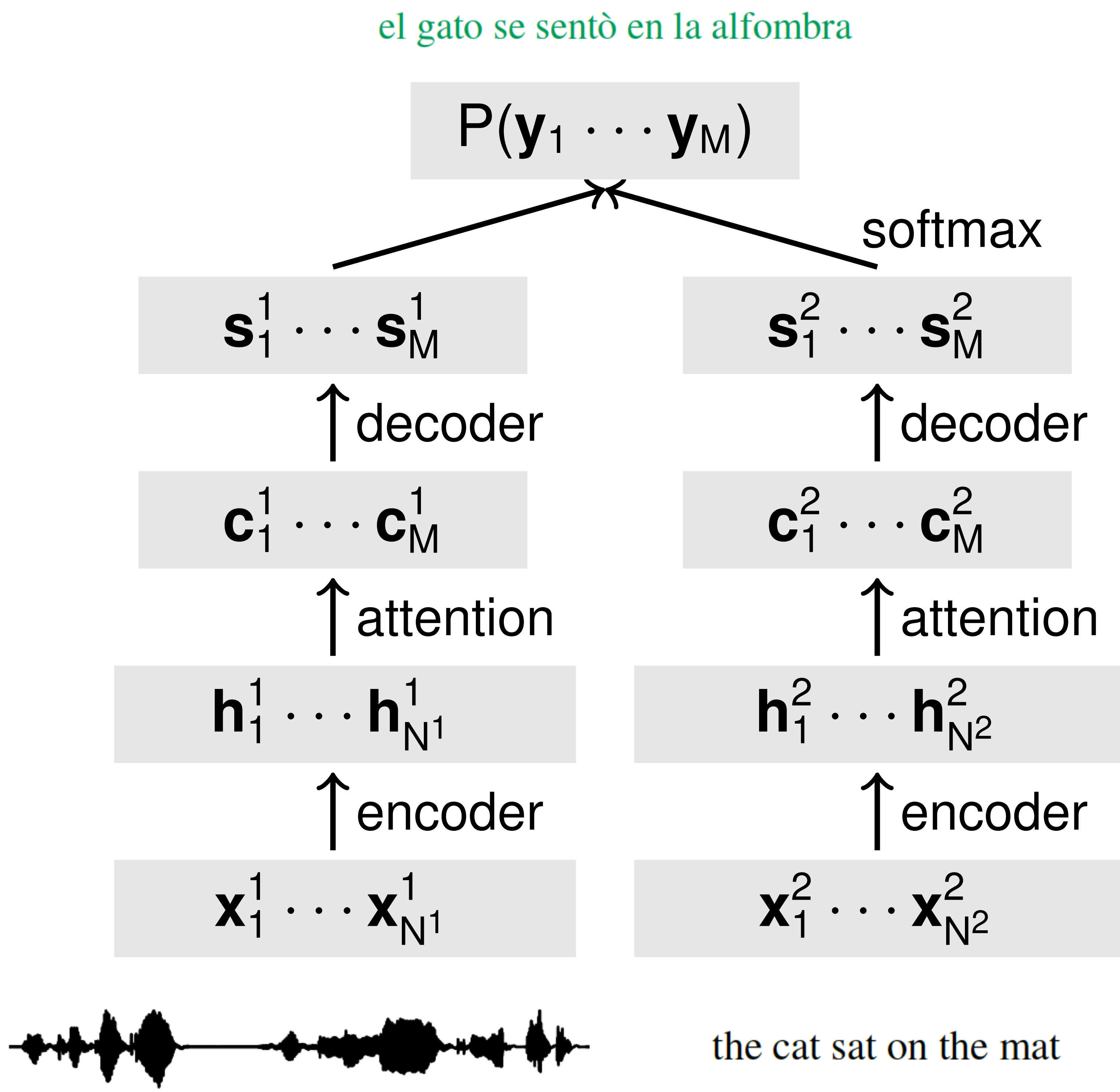






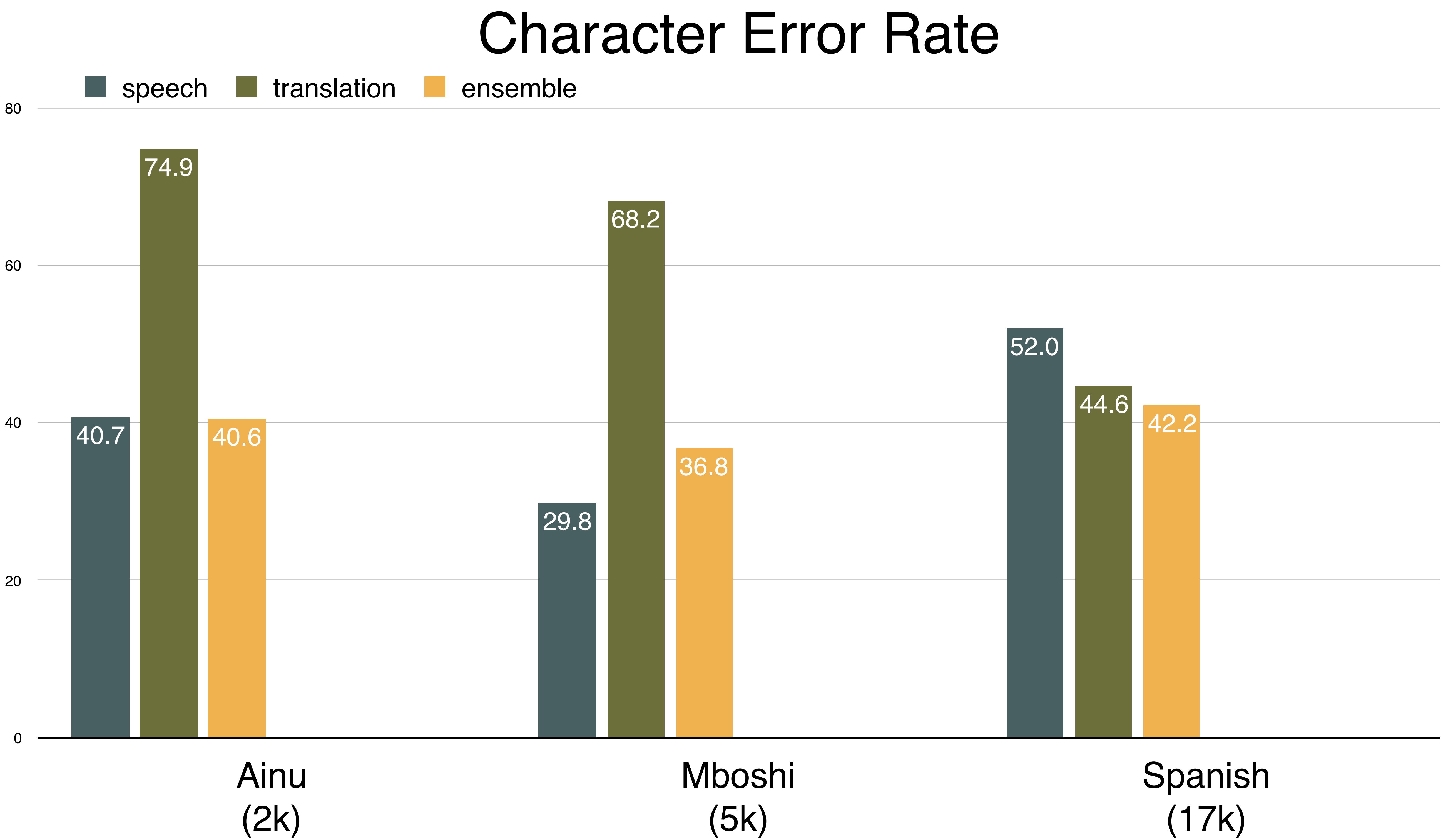
multi-source models: encoder-decoder model

multi-source models: results

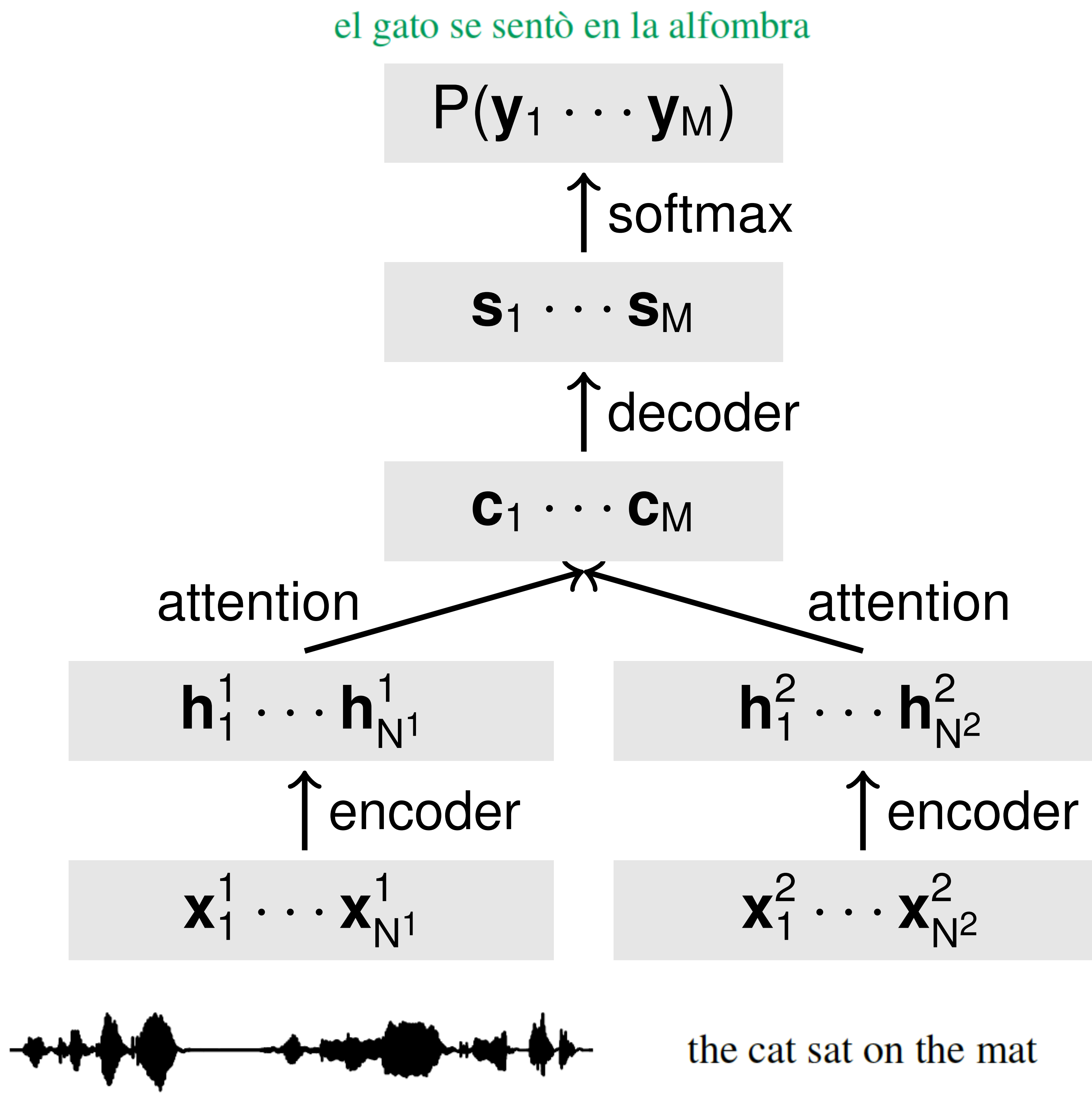


[Dutt et al. 2017]

