Multilingual Models

Dan Klein, John DeNero
UC Berkeley

Linguistic Typology

Constituent Order

Quoting Wikipedia...

SOV is the order used by the largest number of distinct languages... [including] Japanese, Korean, Mongolian, Turkish...
"She him loves."

SVO languages include English, Bulgarian, Macedonian, Serbo-Croatian, the Chinese languages and Swahili, among others.
"She loves him."

German word order example:

Clause 1: Ich/I werde/will Ihnen/to you die/the entsprechenden/corresponding Anmerkungen/comments aushaendigen/pass on

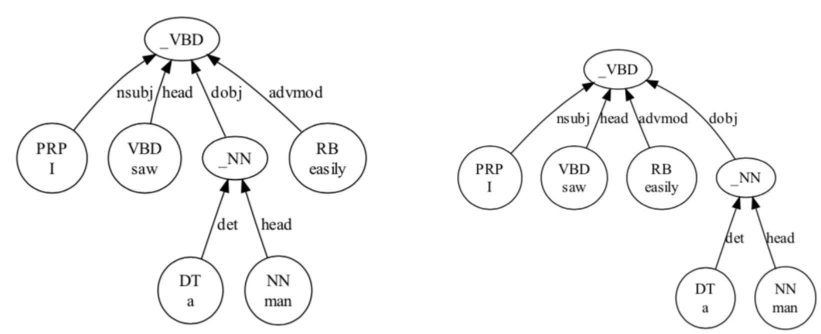
Clause 2: damit/so that Sie/you das/them eventuell/perhaps bei/in der/the Abstimmung/vote uebernehmen/adopt koennen/can

German example from Collins et al., 2005, "Clause Restructuring for Statistical Machine Translation"

Aside: Pre-Ordering for Statistical Machine Translation

2010-2016 Google Translate used a pipeline involving syntactic parser for many language pairs (starting with en-ja):

source ▷ parsed source ▷ reordered source ▷ target



(a) A sample parse tree

(b) After reordering (moving RB over _NN)

Genzel, 2010, "Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation"

Aside: Pre-Ordering for Statistical Machine Translation

2010–2016 Google Translate used a pipeline involving syntactic parser for many language pairs (starting with en-ja):

source ▷ parsed source ▷ reordered source ▷ target

Table 6: Examples of top rules and their application

Languages	Context	Order	Example
Hindi	1L:head 3L:none	2,1,3	*I see him* → *I him see*
Japanese, Korean	2L:prep	2,1	*eat with a spoon* → *eat a spoon with*
German	1T:VBN 2L:prep	2,1	*struck with a ball* → *with a ball struck*
Russian, Czech	1L:nn 2L:head	2,1	*a building entrance* → *a entrance building*
Welsh	1L:amod 2L:head	2,1	*blue ball* → *ball blue*

Label of the first child

Genzel, 2010, "Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation"

Aside: Pre-Ordering for Statistical Machine Translation

2010–2016 Google Translate used a pipeline involving syntactic parser for many language pairs (starting with en-ja):

source ▷ parsed source ▷ reordered source ▷ target

(Genzel, 2010): hand-crafted rules transform a dependency parse
(Lerner & Petrov, 2013): classifier permutes a phrase structure parse
- 1-step: predict a permutation for the children of each node
- 2-step: first predict whether each child should be placed before or after the head constituent, then permute each side.

	base	rule	1-step	2-step
en-ar	11.4	12.3	**12.5**	**12.6**
en-cy	29.3	31.1	31.9♧	**32.4**♣
en-ga	17.0	18.5	18.8♧	**19.1**♣
en-iw	18.8	19.7	**20.2**	**20.2**
en-id	31.0	33.4	**34.0**♧	**34.3**♧
en-ja	10.4	16.4	**17.5**♧	**18.0**♣
en-ja*	14.9	18.0	**18.2**♧	**18.6**♣
en-ko	24.1	31.8	31.8♧	**32.7**♣
en-ms	20.4	22.5	**22.9**	**22.9**

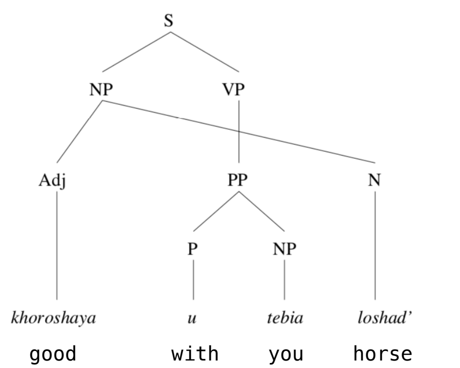
Table 3: BLEU scores for language from various language families: Arabic (ar), Welsh (cy), Irish (ga), Indonesian (id), Hebrew (iw), Japanese (ja), Korean (ko), and Malay (ms). Lexical reordering is not included in any of the systems. Bolded results are significant at 99%. ♣ is significantly better than ♧ in a human eval at 95%.

