Harnessing Multilingual Models for Ancient Language Processing

Gabriel Stanovsky

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY of JERUSALEM
الجامعة العبرية في القدس

ALP2023

Disclaimer: I’m not an Historian / Epigraph / Philologist

- I’m an NLP researcher interested in real-world applications
  - Medicine
  - Law
  - **Archeology and ancient languages**

- **What challenges do they raise?**
  - **Low resource**, extinct languages

- **What computational and linguistic observations can we draw?**
  - **Multilingual** and few or zero shot transfer

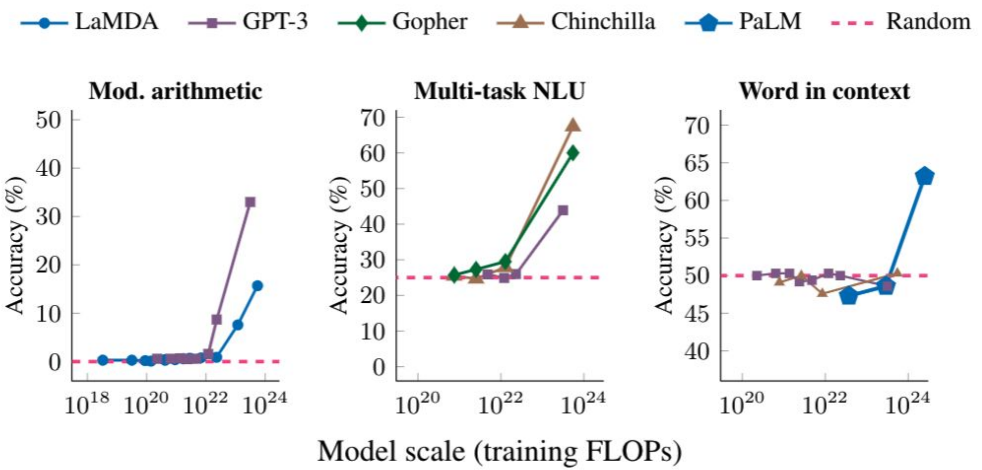
Large Language Models (LLMs)

- Trained to to predict the next word in **naturally occurring** texts
  - News reports
  - Blogs
  - Medical texts

- Form the **foundation** for most NLP models for high-resource languages
  - Text classification, named entity recognition, sentiment analysis, author attribution, dating …

- These tasks are **relevant for ancient language processing**

LLMs for Ancient Language Processing?



- LLMs require large scale data, yet **ancient language data is limited**
- **How can we still leverage LLMs?**

blog.research.google/2022/11/characterizing-emergent-phenomena-in

Agenda: Harnessing **Multilingual Signal**

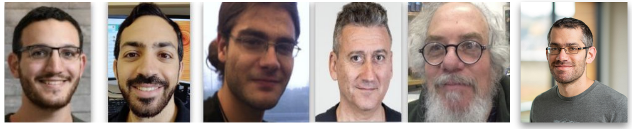
- State-of-the-art language modelling in Akkadian (EMNLP 2021)
  - By adding signal from 100 different languages
- Selective language combinations improves performance (NAACL 2022)
  - Mapping the linguistic blood bank
- Speculative recipe for future work
  - Train multilingual LLM with a downstream objective in mind

Agenda: Harnessing **Multilingual Signal**

- **State-of-the-art language modelling in Akkadian**
  - **By adding signal from 100 different languages**

- Selective language combinations improves performance (NAACL 2022)
  - Mapping the linguistic blood bank

- Speculative recipe for future work
  - Train multilingual LLM with a downstream objective in mind

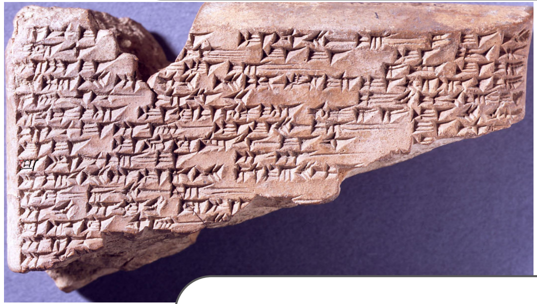


Lazar et al., EMNLP 2021

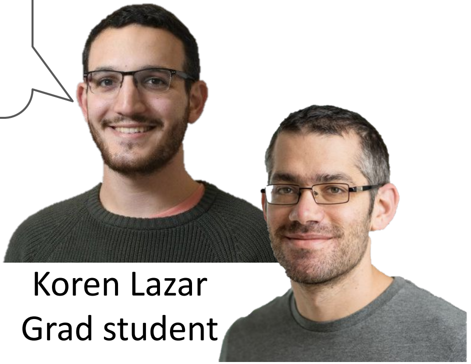
When transcribing ancient tablets found in archeological sites we need to **fill in gaps** formed in the clay due to **erosion over 1000s of years**

Prof. Wayne Horowitz

Prof. Nathan Wasserman



and **how do you know** how to fill in those missing parts?



Koren Lazar
Grad student

me

Well, we look at the symbols we recognize in the surrounding context, and try to **guess** the most **probable sequence**

The Institute of Archaeology, Hebrew University

That sounds awfully familiar...

Data Size: Low-Resource Setting

	# Texts	# Words	# Signs
Akkadian Train	8K	950K	1.8M
Akkadian Test	2K	250K	500K
English Train	7K	950K	–
English Test	2K	250K	–

Data Size: Low-Resource Setting (ORACC corpus)

	# Texts	# Words	# Signs
Akkadian Train	8K	950K	1.8M
Akkadian Test	2K	250K	500K
English Train	7K	950K	–
English Test	2K	250K	–

PHILBERTA trained on 185M tokens
BERT-base on 3.3B words
ChatGPT on 300B

From Cuneiform to Latin Transliteration

From Cuneiform to Latin Transliteration

From Cuneiform to Latin Transliteration

Digitization

From Cuneiform to Latin Transliteration

Digitization

Latin transliteration

ša ina É.GAL *áš-pur-an-ni*

nu-uk ÍD-MEŠ *lu-⌜x⌝ [x]-⌜ru*⌝*

From Cuneiform to Latin Transliteration

Digitization

Latin transliteration

ša ina É.GAL *áš-pur-an-ni*
nu-uk ÍD-MEŠ *lu*-⸢*x*⸣ [*x*]-⸢*ru**⸣

Fill missing signs based on context

ša ina É.GAL *áš-pur-an-ni*
nu-uk ÍD-MEŠ *lu*-⸢a⸣-[UGU]-⸢*ru**⸣

From Cuneiform to Latin Transliteration

Digitization

Latin transliteration

ša ina É.GAL *áš-pur-an-ni*
nu-uk ÍD-MEŠ *lu-*⸢*x*⸣ [*x*]-⸢*ru**⸣

Predicting missing signs

ša ina É.GAL *áš-pur-an-ni*
nu-uk ÍD-MEŠ *lu-*⸢a⸣-[UGU]-⸢*ru**⸣

Task Definition

- **Input**:

  ša ina É.GAL áš-pur-an-ni
  nu-uk ÍD-MEŠ lu-⸢x⸣ [x]-⸢ru*⸣

Task Definition

- **Input**:

