Advanced NLP

Lecture 8?: Multilinguality

Dr. Gabriel Stanovsky

NLP = English Processing?

- Vast majority of NLP is done on **English**

- **Why?**
  - Lingua franca
  - Easier to publish
  - Hard to work on other languages

Emily M. Bender
@emilymbender

Dear computer scientists, “Natural Language” is *not* a synonym for “English”

Rada Mihalcea
@radamihalcea

Much of #NLProc targets English **yet only 20% of people speak English.**

Yes—we need NLP for the remaining 80%!

Agenda

- Why you should do NLP Beyond English?
- The State of Linguistic Diversity in NLP
- Multilingual Datasets and Benchmarks
- Cross-Lingual Transfer

Agenda

- **Why you should do NLP Beyond English?**
- The State of Linguistic Diversity in NLP
- Multilingual Datasets and Benchmarks
- Cross-Lingual Transfer

Why you should do NLP Beyond English? (Ruder, 2020)

- **The linguistic perspective - Are we focusing on English-specific properties?**
  - Indo-European
  - Latin alphabet
  - Subject-verb-object word order
  - Relatively limited morphology
  - Relatively fixed word order
- The world atlas of languages lists 192 ***typological features***
  - With average of ~6 possible values for each
  - I.e., 6^192 linguistic combinations
- NLP focuses on ~1 of these combinations

Why you should do NLP Beyond English?

- **The ML perspective**
  - English has orders of magnitude more data than most other languages
  - Both supervised and unsupervised
- NLP architecture biased towards English properties
  - E.g., tokenization for agglutinative, fusional languages affects token boundaries
- **Can we get good performance with much less data?**
- **Can training data in English help other languages?**
  - May in turn lead to more efficient models in English as well

Why you should do NLP Beyond English?

- **The cultural and normative perspective**
  - Culture is inherently tied to language
  - English text will amplify certain set of beliefs and common-sense reasoning
  - E.g., western, capitalistic, patriarchic, Christian, etc.

- Such assumptions **do not hold** for most of the world
  - And hence most of the potential users

Why you should do NLP Beyond English?

- **The cognitive perspective**
  - Current concesus is that language *does not* influence cognition
  - Contrary to the Sapir–Whorf hypothesis
  - E.g., 50 Inuit words for snow, no word for X, no tense in Mandarin, …

- **Implies a language-agnostic abstraction?**
  - Requires multilingual analysis

Agenda

- Why you should do NLP Beyond English?
- **The State of Linguistic Diversity in NLP**
- Multilingual Datasets and Benchmarks
- Cross-Lingual Transfer

The State of Linguistic Diversity in NLP (Joshi et al., 2020)

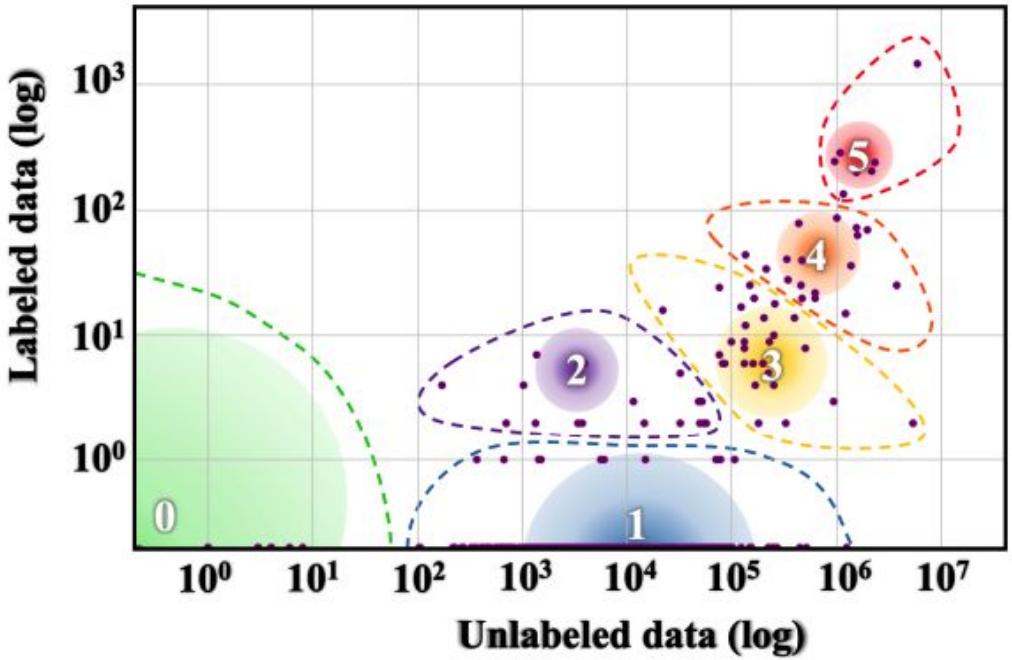
- Over **7,000 languages** spoken world-wide!

- Representing a combination of the **various aspects** we discussed
  - Morphology
  - Culture
  - Low-resource

- **How does NLP work distribute around them?**

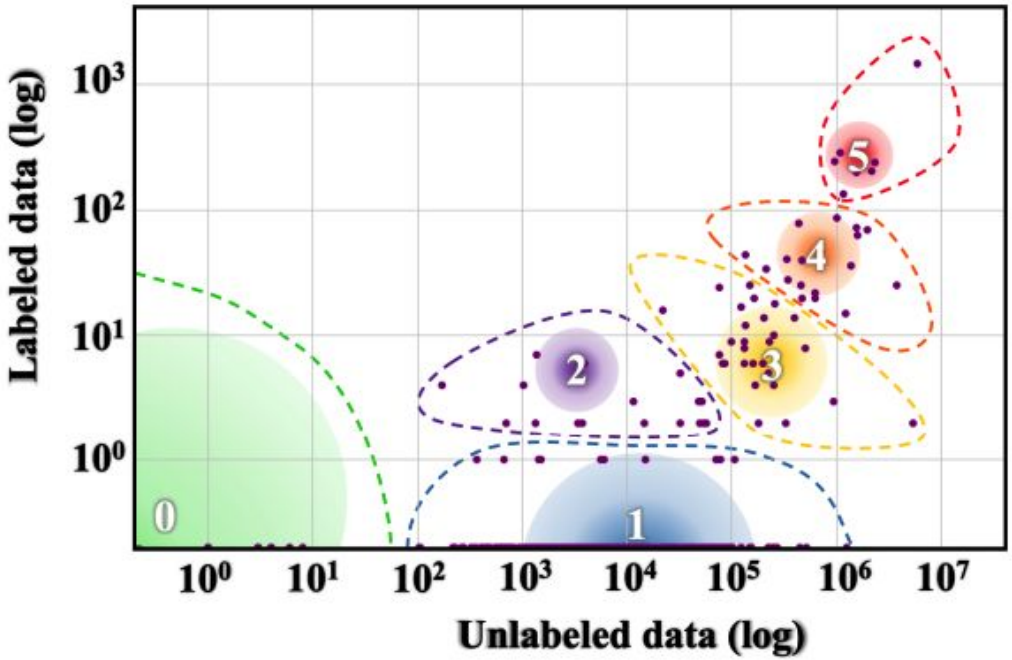
The State of Linguistic Diversity in NLP (Joshi et al., 2020)

- **5: The winners**
  - Dominant online presence
  - Massive industry & gov. investment



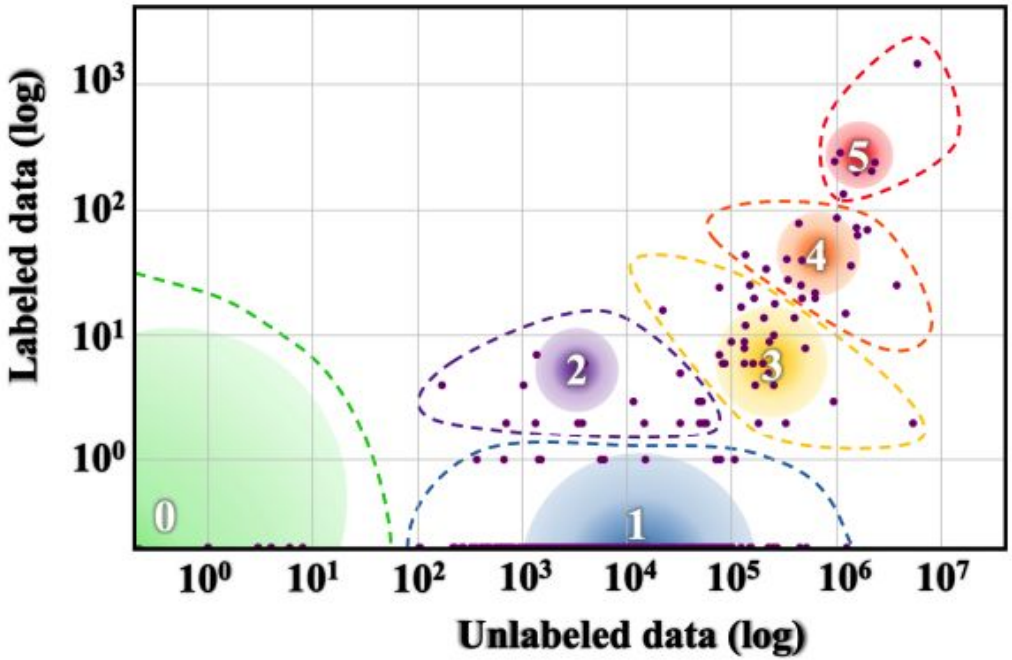
The State of Linguistic Diversity in NLP (Joshi et al., 2020)

- **4: The Underdogs**
  - Distinguished from (5) in amount of supervision



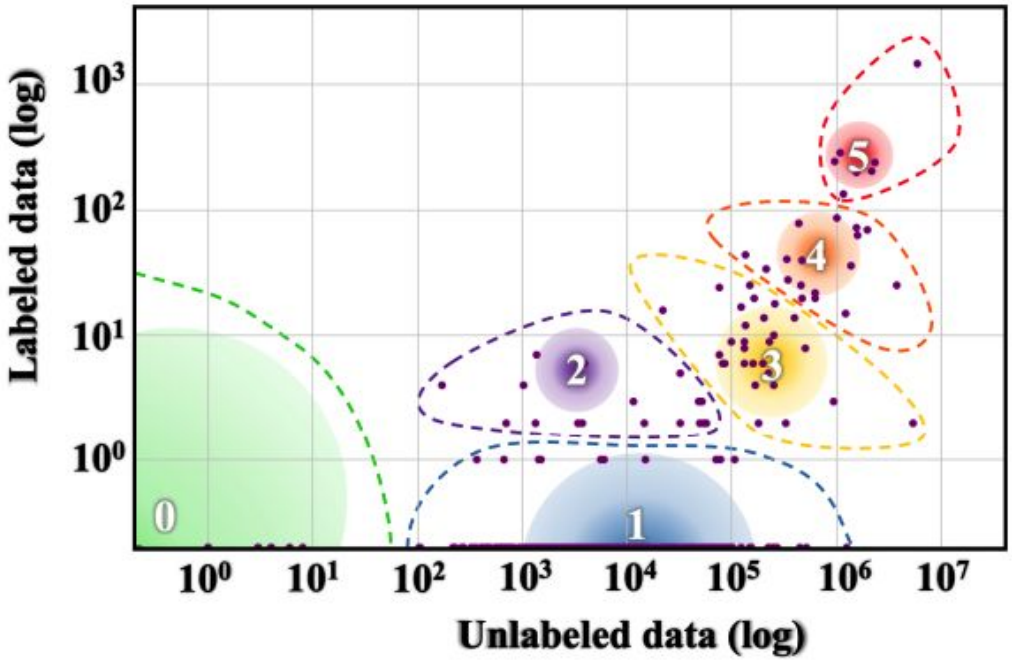
The State of Linguistic Diversity in NLP (Joshi et al., 2020)

- **3: The Rising Stars**
  - Benefit from self-supervision
  - Few annotated resources