Lerner & Petrov, 2013, "Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation"

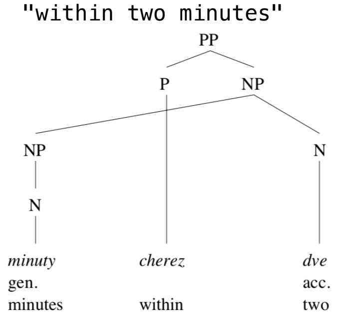
Free Word Order and Syntactic Structure

In Russian, "The dog sees the cat" can be translated as:
Sobaka vidit koshku
Sobaka koshku vidit
Vidit sobaka koshku
Vidit koshku sobaka
Koshku vidit sobaka
Koshku sobaka vidit

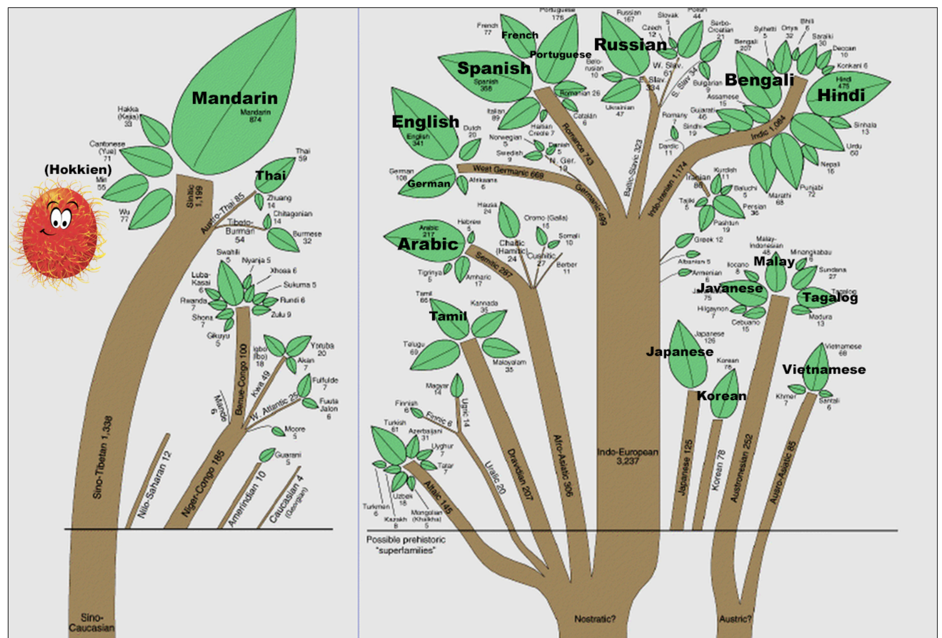
"You have a good horse"
(literally, "A good horse is with you")