multi-source models: coupled ensemble

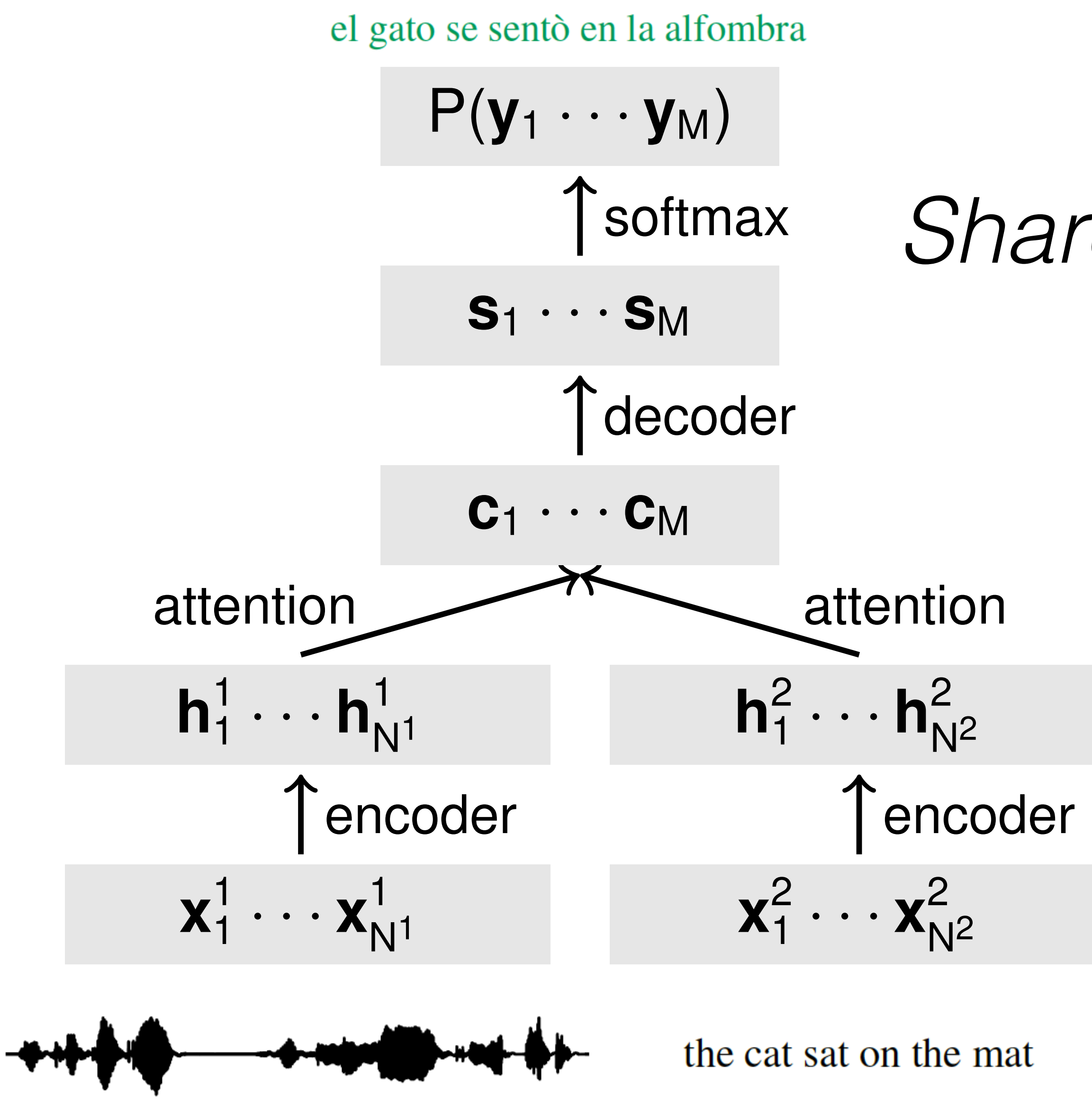


multi-source models: results



multi-source models: multisource

multi-source models: results



Standard:

$$\alpha^1_{kn} = \mathrm{softmax}(\mathbf{v}^1 \tanh(\left[\mathbf{W}^s_{\alpha^1}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^1}\mathbf{h}^1_n\right]))$$

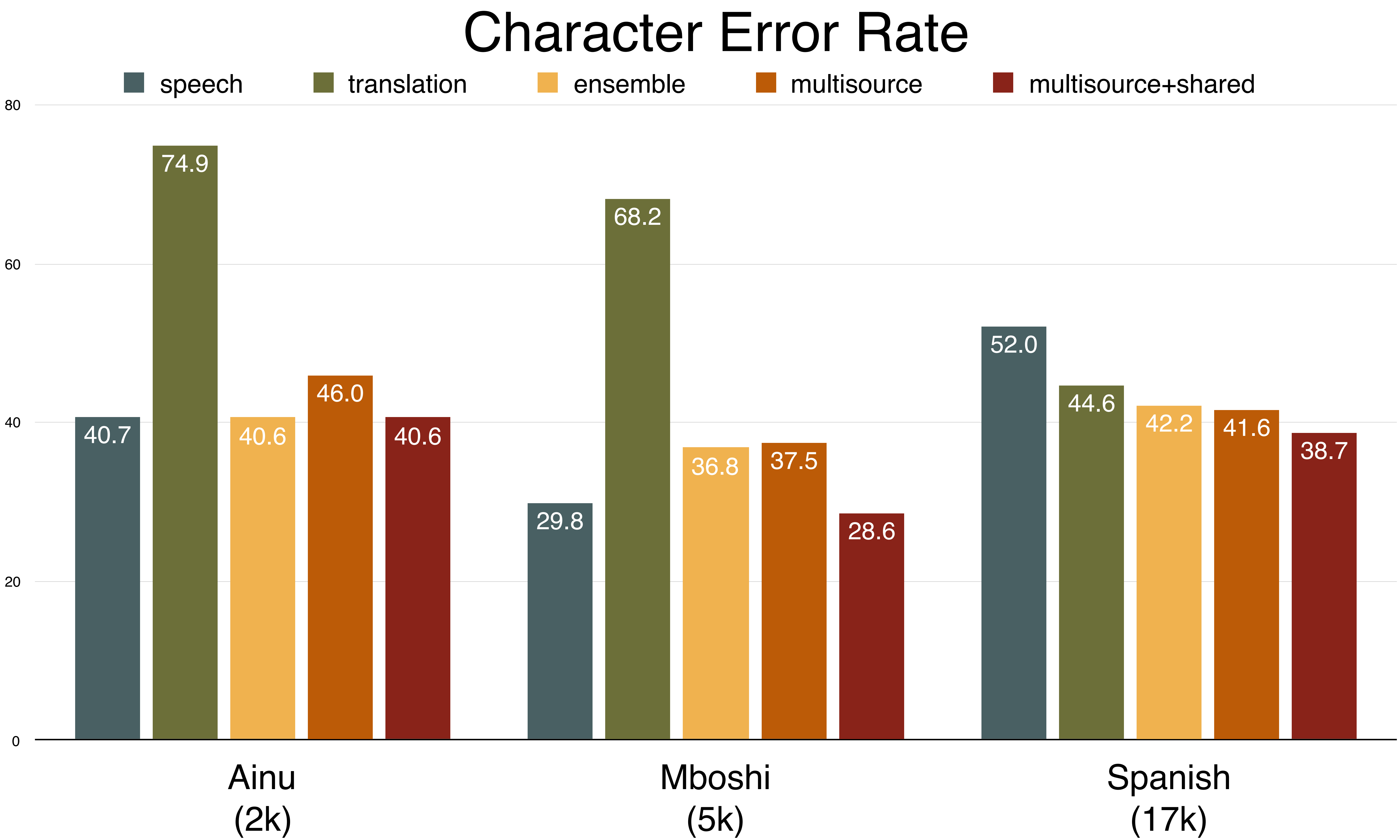
$$\alpha^2_{km} = \mathrm{softmax}(\mathbf{v}^2 \tanh(\left[\mathbf{W}^s_{\alpha^2}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^2}\mathbf{h}^2_m\right]))$$

