Computational Multilingualism in the Era of Large Language Models

Freda Shi

Toyota Technological Institute at Chicago

University of Waterloo & Vector Institute (starting July 2024)

My Research in the Context of Science

Computation

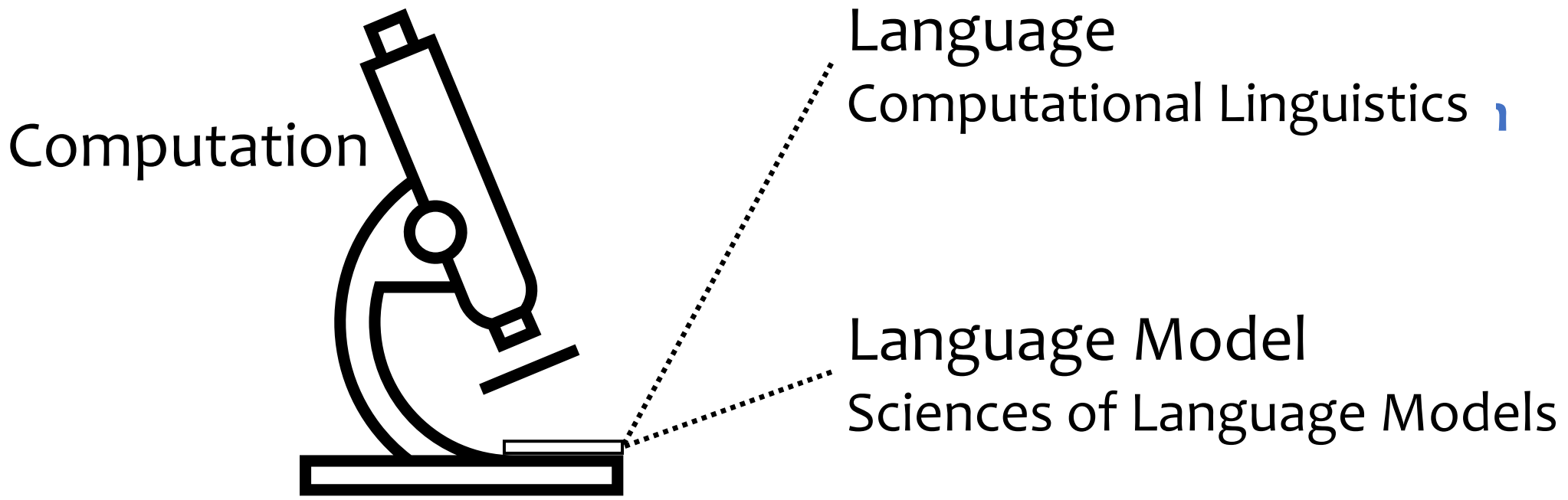
Language

Computational Linguistics

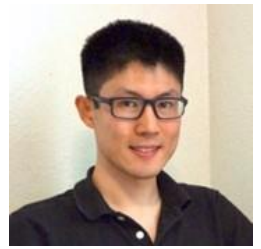
Language Model

Sciences of Language Models

This Talk: Computational Multilingualism



Bilingual Lexicon Induction