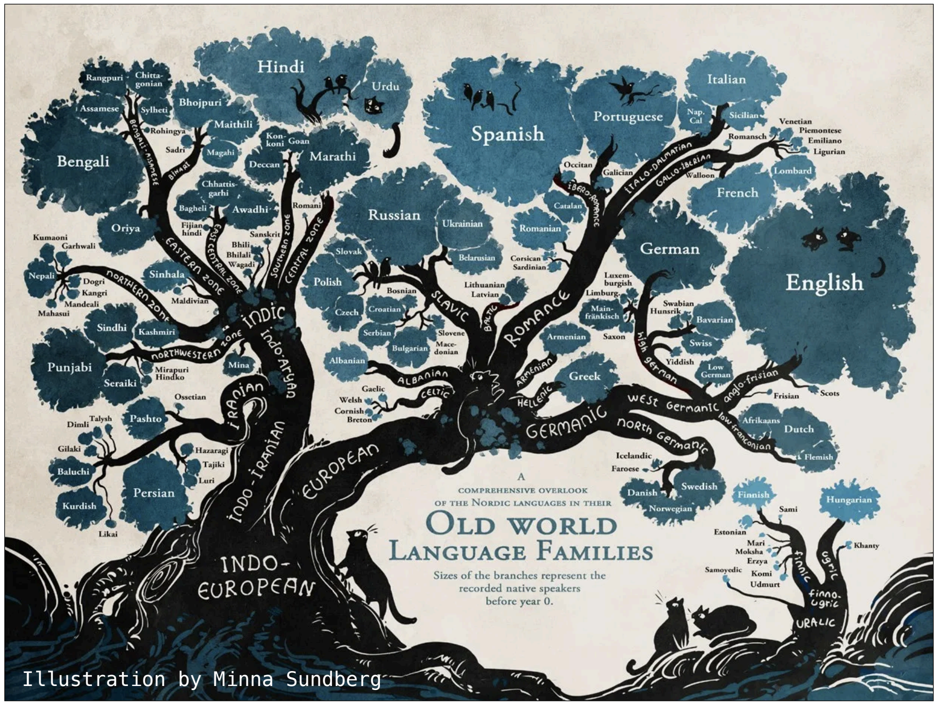
"within two minutes"



Examples from Covington, 1990, "A Dependency Parser for Variable Word Order Languages"



https://www.angmohdan.com/wp-content/uploads/2014/10/FullTree.jpg



Morphology

Morphological Variation

Morphology: how words are formed

Derivational morphology: constructing new lexemes
- estrange (v) => estrangement (n)
- become (v) => unbecoming (adj)

Inflectional morphology: build surface forms of a lexeme

		singular			plural		
		first	second	third	first	second	third
indicative		je (j')	tu	il, elle	nous	vous	ils, elles
(simple tenses)	present	arrive /a.ʁiv/	arrives /a.ʁiv/	arrive /a.ʁiv/	arrivons /a.ʁi.vɔ̃/	arrivez /a.ʁi.ve/	arrivent /a.ʁiv/
	imperfect	arrivais /a.ʁi.vɛ/	arrivais /a.ʁi.vɛ/	arrivait /a.ʁi.vɛ/	arrivions /a.ʁi.vjɔ̃/	arriviez /a.ʁi.vje/	arrivaient /a.ʁi.vɛ/
	past historic[2]	arrivai /a.ʁi.vɛ/	arrivas /a.ʁi.va/	arriva /a.ʁi.va/	arrivâmes /a.ʁi.vam/	arrivâtes /a.ʁi.vat/	arrivèrent /a.ʁi.vɛʁ/
	future	arriverai /a.ʁi.vʁɛ/	arriveras /a.ʁi.vʁa/	arrivera /a.ʁi.vʁa/	arriverons /a.ʁi.vʁɔ̃/	arriverez /a.ʁi.vʁe/	arriveront /a.ʁi.vʁɔ̃/
	conditional	arriverais /a.ʁi.vʁɛ/	arriverais /a.ʁi.vʁɛ/	arriverait /a.ʁi.vʁɛ/	arriverions /a.ʁi.və.ʁjɔ̃/	arriveriez /a.ʁi.və.ʁje/	arriveraient /a.ʁi.vʁɛ/

Examples from Greg Durrett

Noun Declension

Declension of Kind [hide ▲]

	singular			plural	
	indef.	def.	noun	def.	noun
nominative	ein	das	Kind	die	Kinder
genitive	eines	des	Kindes, Kinds	der	Kinder
dative	einem	dem	Kind, Kinde[1]	den	Kindern
accusative	ein	das	Kind	die	Kinder

- Nominative: I/he/she, accusative: me/him/her, genitive: mine/his/hers
- Dative: merged with accusative in English, shows recipient of something

I taught the children <=> Ich unterrichte die Kinder

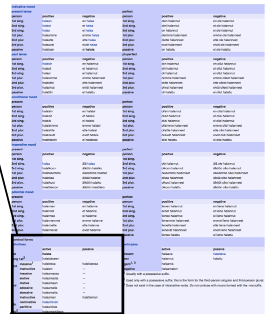
I give the children a book <=> Ich gebe den Kindern ein Buch