Shared:

$$\alpha^1_{kn} = \mathrm{softmax}(\boxed{\mathbf{v}} \tanh(\left[\boxed{\mathbf{W}^s_{\alpha}}\mathbf{s}_{k-1}; \boxed{\mathbf{W}^h_{\alpha}}\mathbf{h}^1_n\right]))$$

$$\alpha^2_{km} = \mathrm{softmax}(\boxed{\mathbf{v}} \tanh(\left[\boxed{\mathbf{W}^s_{\alpha}}\mathbf{s}_{k-1}; \boxed{\mathbf{W}^h_{\alpha}}\mathbf{h}^2_m\right]))$$

multi-source models: attention parameter sharing



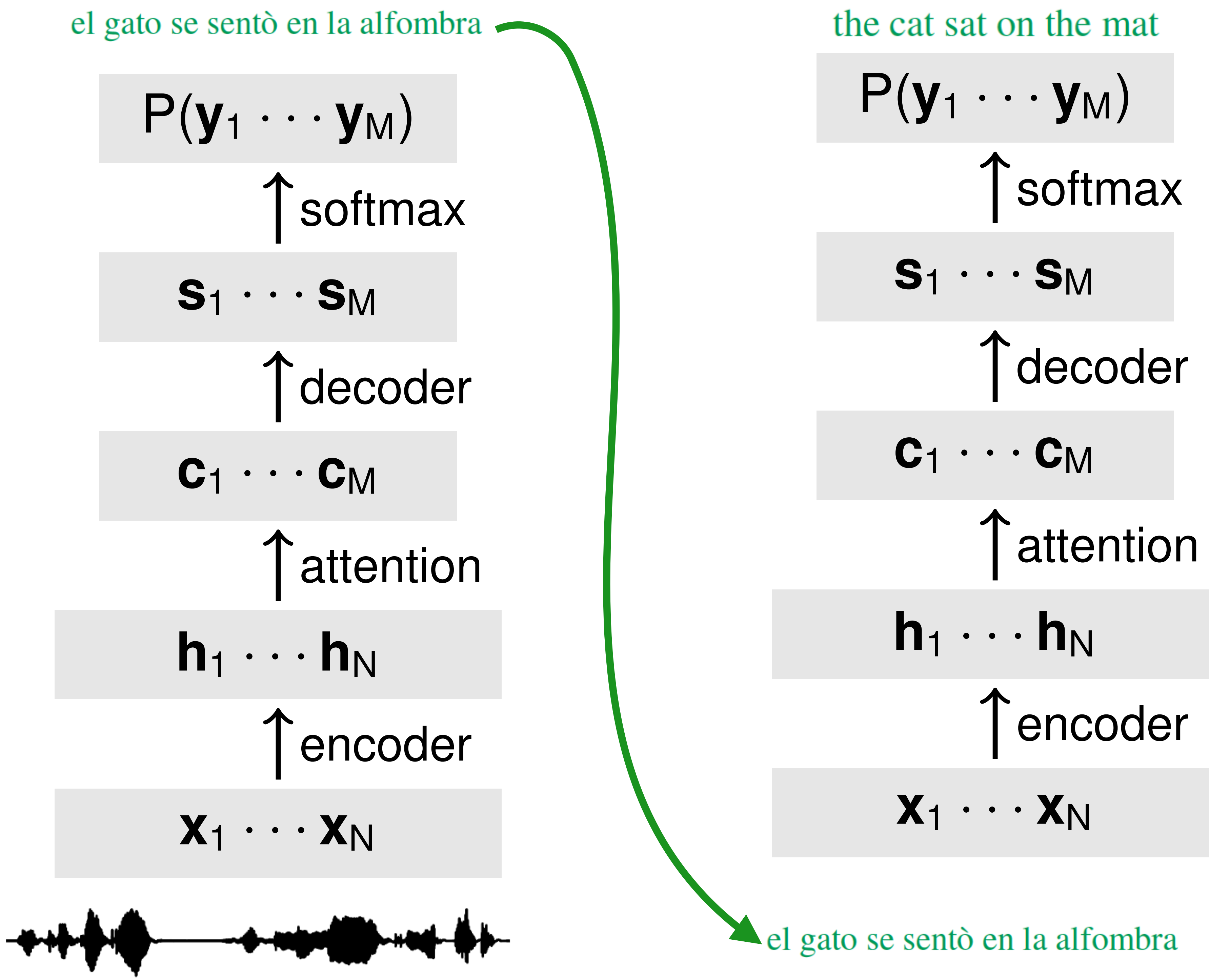
multi-source models: results

Speech Transcription and Translation

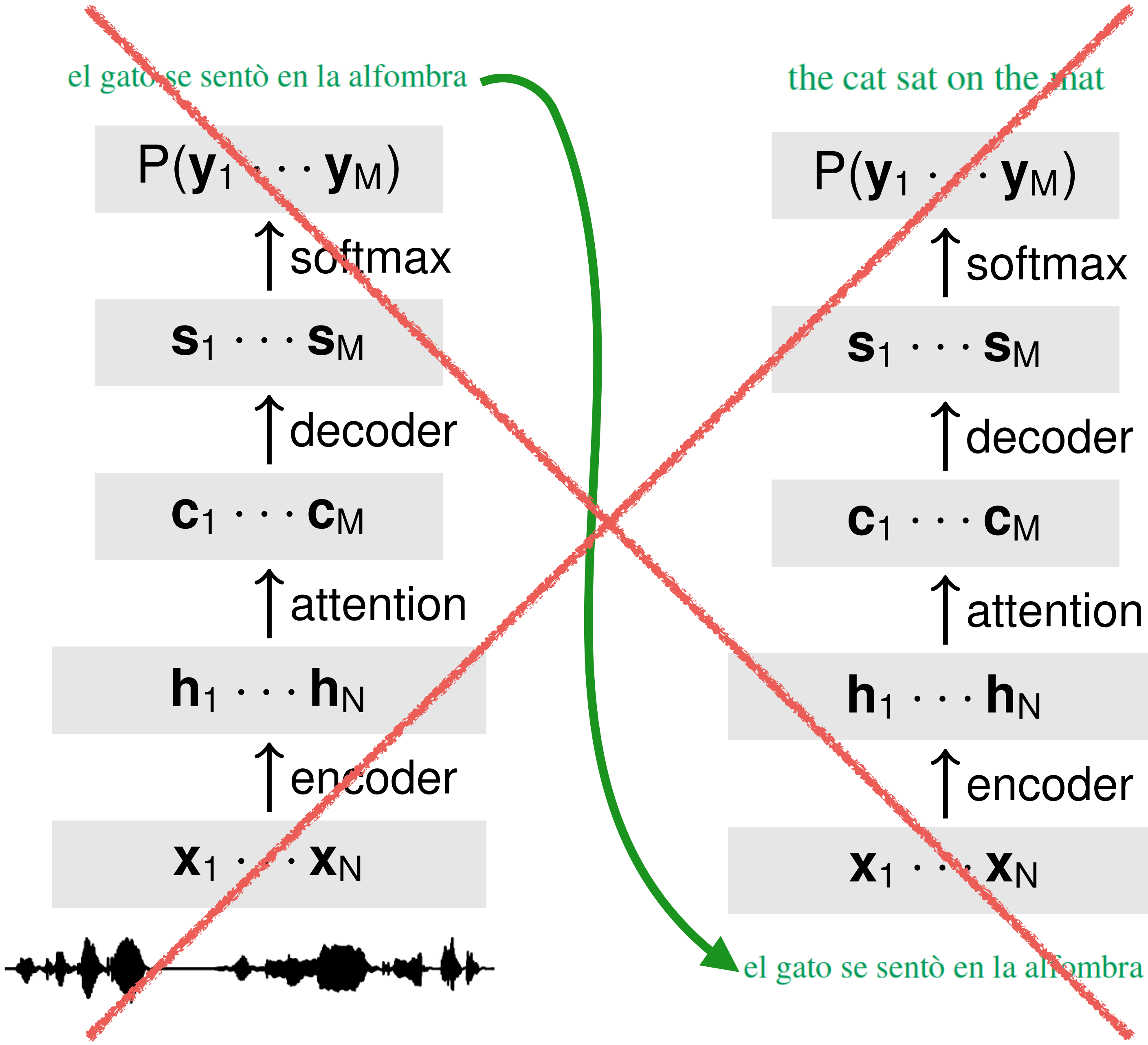
el gato se sentò en la alfombra

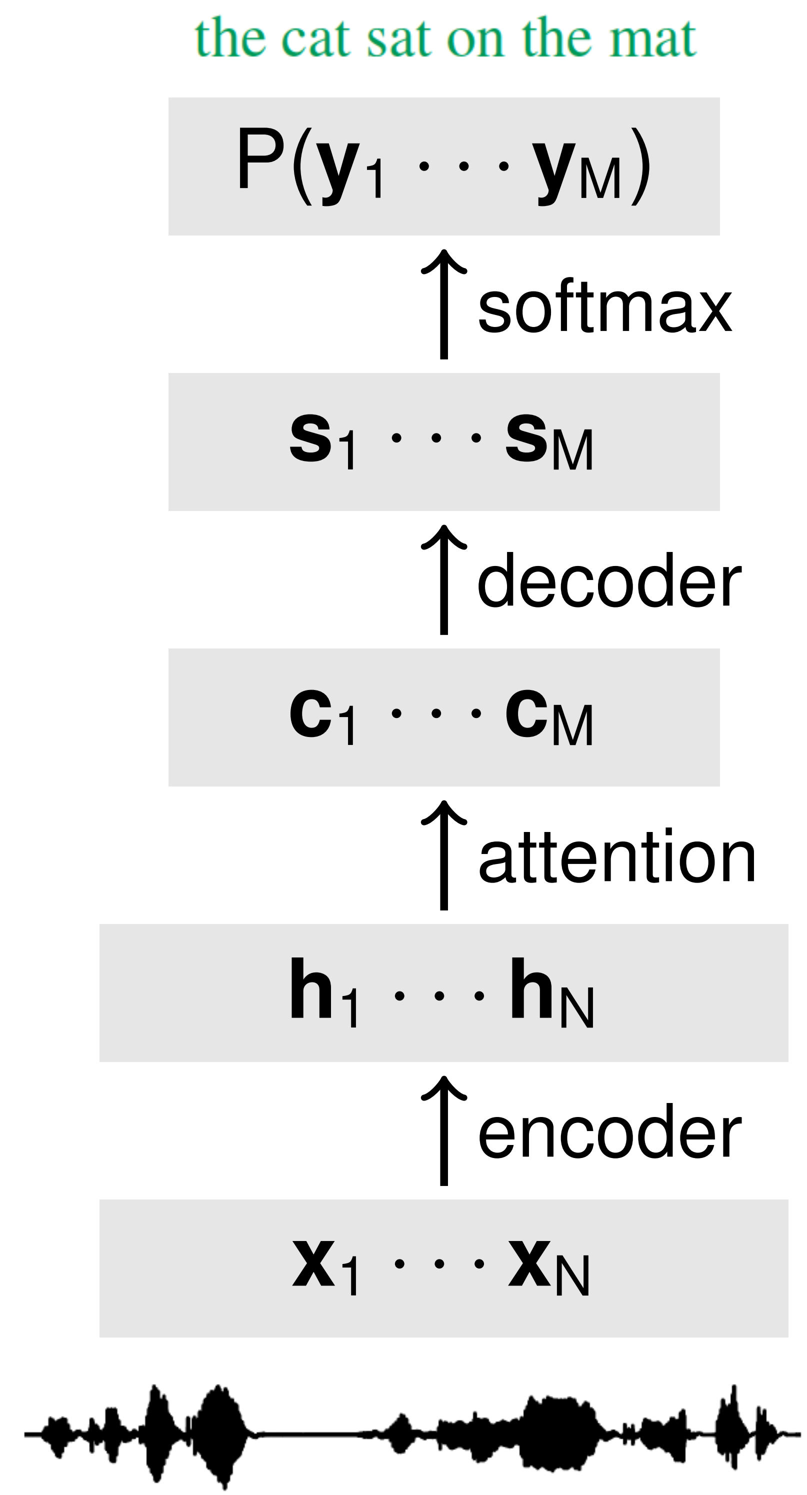
the cat sat on the mat

Tied Multitask Models for Speech Transcription and Translation
Antonios Anastasopoulos and David Chiang.
NAACL 2018.