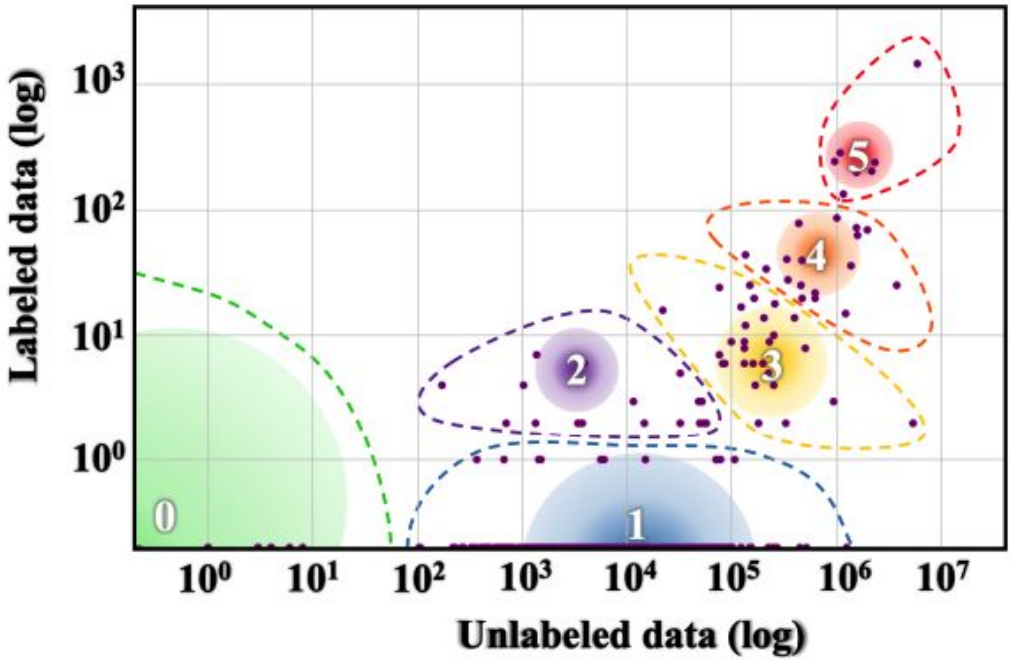
The State of Linguistic Diversity in NLP (Joshi et al., 2020)

- **2: The Hopefuls**
  - Mostly less raw data
  - I.e., less online presence



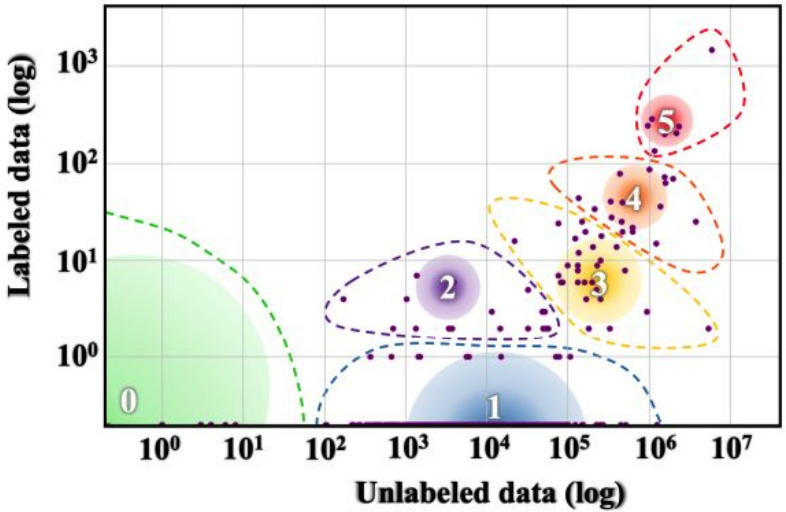
The State of Linguistic Diversity in NLP (Joshi et al., 2020)

- **1: The Scraping-Bys**
  - Some unlabeled data
  - Little-to-no annotations



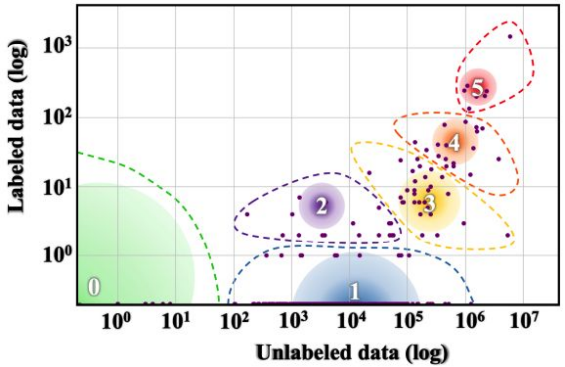
The State of Linguistic Diversity in NLP (Joshi et al., 2020)

1. **Who are the Winners?**
2. **What % of native speakers are in category 0?**
3. **Where do you think Hebrew / Arabic / Russian are?**



The State of Linguistic Diversity in NLP (Joshi et al., 2020)

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.0B	88.17%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	1.0B	8.93%
2	Zulu, Konkani, Lao, Maltese, Irish	19	300M	0.76%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.1B	1.13%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	1.6B	0.72%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%



Taxonomy of all languages

Agenda

- Why you should do NLP Beyond English?
- The State of Linguistic Diversity in NLP
- **Multilingual Datasets and Resources**
- Cross-Lingual Transfer

Multilingual Datasets

- Much of the progress in English NLP enabled by **large-scale annotations**
  - Penn Tree Bank
  - SQuAD
  - SNLI
  - Natural Questions

- Used for training, fine-tuning, few-shot, evaluation

- Such resources are **required in other languages**

XTREME (Hu et al., 2020)

- A multilingual benchmark covering 9 diverse tasks and 40 languages

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517–11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

XTREME

Language	ISO 639-1 code	# Wikipedia articles (in millions)	Script	Language family	Diacritics / special characters	Extensive compound-ing	Bound words / clitics	Inflec-tion	Deriva-tion	# datasets with language
Afrikaans	af	0.09	Latin	IE: Germanic		X				3
Arabic	ar	1.02	Arabic	Afro-Asiatic	X		X	X		7
Basque	eu	0.34	Latin	Basque	X		X	X	X	3
Bengali	bn	0.08	Brahmic	IE: Indo-Aryan	X	X	X	X	X	3
Bulgarian	bg	0.26	Cyrillic	IE: Slavic	X		X	X		4
Burmese	my	0.05	Brahmic	Sino-Tibetan	X	X				1
Dutch	nl	1.99	Latin	IE: Germanic		X				3
English	en	5.98	Latin	IE: Germanic						9
Estonian	et	0.20	Latin	Uralic	X	X		X	X	3
Finnish	fi	0.47	Latin	Uralic				X	X	3
French	fr	2.16	Latin	IE: Romance	X		X			6
Georgian	ka	0.13	Georgian	Kartvelian				X	X	2
German	de	2.37	Latin	IE: Germanic		X		X		8
Greek	el	0.17	Greek	IE: Greek	X	X		X		5
Hebrew	he	0.25	Hebrew	Afro-Asiatic				X		3
Hindi	hi	0.13	Devanagari	IE: Indo-Aryan	X	X	X	X	X	6
Hungarian	hu	0.46	Latin	Uralic	X	X		X	X	4
Indonesian	id	0.51	Latin	Austronesian			X	X	X	4
Italian	it	1.57	Latin	IE: Romance	X		X			3
Japanese	ja	1.18	Ideograms	Japonic			X	X		4
Javanese	jv	0.06	Brahmic	Austronesian	X		X			1
Kazakh	kk	0.23	Arabic	Turkic	X			X	X	1
Korean	ko	0.47	Hangul	Koreanic		X		X	X	5
Malay	ms	0.33	Latin	Austronesian			X	X		2
Malayalam	ml	0.07	Brahmic	Dravidian	X	X	X	X		2
Mandarin	zh	1.09	Chinese ideograms	Sino-Tibetan		X				8
Marathi	mr	0.06	Devanagari	IE: Indo-Aryan			X	X		3
Persian	fa	0.70	Perso-Arabic	IE: Iranian		X				2
Portuguese	pt	1.02	Latin	IE: Romance	X		X			3
Russian	ru	1.58	Cyrillic	IE: Slavic				X		7
Spanish	es	1.56	Latin	IE: Romance	X		X			7
Swahili	sw	0.05	Latin	Niger-Congo			X	X	X	3
Tagalog	tl	0.08	Brahmic	Austronesian	X		X	X		1
Tamil	ta	0.12	Brahmic	Dravidian	X	X	X	X	X	3
Telugu	te	0.07	Brahmic	Dravidian	X	X	X	X	X	4
Thai	th	0.13	Brahmic	Kra-Dai	X					4
Turkish	tr	0.34	Latin	Turkic	X	X		X	X	5
Urdu	ur	0.15	Perso-Arabic	IE: Indo-Aryan	X	X	X	X	X	4
Vietnamese	vi	1.24	Latin	Austro-Asiatic	X					6
Yoruba	yo	0.03	Arabic	Niger-Congo	X					1

Hebrew QA (ParaShoot; Keren & Levy, 2021)

- Crowdsourced over Wikipedia articles in SQuAD format
- Used the Prolific platform

	#Articles	#Paragraphs	#Questions
Train	295	565	1792
Validation	33	63	221
Test	165	319	1025
Total	493	947	3038

Q: מה היה שטחו של כפר שמריהו כשהוקם?

ma haya shitkho shel kfar shmaryahu
what was area-of-it of Kfar Shmaryahu
kshe-hukam
when-was.established

'What was Kfar Shmaryahu's area when it was established?'

A: ... הישוב הוקם על שטח של

ha-yeshuv hukam al shetakh
the-village was.established on area
shel ...
of ...

'The village was established on an area of ...'

Hebrew NER (NEMO; Bareket & Tsarfaty, 2021)

- Annotation scheme, data and model for Hebrew NER
- Over the Hebrew Tree Bank (news)

(a) השרה לבני
hasara *livni*
the-minister [Livni]$_{PER}$
'Minister [Livni]$_{PER}$'
(b) לבני גנץ
le-beny *gantz*
for-[Benny Gantz]$_{PER}$
'for [Benny Gantz]$_{PER}$'
(c) לבני היקר
li-bni *hayakar*
for-son.POSS.1SG the-dear
'for my dear son'
(d) לבני חימר
livney *kheymar*
brick.CS clay
'clay bricks'

	train	**dev**	**test**
Sentences	4,937	500	706
Tokens	93,504	8,531	12,619
Morphemes	127,031	11,301	16,828
All mentions	6,282	499	932
Type: Person (PER)	2,128	193	267
Type: Organization (ORG)	2,043	119	408
Type: Geo-Political (GPE)	1,377	121	195
Type: Location (LOC)	331	28	41
Type: Facility (FAC)	163	12	11
Type: Work-of-Art (WOA)	114	9	6
Type: Event (EVE)	57	12	0
Type: Product (DUC)	36	2	3
Type: Language (ANG)	33	3	1

Hebrew Resources Getting Recent Attention

קול קורא – הקמת מאגרי מידע בעברית ו/או בערבית מדוברת

רוצים לצפות בוובינר שנערך בנושא? לחצו כאן

תאריך אחרון להגשה 15/05/2022

חטיבות תשתיות חדשנות

תמצית הקול קורא

הרשות הלאומית לחדשנות טכנולוגית (להלן: "רשות החדשנות") מודיעה על פתיחת שני הליכים (כמפורט להלן) להגשת בקשות לקבלת מענק לצורך הקמת מאגרי מידע בעברית ו/או בערבית מדוברת. מאגרי מידע אלו יהוו תשתית לתאגידים ישראליים ו/או למוסדות מחקר ישראליים, העוסקים במחקר ופיתוח ואשר נדרשים ליכולות עיבוד שפה למטרות שונות.

הקול הקורא פתוח למאגרים בעברית ובערבית ואינו מוגבל לערבית מדוברת בלבד (*עדיפות תינתן לניב פלסטיני, מאגרים עדכניים וערבית מדוברת)

במסגרת הקול הקורא תתאפשר הגשת בקשות באחד משני ההליכים:

1. מבקש שהינו תאגיד תעשייתי: הגשת בקשה תיעשה מכוח מסלול משנה ב' – תשתיות מו"פ לתעשייה של מסלול הטבה מס' 5 של רשות החדשנות – פיתוח תשתיות טכנולוגיות, מסחור ומחקר יישומי באקדמיה ובתעשייה (להלן: "מסלול משנה תשתיות מו"פ לתעשייה").
2. מבקש שהינו חברת יישום: הגשת בקשה תיעשה מכוח מסלול משנה ג' – מחקר יישומי באקדמיה להטמעה בתעשייה של מסלול הטבה מס' 5 של רשות החדשנות – פיתוח תשתיות טכנולוגיות, מסחור ומחקר יישומי באקדמיה ובתעשייה (להלן: "מסלול משנה מחקר יישומי באקדמיה").