Examples from Greg Durrett

Agglutinative Languages

Finnish/Hungarian (Finno-Ugric), and Turkish: what a preposition would do in English is instead part of the verb

		active	passive
1st		halata	
long 1st[2]		halatakseen	
2nd	inessive[1]	halatessa	halattaessa
	instructive	halaten	—
3rd	inessive	halaamassa	—
	elative	halaamasta	—
	illative	halaamaan	—
	adessive	halaamalla	—
	abessive	halaamatta	—
	instructive	halaaman	halattaman
4th	nominative	halaaminen	
	partitive	halaamista	
5th[2]		halaamaisillaan	



halata: "hug"

illative: "into" adessive: "on"

Examples from Greg Durrett

Writing Systems

Characteristics of Scripts

Cyrillic, Arabic, and Roman alphabets are (mostly) phonetic.

- The Serbian language is commonly written in both Gaj's Latin and Serbian Cyrillic scripts.
- Urdu and Hindi are (mostly) mutually intelligible, but Urdu is written in Arabic script, while Hindi is written in Devanagari.
- Arabic can be written with short vowels and consonant length annotated by diacritics (accents and such), but these are typically omitted in printed text.
- The Korean writing system builds syllabic blocks out of phonetics glyphs.

In logographic writing systems (e.g., Chinese), glyphs represent words or morphemes.

- Japanese script uses adopted Chinese characters (Kanji) alongside syllabic scripts (Hiragana for ordinary words & Katakana for loan words).

Transliteration

Transliteration is the process of rendering phrases (typically proper names or scientific terminology) in another script.

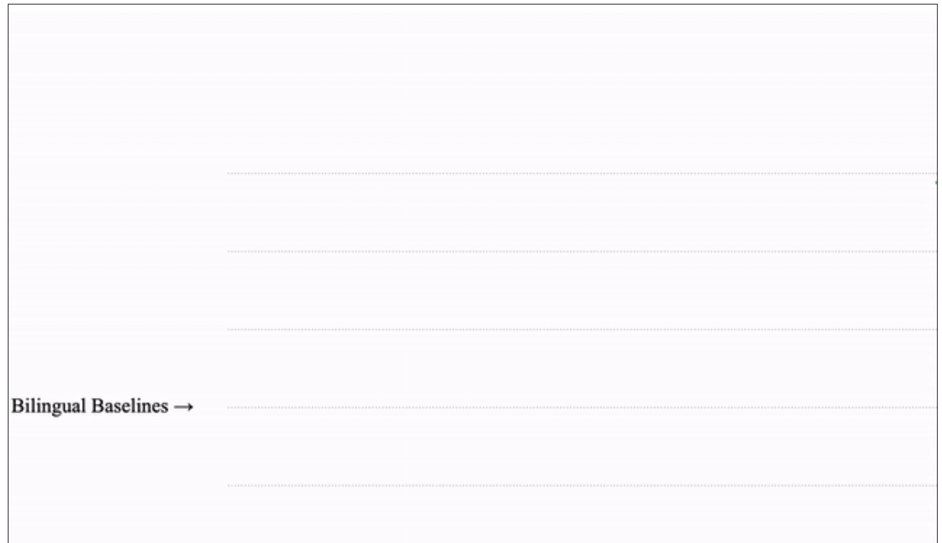
- Rule-based systems are effective in some cases.
- When English names are transliterated into Chinese, the choice of characters is often based on both phonetic similarity and meaning: E.g., "Yosemite" is often transliterated as 优山美地 Yōushānměidì (excellent, mountain, beautiful, land).
- A word's language of origin can affect its transliteration.

System	EnTh	ThEn	EnPe	PeEn	EnCh	ChEn	EnVi	EnHi	EnTa	EnKa	EnBa	EnHe	HeEn
No dropouts	0.434	0.467	0.566	0.365	0.754	0.306	0.390	0.466	0.451	0.387	0.450	0.616	0.286
Baseline model	0.467	0.503	0.594	0.390	0.739	0.347	0.458	0.481	0.455	0.418	0.465	0.632	0.284
Right-left model	0.462	0.502	0.598	0.402	0.751	0.351	0.458	0.476	0.446	0.403	0.476	0.606	0.287
Ensemble ×4	0.477	0.526	0.605	0.407	0.752	0.366	0.478	0.504	0.469	0.438	0.489	0.633	0.291
+ Re-ranking	0.475	0.534	0.606	0.436	**0.765**	0.365	0.494	0.515	**0.483**	0.441	**0.488**	**0.638**	0.294
+ Synthetic data	**0.484**	**0.728**	**0.610**	**0.585**	0.760	**0.759**	**0.496**	**0.519**	0.471	**0.455**	0.484	0.626	**0.615**
Test set	0.167	0.328	—	—	0.304	0.276	0.502	0.333	0.237	0.340	0.461	0.187	0.153

Table 3: Results (Acc) on the official NEWS 2018 development set. Bolded systems have been evaluated on the official test set (last row).