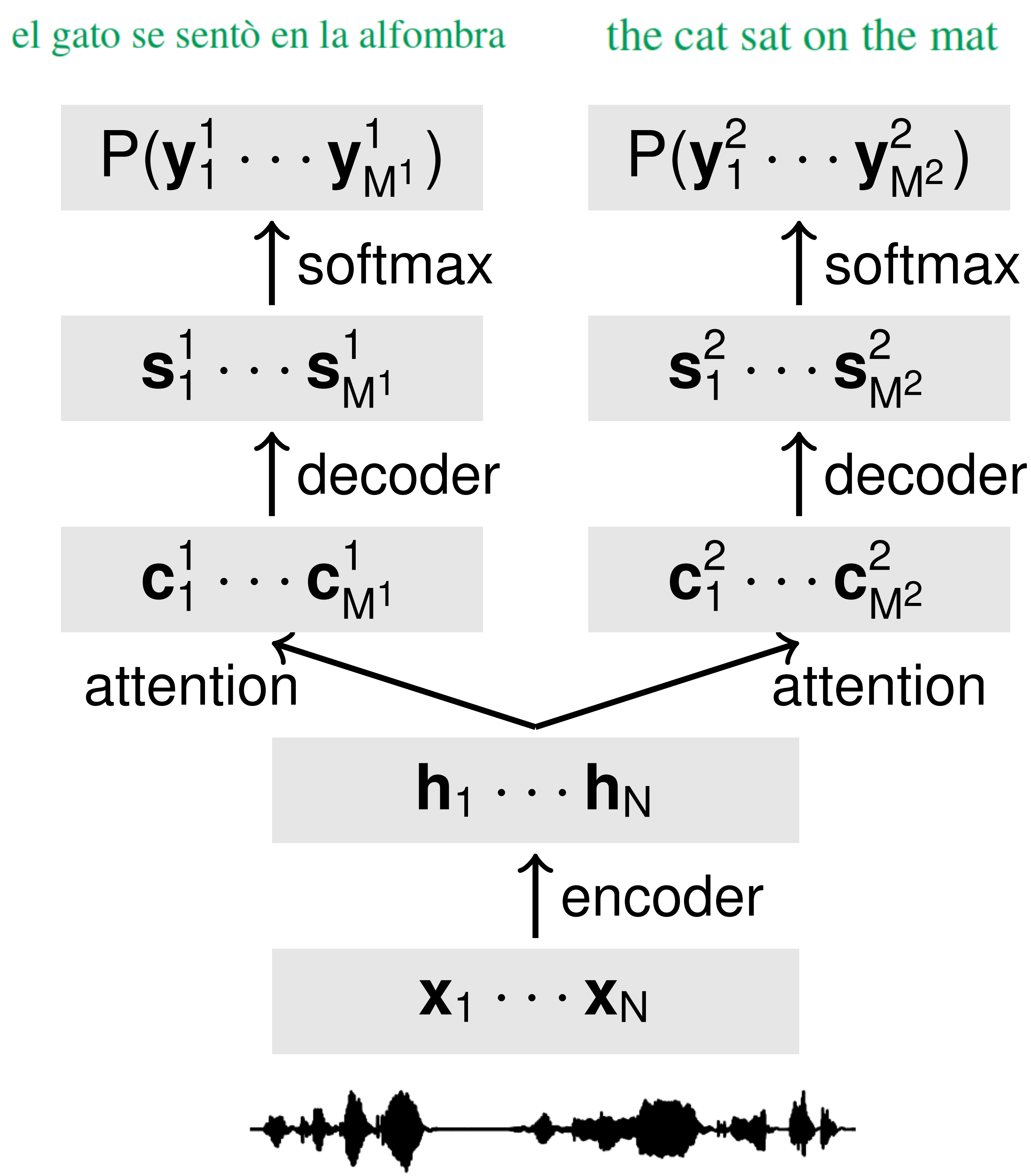


multi-task models: pivot

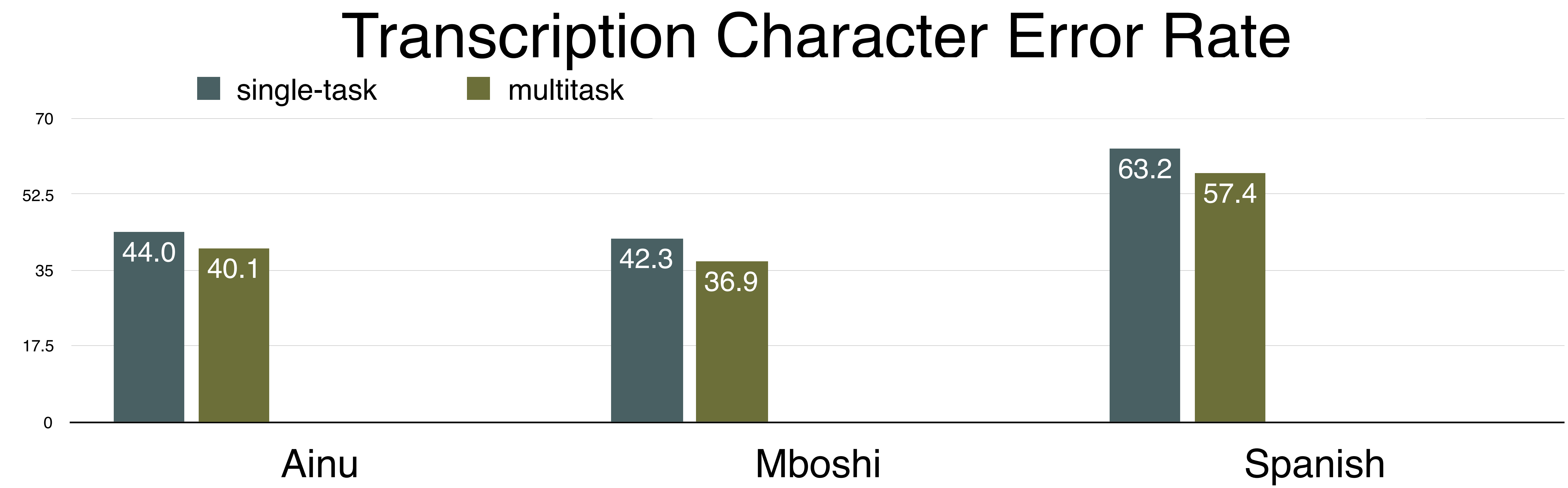


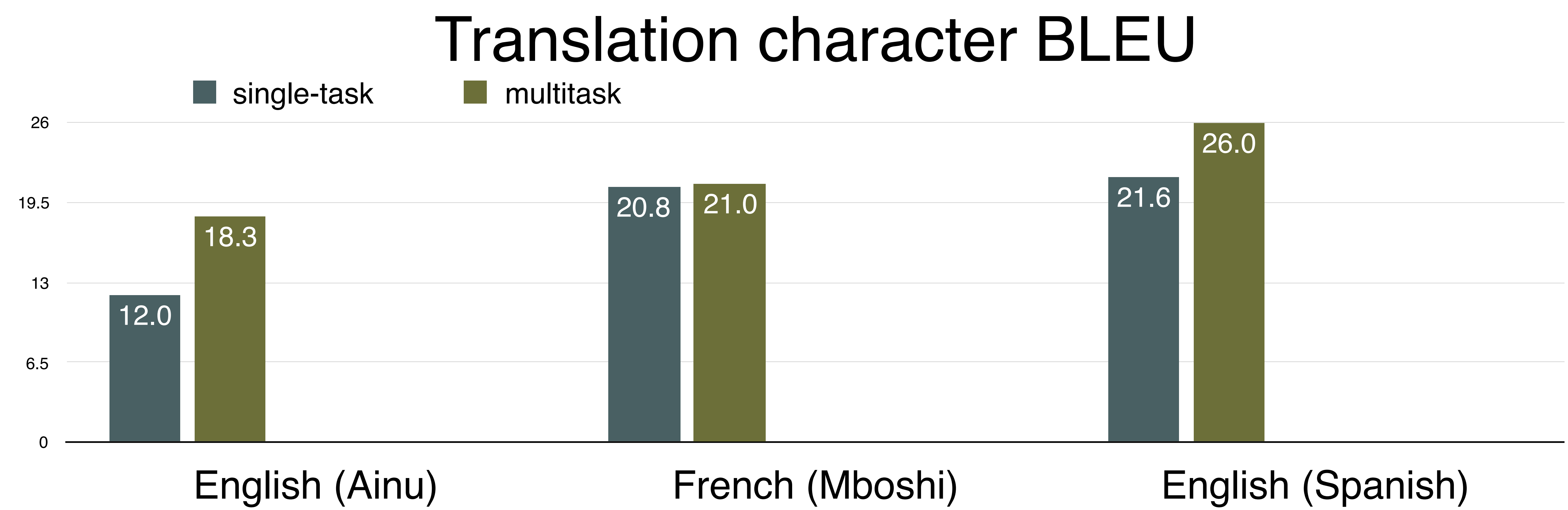


An Attentional Model for *Speech Translation without Transcription*
Long Duong, Antonios Anastasopoulos,
Trevor Cohn, Steven Bird, and David Chiang.
NAACL 2016.