מובהר, כי כל אחת מהחלופות הנ"ל מהווה הליך נפרד ובלתי תלוי, כאשר הבקשות אשר תוגשנה במסגרת מסלול משנה תשתיות מו"פ לתעשייה תיבחנה זו מול זו והבקשות שתוגשנה במסגרת מסלול משנה מחקר יישומי באקדמיה תיבחנה זו מול זו.

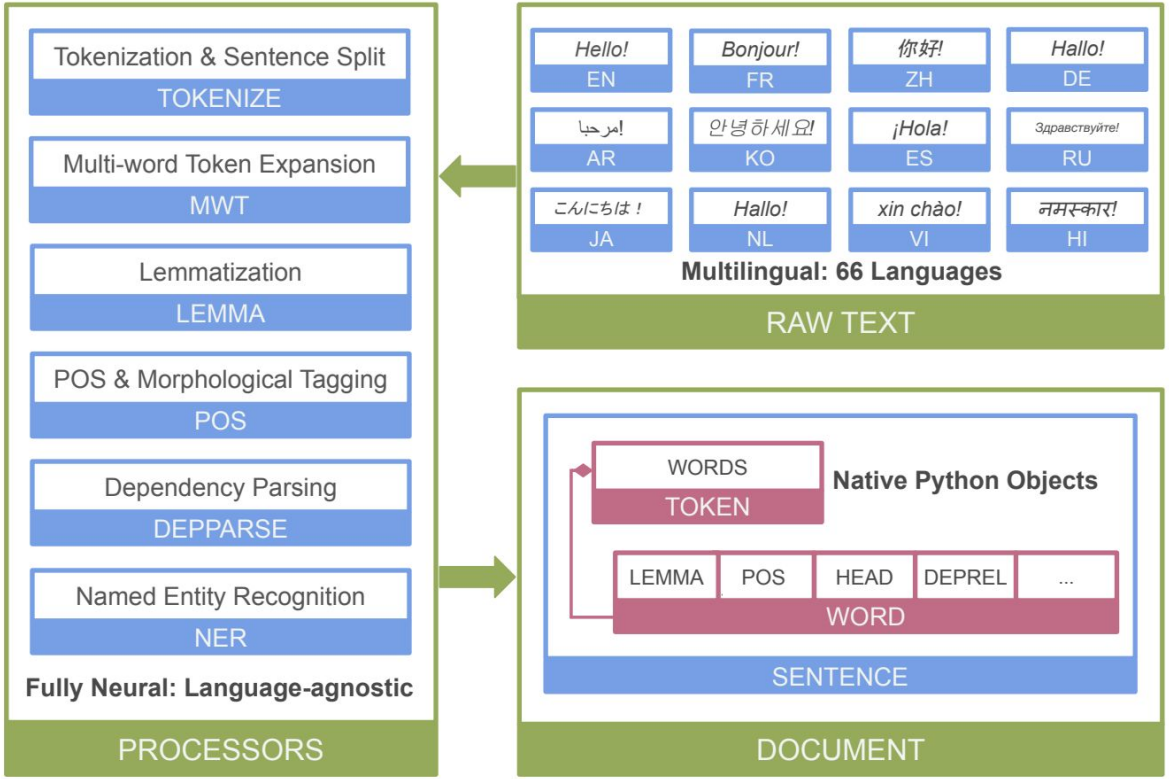
Multilingual Resources

- “Daily” NLP work requires parsers for **intermediate structures**
  - Word / sentence splitting
  - POS tagging
  - Dependency parsing
  - Coreference resolution
  - And many more

- **Off-the-shelf** English solutions are decent for some domains (e.g., spaCy)

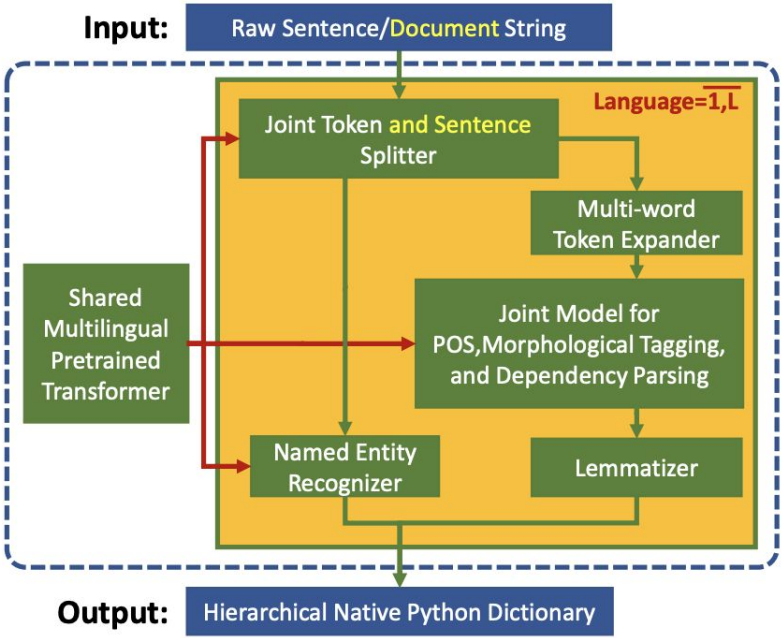
- Similar solutions are needed for **other languages**

Stanza (Qi et al., 2020)



Trankit (Van Nguyen et al., 2021)

- Joint training for 56 languages
- Best performance I saw for Hebrew?
  - Still lots to improve



Agenda

- Why you should do NLP Beyond English?
- The State of Linguistic Diversity in NLP
- Multilingual Datasets and Resources
- **Cross-Lingual Transfer**

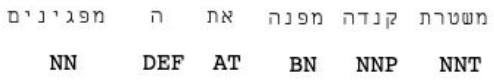
Cross-Lingual Transfer

Multilingual Pretraining on L = {l-1,..., l-k}

1

Task-specific finetuning in L’ \subseteq L

2



Zero-shot evaluation in l’

3

Multilingual Pretraining

- Circa BERT (2019) LLMs start pretraining on **multilingual raw datasets**
- E.g., MBERT trains on the “top” **100 languages** from Wikipedia
- Does so by **interleaving batches** from different languages
  - Using a single shared tokenizer for all languages

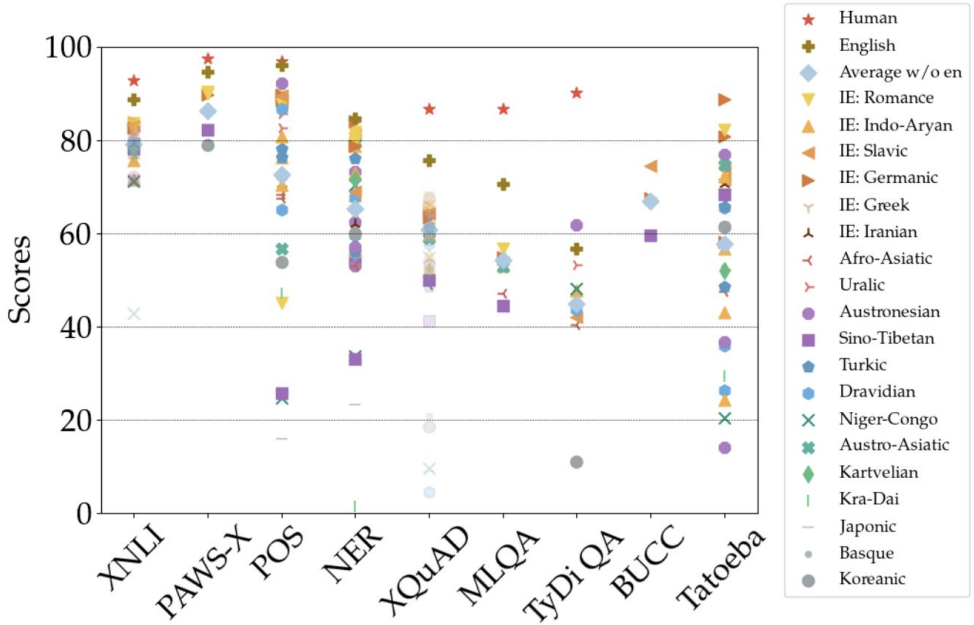


Figure 1. An overview of XLM-R's performance on the XTREME tasks across all languages in each task. We highlight an estimate of human performance, performance on the English test set, the average of all languages excluding English, and the family of each language. Performance on pseudo test sets for XNLI and XQuAD is shown with slightly transparent markers.

Benefits of Cross-Lingual Transfer

- Low resource languages benefit from supervision in other languages
- In the extreme case (zero-shot), **no need for training signal at all**

Multilingual
Pretraining on
L = {l-1,..., l-k}

1

Russia	stokes	Ukraine	tensions
PROPN	VERB	PROPN	NOUN

Task-specific finetuning
in L' \subseteq L

2

מפגינים	את	ה	מפנה	קנדה	משטרת
NN	AT	DEF	BN	NNP	NNT

Zero-shot evaluation
in l'

3

Why and how does multilingual transfer work?

- Pretraining languages X Finetuning languages X Evaluation languages
  - Tokenization
  - Linguistic properties
  - Domain
  - Size
- Many works trying to find **explaining variables**

Multilingual Pretraining on L = {l-1,..., l-k}

1

Task-specific finetuning in L' \subseteq L

2

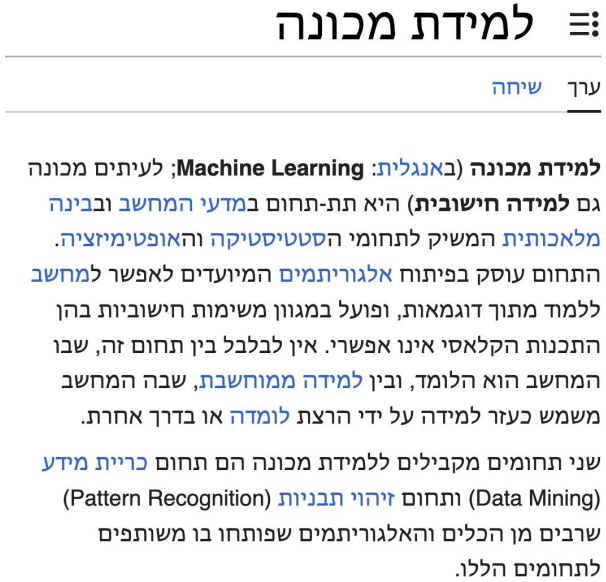
Zero-shot evaluation in l'

3

How much does a shared vocabulary contribute?

- Do **shared vocabulary items** “align” embedding spaces?

- Texts in different languages **tend to overlap**
  - Named entities
  - Loan words
  - Data contamination (Blevins and Zettlemoyer, 2022)



למידת מכונה

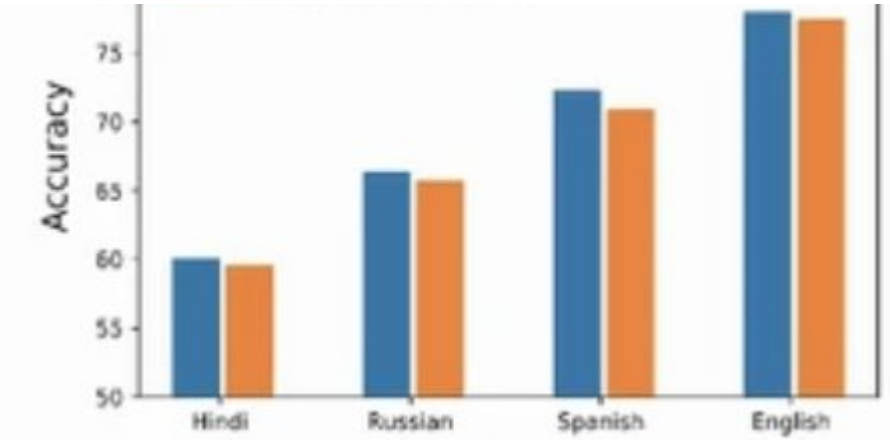
ערך שיחה

למידת מכונה (באנגלית: **Machine Learning**; לעיתים מכונה גם **למידה חישובית**) היא תת-תחום במדעי המחשב ובבינה מלאכותית המשיק לתחומי הסטטיסטיקה והאופטימיזציה. התחום עוסק בפיתוח אלגוריתמים המיועדים לאפשר למחשב ללמוד מתוך דוגמאות, ופועל במגוון משימות חישוביות בהן התכנות הקלאסי אינו אפשרי. אין לבלבל בין תחום זה, שבו המחשב הוא הלומד, ובין למידה ממוחשבת, שבה המחשב משמש כעזר למידה על ידי הרצת לומדה או בדרך אחרת.

