CS11-731
Machine Translation and
Sequence-to-Sequence Models

Languages of the World

Antonis Anastasopoulos

Carnegie Mellon University
Language Technologies Institute

Site
https://phontron.com/class/mtandseq2seq2019/

The state-of-the-art in German-English MT on News translation is around 42 BLEU.

	→	←
What is it for English-German?	~45	
What is it for Chinese-English?	~39	~45
What is it for French-German?	~35	~37
What is it for Gujarati-English?	~25	~28
What is it for Greek-Swahili?	???	???

What do the different languages of the world look like?

Mitä tämä lause sanoo?

ماذا تقول هذه الجملة؟

Энэ өгүүлбэрт юу гэж хэлдэг вэ?

О чем говорит это предложение?

이 문장은 무엇을 말합니까?

Ի՞նչ է ասում այս նախադասությունը:

Case Study: Kazakh-English

бұл сөйлем нені білдіреді? what does this sentence mean?

Only 97k parallel sentences

+3.7M more by pivoting through Russian

+back-translation

+distillation, ensembling

System	EN-KK 19dev	EN-KK 19test	KK-EN 19dev	KK-EN 19test
Big	2.6	1.9	10.1	11.5
+Pivot	14.9	7.8	23.4	19.8
+Sampling	19.7	10.3	26.2	28.8
DLCL25	20.5	10.7	26.3	29.0
+RPR	-	-	26.6	30.1
+Ensemble	21.3	11.1	26.8	30.5

Results from: *The NiuTrans Machine Translation Systems for WMT19, Li et al. 2019*

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

Let’s look at the "similar languages" shared task results

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, tokenisation

http://anoopkunchukuttan.github.io/indic_nlp_library/

Case Study: English-Chinese

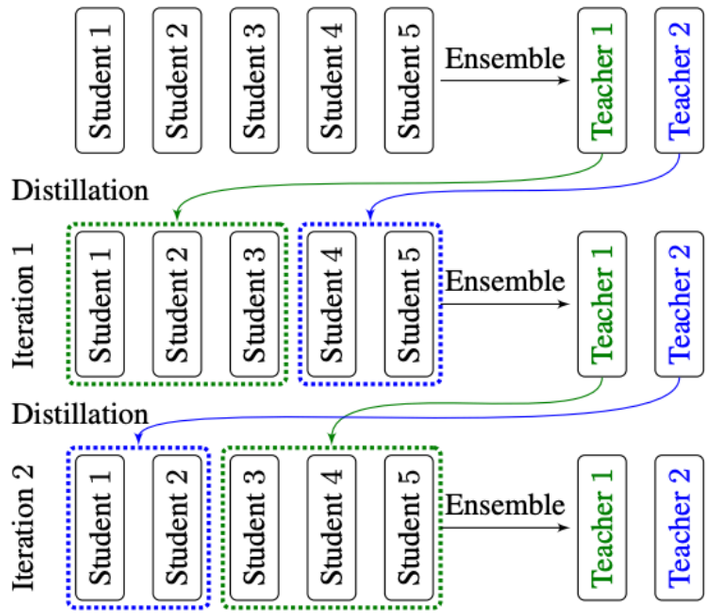
what does this sentence mean?

這句話是什麼意思?
这句话是什么意思?

Very high resource, but:

logographic writing system —> hug
tokenization?

Filtering, ensembling, distillation



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?

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	chʜ
池 pool	水(氵) water	也	chʜ
施 impose	方 direction	也	sh
弛 loosen	弓 bow	也	chʜ
地 land	土 soil	也	dᴍ
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