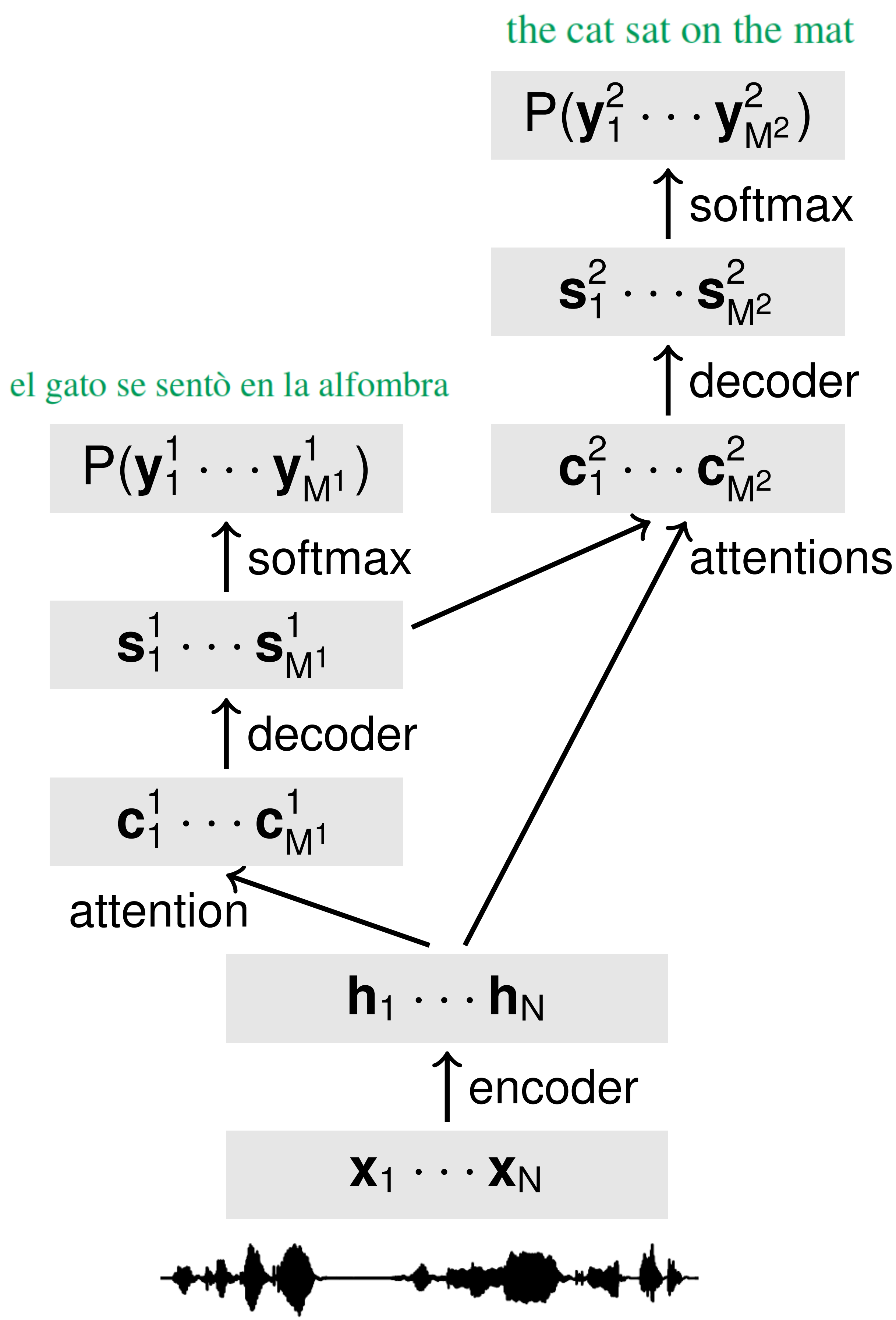


multi-task models: simple multitask

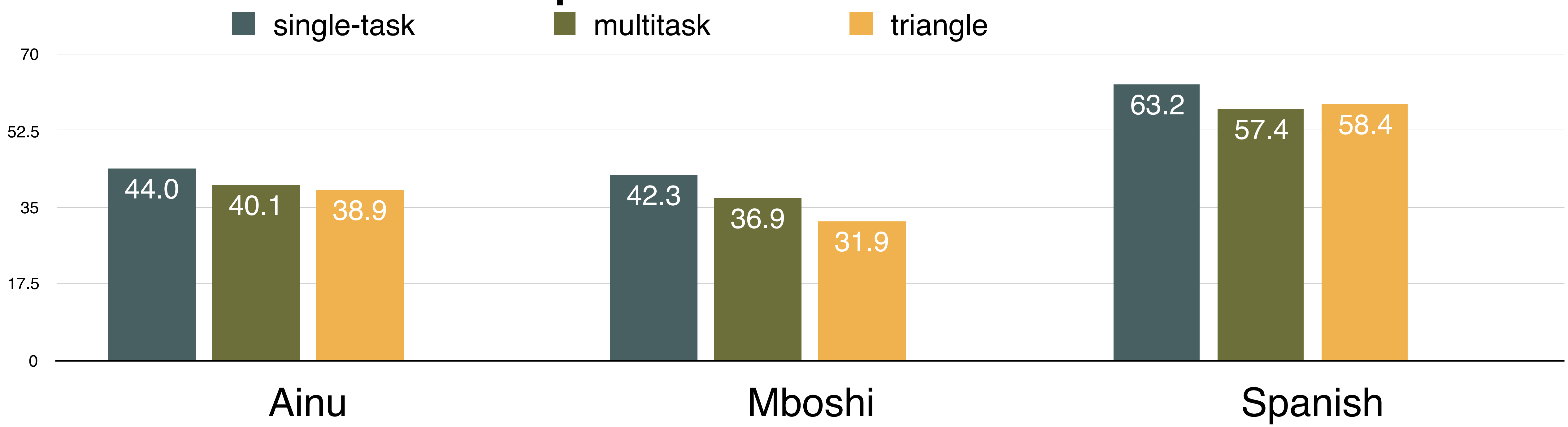




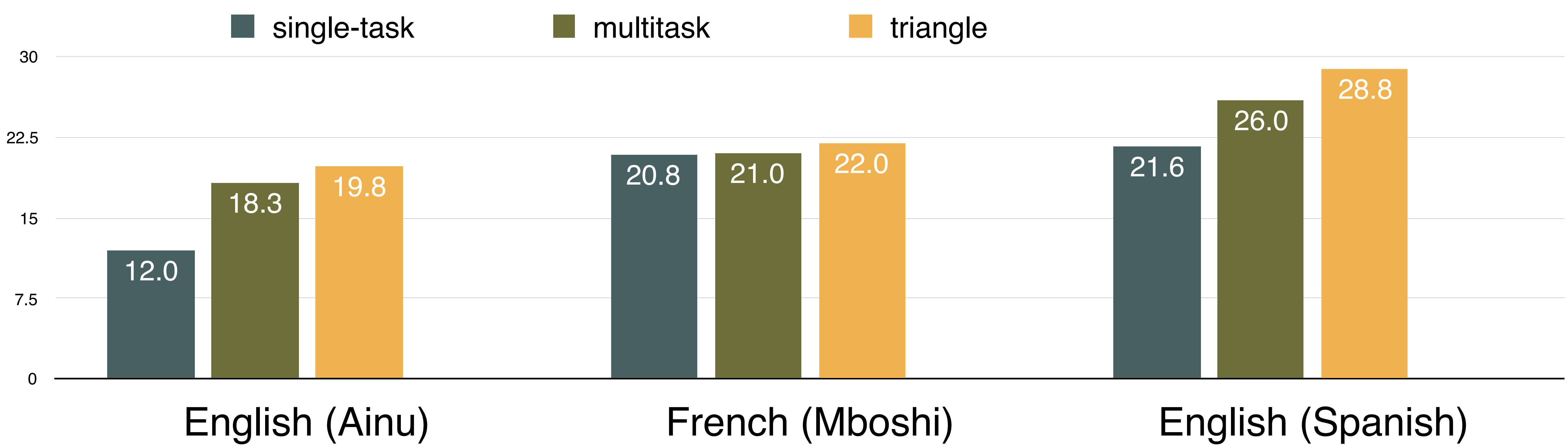
multi-task models: results


the cat sat on the mat
$P(\mathbf{y}^2_1 \cdots \mathbf{y}^2_{M^2})$
softmax
$\mathbf{s}^2_1 \cdots \mathbf{s}^2_{M^2}$
decoder
el gato se sentò en la alfombra
$P(\mathbf{y}^1_1 \cdots \mathbf{y}^1_{M^1})$
$\mathbf{c}^2_1 \cdots \mathbf{c}^2_{M^2}$
softmax
attentions
$\mathbf{s}^1_1 \cdots \mathbf{s}^1_{M^1}$
decoder
$\mathbf{c}^1_1 \cdots \mathbf{c}^1_{M^1}$
attention
$\mathbf{h}_1 \cdots \mathbf{h}_N$
encoder
$\mathbf{x}_1 \cdots \mathbf{x}_N$