Roman Grundkiewicz, Kenneth Heafield, 2018, "Neural Machine Translation Techniques for Named Entity Transliteration"

Multilingual Neural Machine Translations



Translation quality improvement of a single massively multilingual model as we increase the capacity (number of parameters) compared to 103 individual bilingual baselines.

https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html

First Large-Scale Massively Multilingual Experiment

Trained on Google-internal corpora for 103 languages.

1M or fewer sentence pairs per language; 95M examples total.

Evaluated on "10 languages from different typological families: Semitic – Arabic (Ar), Hebrew (He), Romance – Galician (Gl), Italian (It), Romanian (Ro), Germanic – German (De), Dutch (Nl), Slavic – Belarusian (Be), Slovak (Sk) and Turkic – Azerbaijani (Az) and Turk- ish (Tr)."

Model architecture: Sequence-to-sequence Transformer with a target-language indicator token prepended to each source sentence to enable multiple output languages.

- 6 layer encoder & decoder; 1024/8192 layer sizes; 16 heads
- 473 million trainable model parameters
- 64k subwords shared across 103 languages

Baseline: Same model architecture trained on bilingual examples.

Roee Aharoni, Melvin Johnson, Orhan Firat, 2019, "Massively Multilingual Neural Machine Translation"

First Large-Scale Massively Multilingual Experiment

Evaluated on "10 languages from different typological families: Semitic – Arabic (Ar), Hebrew (He), Romance – Galician (Gl), Italian (It), Romanian (Ro), Germanic – German (De), Dutch (Nl), Slavic – Belarusian (Be), Slovak (Sk) and Turkic – Azerbaijani (Az) and Turk- ish (Tr)."

	Ar	Az	Be	De	He	It	Nl	Ro	Sk	Tr	Avg.
baselines	23.34	16.3	21.93	30.18	31.83	**36.47**	36.12	34.59	25.39	27.13	28.33
many-to-one	**26.04**	**23.68**	**25.36**	35.05	**33.61**	35.69	**36.28**	36.33	28.35	**29.75**	**31.01**
many-to-many	22.17	21.45	23.03	**37.06**	30.71	35.0	36.18	**36.57**	**29.87**	27.64	29.97

Table 5: X→En test BLEU on the 103-language corpus

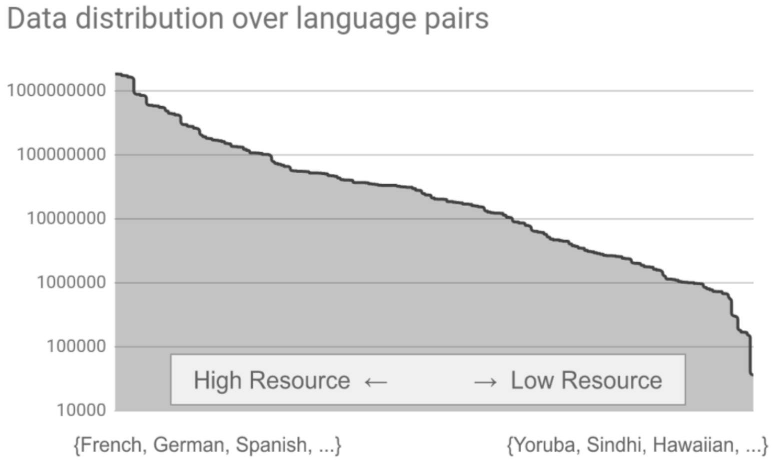
	Ar	Az	Be	De	He	It	Nl	Ro	Sk	Tr	Avg.
baselines	10.57	8.07	15.3	23.24	19.47	31.42	28.68	27.92	11.08	15.54	19.13
one-to-many	**12.08**	**9.92**	**15.6**	**31.39**	**20.01**	**33**	**31.06**	**28.43**	**17.67**	**17.68**	**21.68**
many-to-many	10.57	9.84	14.3	28.48	17.91	30.39	29.67	26.23	18.15	15.58	20.11

Table 6: En→X test BLEU on the 103-language corpus

Roee Aharoni, Melvin Johnson, Orhan Firat, 2019, "Massively Multilingual Neural Machine Translation"

Full-Scale Massively Multilingual Experiment

25 billion parallel sentences in 103 languages.



