Multilingual Natural Language Processing (NLP)

Derry Wijaya

BOSTON UNIVERSITY

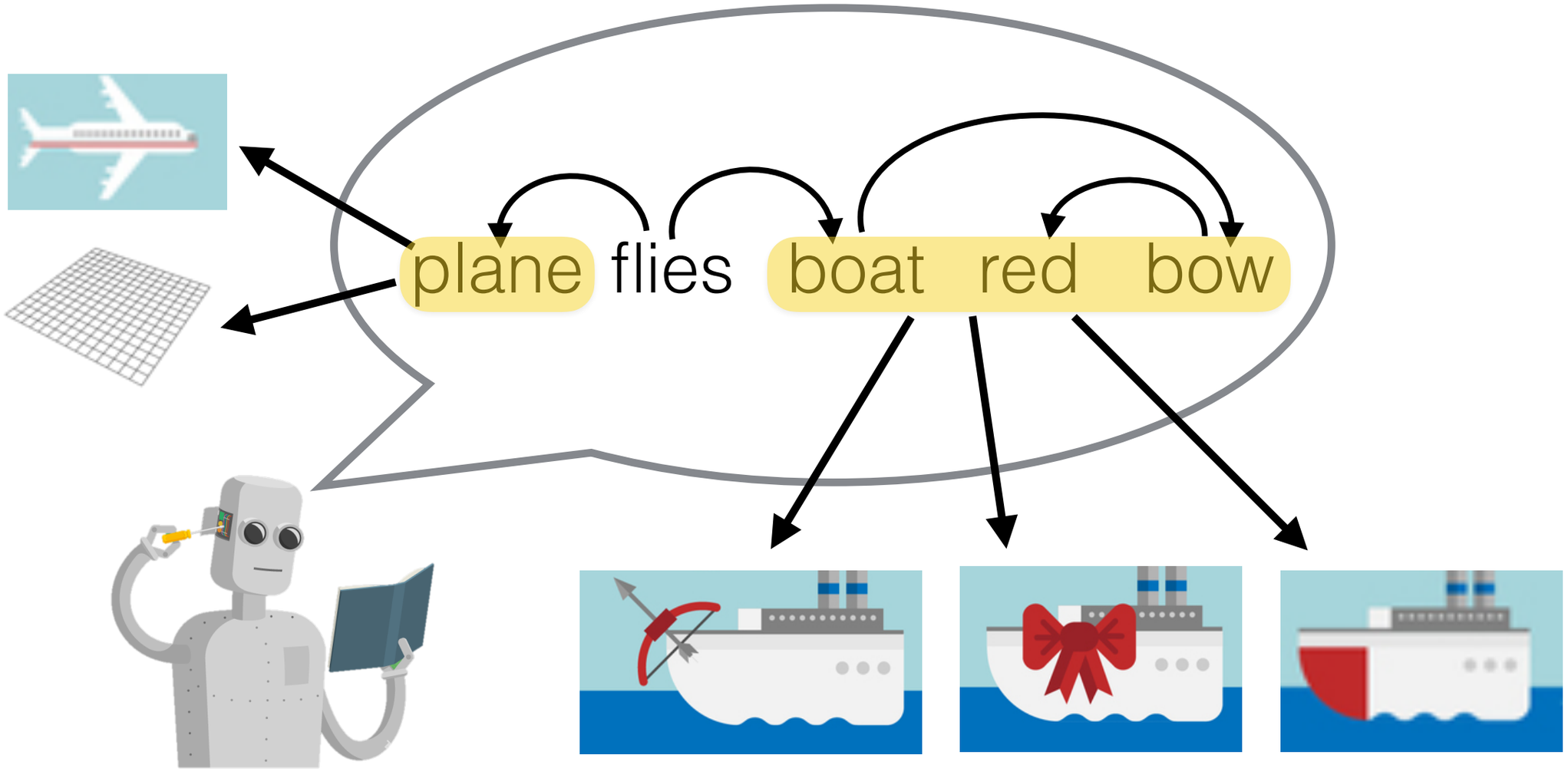
About Me

- Bachelors from National University of Singapore
- PhD from Carnegie Mellon University
- Postdoc at University of Pennsylvania
- Now Asst. Professor at Boston University (CS)
- Developed interest in research as an undergraduate doing UROP