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

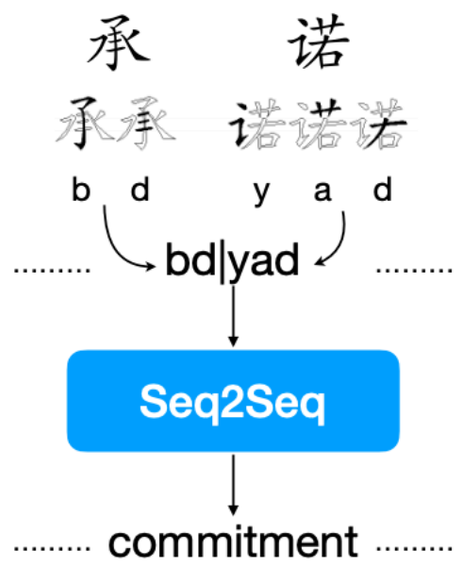


Figure 1: Overview of the **wubi2en** approach to Chinese-to-English translation. A raw Chinese word (‘承诺’) is encoded into ASCII characters (‘bd|yad’), using the Wubi encoding method, before passing it to a Seq2Seq network. The network generates the English translation ‘commitment’, processing one ASCII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

Case Study: 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

or even strokes:

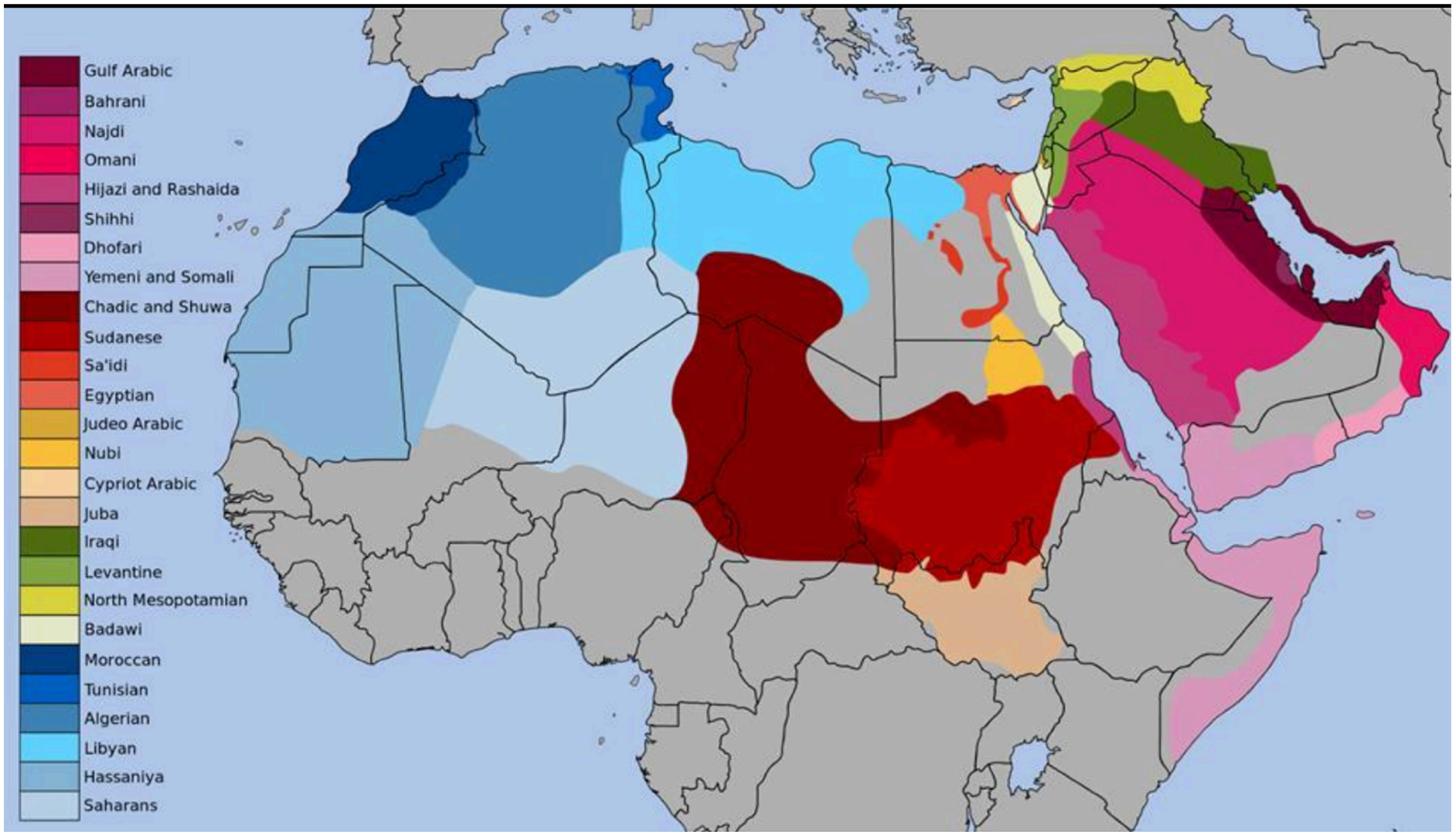
Language	Word
JP-character	風 景
JP-ideograph	几一虫 日亠口小_1
JP-stroke	丿乁一丨フ一丨ノ丶 丨フ一一、一丨フ一丨ノ丶_1
CN-character	风 景
CN-ideograph	几乂 日亠口小_1
CN-stroke	丿乁ノ丶 丨フ一一、一丨フ一丨ノ丶_1
EN	landscape