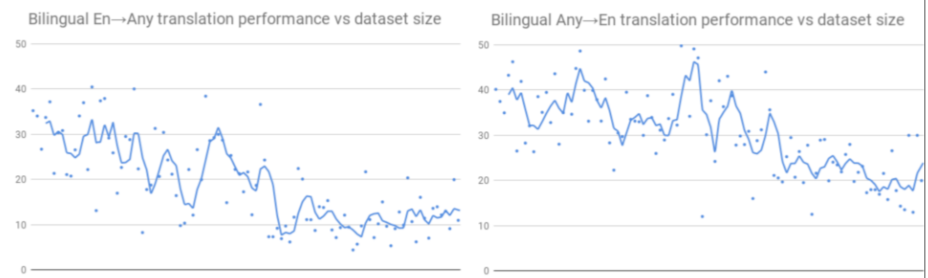
Arivazhagan, Bapna, Firat, et al. (2019) "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges"

Full-Scale Massively Multilingual Experiment

25 billion parallel sentences in 103 languages.

Baselines: Bilingual Transformer Big w/ 32k Vocab (~375M params) for most languages; Transformer Base for low-resource languages.

Evaluation: Constructed multi-way dataset of 3k-5k translated English sentences.



"Performance on individual language pairs is reported using dots and a trailing average is used to show the trend."

Arivazhagan, Bapna, Firat, et al. (2019) "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges"

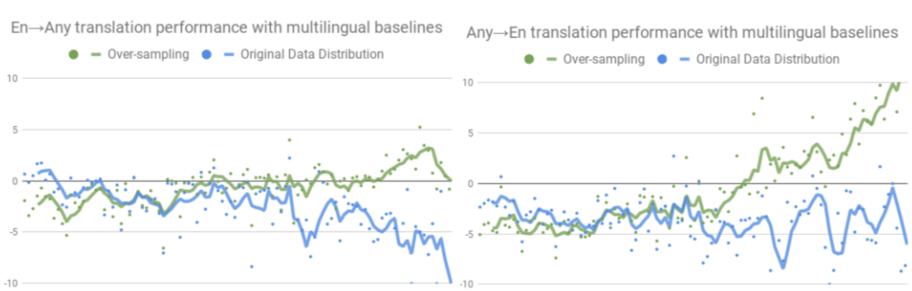
Full-Scale Massively Multilingual Experiment

25 billion parallel sentences in 103 languages.

Baselines: Bilingual Transformer Big w/ 32k Vocab (~375M params) for most languages; Transformer Base for low-resource languages.

Multilingual system: Transformer Big w/ 64k Vocab trained 2 ways:

- "All the available training data is combined as it is."
- "We over-sample (up-sample) low-resource languages so that they appear with equal probability in the combined dataset."



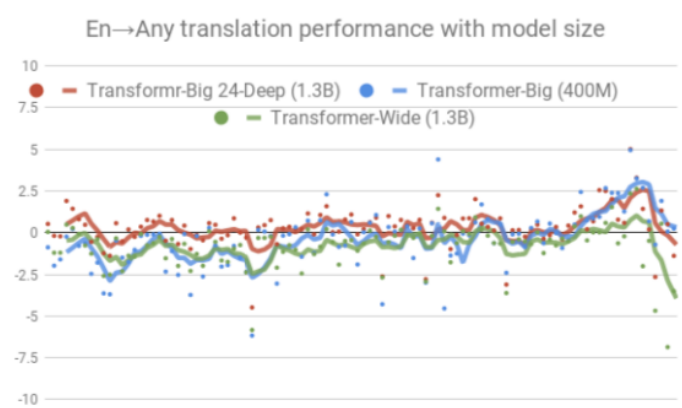
Arivazhagan, Bapna, Firat, et al. (2019) "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges"

Full-Scale Massively Multilingual Experiment

25 billion parallel sentences in 103 languages.

Baselines: Bilingual Transformer Big w/ 32k Vocab (~375M params) for most languages; Transformer Base for low-resource languages.