  ša ina É.GAL *áš-pur-an-ni*
  nu-uk ÍD-MEŠ lu-⸢x⸣ [x]-⸢ru*⸣

- **Assumption**: Number of missing signs is estimated by a human editor

Task Definition

- **Input**:

  ša ina É.GAL *áš-pur-an-ni*
  nu-uk ÍD-MEŠ *lu*-⸢*x*⸣ [*x*]-⸢*ru**⸣

- **Assumption**: Number of missing signs is estimated by a human editor

- **Output**:

  ša ina É.GAL *áš-pur-an-ni*
  nu-uk ÍD-MEŠ *lu*-⸢a⸣-[UGU]-⸢*ru**⸣

LLM Results on Akkadian

Genre	LSTM (Fetaya et al, 2020)	Akkadian Transformer
Royal Inscription	52%	57%
Royal or Monumental	51%	61%
Astrological Report	53%	55%
Lexical	10%	69%
Decree	49%	39%
Overall	**52%**	**50%**

LLM Results on Akkadian

Genre	LSTM (Fetaya et al, 2020)	Akkadian Transformer
Royal Inscription	52%	57%
Royal or Monumental	51%	61%
Astrological Report	53%	55%
Lexical	10%	69%
Decree	49%	39%
Overall	**52%**	**50%**

Similar results between transformers and LSTM

Adding Multilingual Signal with Multilingual BERT

- Finetune Akkadian together **with 100 popular languages from Wikipedia**

Genre	LSTM (Fetaya et al, 2020)	Akkadian Transformer	Multilingual Akkadian Transformer
Royal Inscription	52%	57%	83%
Royal or Monumental	51%	61%	84%
Astrological Report	53%	55%	81%
Lexical	10%	69%	69%
Decree	49%	39%	71%
Overall	**52%**	**50%**	**83%**

A Quick Aside: Human Evaluation

Scheme for Manual Evaluation: Desiderata

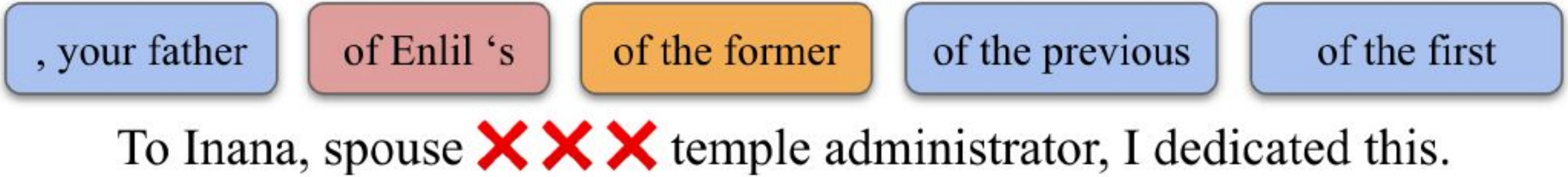
- Allow for **multiple correct** predictions
- Account for **inherent noise** in estimation
- Account for the annotators being **non-native Akkadian**

Manual Evaluation Scheme

, your father | of Enlil 's | of the former | of the previous | of the first

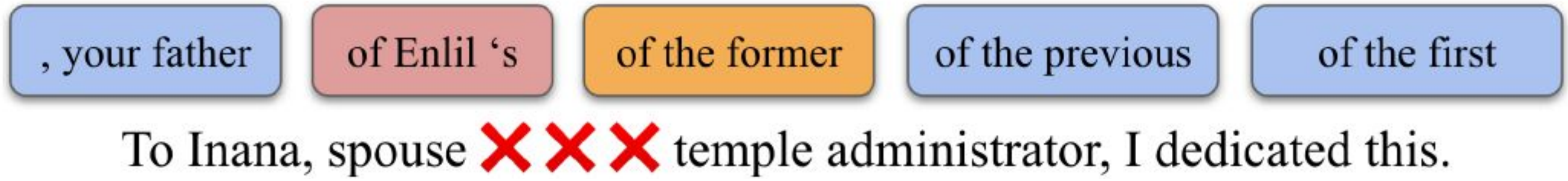
To Inana, spouse temple administrator, I dedicated this.

Manual Evaluation Scheme: Behind the Scenes

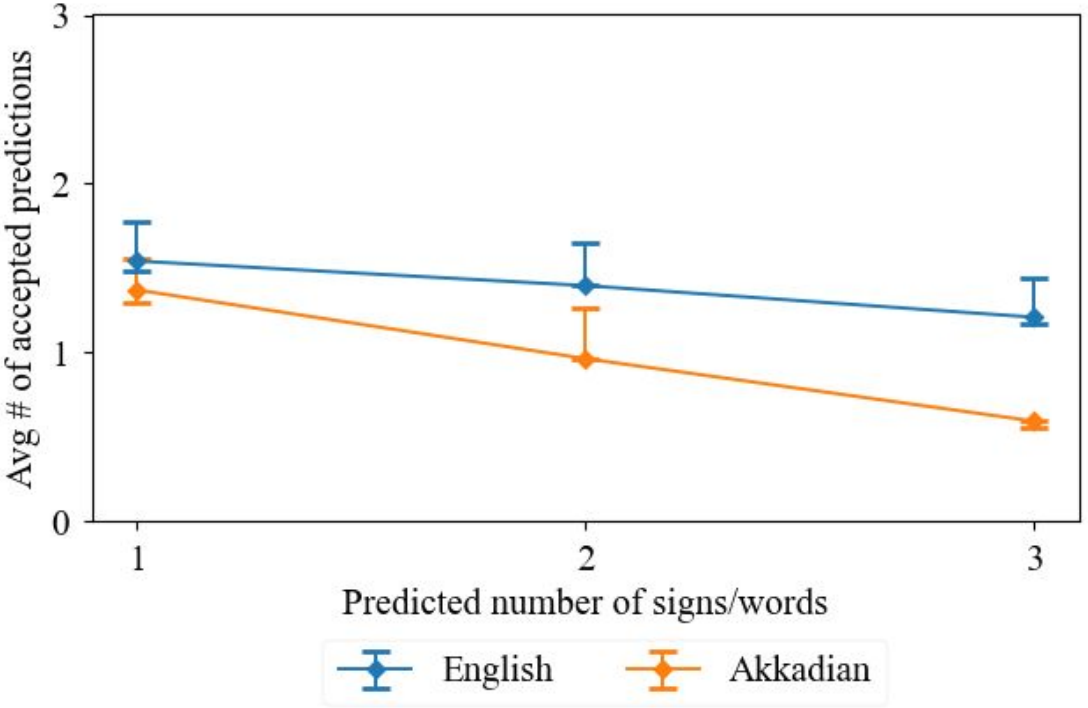


Manual Evaluation Scheme: Behind the Scenes

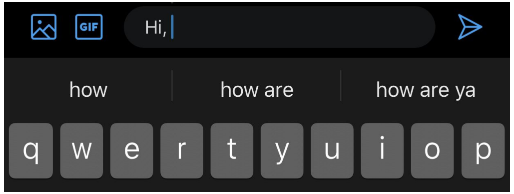
model prediction **noise** **gold** **model prediction** **model prediction**