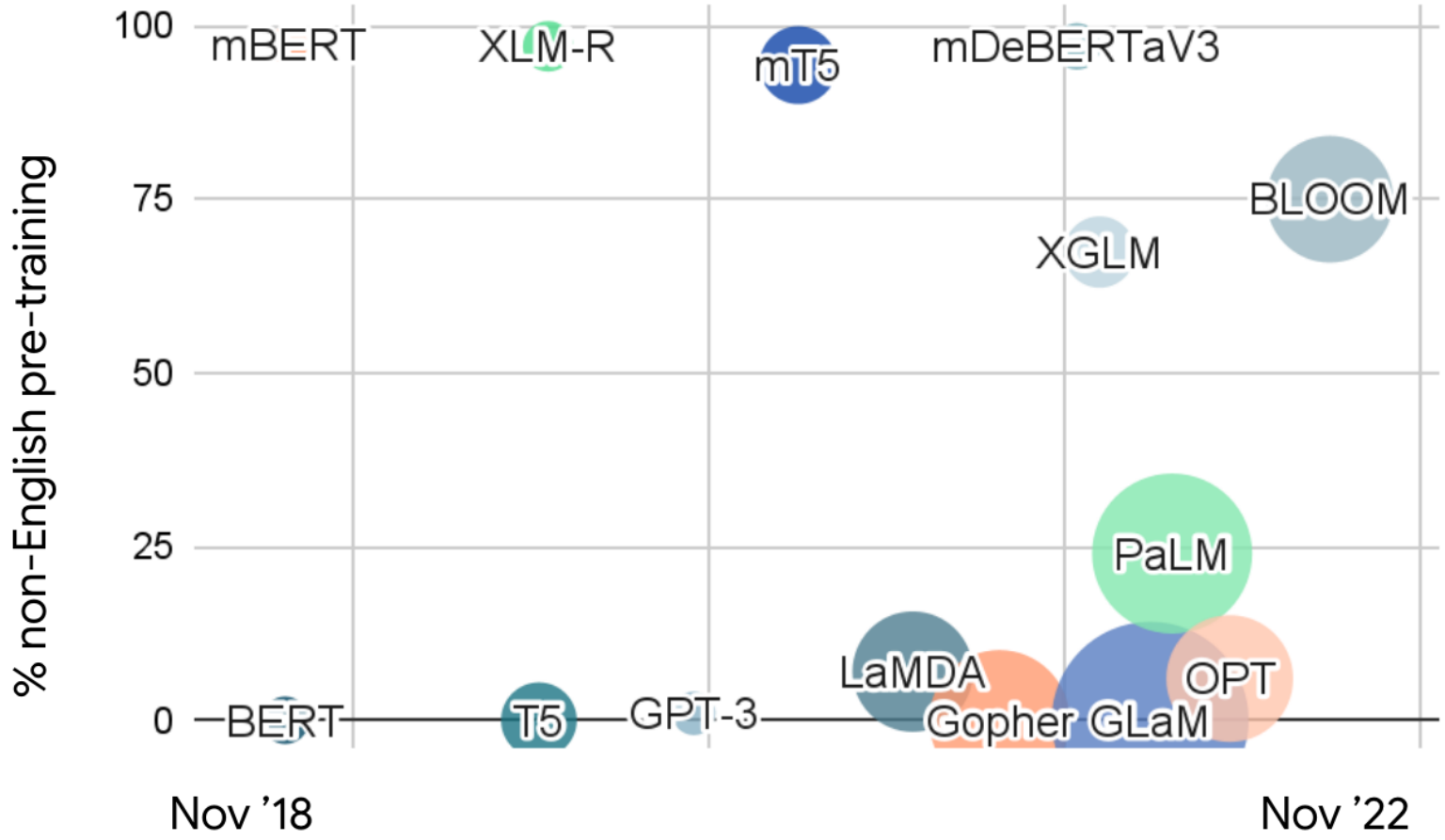
w/ Luke Zettlemoyer and Sida Wang

Research Question:

- How much translation exists in monolingual corpora?
  Can we do unsupervised translation?
- How many mutually translatable word pairs can we find given only monolingual corpora, and optionally a small seed lexicon?

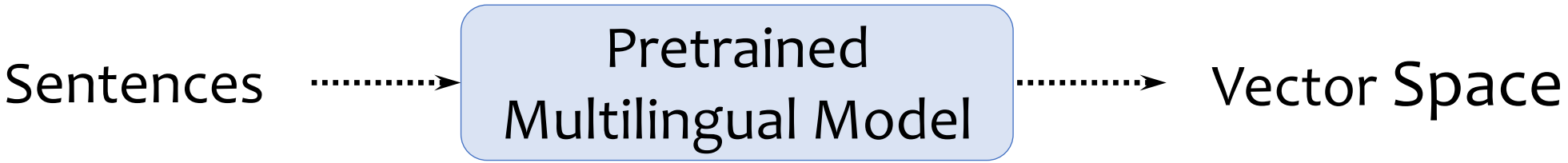
English	Spanish
bank	*orilla*
shore	*orilla*
bank	*banco*

Background: Pretrained Multilingual Models



[Figure credit: Sebastian Ruder]