Multilingual systems: Transformers of varying sizes.



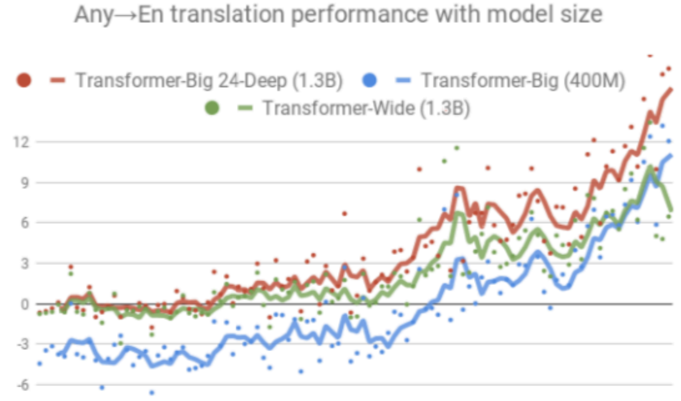
Arivazhagan, Bapna, Firat, et al. (2019) "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges"

Full-Scale Massively Multilingual Experiment

25 billion parallel sentences in 103 languages.

Baselines: Bilingual Transformer Big w/ 32k Vocab (~375M params) for most languages; Transformer Base for low-resource languages.

Multilingual systems: Transformers of varying sizes.



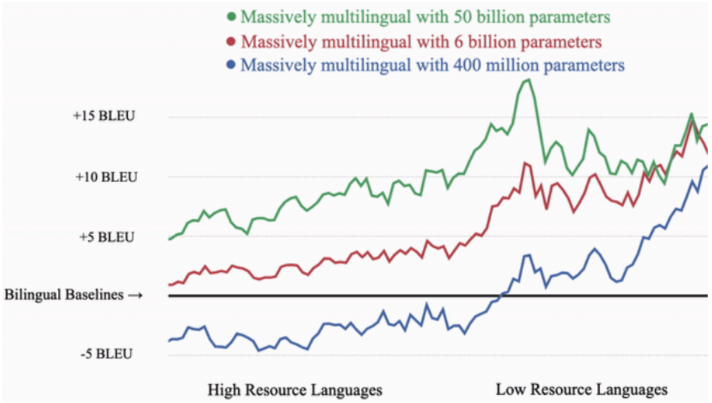
Arivazhagan, Bapna, Firat, et al. (2019) "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges"

Full-Scale Massively Multilingual Experiment

25 billion parallel sentences in 103 languages.

Baselines: Bilingual Transformer Big w/ 32k Vocab (~375M params) for most languages; Transformer Base for low-resource languages.

Multilingual systems: Transformers of varying sizes.