The Quest

Build machines that understand natural (human) language

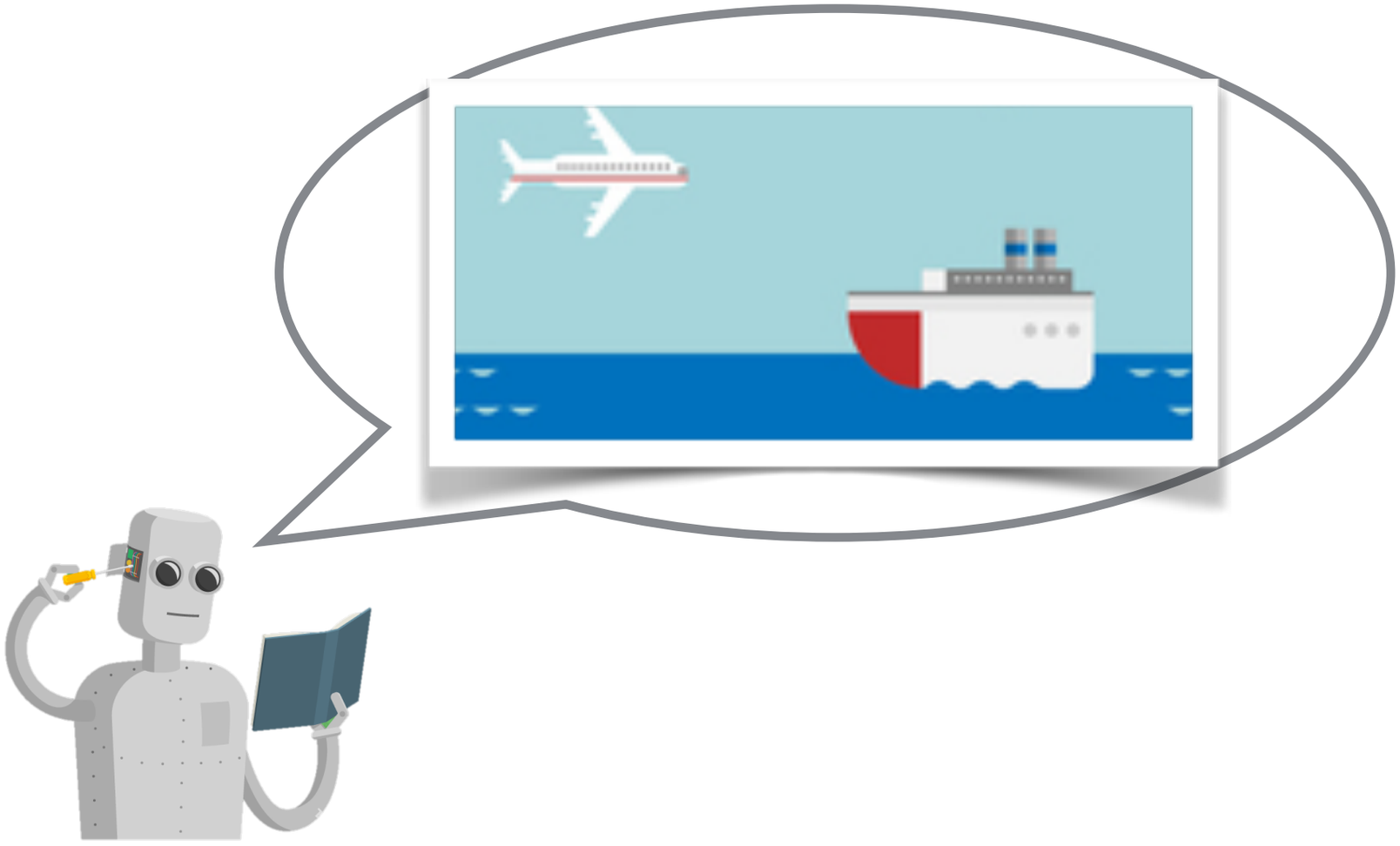
the plane flies over the boat with the red bow