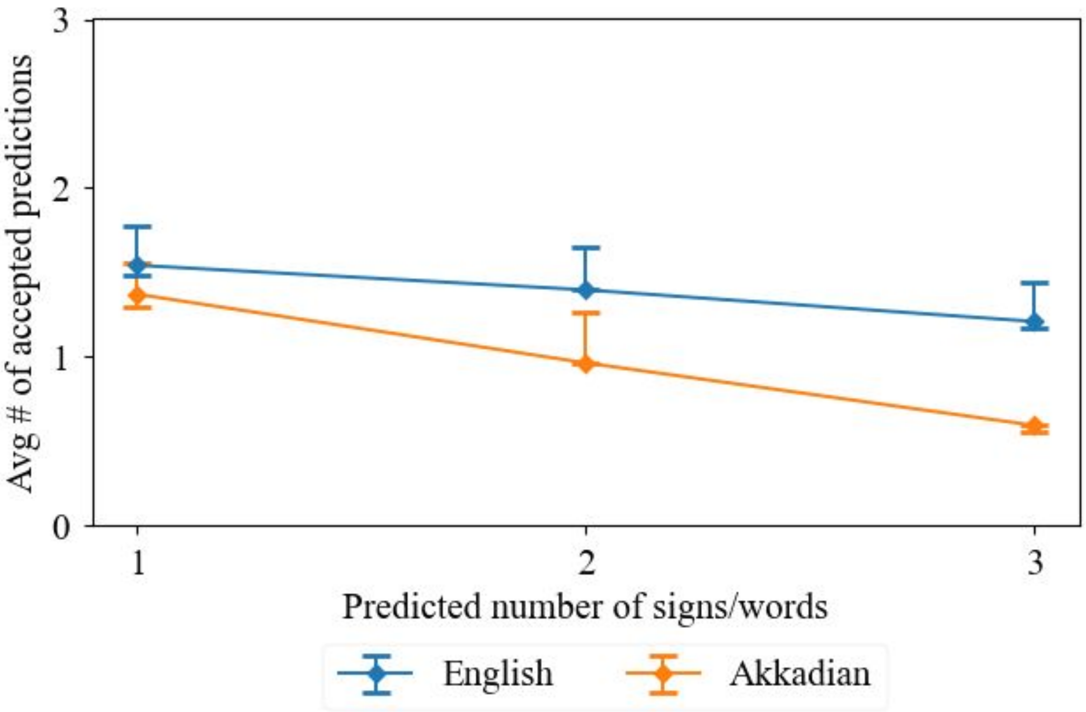


Manual Evaluation Results



Manual Evaluation Results





Adding Multilingual Signal with Multilingual BERT

	LSTM (Fetaya et al, 2020)	Akkadian Transformer	Multilingual Akkadian Transformer
Overall	52%	50%	83%

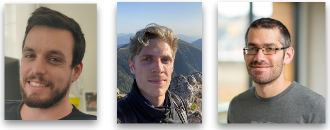
- Adding **modern languages** to training vastly improves results
  - Producing an **Akkadian LLM** which can benefit various downstream tasks
- **Why is this happening?**
  - Perhaps due to related languages in Wikipedia? (Hebrew, Arabic, etc.)
- **Can selective language choice further improve results?**

Agenda: Harnessing **Multilingual Signal**

- State-of-the-art language modelling in Akkadian
  - By adding signal from 100 different languages

- **Selective language combinations improves performance (NAACL 2022)**
  - **Mapping the linguistic blood bank**

- Speculative recipe for future work
  - Train multilingual LLM with a downstream objective in mind



Malkin et al.
🎉 **Outstanding Paper @ NAACL 2022**

Zero-Shot Pretraining Language Graph

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

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

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

Performance of a monolingual model on t

Zero-Shot Pretraining Language Graph

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

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

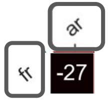
$$\mathcal{F}(s \to t) := \frac{\boxed{\varepsilon(M^{s,t}, t)} - \boxed{\varepsilon(M^{t}, t)}}{\boxed{\varepsilon(M^{t}, t)}}$$

Performance of a monolingual model on t

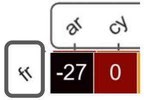
- In other words, ***F* measures how much *t* gains from *s*** ar

fr

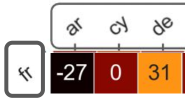
$$\mathcal{F}(\mathrm{fr} \rightarrow \mathrm{ar}) := \frac{\varepsilon(\boxed{23.4}\, t) - \varepsilon(\boxed{32.1}\, t)}{\varepsilon(\boxed{32.1}\, t)}$$



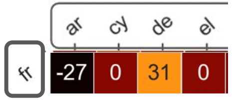
$$\mathcal{F}(\text{fr} \rightarrow \text{ar}) := \frac{\varepsilon(\text{23.4}, t) - \varepsilon(\text{32.1}\ t)}{\varepsilon(\text{32.1}\ t)}$$



$$\mathcal{F}(\text{fr} \rightarrow \text{cy}) := \frac{\varepsilon(1\ 39.9\ , t) - \varepsilon(\ 39.89\ t)}{\varepsilon(\ 39.89\ t)}$$



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



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

	ar	cy	de	el	en
fr	-27	0	31	0	-3

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

	ar	cy	de	el	en	fi
fr	-27	0	31	0	-3	-1

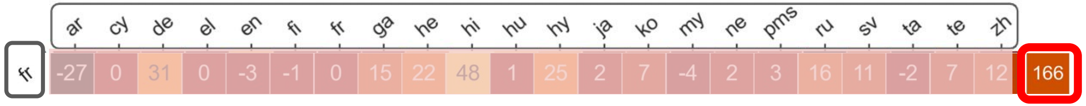
$$\mathcal{F}(\textbf{fr} \rightarrow \textbf{fi}) := \frac{\varepsilon(M^{\textbf{fr},t}, t) - \varepsilon(M^{t}, t)}{\varepsilon(M^{t}, t)}$$

	ar	cy	de	el	en	fi	fr	ga	he	hi	hu	hy	ja	ko	my	ne	pms	ru	sv	ta	te	zh
fr	-27	0	31	0	-3	-1	0	15	22	48	1	25	2	7	-4	2	3	16	11	-2	7	12

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

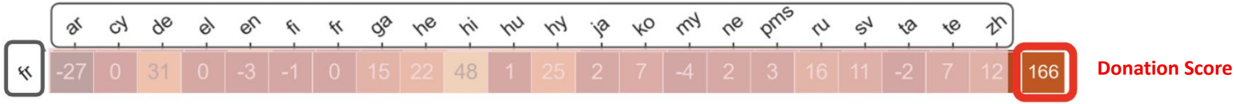
	ar	cy	de	el	en	fi	fr	ga	he	hi	hu	hy	ja	ko	my	ne	pms	ru	sv	ta	te	zh
fr	-27	0	31	0	-3	-1	0	15	22	48	1	25	2	7	-4	2	3	16	11	-2	7	12