שני תחומים מקבילים ללמידת מכונה הם תחום כריית מידע (Data Mining) ותחום זיהוי תבניות (Pattern Recognition) שרבים מן הכלים והאלגוריתמים שפותחו בו משותפים לתחומים הללו.

Token Overlap doesn't Explain Cross-Lingual Transfer

- K et al. (2020) train BERT where **English tokens are shifted**
- Cross-lingual transfer over shifted-BERT (Orange) **decreases only slightly**



How does the finetuning language affect transfer?

- Turc et al., (2021) formulate metrics for assessing cross-lingual transfer

$$\mathcal{Z}(S \to T) = \frac{\mathcal{E}(M^S, T)}{\mathcal{E}(M^T, T)} \qquad \mathcal{Z}(S \to T) = \frac{\mathcal{E}(M^S, T)}{\mathcal{E}(M^T, T)}$$

$$\mathcal{Z}(S \to \mathcal{L}) - \mathcal{Z}(\text{en} \to \mathcal{L})$$

Train Data	Latin–High Resource				Latin–Low Res.			Miscellaneous								Averages			
	en$^{\circ}$	de^{HT}	es^{HT}	fr^{HT}	sw^{HT}	tr^{HT}	vi^{HT}	ar^{HT}	bg^{HT}	el^{HT}	hi^{HT}	ru^{HT}	ur^{HT}	th^{HT}	zh^{HT}	→LH	→LL	→M	→All
mBERT																			
en$^{\circ}$	100.0	92.2	95.7	93.4	81.2	89.0	92.1	94.8	90.5	91.4	89.1	93.8	94.7	80.5	90.5	95.3	87.4	90.7	91.3
de^{MT}	-4.2	+7.8	+0.4	+2.1	-4.5	+2.3	+0.7	+2.5	+2.5	+2.0	+4.4	+1.5	+4.0	+5.8	+3.0	+1.5	-0.5	+3.2	+2.0
es^{MT}	-3.6	+2.8	+4.3	+2.3	-1.7	-1.0	+2.5	+1.9	+1.9	+0.1	+3.8	+2.6	+3.4	+4.5	+3.2	+1.4	-0.1	+2.7	+1.8
fr^{MT}	-3.0	+2.9	+1.4	+6.6	-1.9	-1.8	+1.1	+3.0	+1.5	-1.2	+1.2	+2.2	+3.7	+2.7	+3.5	+2.0	-0.9	+2.1	+1.5
sw^{MT}	-9.7	-3.9	-8.0	-5.1	+18.8	-5.7	-4.3	-4.2	-4.7	-2.7	-1.1	-5.0	-3.0	+0.1	-5.5	-6.7	+2.9	-3.3	-2.9
tr^{MT}	-14.2	-2.9	-4.6	-2.5	-1.7	+11.0	-2.9	-0.5	-0.1	-1.7	+4.6	-0.3	+2.6	+2.5	+0.4	-6.1	+2.2	+0.9	-0.7
vi^{MT}	-8.4	-1.0	-2.0	+0.6	-0.9	-2.7	+7.9	+0.6	+0.7	-0.1	+3.4	+0.0	+1.6	+6.5	+1.5	-2.7	+1.4	+1.8	+0.5
ar^{MT}	-9.3	-0.5	-2.8	+0.2	-2.0	-1.8	-0.7	+5.2	+1.6	+0.3	+2.7	+0.4	+2.6	+1.5	-0.2	-3.1	-1.5	+1.8	-0.2
bg^{MT}	-7.6	+0.8	-2.1	+0.5	-3.4	-1.4	-0.1	+1.5	+9.5	+0.7	+3.5	+1.5	+1.8	+2.2	+2.2	-2.1	-1.6	+2.9	+0.6
el^{MT}	-9.4	-1.6	-3.4	-0.9	-1.2	-0.5	-1.8	-0.2	+0.7	+8.6	+3.5	-0.6	+0.4	+5.7	-0.3	-3.8	-1.2	+2.2	-0.1
hi^{MT}	-15.5	-3.3	-8.4	-3.6	-4.2	-2.1	-3.4	-2.0	-1.9	-3.5	+10.9	-2.4	+7.5	+2.0	-0.3	-7.7	-3.2	+1.3	-2.0
ru^{MT}	-6.2	+2.1	-0.1	+1.8	-4.3	-0.6	+2.0	+1.5	+4.8	+2.1	+3.7	+6.2	+4.3	+4.5	+2.9	-0.6	-1.0	+3.7	+1.6
ur^{MT}	-24.2	-12.9	-16.7	-13.1	-16.1	-12.4	-14.6	-9.8	-9.9	-11.8	+1.5	-9.8	+5.3	-17.0	-9.4	-16.7	-14.3	-7.6	-11.4
th^{MT}	-24.1	-11.3	-13.8	-11.3	-4.8	-12.9	-9.8	-10.6	-9.3	-8.6	-10.0	-11.4	-12.6	+19.5	-9.7	-15.1	-9.2	-6.6	-9.4
zh^{MT}	-7.0	-0.9	-2.6	+0.1	-9.0	-0.1	+1.6	+0.5	+0.6	-1.4	+3.1	+0.7	+3.6	-0.2	+9.5	-2.6	-2.5	+2.0	-0.1
mT5																			
en$^{\circ}$	100.0	96.0	98.4	99.1	94.0	92.8	96.2	95.0	96.7	97.0	93.0	96.1	93.8	94.3	92.1	98.4	94.3	94.8	95.6
de^{MT}	-1.6	+4.0	+0.6	+1.3	+2.7	+2.6	+1.2	+2.7	+1.5	+1.3	+4.5	+2.5	+4.1	+3.2	+2.5	+1.1	+2.2	+2.8	+2.2
es^{MT}	-2.0	+0.4	+1.6	+0.8	+1.8	+1.3	+0.9	+2.4	+0.9	+1.3	+3.0	+1.8	+2.3	+2.8	+1.9	+0.2	+1.3	+2.1	+1.4
fr^{MT}	-2.7	-0.7	-0.5	+0.9	+0.3	-0.4	-2.0	+1.1	-0.9	-0.9	+0.5	-0.1	-0.4	+0.1	+1.9	-0.8	-0.7	+0.1	-0.3
sw^{MT}	-4.3	-0.8	-1.8	-1.1	+6.0	+1.4	-1.3	+2.9	-1.1	-0.8	+2.8	-0.8	+1.2	+2.6	+2.9	-2.0	+2.0	+1.2	+0.5
tr^{MT}	-4.8	+0.1	-1.6	-0.7	+1.2	+7.2	-0.4	+1.4	+0.0	+0.2	+4.7	+0.8	+4.6	+2.4	+2.4	-1.7	+2.7	+2.1	+1.2
vi^{MT}	-6.6	-1.9	-2.1	-1.1	+2.7	-1.0	+3.8	+2.0	-1.1	-0.9	+2.0	-0.8	+2.4	+3.1	+0.5	-2.9	+1.8	+0.9	+0.1
ar^{MT}	-3.5	-0.6	-0.7	-0.4	-2.5	+0.4	-1.0	+5.0	-0.0	+0.5	+2.0	+0.5	+2.2	+2.8	+2.6	-1.3	-1.0	+2.0	+0.5
bg^{MT}	-1.8	+2.2	+0.7	+1.1	+4.3	+3.6	+1.0	+4.4	+3.3	+2.7	+4.9	+2.6	+4.8	+5.6	+4.8	+0.6	+3.0	+4.1	+3.0
el^{MT}	-3.1	+0.6	+0.5	+0.4	+3.7	+1.2	+0.7	+3.2	+1.2	+3.0	+3.9	+1.8	+3.4	+3.6	+2.0	-0.4	+1.8	+2.7	+1.7
hi^{MT}	-7.0	-1.0	-3.3	-2.2	+1.3	+2.7	-0.6	+1.6	-1.8	-0.2	+7.0	-0.6	+5.7	+0.9	+0.4	-3.4	+1.1	+1.6	+0.2
ru^{MT}	-1.7	+0.8	+1.2	+1.7	+4.0	+2.2	+1.4	+3.0	+2.0	+2.1	+4.8	+3.9	+4.0	+4.7	+4.0	+0.5	+2.5	+3.6	+2.5
ur^{MT}	-8.5	-2.0	-4.5	-3.5	+1.0	+2.6	-1.4	-1.0	-2.0	-0.9	+4.4	-1.2	+6.2	+0.7	+0.1	-4.6	+0.7	+0.8	-0.7
th^{MT}	-3.9	-0.3	-1.6	-0.9	+1.4	+0.4	-0.2	+1.8	-0.2	+0.0	+1.0	+0.1	+1.0	+5.7	+3.0	-1.7	+0.6	+1.5	+0.5
zh^{MT}	-2.5	+1.5	-0.2	+0.9	+4.8	+3.9	+1.5	+3.7	+1.3	+1.3	+5.1	+2.0	+4.5	+5.7	+7.9	-0.1	+3.4	+3.9	+2.8

Table 2: **XNLI**: zero-shot transfer. Values for **en**$^{\circ}$ are *relative* zero-shot abilities (Equation 1). Values for other languages (**xx**^{MT}) are zero-shot *advantages* over English (Equation 2). Surprisingly, some machine-translated datasets (such as German and Russian) are more transferable across the board than the original English set.

How does the pretraining language affect transfer?

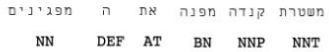
- Pretraining composition is may affect cross-lingual transfer
- Malkin et al (2022) extend these metrics to take pretraining into account

Multilingual Pretraining on L = {l-1,..., l-k}

1

Task-specific finetuning in L' \subseteq L

2

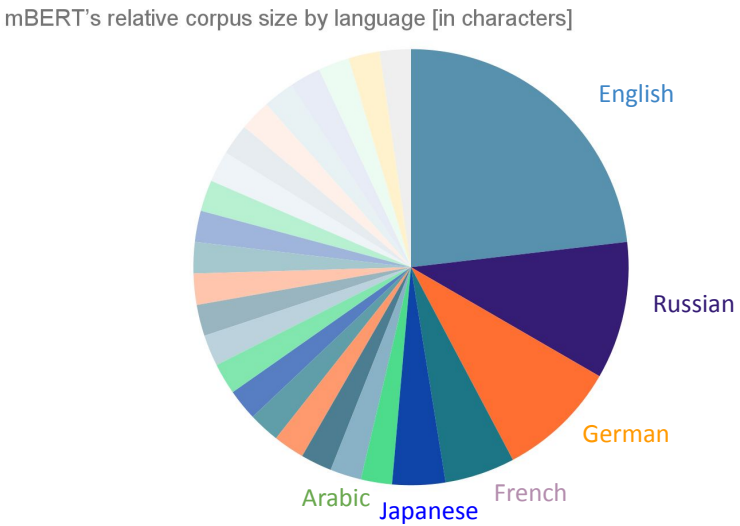


Zero-shot evaluation in l'