Transcription Character Error Rate



Translation character BLEU



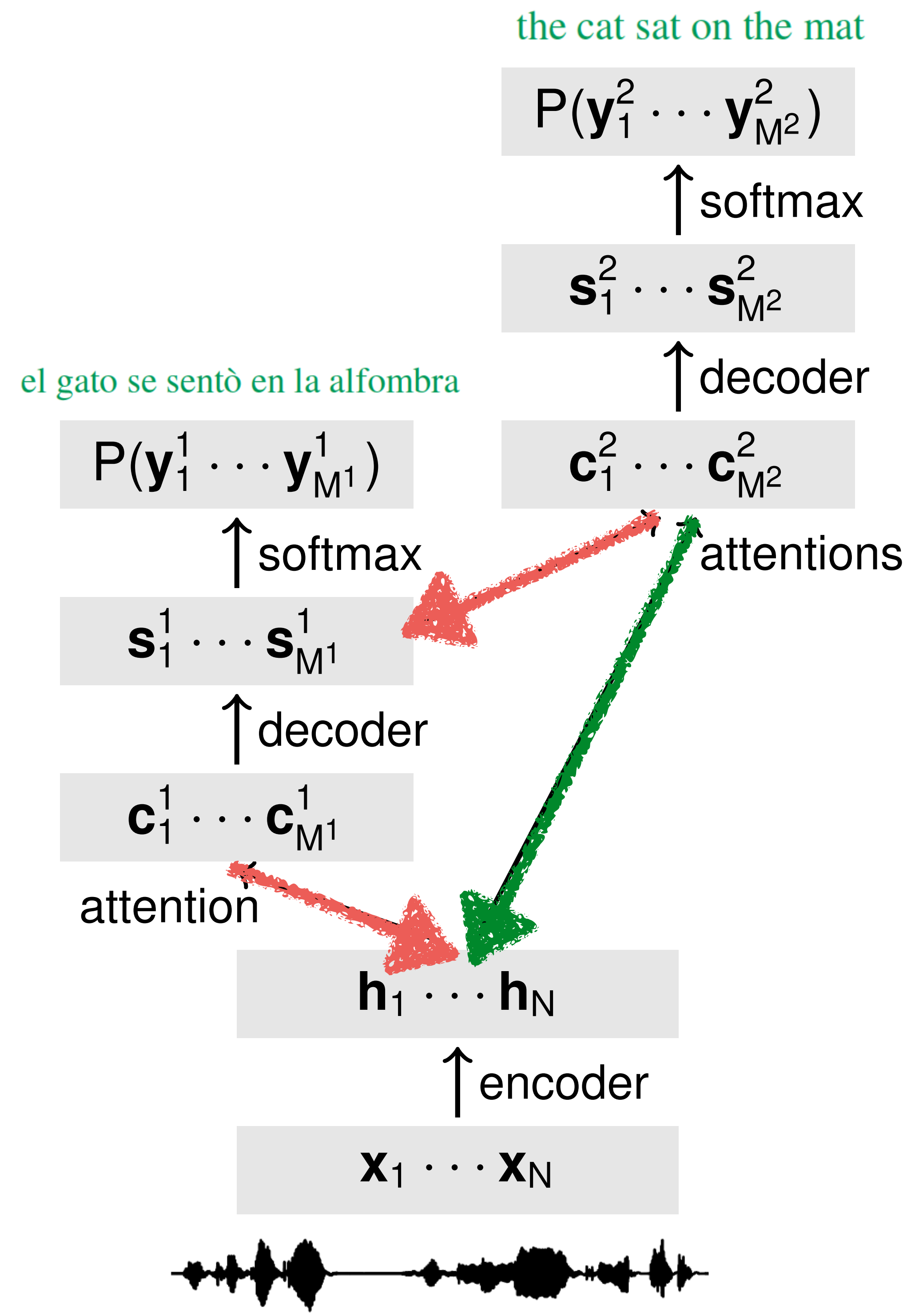
multi-task models: results

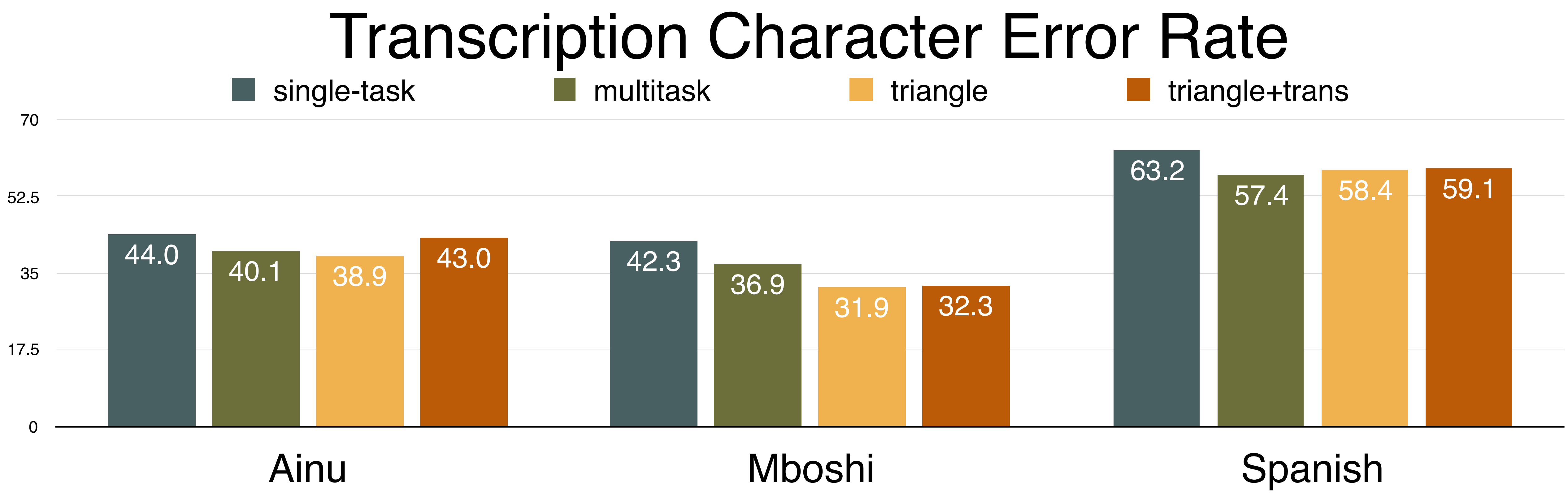
If A attends over B…

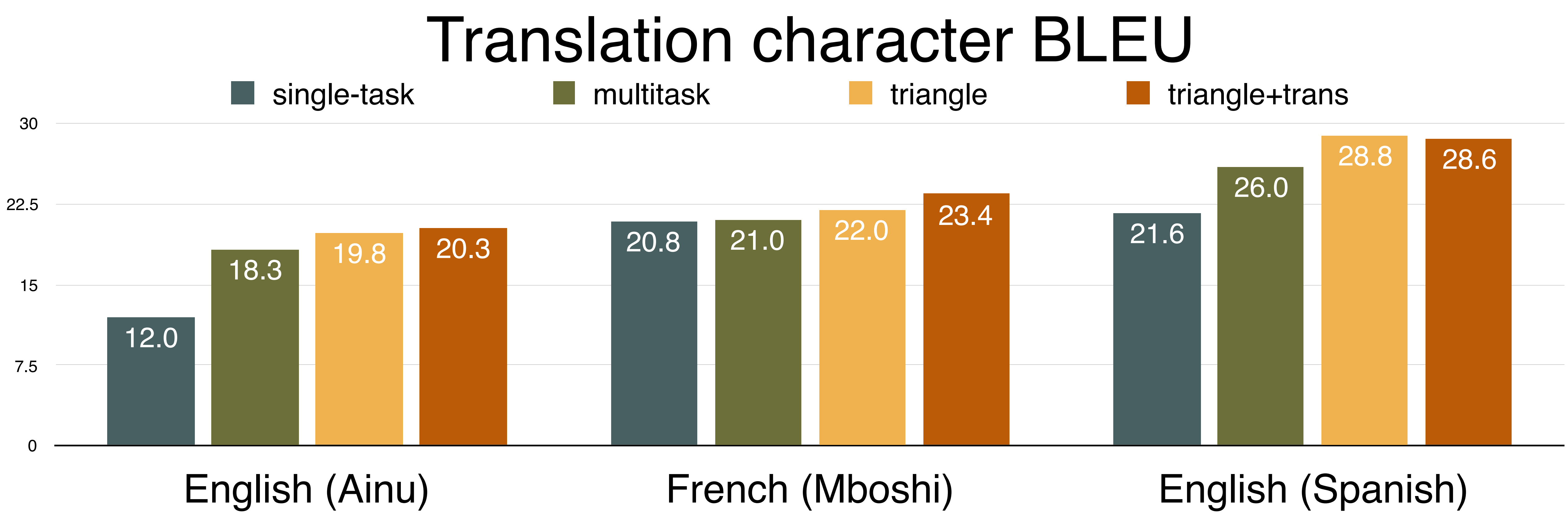
and B attends over C…

this should be similar to A attending directly over C.

$$\mathcal{R}_{\text{trans}} = -\lambda_{\text{trans}} \left\| \mathbf{A}^{12}\mathbf{A}^{1} - \mathbf{A}^{2} \right\|_2^2.$$







multi-task models: results

We can improve translation and transcription accuracy by jointly performing the two tasks.

Translation can be further improved by using intermediate representations and transitivity.

Other (relevant and ongoing) work

Build a tool for linguists that uses ML in its backend to aid annotation:

work with Graham Neubig

Data Augmentation, Cross-Lingual Transfer, and other nice things

Using related languages for MT

"Generalized Data Augmentation for Low-Resource Translation"

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig
ACL 2019

Transfer for MT

Typical scenario: continued training

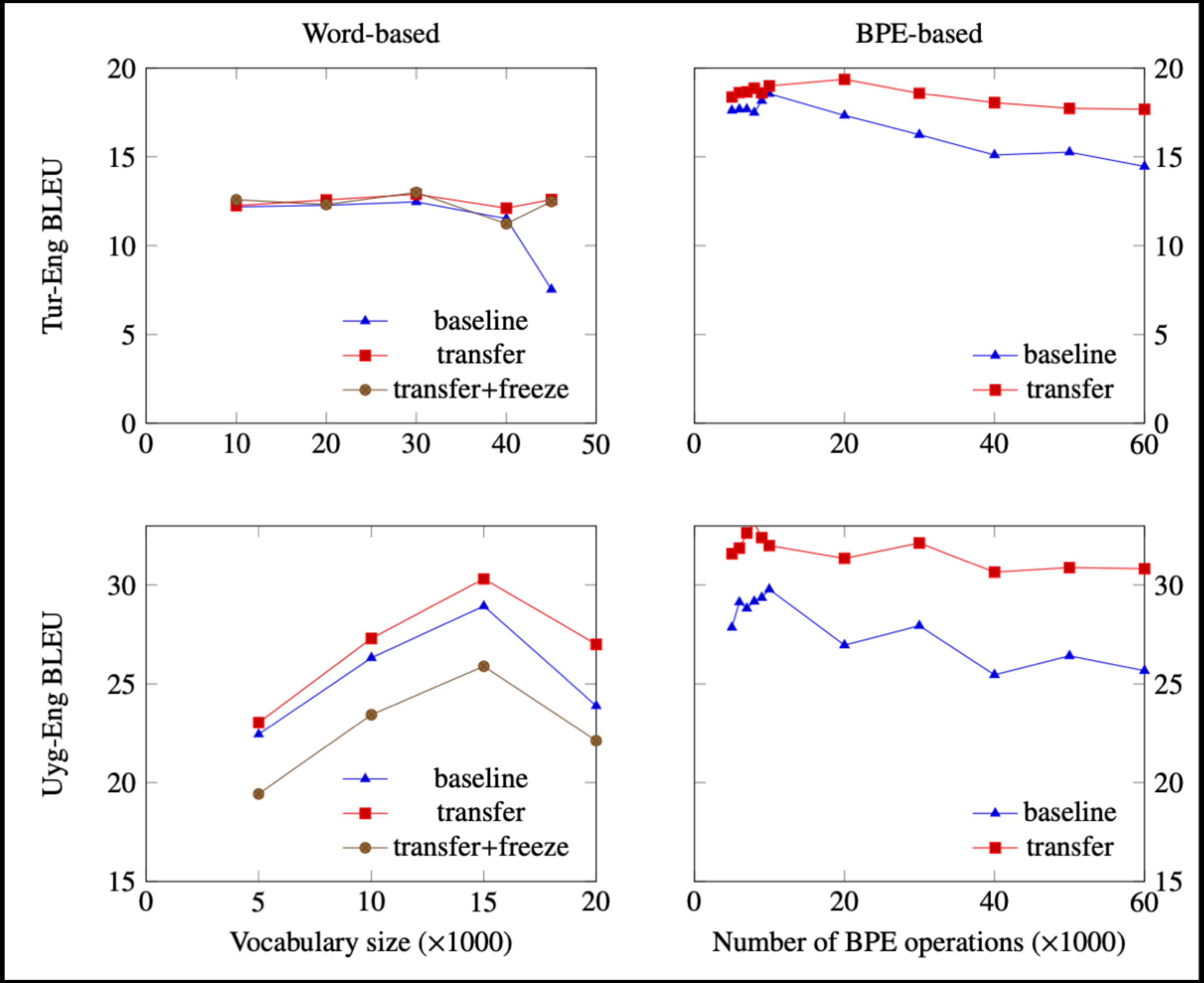


figure from "Transfer Learning across Low-Resource, Related Languages for NMT". Nguyen and Chiang, 2018.