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



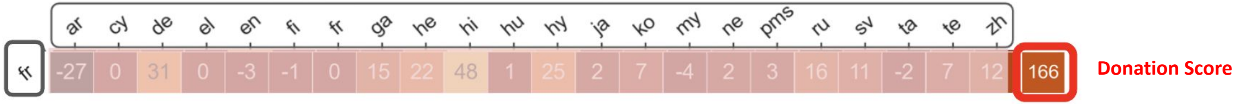
$$\mathcal{D}(l) := \boxed{\sum_{\substack{t \in P \\ t \neq l}} \mathcal{F}(l \rightarrow t)}$$

Donation Score



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

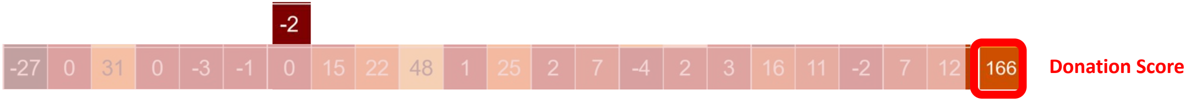
$$\mathcal{F}(\mathbf{fr} \rightarrow t)$$



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

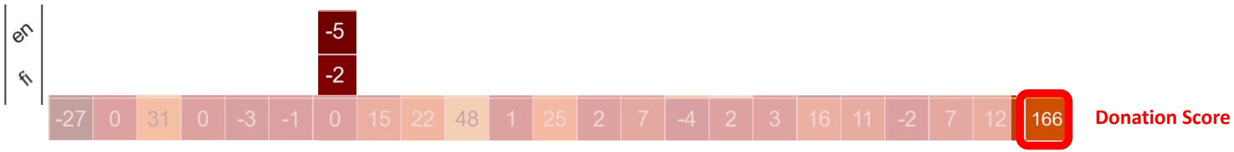
$$\mathcal{F}(s \rightarrow \textbf{fr})$$

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

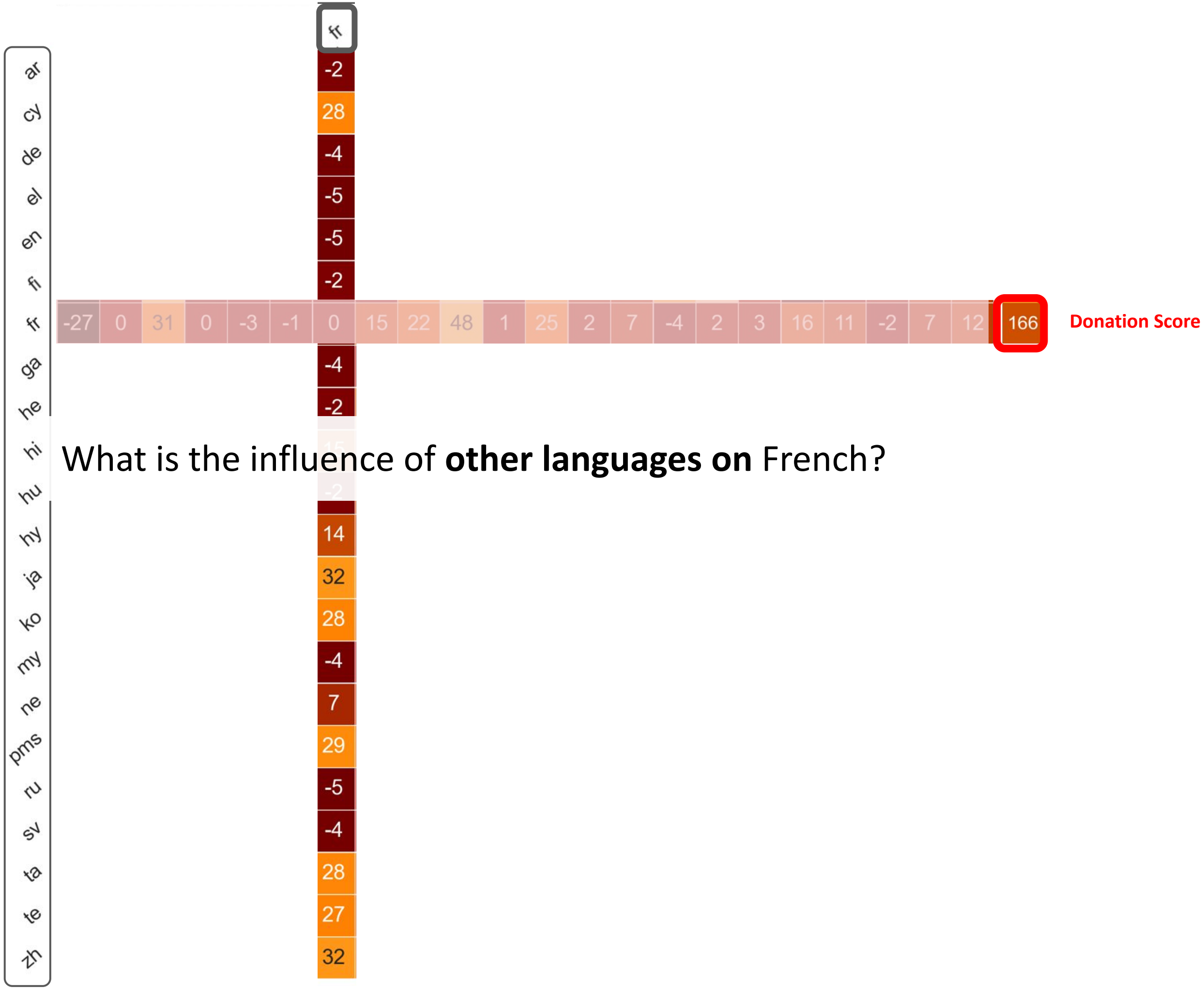


What is the influence of **other languages on** French?