3

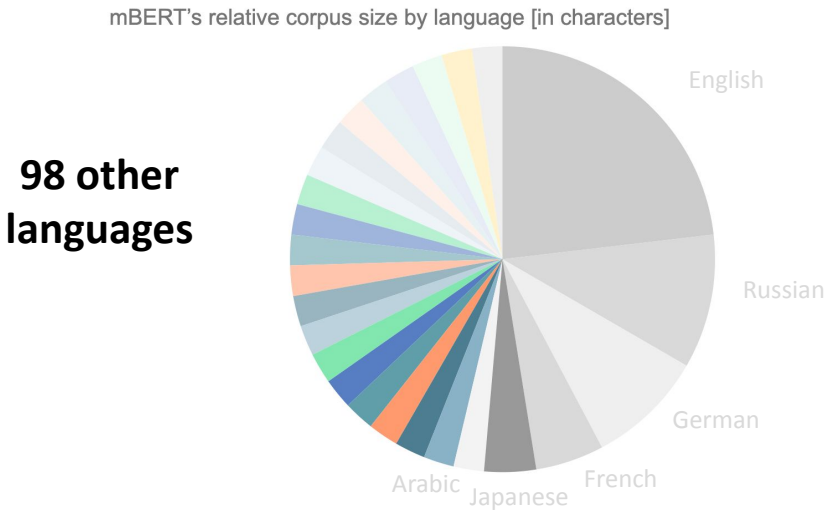
Confound: Unbalanced Corpus Size

- mBERT is composed of 104 languages, **but is far from balanced**



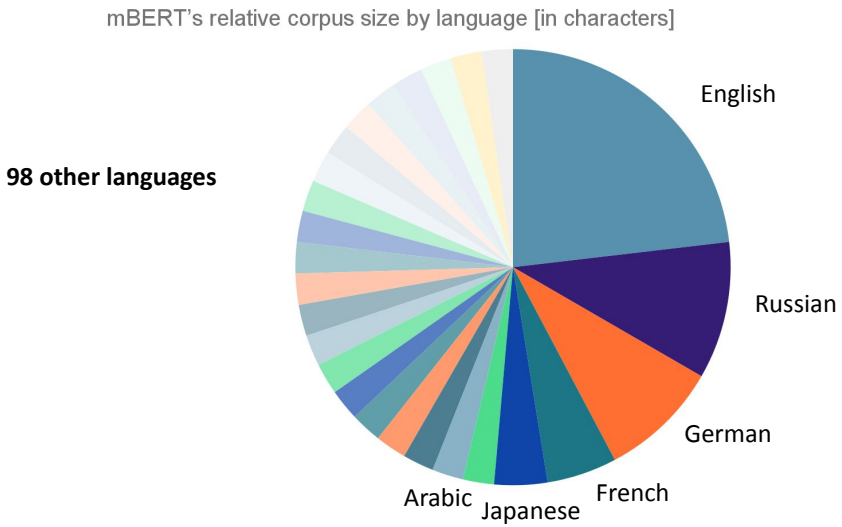
Confound: Unbalanced Corpus Size

- mBERT is composed of 104 languages, **but is far from balanced**



Confound: Unbalanced Corpus Size

- mBERT is composed of 104 languages, **but is far from balanced**



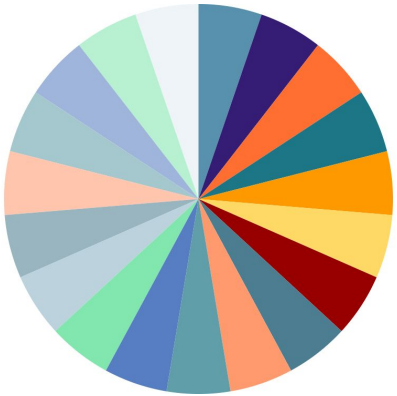
Let's **balance** the pretraining data to get closer to saying something about *language-inherent* properties

A Balanced Pretraining Corpus?

- Subsample 10M characters from each language (in consecutive sentences)

A diverse set of 22 languages

Language	Code	Family	Size [M chars] Wiki	Sample
Piedmontese	pms	Indoeuropean	14	10
Irish	ga	Indoeuropean	38	10
Nepali	ne	Indoeuropean	78	10
Welsh	cy	Indoeuropean	85	10
Finnish	fi	Uralic	131	10
Armenian	hy	Indoeuropean	174	10
Burmese	my	Sino-Tibetian	229	10
Hindi	hi	Indoeuropean	473	10
Telugu	te	Dravidian	533	10
Tamil	ta	Dravidian	573	10
Korean	ko	Korean	756	10
Greek	el	Indoeuropean	906	10
Hungarian	hu	Uralic	962	10
Hebrew	he	Afroasiatic	1,261	10
Chinese	zh	Sino-Tibetian	1,546	10
Arabic	ar	Afroasiatic	1,695	10
Slovak	sv	Indoeuropean	1,744	10
Japanese	ja	Japonese	3,288	10
French	fr	Indoeuropean	4,958	10
German	de	Indoeuropean	6,141	10
Russian	ru	Indoeuropean	6,467	10
English	en	Indoeuropean	14,433	10



Zero-Shot Pretraining Language Graph

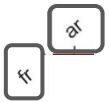
- We define a *directed* bilingual MLM finetune score:

Performance of a model on t after pretraining on s, t

$$\mathcal{F}(s \to t) := \frac{\varepsilon(M^{s,t}, t) - \varepsilon(M^{t}, t)}{\varepsilon(M^{t}, t)}$$

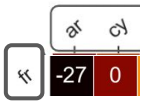
Performance of a monolingual model on t

- In other words, ***F* measures how much *t* gains from *s***

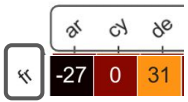


$$\mathcal{F}(\text{fr} \to \text{ar}) := \frac{\varepsilon(\boxed{23.4}\, t) - \varepsilon(\boxed{32.1}\, t)}{\varepsilon(\boxed{32.1}\, t)}$$

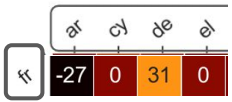
$$\mathcal{F}(\mathrm{fr} \to \mathrm{ar}) := \frac{\varepsilon(\mathrm{l}\ 23.4\ , t) - \varepsilon(\mathrm{.}\ 32.1\ t)}{\varepsilon(\mathrm{.}\ 32.1\ t)}$$



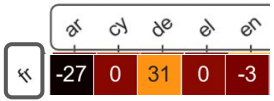
$$\mathcal{F}(\mathrm{fr} \to \mathrm{cy}) := \frac{\varepsilon(39.9, t) - \varepsilon(39.89, t)}{\varepsilon(39.89, t)}$$



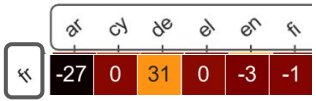
$$\mathcal{F}(\text{fr} \rightarrow \text{de}) := \frac{\varepsilon(M^{\text{fr},t}, t) - \varepsilon(M^t, t)}{\varepsilon(M^t, t)}$$



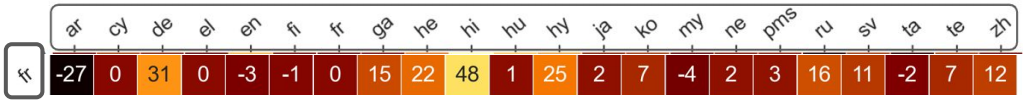
$$\mathcal{F}(\text{fr} \rightarrow \text{el}) := \frac{\varepsilon(M^{\text{fr},t}, t) - \varepsilon(M^t, t)}{\varepsilon(M^t, t)}$$



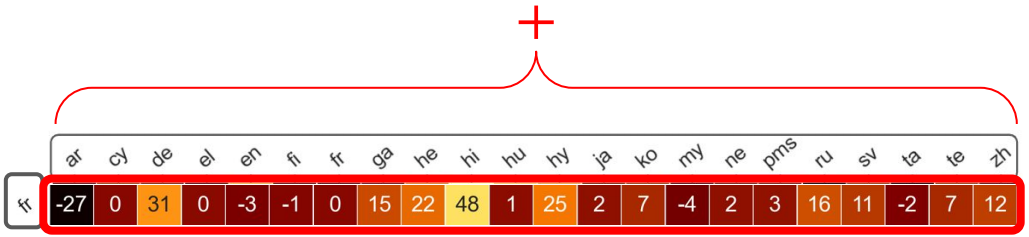
$$\mathcal{F}(\text{fr} \rightarrow \text{en}) := \frac{\varepsilon(M^{\text{fr},t}, t) - \varepsilon(M^t, t)}{\varepsilon(M^t, t)}$$



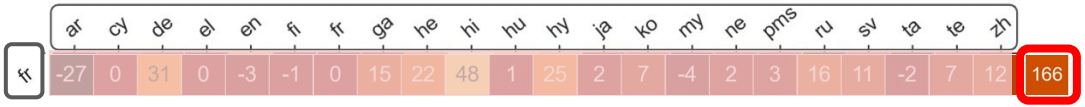
$$\mathcal{F}(\text{fr} \rightarrow \text{fi}) := \frac{\varepsilon(M^{\text{fr},t}, t) - \varepsilon(M^t, t)}{\varepsilon(M^t, t)}$$



$$\mathcal{F}(\text{fr} \to t) := \frac{\varepsilon(M^{\text{fr},t}, t) - \varepsilon(M^t, t)}{\varepsilon(M^t, t)}$$

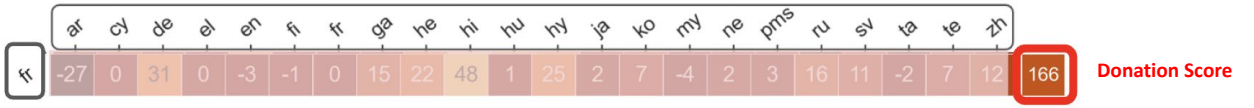


$$\mathcal{D}(\text{fr}) := \sum_{\substack{t \in P \\ t \neq \text{fr}}} \mathcal{F}(\text{fr} \rightarrow t)$$



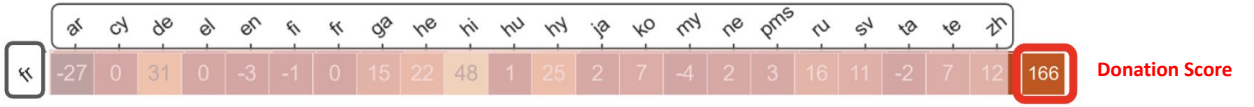
$$\mathcal{D}(l) := \sum_{\substack{t \in P \\ t \neq l}} \mathcal{F}(l \to t)$$

Donation Score



What is the influence of **other languages on** French?

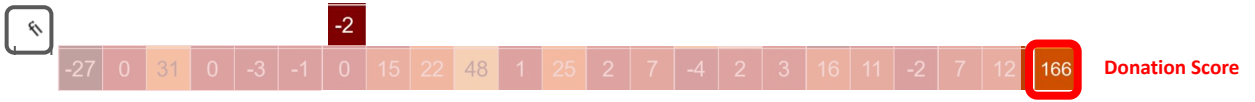
$$\mathcal{F}(\mathbf{fr} \rightarrow t)$$



What is the influence of **other languages on** French?

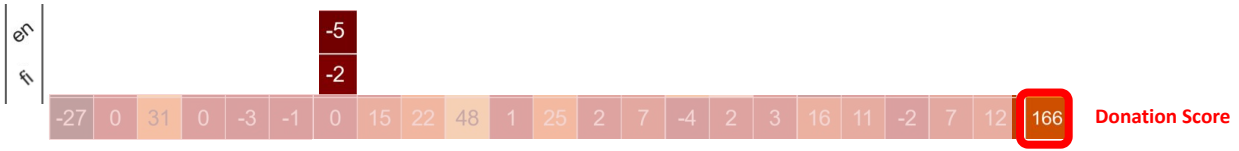
$$\mathcal{F}(s \rightarrow \mathbf{fr})$$

$$\mathcal{F}(\textbf{fi} \rightarrow \textbf{fr}) := \frac{\varepsilon(M^{s,t}, t) - \varepsilon(M^t, t)}{\varepsilon(M^t, t)}$$