Table 3: The Japanese word 風景 and Chinese word 风景 both mean "landscape" in English, and they only differ in the middle part of the first character. Note that there are "_1" tags at the ends of some decomposed sequences to distinguish between possible duplications.

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$_{NNP}$.

Arabic Preprocessing Schemes for Statistical Machine Translation, Habash and Sadat (2006)

Case Study: Arabic

what does this sentence mean? ماذا تعني هذه الجمله؟

Preprocessing (tokenization+segmentation):

Setting	Sentence

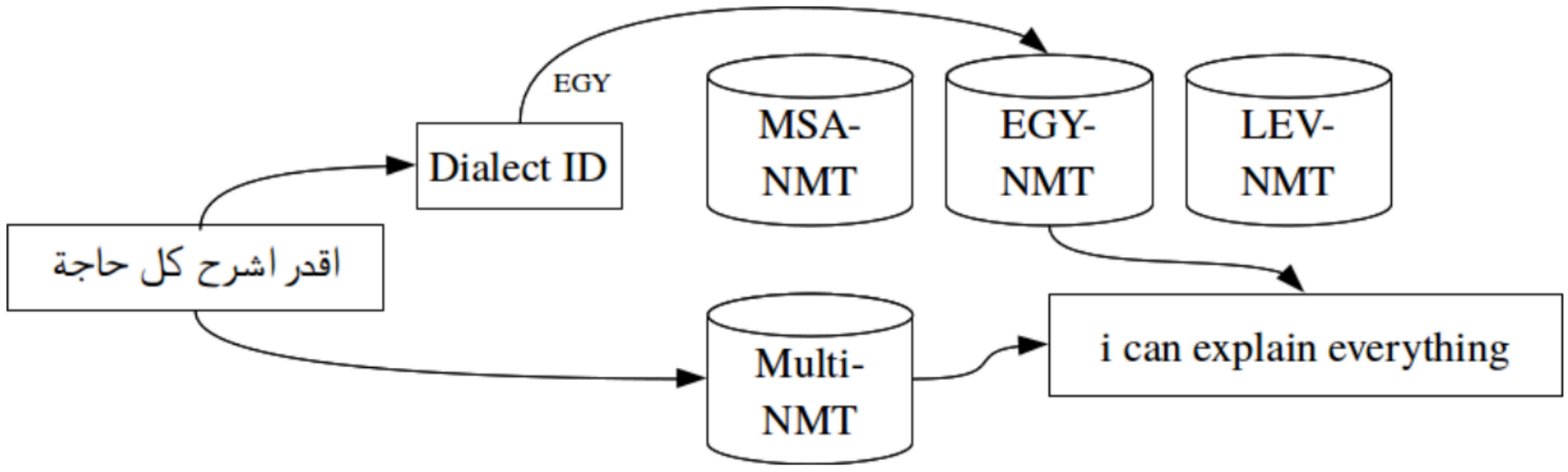
	#Vocab	SMT_{tgt++}	CI	$NMT_{scr/tgt++}$	CI	P-value
Raw	331K	52.78	± 0.98	52.76	± 1.24	0.412
ATB	208K	55.42	± 1.07	**53.54**	± 1.20	0.002
D3	190K	54.66	± 1.02	53.51	± 1.20	0.027
Raw+BPE	20K	53.78	± 1.10	52.41	± 1.17	0.003
ATB+BPE	20K	**55.64**	± 1.11	53.18	± 1.15	0.001
D3+BPE	20K	54.59	± 1.07	53.38	± 1.16	0.018