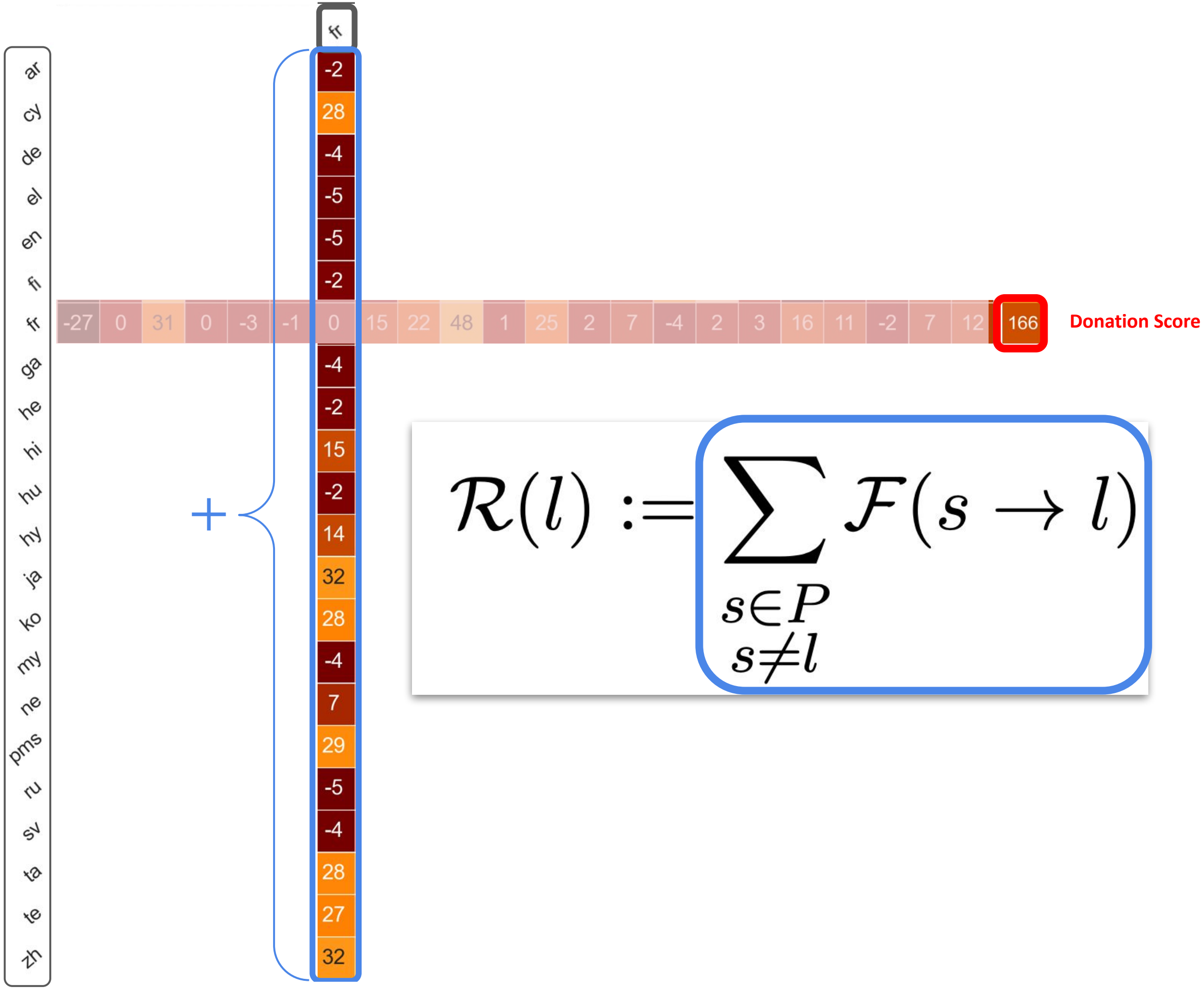
$$\mathcal{F}(\textbf{en}\rightarrow\textbf{fr}) := \frac{\varepsilon(M^{s,t},t) - \varepsilon(M^{t},t)}{\varepsilon(M^{t},t)}$$



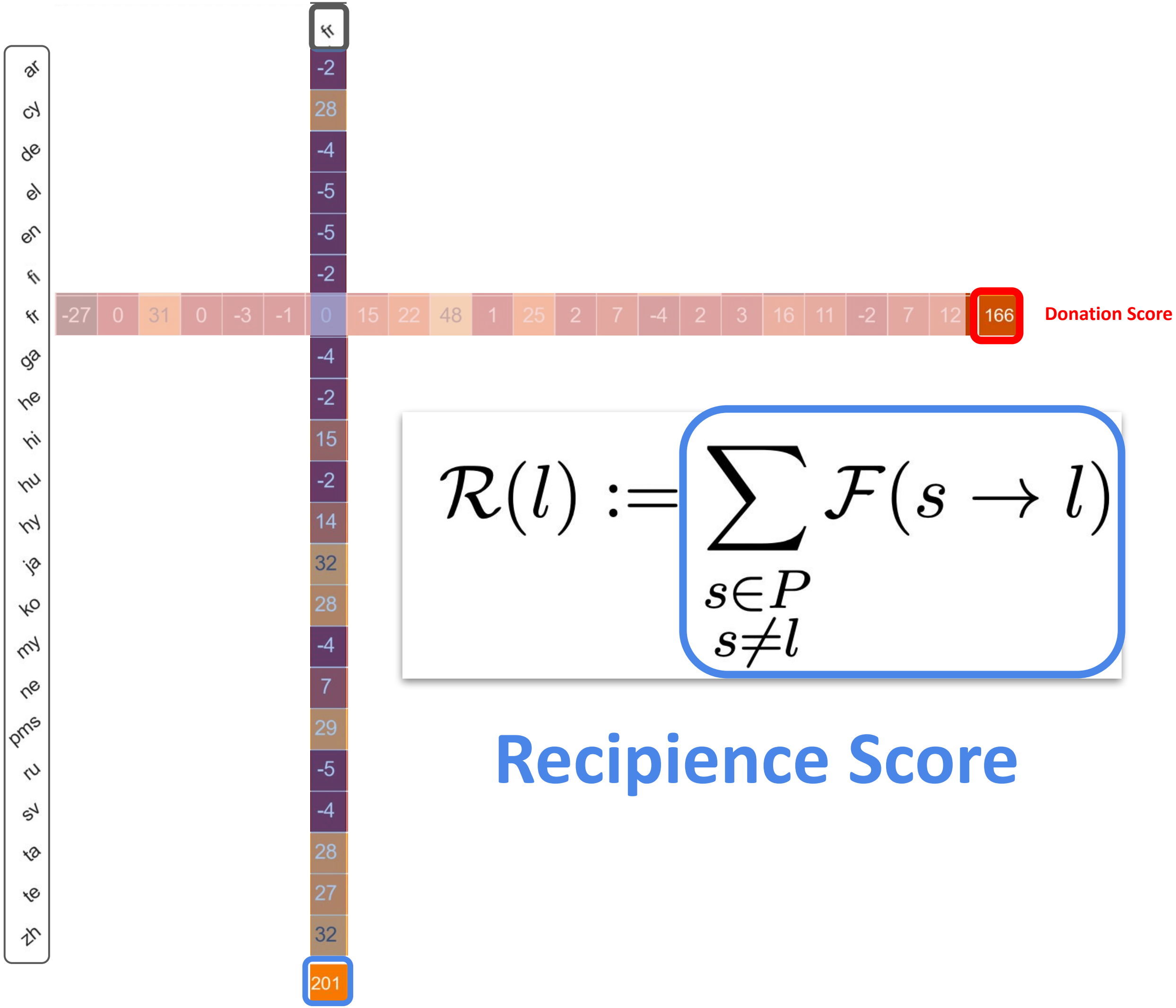
What is the influence of **other languages on** French?