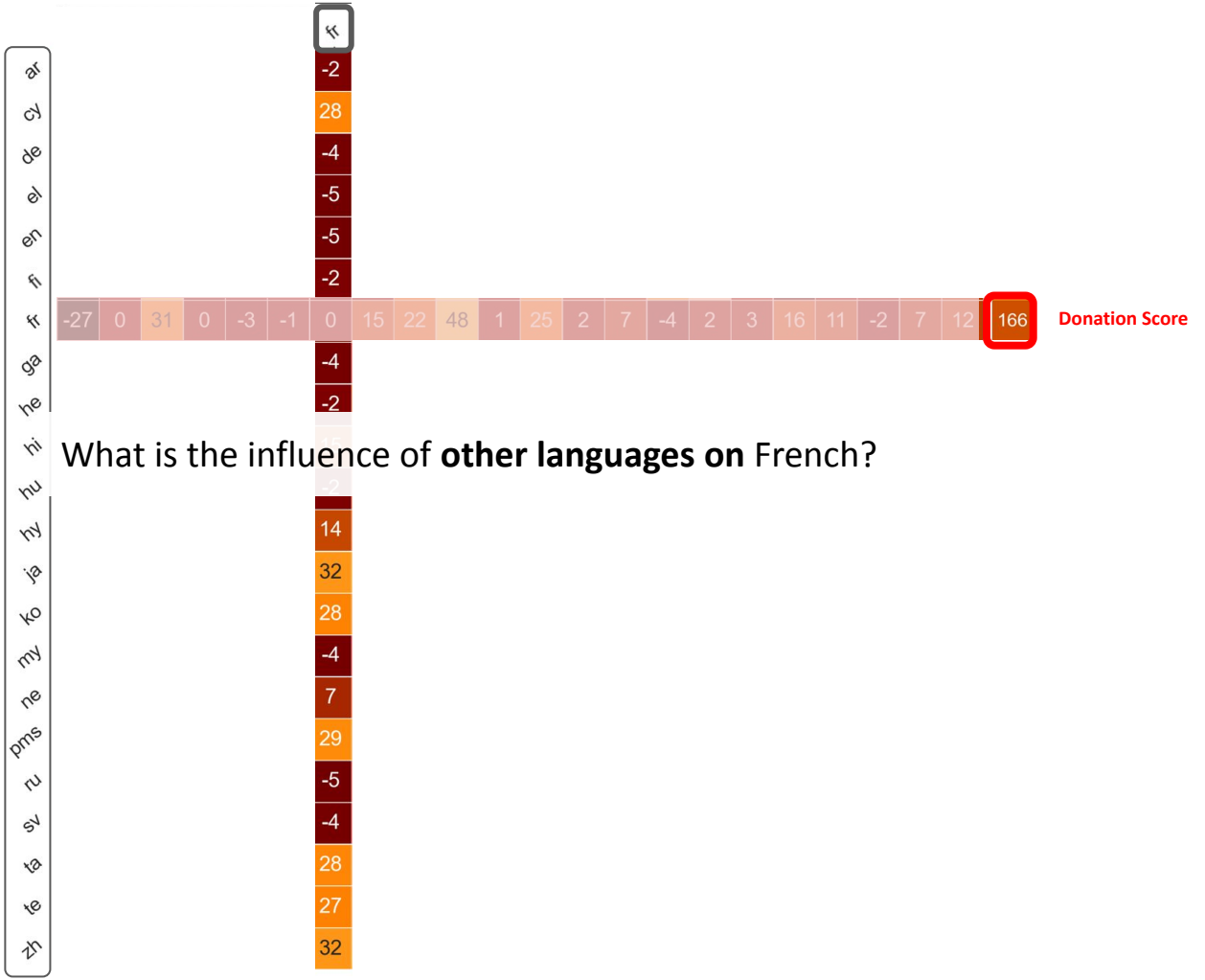


What is the influence of **other languages on** French?

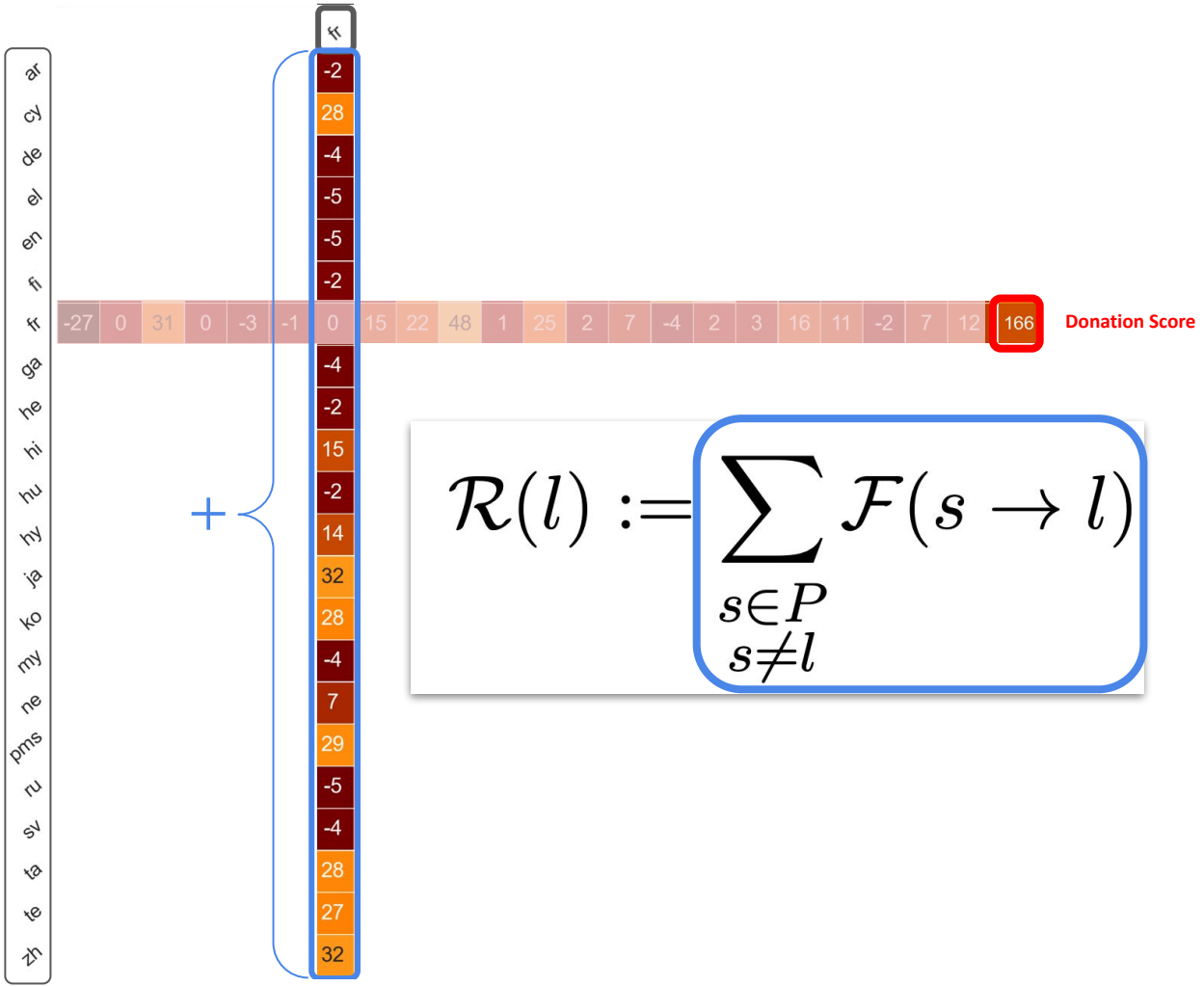
$$\mathcal{F}(\textbf{en}\rightarrow\textbf{fr}) := \frac{\varepsilon(M^{s,t},t) - \varepsilon(M^{t},t)}{\varepsilon(M^{t},t)}$$

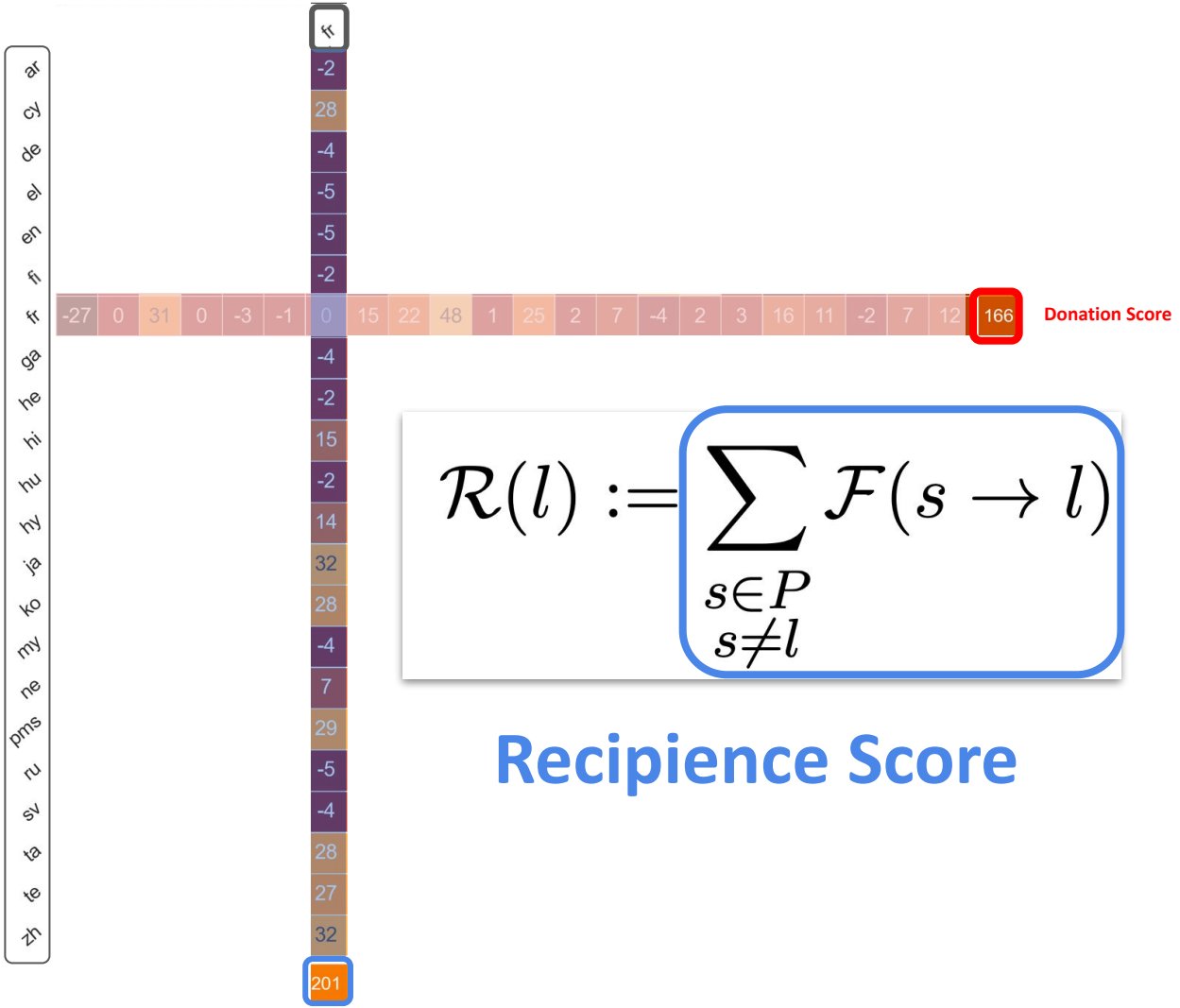


What is the influence of **other languages on** French?