http://www.minettadesignco.com/ibm-think-academy-watson/
https://commons.wikimedia.org/wiki/File:Flat_Surface.svg

Natural Language Understanding

the plane flies over the boat with the red bow



NLP and the Measure of Intelligence

using language as humans do
== *truly intelligent* machines?

Turing Test

by responding as a person to the examiner's questions, the machine wins if it can convince the examiner into believing that it is a person

https://xkcd.com/329/

Natural Language Understanding

the plane flies over the boat with the red bow

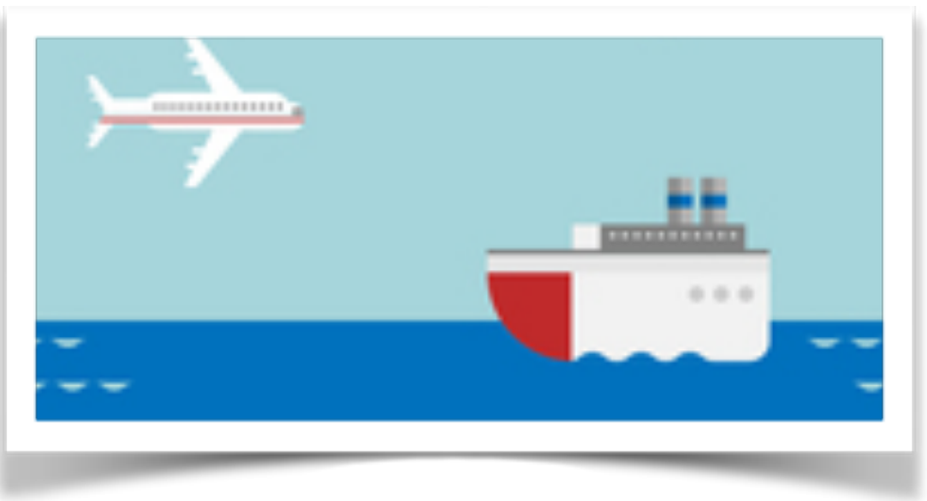
the airplane flies above the ship with the red bow

the jet flies over the ship with the red front

the aircraft flies above the ship with the red bow

pesawat terbang di atas kapal berhaluan merah

သဘော်အထက်တွင်ပျံသန်းလေယာဉ်



ເຮືອບິນບິນເຫນືອເຮືອສີແດງ

Natural Language Understanding

- More than 7,000 languages in the world
  - Few have large annotated corpora for training
  - How can we extend the coverage of current NLP systems?

State of the Art

- However, most NLP resources and systems are **available only for high resource languages**
  - Many low resource languages are spoken by millions of people e.g., Bengali, Indonesian, Swahili, …
  - The challenge is how to develop resources and tools for thousands of languages, not just a few