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

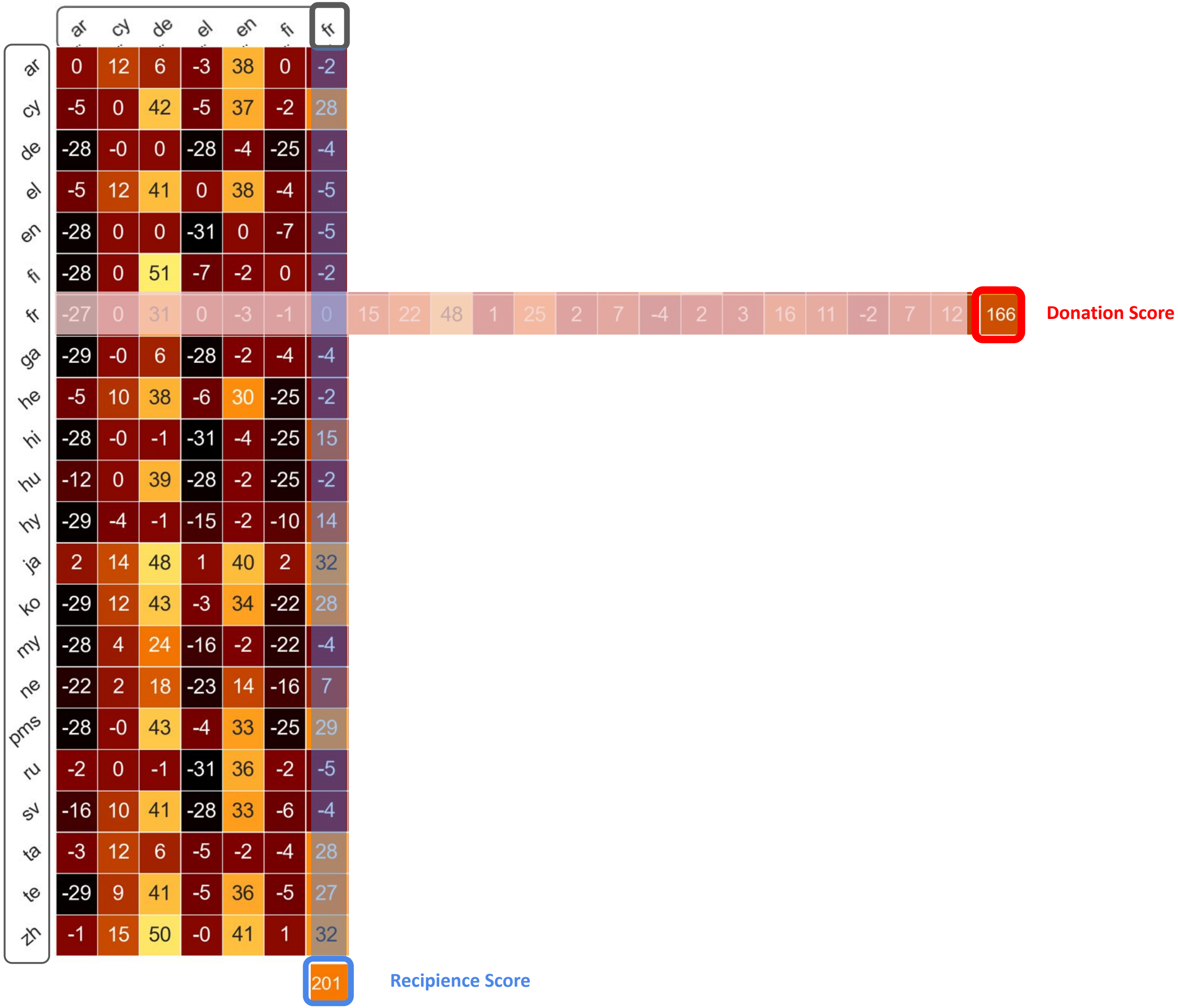


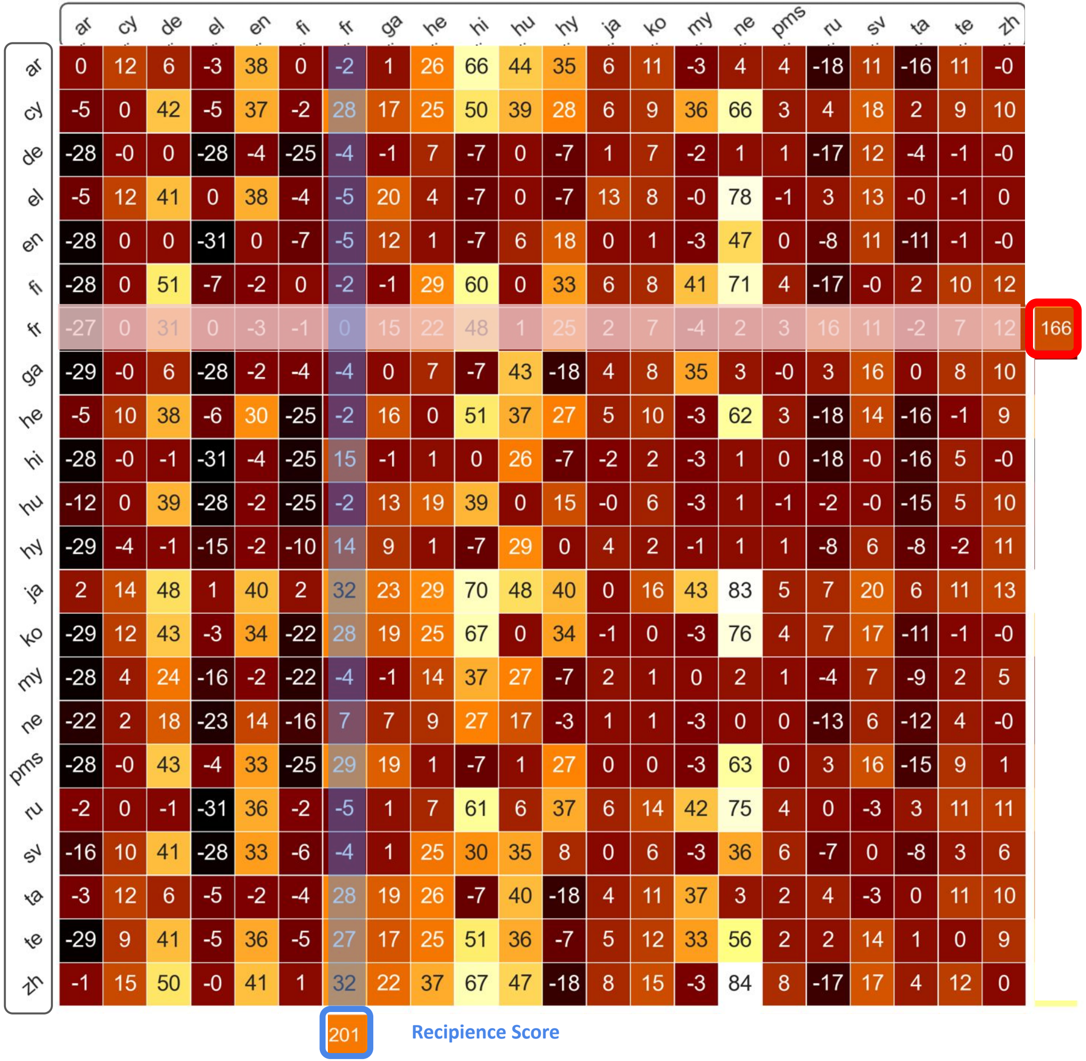
$$\mathcal{R}(l) := \boxed{\sum_{\substack{s \in P \\ s \neq l}} \mathcal{F}(s \to l)}$$