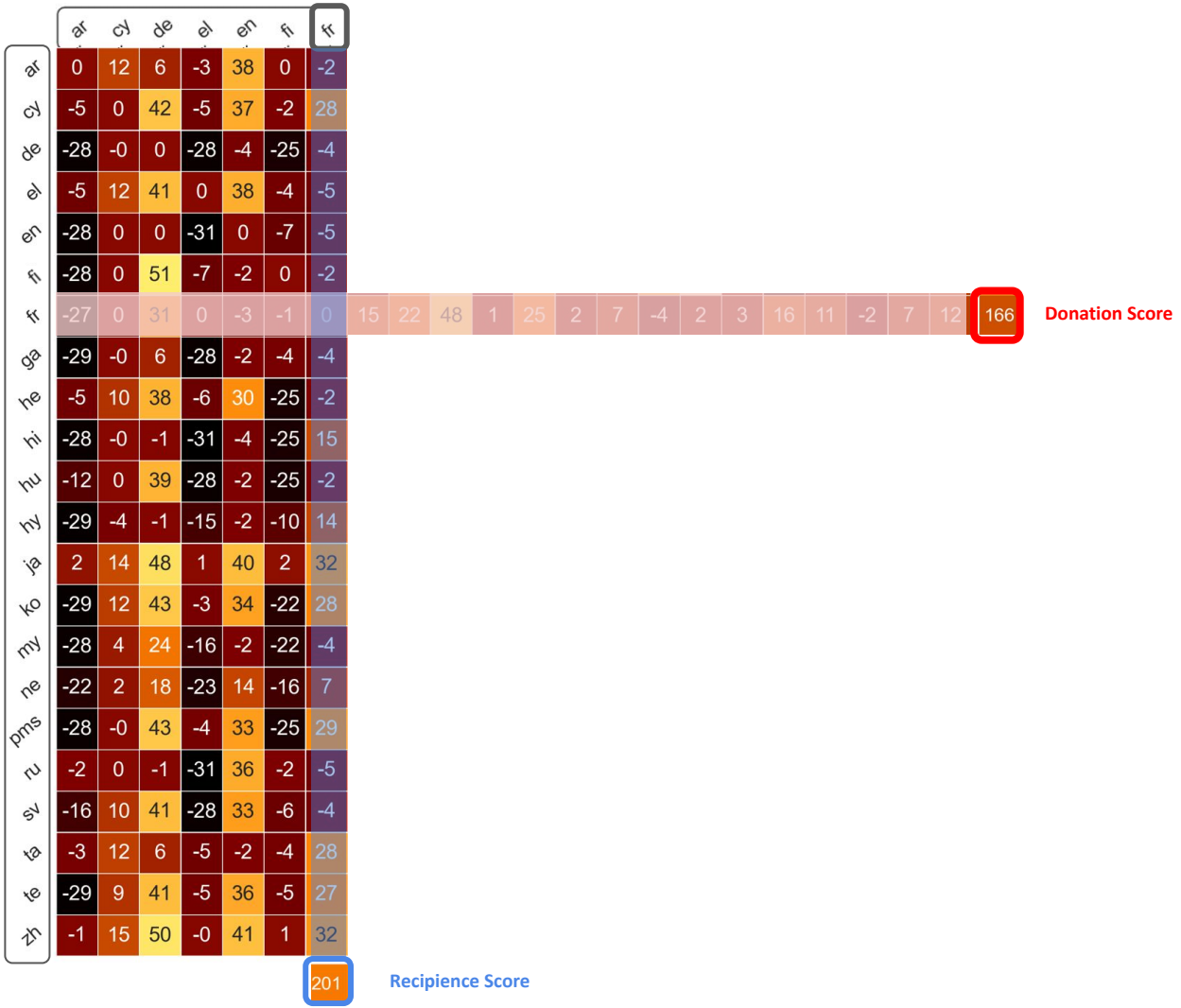
What is the influence of **other languages on** French?

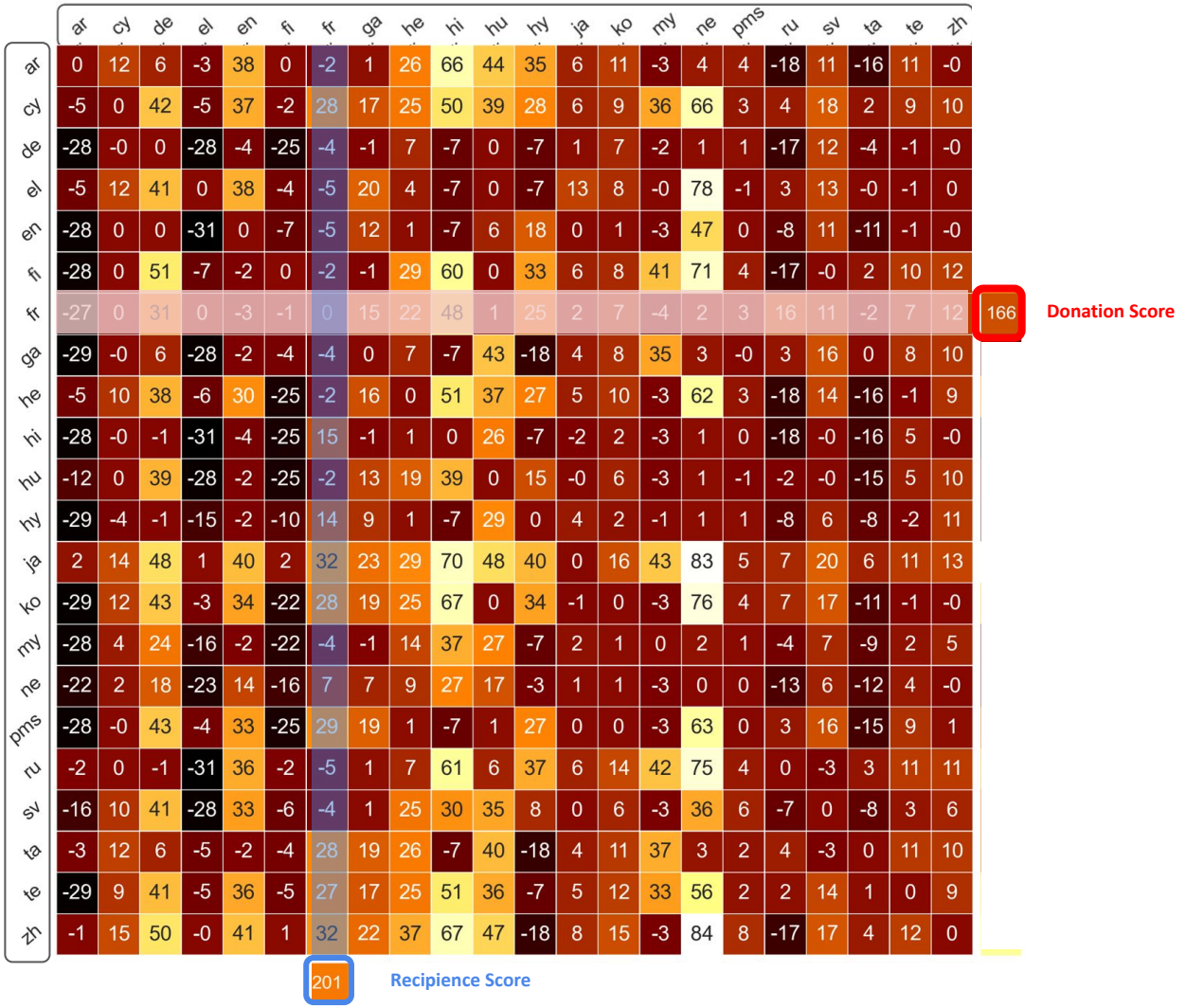

ar
cy
de
el
en
fi
fr
ga
he
hi
hu
hy
ja
ko
my
ne
pms
ru
sv
ta
te
zh
-2
28
-4
-5
-5
-2
-4
-2
15
-2
14
32
28
-4
7
29
-5
-4
28
27
32
-27 0 31 0 -3 -1 0 15 22 48 1 25 2 7 -4 2 3 16 11 -2 7 12 166
Donation Score
$$\mathcal{R}(l) := \sum_{\substack{s \in P \\ s \neq l}} \mathcal{F}(s \to l)$$

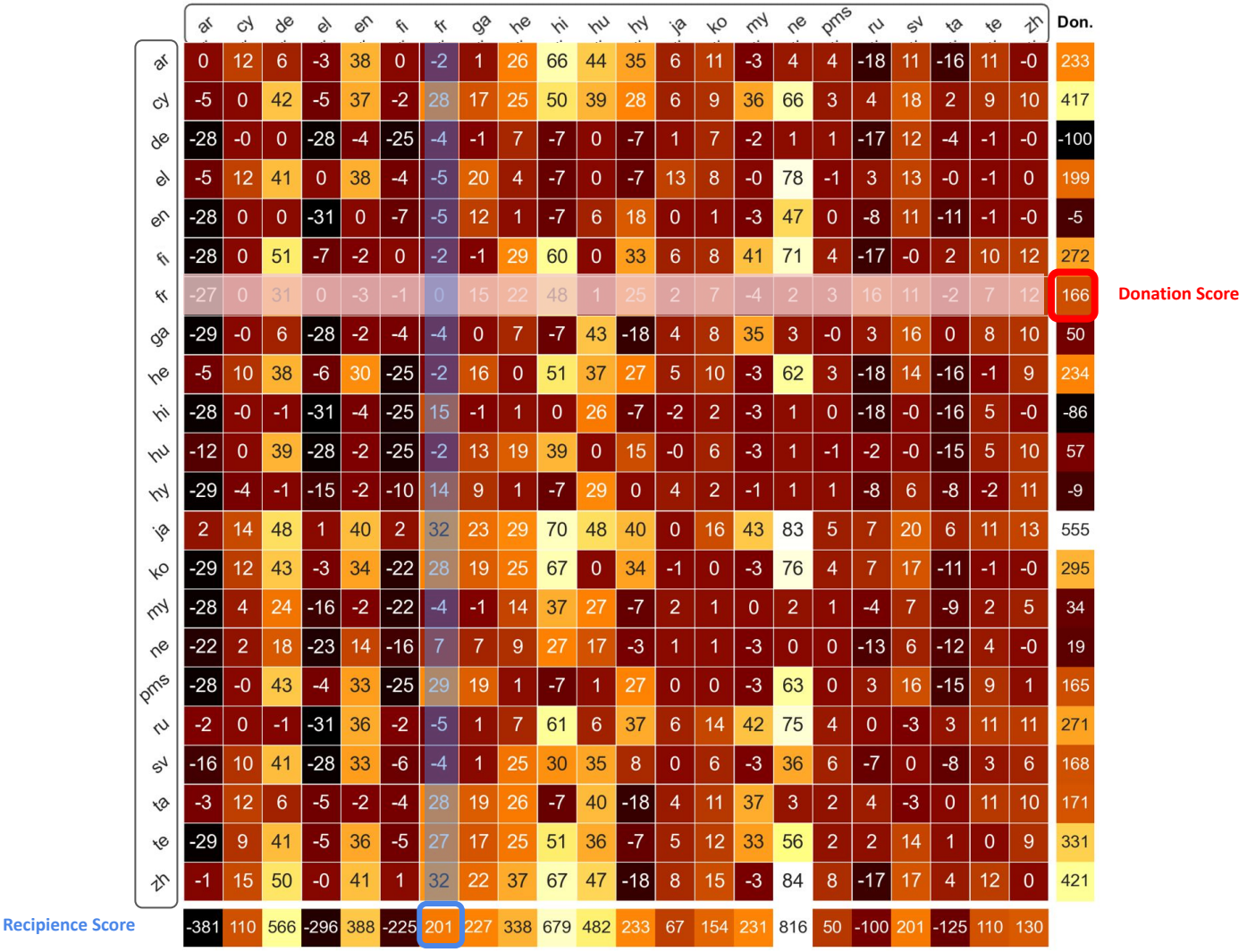


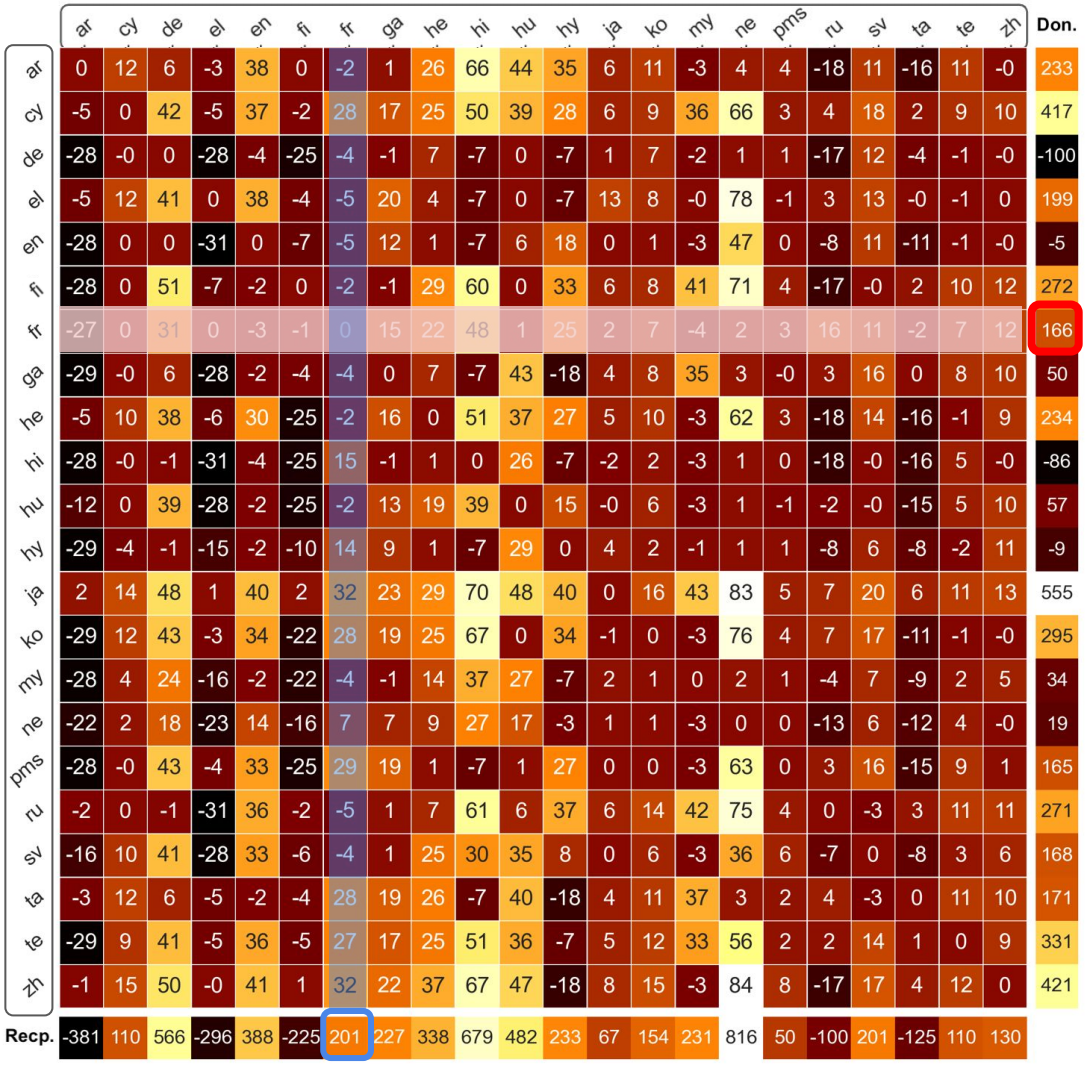
$$\mathcal{R}(l) := \sum_{\substack{s \in P \\ s \neq l}} \mathcal{F}(s \to l)$$