Background: Pretrained Multilingual Models

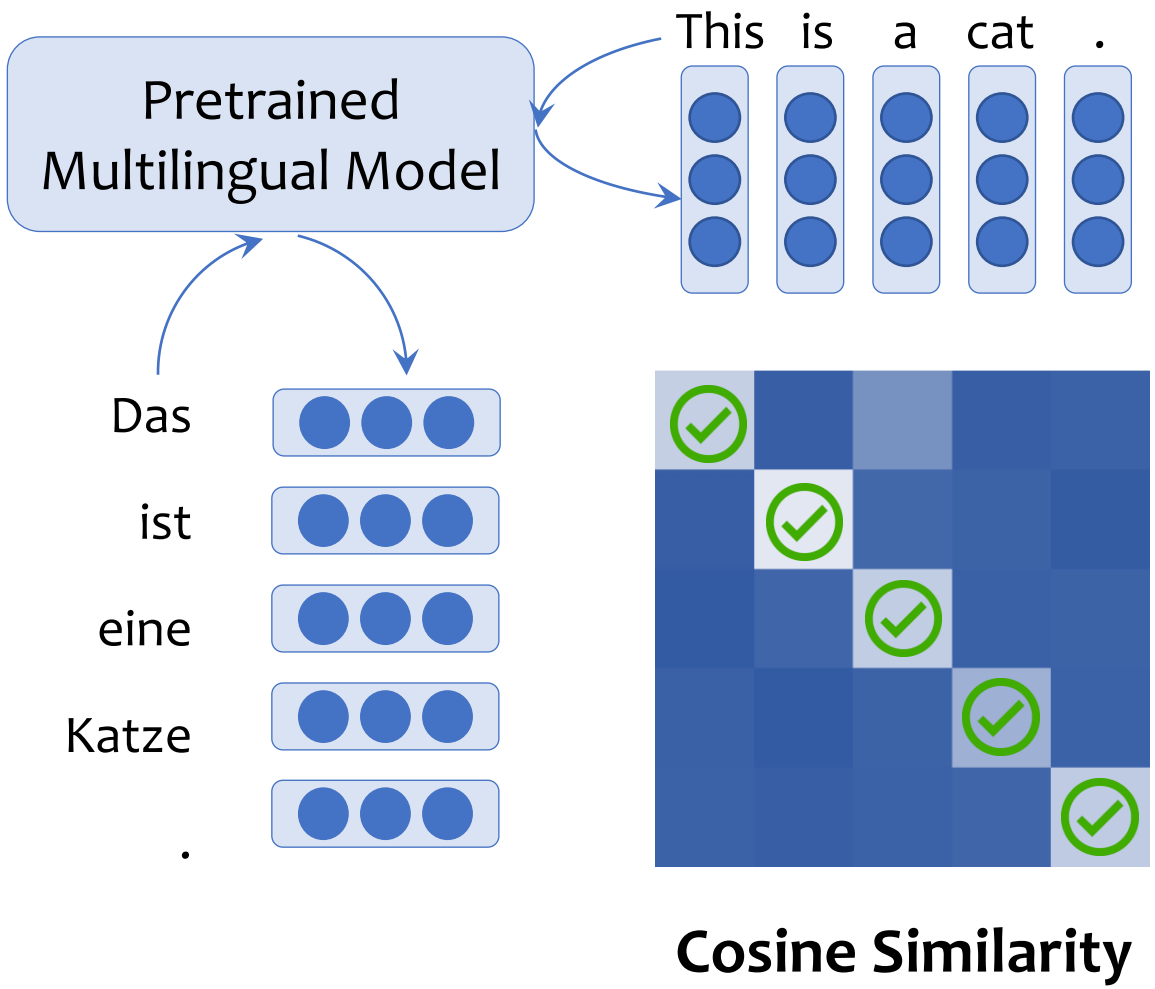


In the vector space, sentences with similar meanings are closer, regardless of their language or writing system.

Hypotheses on the reason: code-switching data, shared vocabulary, …

We can encode sentences and retrieve pseudo bitext in the vector space.

Background: Cross-Lingual Word Aligner



Tokens are subwords in practice.
Many-to-one alignment is possible.

No supervision is required!