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

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
166
Donation Score
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
Recipience Score	-381	110	566	-296	388	-225	201	227	338	679	482	233	67	154	231	816	50	-100	201	-125	110	130	

Donation Score

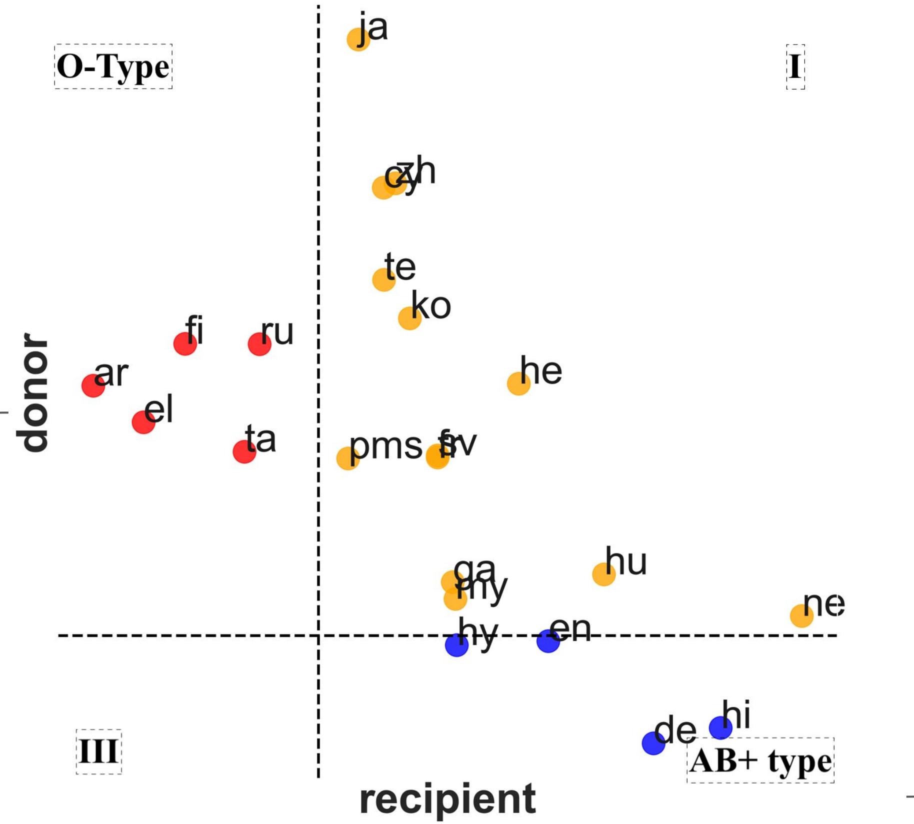
Recipience Score

	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
Recp.	-381	110	566	-296	388	-225	201	227	338	679	482	233	67	154	231	816	50	-100	201	-125	110	130	

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

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

A Linguistic Blood Bank

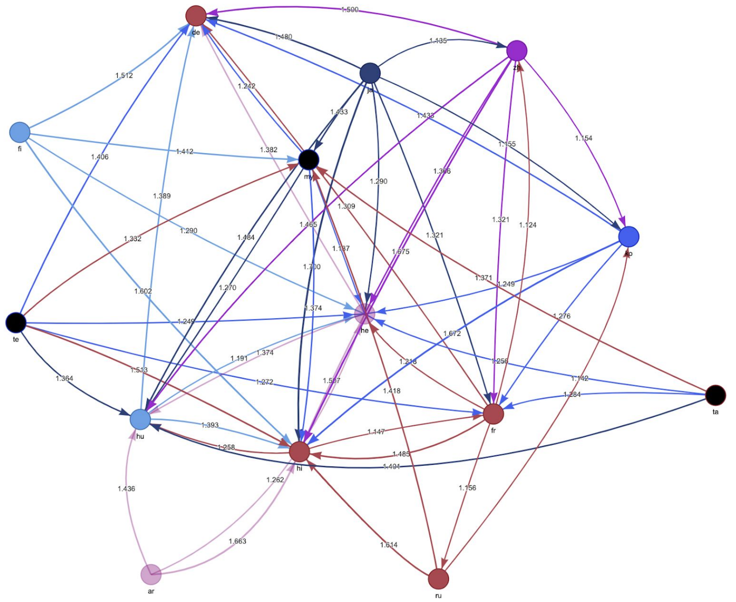


Languages which tend to ***donate*** but **do not receive**

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

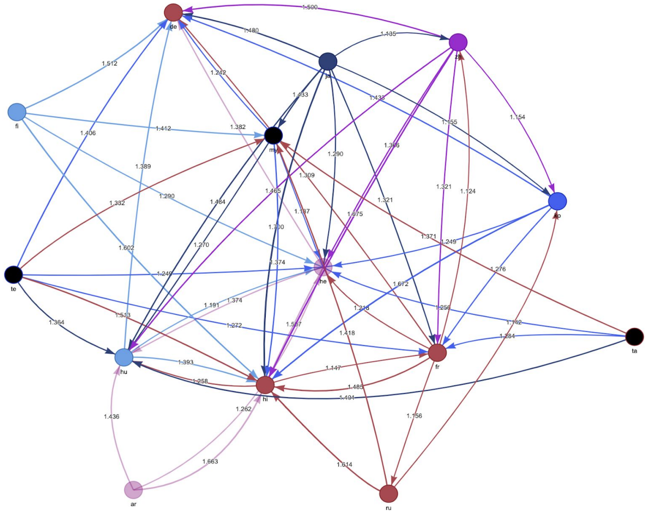
Pretrain Language Graph

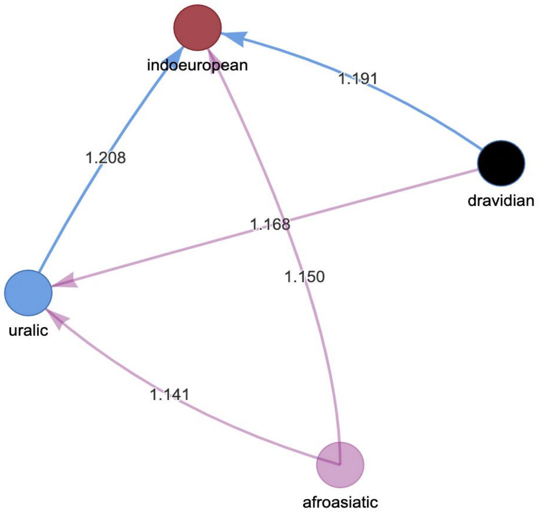
- Interpret this matrix as a weighted adjacency matrix
- Forms a complete, directed, weighted graph



How Language Properties Transfer?