Machine Translation

The current best approach is a semi-supervised one:

- Back-translation of target-side monolingual data

What if we don't have tons of monolingual data for a language?

Does the quality of the back-translated data matter?

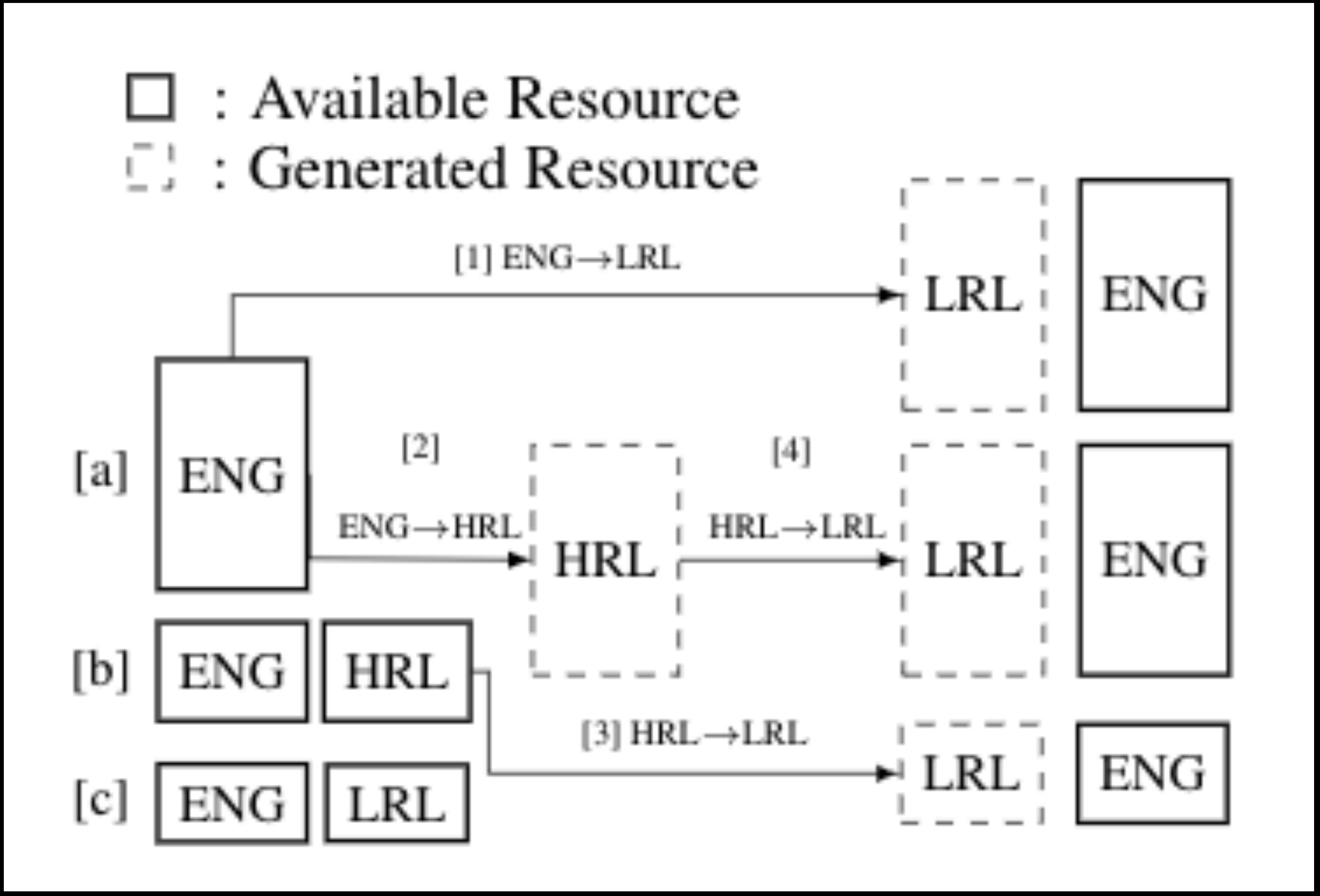
Generalized Back-Translation

For low-resource languages, there maybe exist a related high-resource one e.g.

1. Azerbaijani (Turkish)
2. Belarusian (Russian)
3. Galician (Portuguese)
4. Slovak (Czech)

We should use them!

Generalized Back-Translation



Generalized Back-Translation

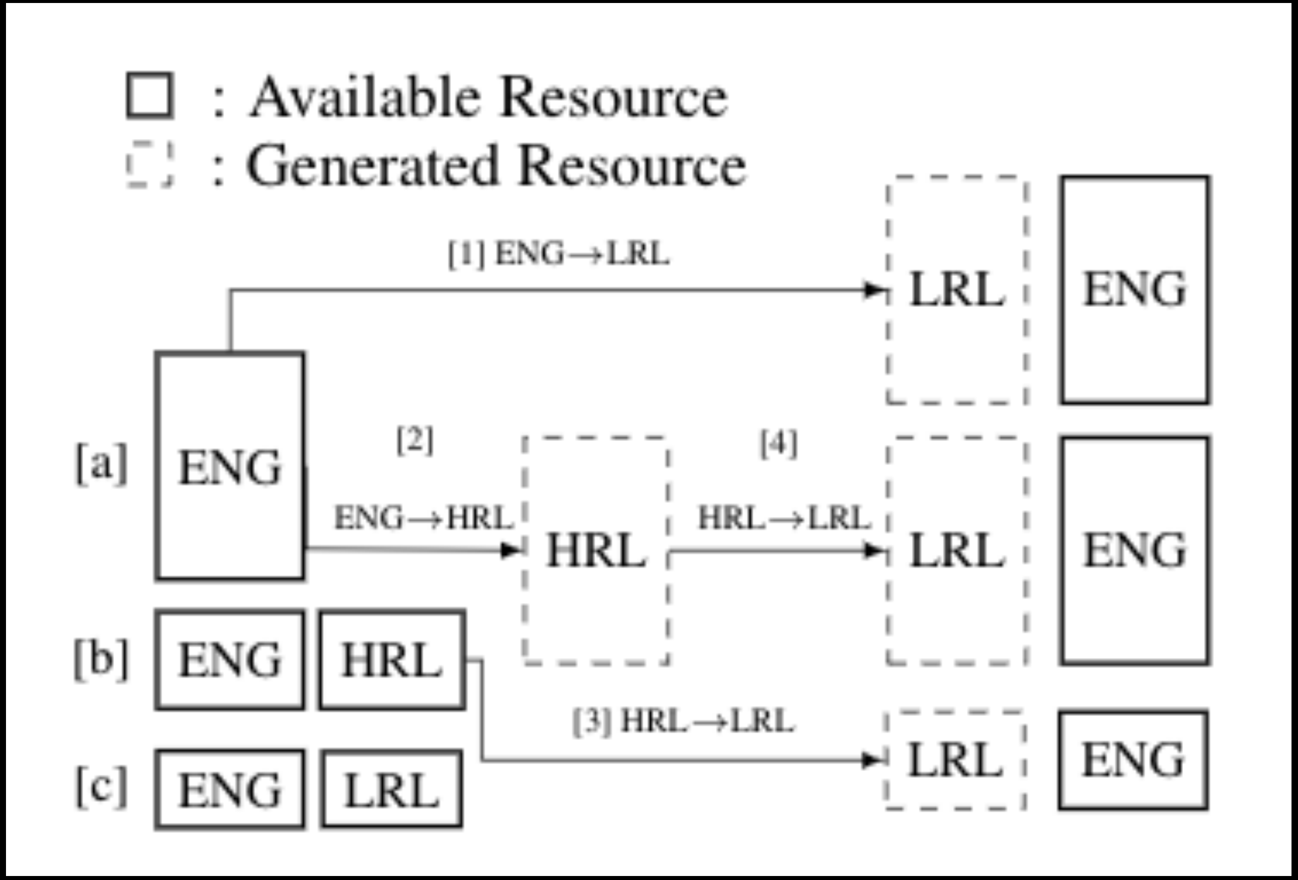
Typical:

only use [1] for data augmentation

OR

add [b] to [c] and train.

But HRL to LRL might be easier!



From HRL to LRL

Assuming a parallel dataset is probably too much.

If the languages are related enough:

1. Get monolingual embeddings
2. Align the embedding space [Lample et al, 2018]
3. Learn a dictionary
4. Word substitution in HRL to create *pseudo*-LRL

From ENG to LRL through HRL

ENG to LRL system would be bad (duh!)

ENG to HRL system would be better…

… and HRL to LRL might be easy-ish (cause related)

Results

	Training Data		BLEU for X→ENG AZE (TUR)	BEL (RUS)	GLG (POR)	SLK (CES)
1	Base Supervised NMT		11.83	16.34	29.51	28.12
2	Base Unsupervised NMT		0.47	0.18	1.15	0.75
Standard Supervised Back-translation						
3	$+\{\hat{\mathcal{S}}^{s}_{\mathrm{E\to L}}\ ,\ \mathcal{M}_{\mathrm{E}}\}$		11.84	15.72	29.19	29.79
4	$+\{\hat{\mathcal{S}}^{s}_{\mathrm{E\to H}}\ ,\ \mathcal{M}_{\mathrm{E}}\}$		12.46	16.40	30.07	30.60
Augmentation from HRL-ENG						
5	$+\{\hat{\mathcal{S}}^{s}_{\mathrm{H\to L}}\ ,\ \mathcal{T}_{\mathrm{HE}}\}$	(supervised MT)	11.92	15.79	29.91	28.52
6	$+\{\hat{\mathcal{S}}^{u}_{\mathrm{H\to L}}\ ,\ \mathcal{T}_{\mathrm{HE}}\}$	(unsupervised MT)	11.86	13.83	29.80	28.69
7	$+\{\hat{\mathcal{S}}^{w}_{\mathrm{H\to L}}\ ,\ \mathcal{T}_{\mathrm{HE}}\}$	(word subst.)	14.87	23.56	32.02	29.60
8	$+\{\hat{\mathcal{S}}^{m}_{\mathrm{H\to L}}\ ,\ \mathcal{T}_{\mathrm{HE}}\}$	(modified UMT)	14.72	23.31	32.27	29.55
9	$+\{\hat{\mathcal{S}}^{w}_{\mathrm{H\to L}}\hat{\mathcal{S}}^{m}_{\mathrm{H\to L}}\ ,\ \mathcal{T}_{\mathrm{HE}}\mathcal{T}_{\mathrm{HE}}\}$		15.24	**24.25**	32.30	30.00
Augmention from ENG by pivoting						
10	$+\{\hat{\mathcal{S}}^{w}_{\mathrm{E\to H\to L}}\ ,\ \mathcal{M}_{\mathrm{E}}\}$	(word subst.)	14.18	21.74	31.72	30.90
11	$+\{\hat{\mathcal{S}}^{m}_{\mathrm{E\to H\to L}}\ ,\ \mathcal{M}_{\mathrm{E}}\}$	(modified UMT)	13.71	19.94	31.39	30.22
Combinations						
12	$+\{\hat{\mathcal{S}}^{w}_{\mathrm{H\to L}}\hat{\mathcal{S}}^{w}_{\mathrm{E\to H\to L}}\ ,\ \mathcal{T}_{\mathrm{HE}}\mathcal{M}_{\mathrm{E}}\}$	(word subst.)	**15.74**	**24.51**	**33.16**	**32.07**
13	$+\{\hat{\mathcal{S}}^{w}_{\mathrm{H\to L}}\hat{\mathcal{S}}^{m}_{\mathrm{H\to L}}\ ,\ \mathcal{T}_{\mathrm{HE}}\mathcal{T}_{\mathrm{HE}}\}$ $+\{\hat{\mathcal{S}}^{w}_{\mathrm{E\to H\to L}}\hat{\mathcal{S}}^{m}_{\mathrm{E\to H\to L}}\ ,\ \mathcal{M}_{\mathrm{E}}\mathcal{M}_{\mathrm{E}}\}$		**15.91**	23.69	32.55	31.58