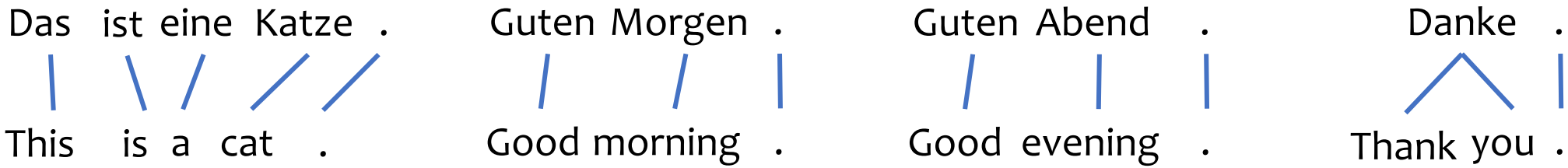
✓ : Alignment

(horizontal and vertical maximum)

[SimAlign: Sabet et al., 2020]

Statistics from Language Models & Pseudo Bitext

Aligned (Pseudo) Bitext



Features for a bilingual pair $\langle s, t \rangle$:

#Cooccurrence(s, t) -- How many times s and t appear in a pair of (pseudo) bitext.

#align-1(s, t) -- How many times s and t are matched in one-to-one alignments.

#align-many(s, t) -- How many times s and t are matched in one-to-many alignments.

Cosine similarity, inner product, #count(s), #count(t), …

Learning with Simple Statistics

Weakly Supervised Induction

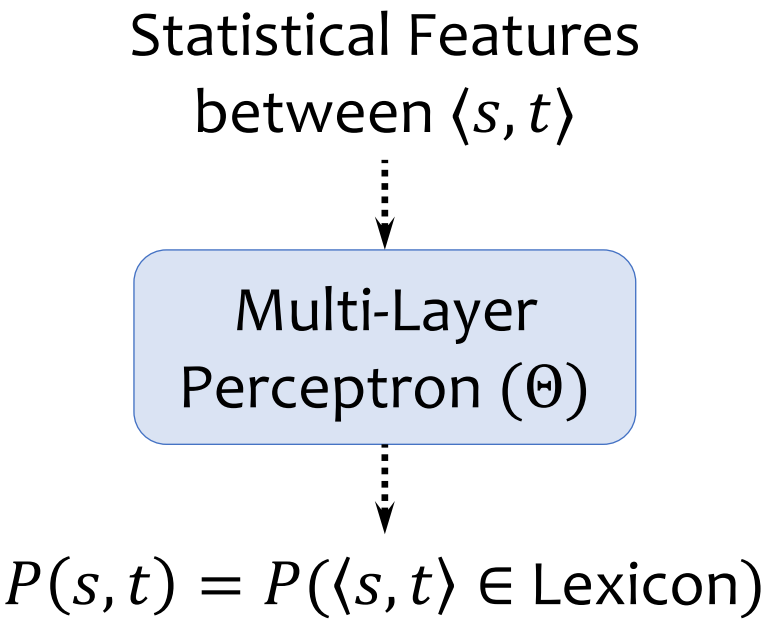
A few bilingual lexicon entries are available.

Positive Examples:

$\mathcal{D}_+ = \{\langle s, t\rangle \in \text{Lexicon}\}$

Negative Examples:

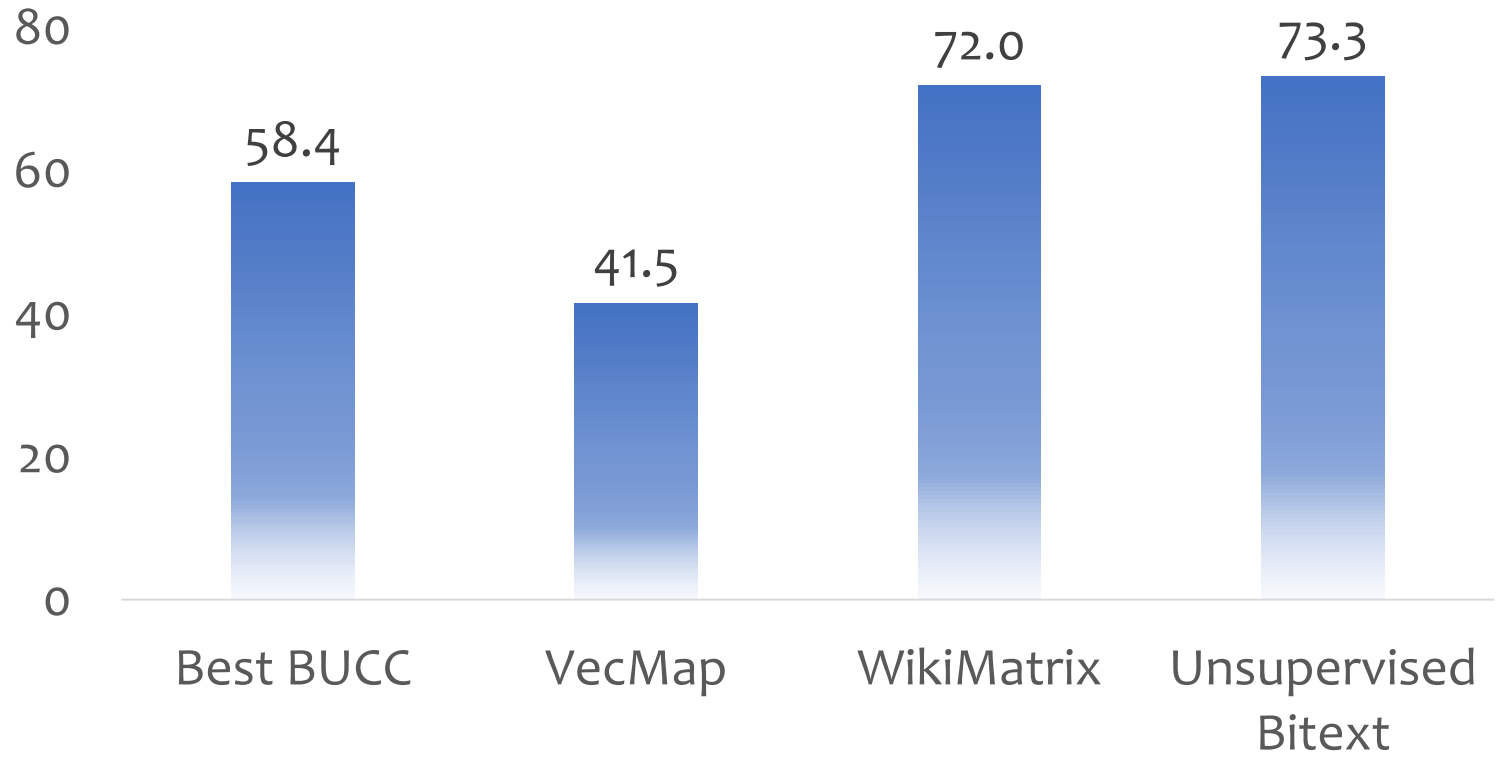
$\mathcal{D}_- = \{\langle s, t\rangle \notin \text{Lexicon},\ \text{cooccurrence} > 0\}$



$$\max_{\Theta} \sum_{\langle s,t\rangle \in \mathcal{D}_+} \log P(s,t) + \sum_{\langle s',t'\rangle \in \mathcal{D}_-} \log(1 - P(s',t'))$$

Results on the BUCC 2020 Shared Task

F_1 score: Harmonic mean between the precision and recall of the predicted lexicon. Averaged across DE-FR, EN-FR, EN-DE, EN-ES, EN-ZH, EN-RU and their reversed language pairs.



Looking into the "*False Positive*" Cases

Ours

- 倉庫 depot
- 浪費 wasting
- 背面 reverse
- 嘴巴 mouths
- 可笑 laughable
- 隱藏 conceal
- 虔誠 devout
- 純淨 purified pure
- 截止 deadline
- 鍾 clocks

VecMap

- 申明 endorsing declare
- 條件 preconditions condition
- 天津 shanghai tianjin
- 個案 cases
- 百合 peony lily
- 申報 filing
- 車廂 carriages
- 海草 seaweed
- 收容所 asylums
- 開幕 soft-opened opening

: Acceptable

: Unacceptable

: Acceptable (in a certain context)

Takeaways

Q: How much translation exists in monolingual corpora?
Can we do unsupervised translation?

A: A large portion of word-level translations may exist in unsupervised data.

Contextualized representations may help on (seemingly) non-contextual tasks.

The simple MLP with statistical feature approach also improves word alignment.

This Talk: Computational Multilingualism

Computation

Multiple Languages

Bilingual Lexicon Induction

Multilingual Model

Multilingual Math Reasoning

Multilingual Math Reasoning

w/ Mirac Suzgun et al.

Research Question:

- ❑ How well can language models do on reasoning with different languages?

We extend GSM8K to the multilingual grade-school math (MGSM) benchmark.