- **Shared script leads to overall better transfer**
- **Shared language family didn’t have visible effect**





Similar Trends in Downstream Zero-Shot

- **Recipient languages perform better in NER and POS tagging**

	NER [$\%F_1$]		**POS** [$\%F_1$]	
	Avg. Monolingual	Avg. Zeroshot	Avg. Monolingual	Avg. Zeroshot
Most Recipient (R_h)	**50.3**$\pm.6$	**18.4**$\pm.6$	**64.1**$\pm.3$	**28.7**$\pm.7$
Least Recipient (R_l)	47.9$\pm.4$	12.4$\pm.4$	58.6$\pm.4$	26.0$\pm.7$

Agenda: Harnessing **Multilingual Signal**

- State-of-the-art language modelling in Akkadian (EMNLP 2021)
  - By adding signal from 100 different languages

- Selective language combinations improves performance (NAACL 2022)
  - Mapping the linguistic blood bank

- **Conclusion and Disucssion: Speculative recipe for future work**
  - **Train multilingual LLM with a downstream objective in mind**

Discussion: When Does Multilingual Signal Help?

- **Low resource** settings
  - Not enough signal to train monolingual LLM
- **Specialized pretraining** after massive multilingual pretraining
  - Maybe helps with the curse of multilinguality?
- **Careful language selection** for downstream tasks?

Conclusion: Recipe for Downstream Applications

1. Build a pretraining graph for your **target corpus and application**

2. Augment ancient training data with **most donating languages**

3. Finetune **multilingual LLMs** on downstream tasks

Conclusion: Recipe for Downstream Applications

1. Build a pretraining graph for your **target corpus and application**
2. Augment ancient training data with **most donating languages**
3. Finetune **multilingual LLMs** on downstream tasks

Thank you!

Code available at github.com/SLAB-NLP

Metrics

$$Hit@k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{[rank_i \le k]} \qquad MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$$

Chance MRR is 0.0001

I ate [MASK] for lunch.

Hit@1 = 0
Hit@2 = 1

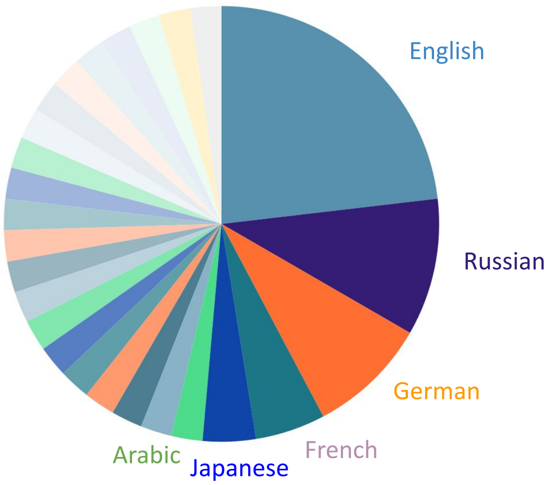
1	50%	hamburger
2	30%	**pasta**
3	10%	pizza
	...	
	...	

MRR=1/**2**

Confound: Unbalanced Corpus Size

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

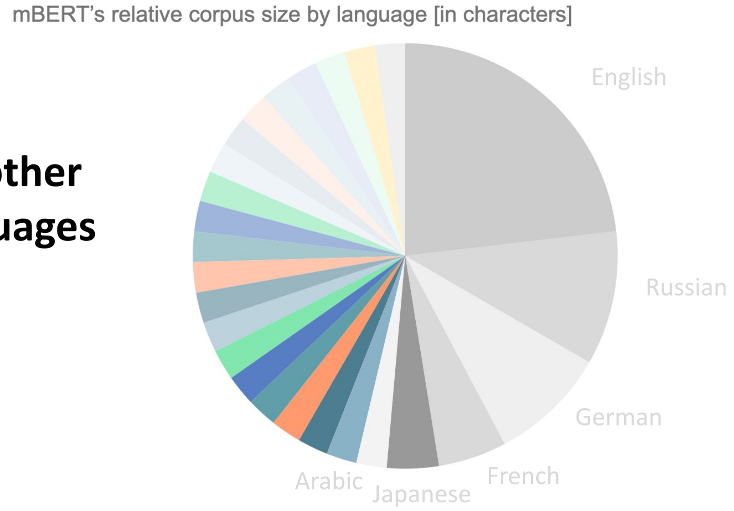
mBERT's relative corpus size by language [in characters]



Confound: Unbalanced Corpus Size

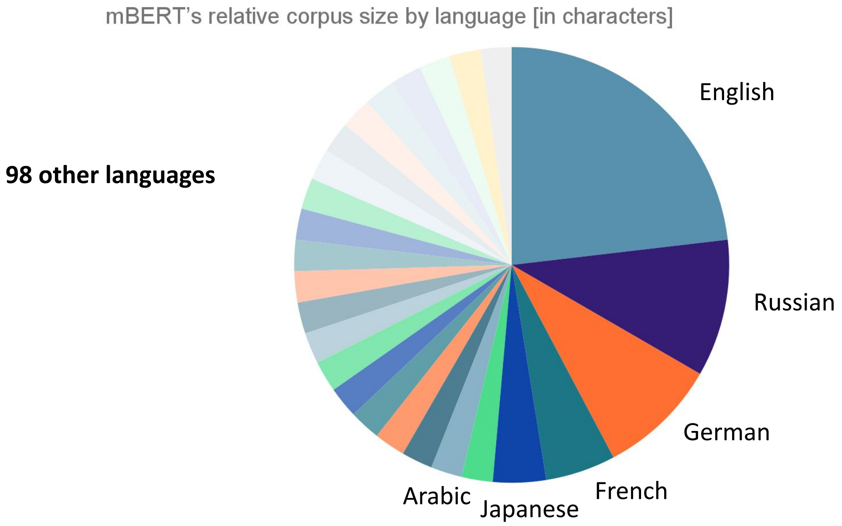
- mBERT is composed of 104 languages, **but is far from balanced**

98 other languages



Confound: Unbalanced Corpus Size

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



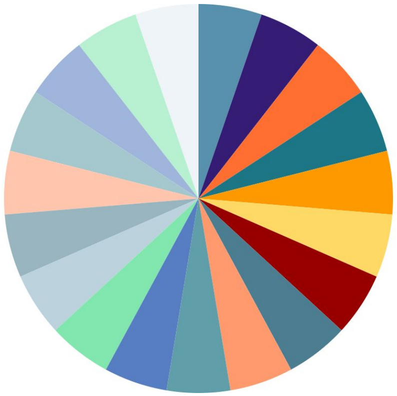
Let's **balance** the pretraining data to get closer to saying something about *language-inherent* properties

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



Caveat: Is the number of characters a good measure?

- Chinese (or Hebrew) may pack more information in a character than English
  - Thus balancing by the number of characters may again skew the results

- Ideally, we would like to balance the data by the *information* it conveys

Caveat: Is the number of characters a good measure?

- Chinese (or Hebrew) may pack more information in a character than English
  - Thus balancing by the number of characters may again skew the results

- Ideally, we would like to balance the data by the *information* it conveys

- **Proposal**: estimate information by $\frac{|\text{tokens}|}{|\textit{unique}\ \text{tokens}|}$

Estimating the Amount of Information

- **Proposal**: estimate information by $\frac{|\text{tokens}|}{|\textit{unique}\ \text{tokens}|}$
  - Our corpus is correlated with this measure (r = 0.73)

- We use a *single* word-piece tokenizer over the entire balanced corpus