https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html

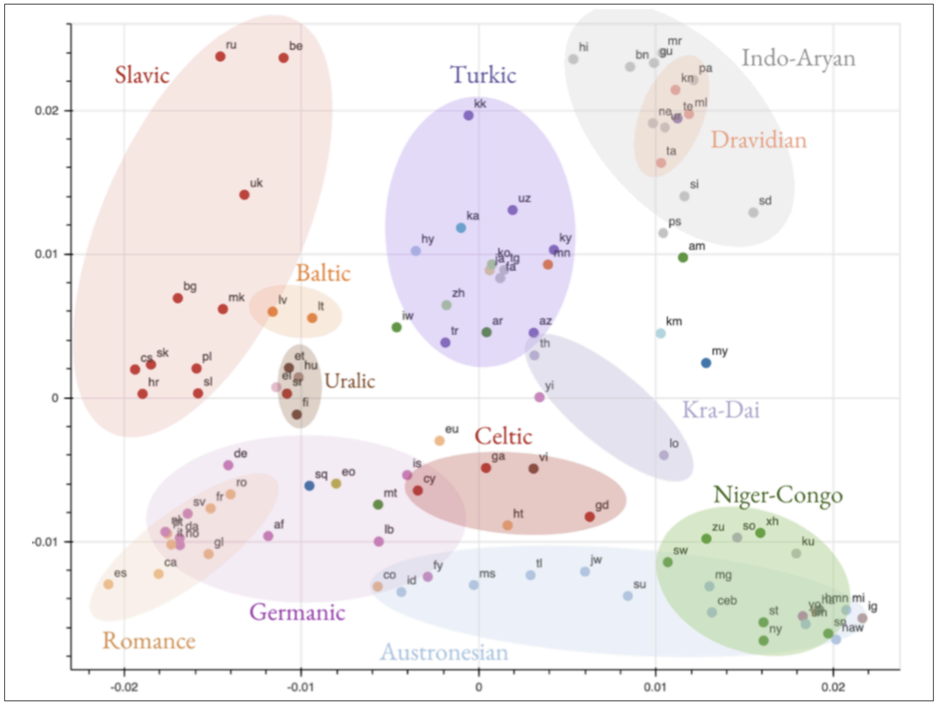
Identifying Language Families

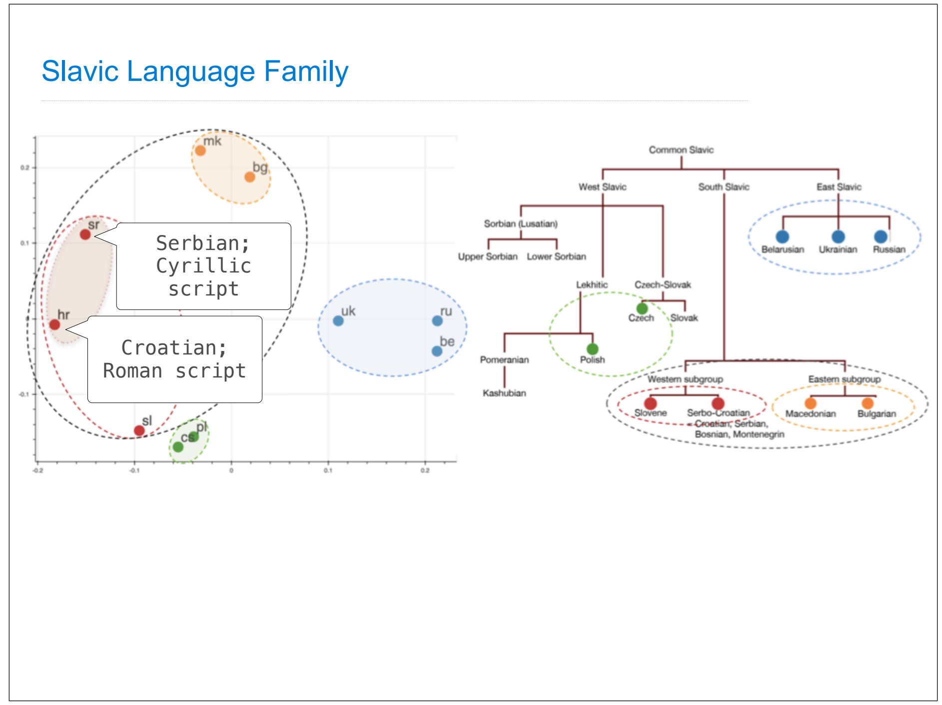
Clustering Language Representations

Measuring similarity between two languages X and Y:

- Translate 3k English sentences to both X and Y.
- For each sentence i, encode both its translation X_i and Y_i.
- Summarize all encoder activations as a low rank vector (SVD).
- Learn linear projections from encoded X_i and encoded Y_i to a shared space in which they are close together (CCA).
- Measure the mean correlation coefficient between projections.
- Result: Similarity matrix with an entry for each language pair.
- Visualization: Reduce each column to a position on a plane (Spectral Embedding).

Kudugunta et al., 2019, "Investigating Multilingual NMT Representations at Scale"


Slavic
Turkic
Indo-Aryan
Dravidian
Baltic
Uralic
Kra-Dai
Celtic
Niger-Congo
Germanic
Romance
Austronesian
0.02
0.01
0
-0.01
-0.02
-0.01
0
0.01
0.02


Slavic Language Family
mk
bg
sr
hr
sl
cs
pl
uk
ru
be
Serbian;
Cyrillic
script
Croatian;
Roman script
Common Slavic
West Slavic
South Slavic
East Slavic
Sorbian (Lusatian)
Upper Sorbian
Lower Sorbian
Lekhitic
Czech-Slovak
Czech
Slovak
Pomeranian
Polish
Kashubian
Belarusian
Ukrainian
Russian
Western subgroup
Eastern subgroup
Slovene
Serbo-Croatian
Croatian, Serbian,
Bosnian, Montenegrin
Macedonian
Bulgarian