MGSM covers 10 languages: Bengali, Mandarin Chinese, French, German, Japanese, Russian, Spanish, Swahili, Telugu and Thai.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

Frage: Shawn hat fünf Spielzeuge. Zu Weihnachten hat er von seiner Mama und seinem Papa jeweils zwei Spielzeuge bekommen. Wie viele Spielzeuge hat er jetzt?

Experiment 1: Native vs. English Reasoning Steps

Native Chains of Thought

Frage: Shawn hat fünf Spielzeuge. Zu Weihnachten hat er von seiner Mama und seinem Papa jeweils zwei Spielzeuge bekommen. Wie viele Spielzeuge hat er jetzt?

Schritt-für-Schritt-Antwort: Er hat 5 Spielzeuge. Er hat 2 von seiner Mama bekommen, sodass er nun 5 + 2 = 7 Spielzeuge hat. Dann hat er noch 2 von seinem Papa bekommen, also hat er insgesamt 7 + 2 = 9 Spielzeuge. Die Antwort lautet 9.

Frage: Roger hat 5 Tennisbälle. Er kauft noch 2 Dosen Tennisbälle. In jeder Dose sind 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt?

Schritt-für-Schritt-Antwort:

English Chains of Thought

Frage: Shawn hat fünf Spielzeuge. Zu Weihnachten hat er von seiner Mama und seinem Papa jeweils zwei Spielzeuge bekommen. Wie viele Spielzeuge hat er jetzt?

Step-by-Step Answer: He has 5 toys. He got 2 from mom, so after that he has 5 + 2 = 7 toys. Then he got 2 more from dad, so in total he has 7 + 2 = 9 toys. The answer is 9.

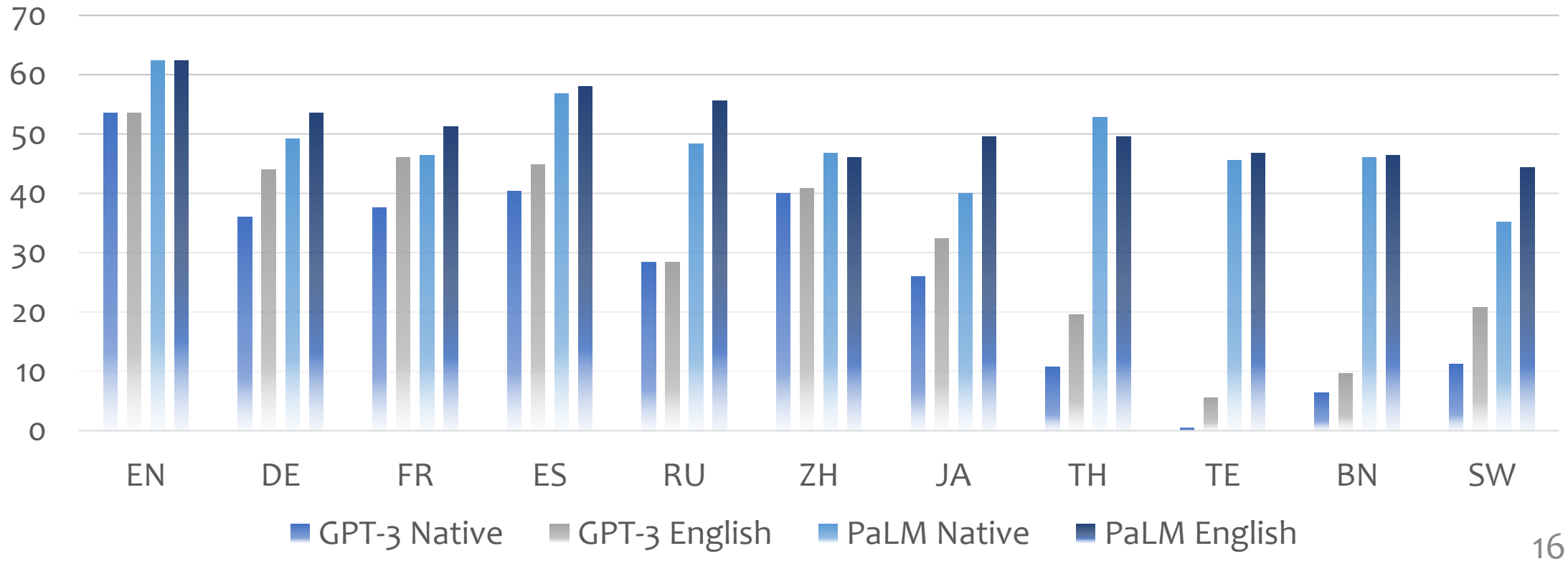
Frage: Roger hat 5 Tennisbälle. Er kauft noch 2 Dosen Tennisbälle. In jeder Dose sind 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt?