from The Impact of Preprocessing on Arabic-English Statistical and Neural Machine Translation, Oudah et al. 2019

Case Study: Arabic

what does this sentence mean? ماذا تعني هذه الجمله؟

Handling dialectal data:



Comparing Pipelined and Integrated Approaches to Dialectal Arabic NMT, Shapiro and Duh, 2019.

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.

One Size Does Not Fit All: Comparing NMT Representations of Different Granularities, Durrani et al., 2019

Case Study: African languages

The most important issue is the
lack of data and standardized evaluation sets.

This is starting to change, but data can be very noisy

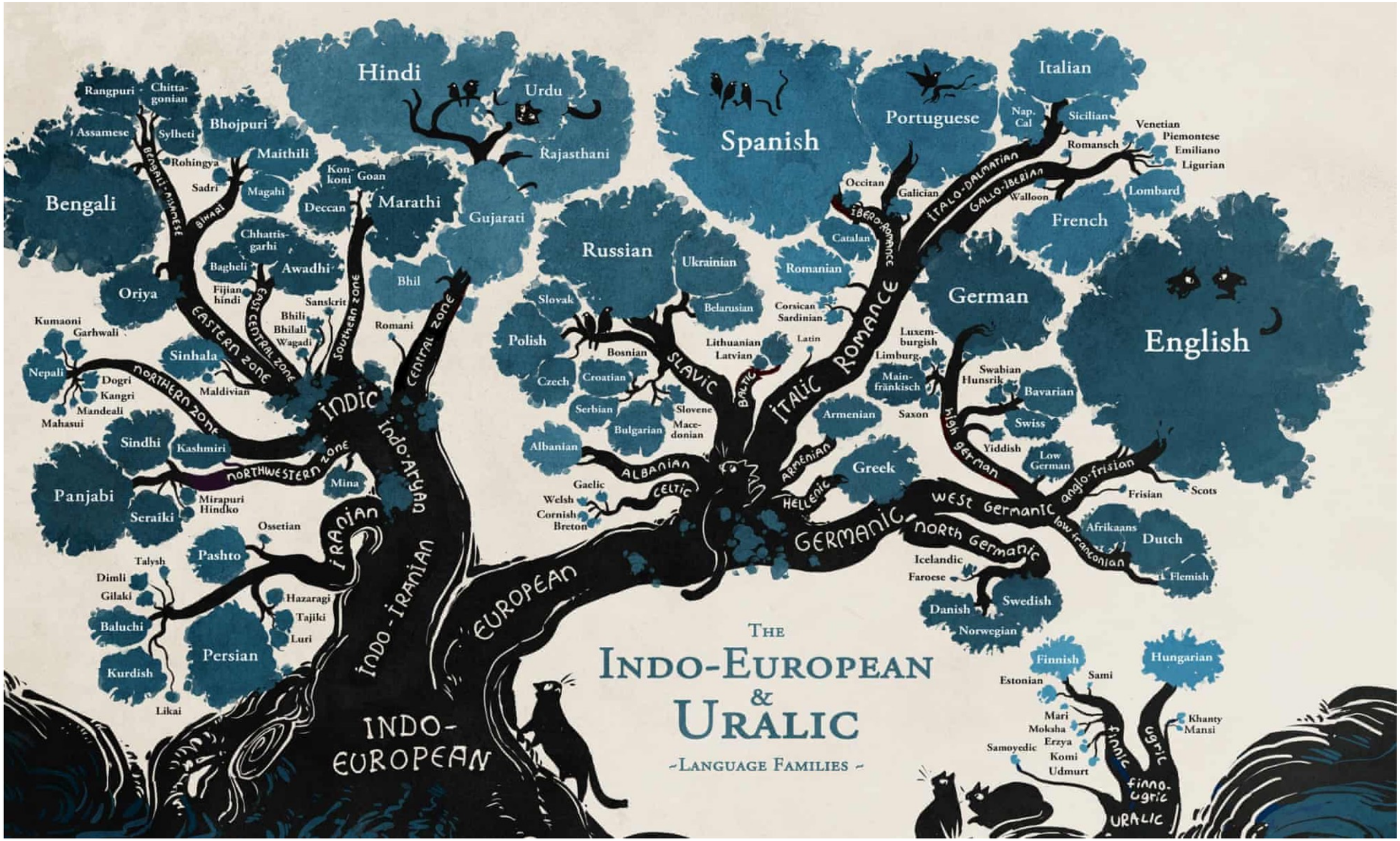
https://github.com/LauraMartinus/ukuxhumana