Takeaways

Translating from HRL to LRL:

- it is better to use word substitution than simple NMT or standard UMT
  (cf lines 5,6 to 7,8,9)

Pivoting from ENG though HRL, improvements but not as much.

Best of both worlds works best (line 12)

- More ENG data, as good as possible LRL data

Using Related Languages for Morphological Inflection

Inflection task and SIGMORPHON

Low-resource target training data (Asturian)

facer	fechu	V;V.PTCP;PST
aguar	aguà	V;PRS;2;PL;IND
...		

High-resource source language training data (Spanish)

tocar	tocando	V;V.PTCP;PRS
bailar	bailaba	V;PST;IPFV;3;SG;IND
mentir	mintió	V;PST;PFV;3;SG;IND

SIGMORPHON challenge:
100 language pairs (43 test languages)

Previous Work

Concatenate tags and lemma, single encoder-decoder

- Issue: inherently different (order, function)

Half task is identifying stem/root and copy characters, so other works focus on copying

- explicit copy mechanism, or
- hard monotonic attention, or
- learn to output the string transduction steps

Augmentation approach: hallucinating data

Most low-resource languages have just 50 or 100 examples.

You can hallucinate more data:

b a i l a r
| | | / /
b a i l a b a

replacing the red parts with random characters

Results on transfer from single language

L1	L2	L1+L2	$+\mathcal{H}$	$\mathcal{H}$
latin	czech	15	71.4	**77.4**
bengali	greek	12.4	70.5	**71.6**
sorani	irish	10.3	**66.3**	65.6
italian	ladin	48	**74**	**74**
latvian	lithuanian	7.1	48.4	**50.5**
english	murrinhpatha	**36**	6	20
italian	neapolitan	70	83	**84**
urdu	old english	13.8	43.4	**44.3**
slovene	old saxon	10.7	**52.3**	50.5
russian	portuguese	34.5	**88.8**	87.7
swahili	quechua	4.2	**92.1**	91.6
portuguese	russian	25.6	**76.3**	74.3
kurmanji	sorani	6.2	**69**	66.7
zulu	swahili	46	**81**	76
kannada	telugu	76	**94**	**94**
Average		27.72	67.77	68.55

If languages are genetically distant, transfer does NOT help.

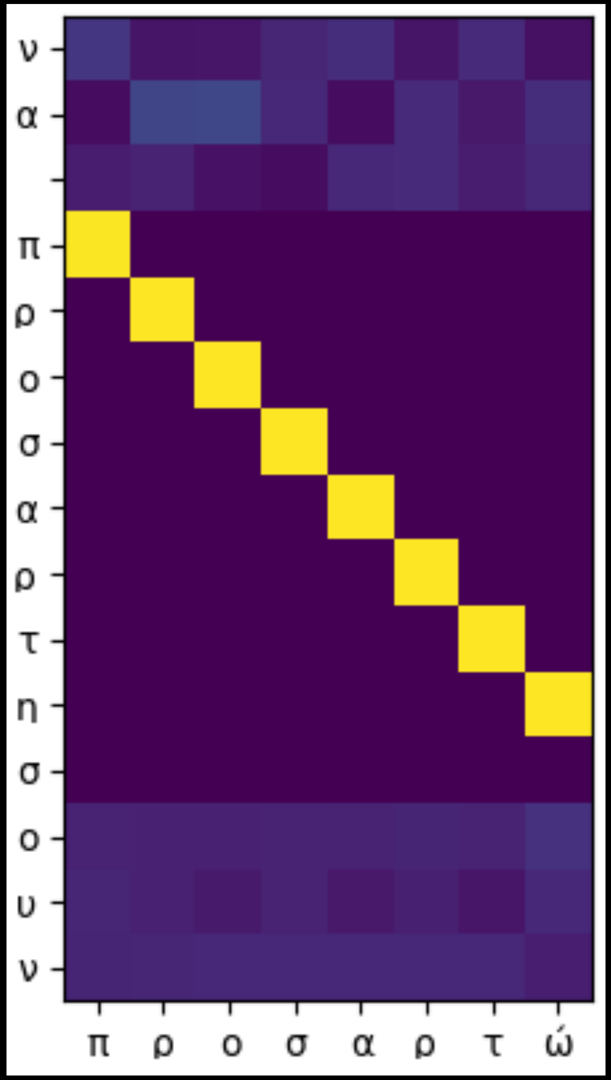
Same alphabet crucial: see Kurmanji-Sorani

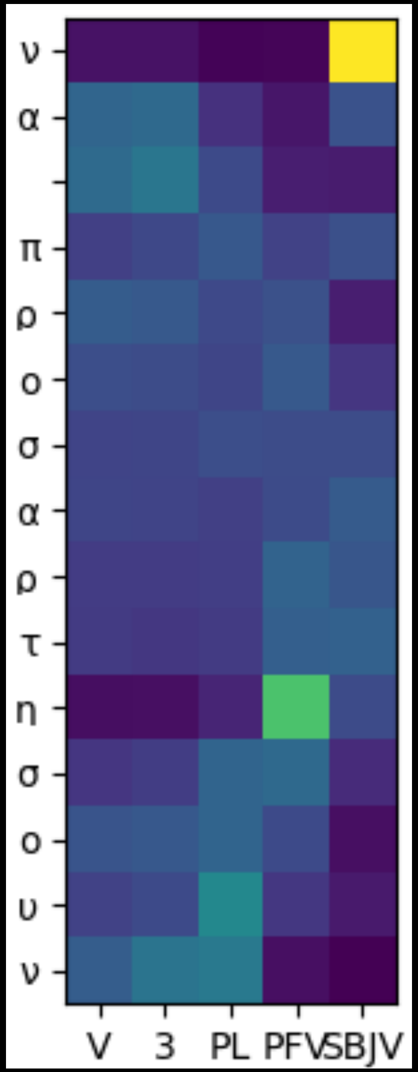
But we can do better if transferring from multiple (related) languages

e.g.

L1	L2	L1+L2	$+\mathcal{H}$	$+\mathcal{L}_l + \mathcal{H}$	$\mathcal{H}$
turkish		81	80	81	
persian		35	74	69	
bashkir	azeri	37	66	67	66.7±0.9
uzbek		27	74	70	
all		84	83	**87**	

Interpreting the model





Takeaways

1. Monolingual data hallucination can take you a long way…

2. … and it's preferable to cross-lingual transfer from distant languages

3. If close enough languages, both data hallucination and cross-lingual transfer should help

4. The closer the languages, the larger the improvements

Main Issues:

- Data Hallucination is language-agnostic. A more informed sampling could probably do better

- Different alphabets really hurt performance (Dutch-Yiddish, Kurmanji-Sorani). Need to find either an a priori mapping between the two, or map the to a common space (IPA?)

What language should you use for cross-lingual transfer?

"*Choosing Transfer Languages for Cross-Lingual Learning*"

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell and Graham Neubig
ACL 2019

Setting

Cross-lingual transfer on 4 tasks:

1. MT: 54x54 pairs (X-Eng TED)
2. POS-tagging: 60x26
3. Entity Linking: 53x9
4. DEP parsing: 30x30

Learning to Rank

For each language pair, extract features:

1. dataset dependent:
   - dataset size
   - type-token ratio
   - word/subword overlap
2. dataset-independent:
   1. typological features (from URIEL)

[plug: check out the lang2vec python library, now with pre-computed distances!]

Learning to Rank

For each test language and a list of potential transfer languages (each pair represented by the features), train a model to rank the candidate languages

Model: tree-based LambdaRank (good in limited feature/data settings)

	Method	MT	EL	POS	DEP
dataset	word overlap o_w	28.6	30.7	13.4	52.3
	subword overlap o_{sw}	29.2	–	–	–
	size ratio s_{tf}/s_{tk}	3.7	0.3	9.5	24.8
	type-token ratio d_{ttr}	2.5	–	7.4	6.4
ling. distance	genetic d_{gen}	24.2	50.9	14.8	32.0
	syntactic d_{syn}	14.8	46.4	4.1	22.9
	featural d_{fea}	10.1	47.5	5.7	13.9
	phonological d_{pho}	3.0	4.0	9.8	43.4
	inventory d_{inv}	8.5	41.3	2.4	23.5
	geographic d_{geo}	15.1	49.5	15.7	46.4
	LANGRANK (all)	51.1	**63.0**	**28.9**	**65.0**
	LANGRANK (dataset)	**53.7**	17.0	26.5	**65.0**
	LANGRANK (URIEL)	32.6	58.1	16.6	59.6

[Available as a python package too: https://github.com/neulab/langrank]

Other Cool Things

Language Technology for Language Documentation and Revitalization

Hackathon-type workshop at CMU, Aug 12-16, 2019

- Language community members
- Documentary linguists
- Computational linguists
- Computer scientists and developers

Example projects:

- Building and using tools for rapid dictionary creation
- Building and using tools for development of speech recognition systems
- Building and using tools to analyze the syntax of language, and extract example sentences for educational materials
- Creating a plugin that incorporates language techology into language documentation software such as ELAN/Praat