Step-by-Step Answer:

Experiment 1: Native vs. English Reasoning Steps

Problem solution accuracy (%) on MGSM.

English, as the CoT solution language, generally outperforms the native language.



Experiment 2: English vs. Multilingual Exemplars

English-Only Exemplars

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

Step-by-Step Answer: He has 5 toys. He got 2 from mom, so after that he has 5 + 2 = 7 toys. Then he got 2 more from dad, so in total he has 7 + 2 = 9 toys. The answer is 9.

… (5 more English examples)

Frage: Roger hat 5 Tennisbälle. Er kauft noch 2 Dosen Tennisbälle. In jeder Dose sind 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt?

Step-by-Step Answer:

Multilingual Exemplars

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

Step-by-Step Answer: He has 5 toys. He got 2 from mom, so after that he has 5 + 2 = 7 toys. Then he got 2 more from dad, so in total he has 7 + 2 = 9 toys. The answer is 9.

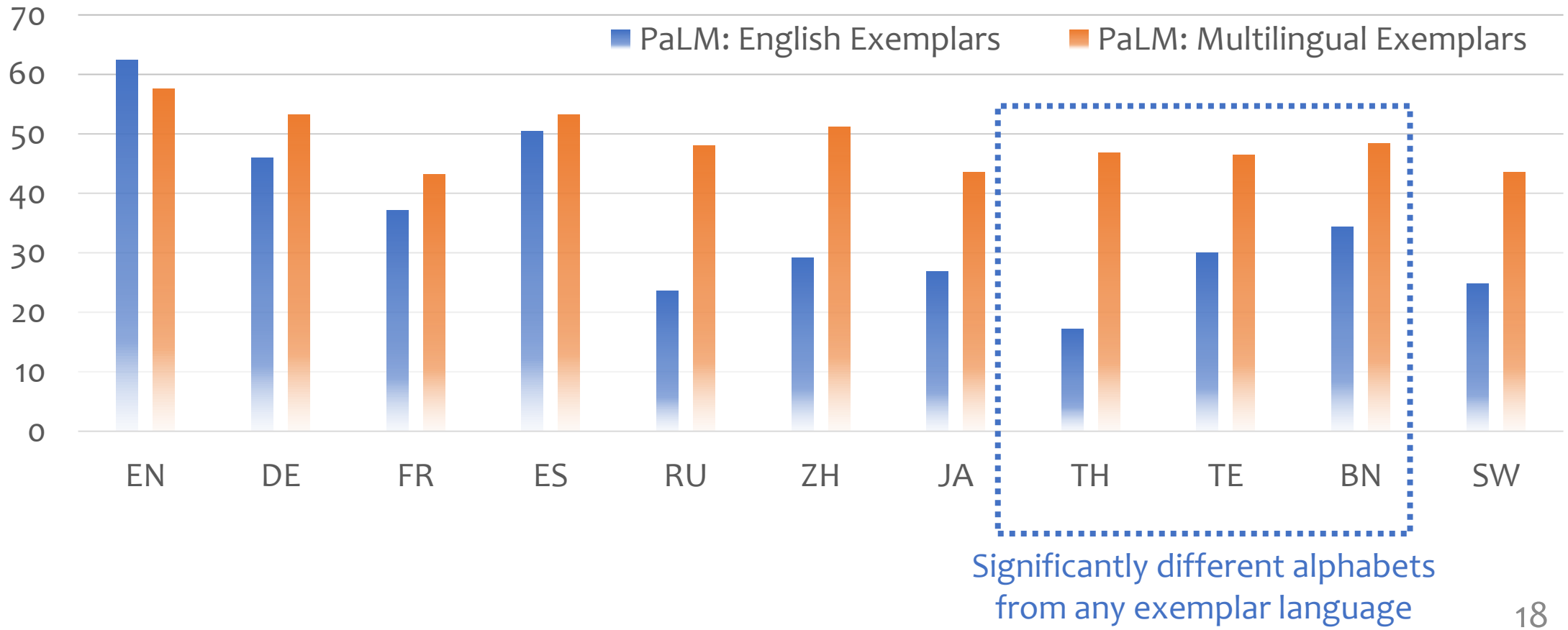
… (5 examples, questions in DE, FR, ES, RU, ZH, respectively; solutions are in English.)