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

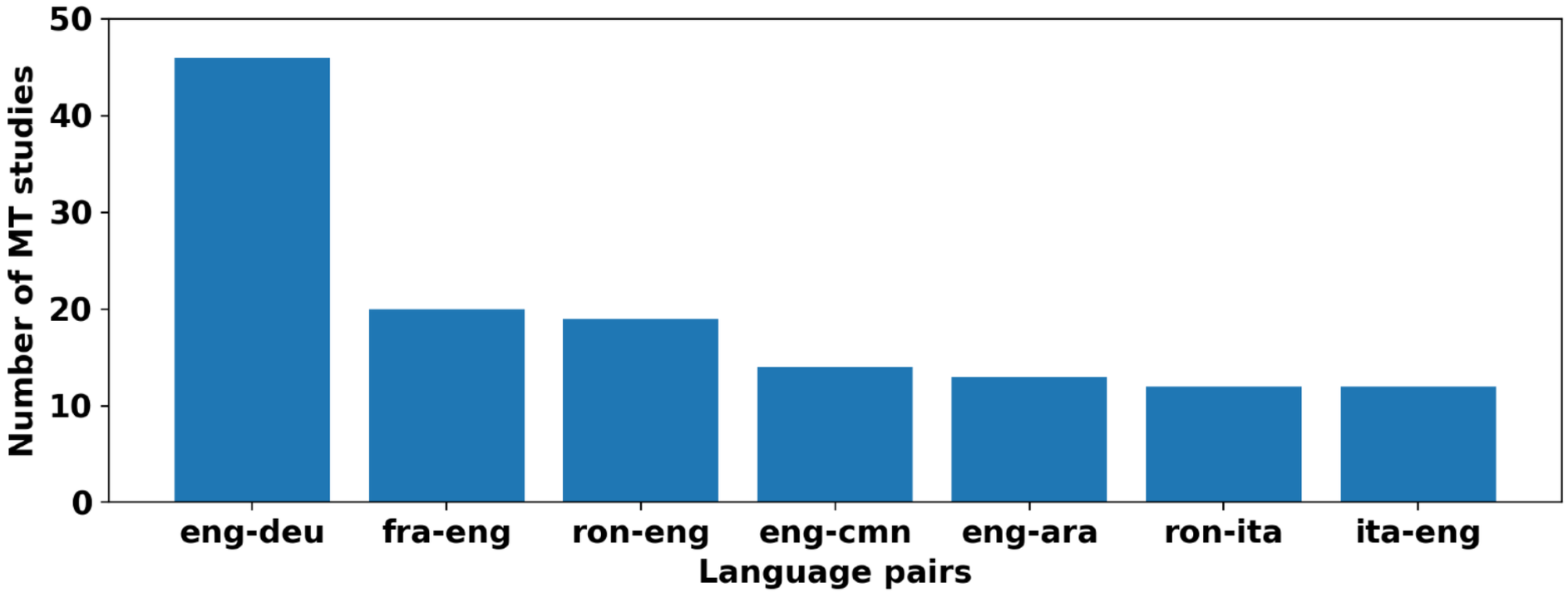
Typological Features



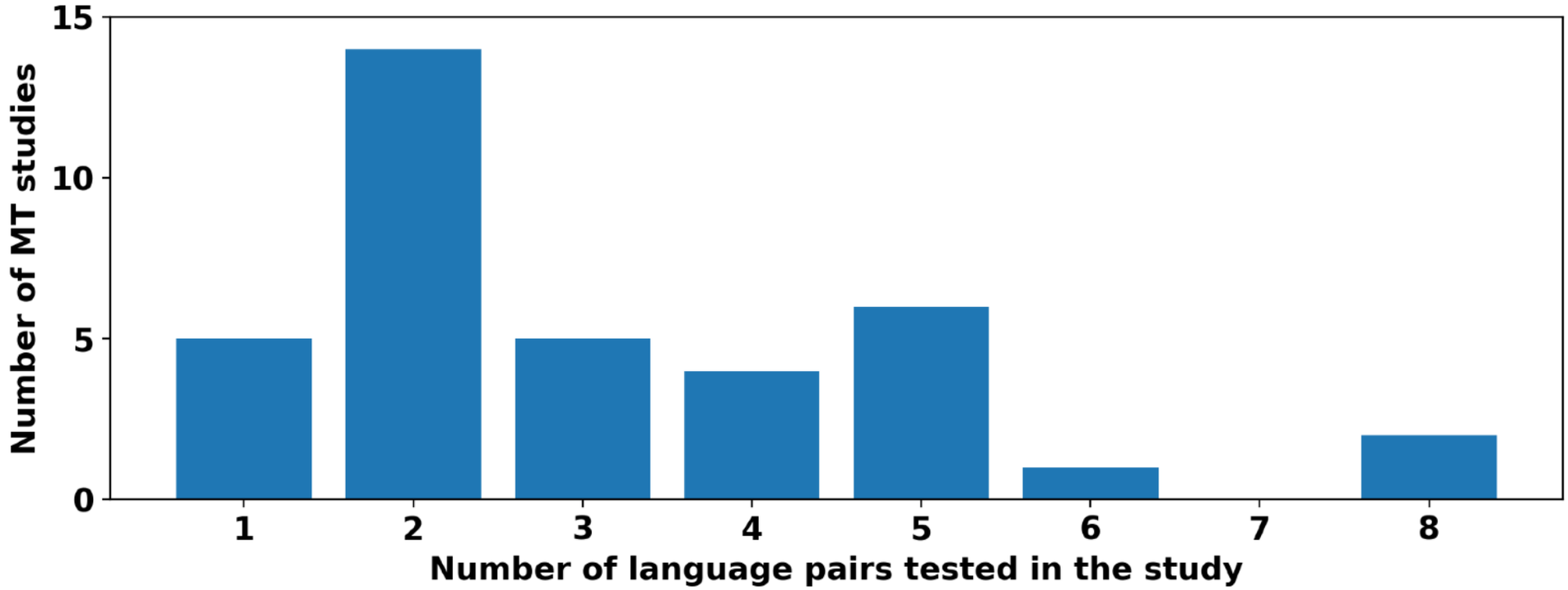
Let’s Try it Out!

lang2vec

How "fairly" is MT technology distributed?



How "fairly" is MT technology distributed?



Latin: We don't need a definite article!

English: We'll just use the. Simple.

French: Le, la, les. For a bit of variety.

Ancient Greek: :)

Latin: Oh no.

Ancient Greek: :) :)

English: Don't do it.

Ancient Greek: :) :) :)

French: Why are you like this?

Ancient Greek: ο, του, τω, τον, οι, των, τοις, τους, η, της, τη, την, αι, ταις, τας, το, τα, τω, τοιν

Ancient Greek: :)