Recipience Score


ar cy de el en fi fr
ar 0 12 6 -3 38 0 -2
cy -5 0 42 -5 37 -2 28
de -28 -0 0 -28 -4 -25 -4
el -5 12 41 0 38 -4 -5
en -28 0 0 -31 0 -7 -5
fi -28 0 51 -7 -2 0 -2
fr -27 0 31 0 -3 -1 0 15 22 48 1 25 2 7 -4 2 3 16 11 -2 7 12 166
ga -29 -0 6 -28 -2 -4 -4
he -5 10 38 -6 30 -25 -2
hi -28 -0 -1 -31 -4 -25 15
hu -12 0 39 -28 -2 -25 -2
hy -29 -4 -1 -15 -2 -10 14
ja 2 14 48 1 40 2 32
ko -29 12 43 -3 34 -22 28
my -28 4 24 -16 -2 -22 -4
ne -22 2 18 -23 14 -16 7
pms -28 -0 43 -4 33 -25 29
ru -2 0 -1 -31 36 -2 -5
sv -16 10 41 -28 33 -6 -4
ta -3 12 6 -5 -2 -4 28
te -29 9 41 -5 36 -5 27
zh -1 15 50 -0 41 1 32
201
Donation Score
Recipience Score


ar cy de el en fi fr ga he hi hu hy ja ko my ne pms ru sv ta te zh
ar 0 12 6 -3 38 0 -2 1 26 66 44 35 6 11 -3 4 4 -18 11 -16 11 -0
cy -5 0 42 -5 37 -2 28 17 25 50 39 28 6 9 36 66 3 4 18 2 9 10
de -28 -0 0 -28 -4 -25 -4 -1 7 -7 0 -7 1 7 -2 1 1 -17 12 -4 -1 -0
el -5 12 41 0 38 -4 -5 20 4 -7 0 -7 13 8 -0 78 -1 3 13 -0 -1 0
en -28 0 0 -31 0 -7 -5 12 1 -7 6 18 0 1 -3 47 0 -8 11 -11 -1 -0
fi -28 0 51 -7 -2 0 -2 -1 29 60 0 33 6 8 41 71 4 -17 -0 2 10 12
fr -27 0 31 0 -3 -1 0 15 22 48 1 25 2 7 -4 2 3 16 11 -2 7 12
166
Donation Score
ga -29 -0 6 -28 -2 -4 -4 0 7 -7 43 -18 4 8 35 3 -0 3 16 0 8 10
he -5 10 38 -6 30 -25 -2 16 0 51 37 27 5 10 -3 62 3 -18 14 -16 -1 9
hi -28 -0 -1 -31 -4 -25 15 -1 1 0 26 -7 -2 2 -3 1 0 -18 -0 -16 5 -0
hu -12 0 39 -28 -2 -25 -2 13 19 39 0 15 -0 6 -3 1 -1 -2 -0 -15 5 10
hy -29 -4 -1 -15 -2 -10 14 9 1 -7 29 0 4 2 -1 1 1 -8 6 -8 -2 11
ja 2 14 48 1 40 2 32 23 29 70 48 40 0 16 43 83 5 7 20 6 11 13
ko -29 12 43 -3 34 -22 28 19 25 67 0 34 -1 0 -3 76 4 7 17 -11 -1 -0
my -28 4 24 -16 -2 -22 -4 -1 14 37 27 -7 2 1 0 2 1 -4 7 -9 2 5
ne -22 2 18 -23 14 -16 7 7 9 27 17 -3 1 1 -3 0 0 -13 6 -12 4 -0
pms -28 -0 43 -4 33 -25 29 19 1 -7 1 27 0 0 -3 63 0 3 16 -15 9 1
ru -2 0 -1 -31 36 -2 -5 1 7 61 6 37 6 14 42 75 4 0 -3 3 11 11
sv -16 10 41 -28 33 -6 -4 1 25 30 35 8 0 6 -3 36 6 -7 0 -8 3 6
ta -3 12 6 -5 -2 -4 28 19 26 -7 40 -18 4 11 37 3 2 4 -3 0 11 10
te -29 9 41 -5 36 -5 27 17 25 51 36 -7 5 12 33 56 2 2 14 1 0 9
zh -1 15 50 -0 41 1 32 22 37 67 47 -18 8 15 -3 84 8 -17 17 4 12 0
201
Recipience Score

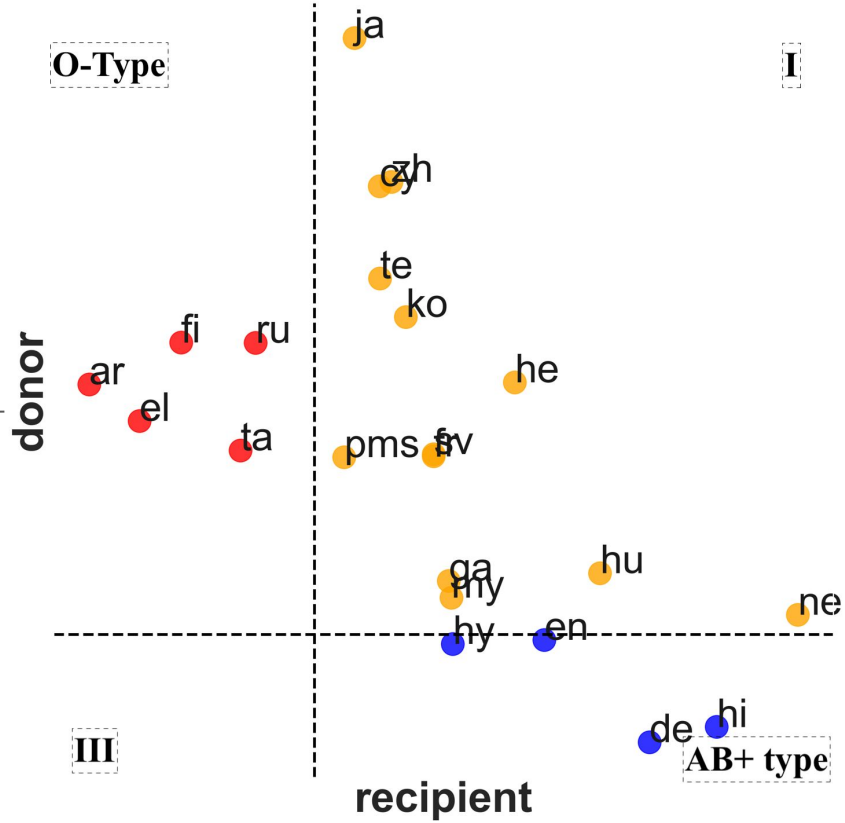

ar cy de el en fi fr ga he hi hu hy ja ko my ne pms ru sv ta te zh Don.
ar 0 12 6 -3 38 0 -2 1 26 66 44 35 6 11 -3 4 4 -18 11 -16 11 -0 233
cy -5 0 42 -5 37 -2 28 17 25 50 39 28 6 9 36 66 3 4 18 2 9 10 417
de -28 -0 0 -28 -4 -25 -4 -1 7 -7 0 -7 1 7 -2 1 1 -17 12 -4 -1 -0 -100
el -5 12 41 0 38 -4 -5 20 4 -7 0 -7 13 8 -0 78 -1 3 13 -0 -1 0 199
en -28 0 0 -31 0 -7 -5 12 1 -7 6 18 0 1 -3 47 0 -8 11 -11 -1 -0 -5
fi -28 0 51 -7 -2 0 -2 -1 29 60 0 33 6 8 41 71 4 -17 -0 2 10 12 272
fr -27 0 31 0 -3 -1 0 15 22 48 1 25 2 7 -4 2 3 16 11 -2 7 12 166
ga -29 -0 6 -28 -2 -4 -4 0 7 -7 43 -18 4 8 35 3 -0 3 16 0 8 10 50
he -5 10 38 -6 30 -25 -2 16 0 51 37 27 5 10 -3 62 3 -18 14 -16 -1 9 234
hi -28 -0 -1 -31 -4 -25 15 -1 1 0 26 -7 -2 2 -3 1 0 -18 -0 -16 5 -0 -86
hu -12 0 39 -28 -2 -25 -2 13 19 39 0 15 -0 6 -3 1 -1 -2 -0 -15 5 10 57
hy -29 -4 -1 -15 -2 -10 14 9 1 -7 29 0 4 2 -1 1 1 -8 6 -8 -2 11 -9
ja 2 14 48 1 40 2 32 23 29 70 48 40 0 16 43 83 5 7 20 6 11 13 555
ko -29 12 43 -3 34 -22 28 19 25 67 0 34 -1 0 -3 76 4 7 17 -11 -1 -0 295
my -28 4 24 -16 -2 -22 -4 -1 14 37 27 -7 2 1 0 2 1 -4 7 -9 2 5 34
ne -22 2 18 -23 14 -16 7 7 9 27 17 -3 1 1 -3 0 0 -13 6 -12 4 -0 19
pms -28 -0 43 -4 33 -25 29 19 1 -7 1 27 0 0 -3 63 0 3 16 -15 9 1 165
ru -2 0 -1 -31 36 -2 -5 1 7 61 6 37 6 14 42 75 4 0 -3 3 11 11 271
sv -16 10 41 -28 33 -6 -4 1 25 30 35 8 0 6 -3 36 6 -7 0 -8 3 6 168
ta -3 12 6 -5 -2 -4 28 19 26 -7 40 -18 4 11 37 3 2 4 -3 0 11 10 171
te -29 9 41 -5 36 -5 27 17 25 51 36 -7 5 12 33 56 2 2 14 1 0 9 331
zh -1 15 50 -0 41 1 32 22 37 67 47 -18 8 15 -3 84 8 -17 17 4 12 0 421
-381 110 566 -296 388 -225 201 227 338 679 482 233 67 154 231 816 50 -100 201 -125 110 130
Donation Score
Recipience Score



How much a language overall ***donates*** *to* other languages

How much a language overall ***receives*** from other languages

A Linguistic Blood Bank



Languages which tend to ***donate*** but **do not receive**

Languages which tend to ***receive*** but **do not donate**

Hard to Draw General Conclusions

- Multilingual analysis often involves many **axes to evaluate**
  - Language family
  - Script
  - Domain
  - Size
  - …

- I think that **we still don’t understand** it well enough

Conclusions

- Broad **multilingual NLP** is important for various reasons
  - Linguistic, ML, cultural, and cognitive
- NLP is focused on a **small subset of languages**
- **Benchmarks in low-resource languages** enables NLP development
  - But require effort and expert knowledge
- **Cross-lingual transfer** facilitates adoption across languages
  - Still lots to understand in how it works