Frage: Roger hat 5 Tennisbälle. Er kauft noch 2 Dosen Tennisbälle. In jeder Dose sind 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt?

Step-by-Step Answer:

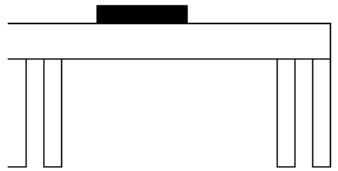
Experiment 2: English vs. Multilingual Exemplars

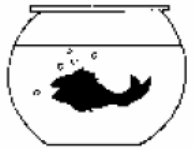
Multilingual exemplars work better than English for all languages except English.



What's Next

- People use different prepositions to describe the same spatial relations [Feist & Gentner, 2003].

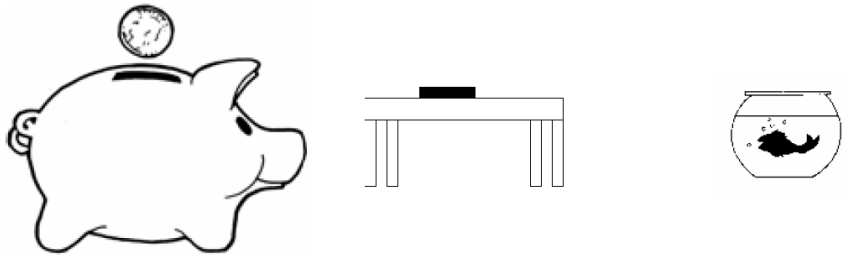




English	over	on	in
Spanish	sobre	en	en
Japanese	ue	ue	naka

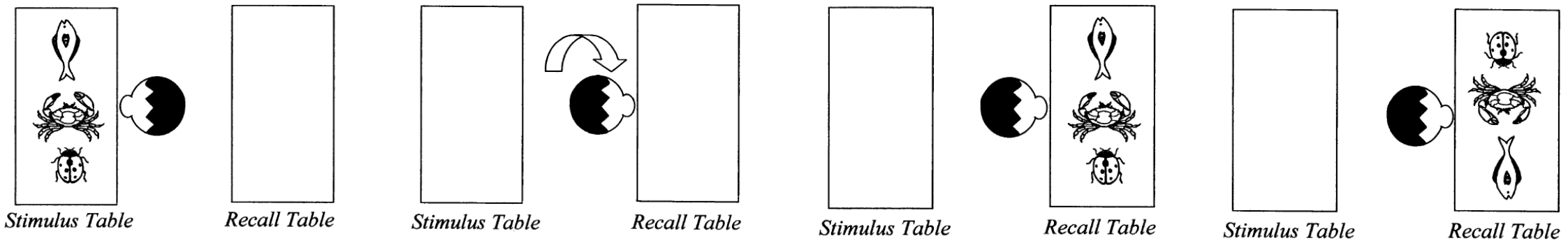
What's Next: Computation for Language Sciences

- People use different prepositions to describe the same spatial relations [Feist & Gentner, 2003].
- Corpus, computation, and large language models provide new toolkits to better discover and understand these phenomena.



What's Next

Language may affect how humans view the world.



[Figure credit: Li & Gleitman, 2002]

Do language models trained on these languages have similar biases?

Can we explain language model behaviors with machine learning theory and/or interpretability techniques?

Can the results shed light on explaining human language acquisition? …

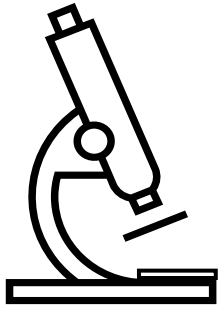
What's Next: Language Model Sciences for Sciences

Human experiments are usually expensive.

Language models are, to some extent, models of humans.

Experiments on faithful language models may even serve as the (pilot) pilot study for human experiments.

Computation
(+X)



Language
Language Models
…

Thanks!