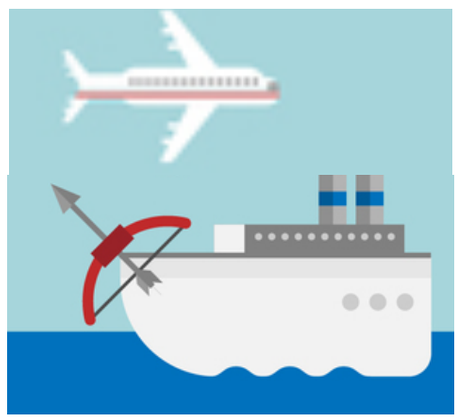
Machine Translation

the plane flies over the boat with the red bow

english

pesawat terbang di atas perahu dengan busur merah

indonesian



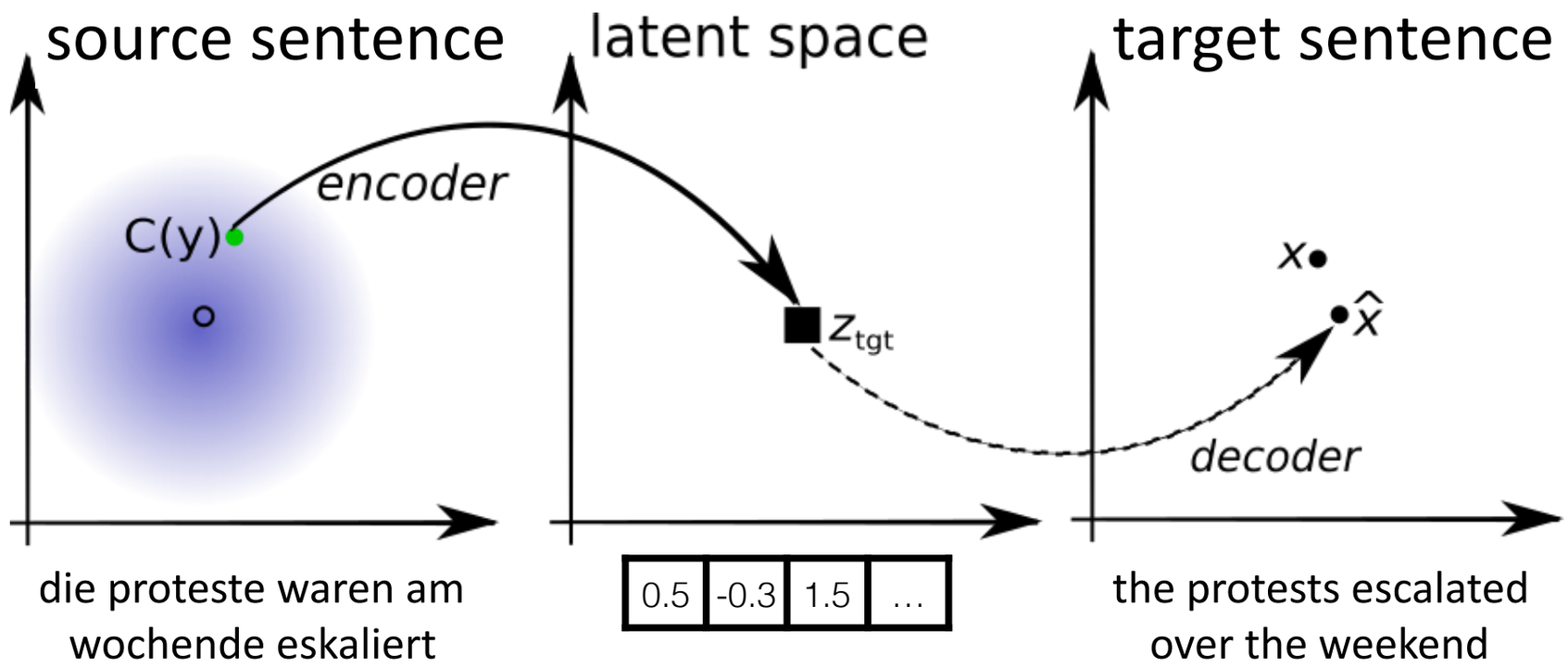
Machine Translation

Machine Translation

- Started with hand-built grammar based systems (limited success)
- Transformed with the availability of **parallel sentences** to collect statistics of word translations and word sequences
    - small word groups often have distinctive translations — phrase based MT, which formed the basis of Google translate

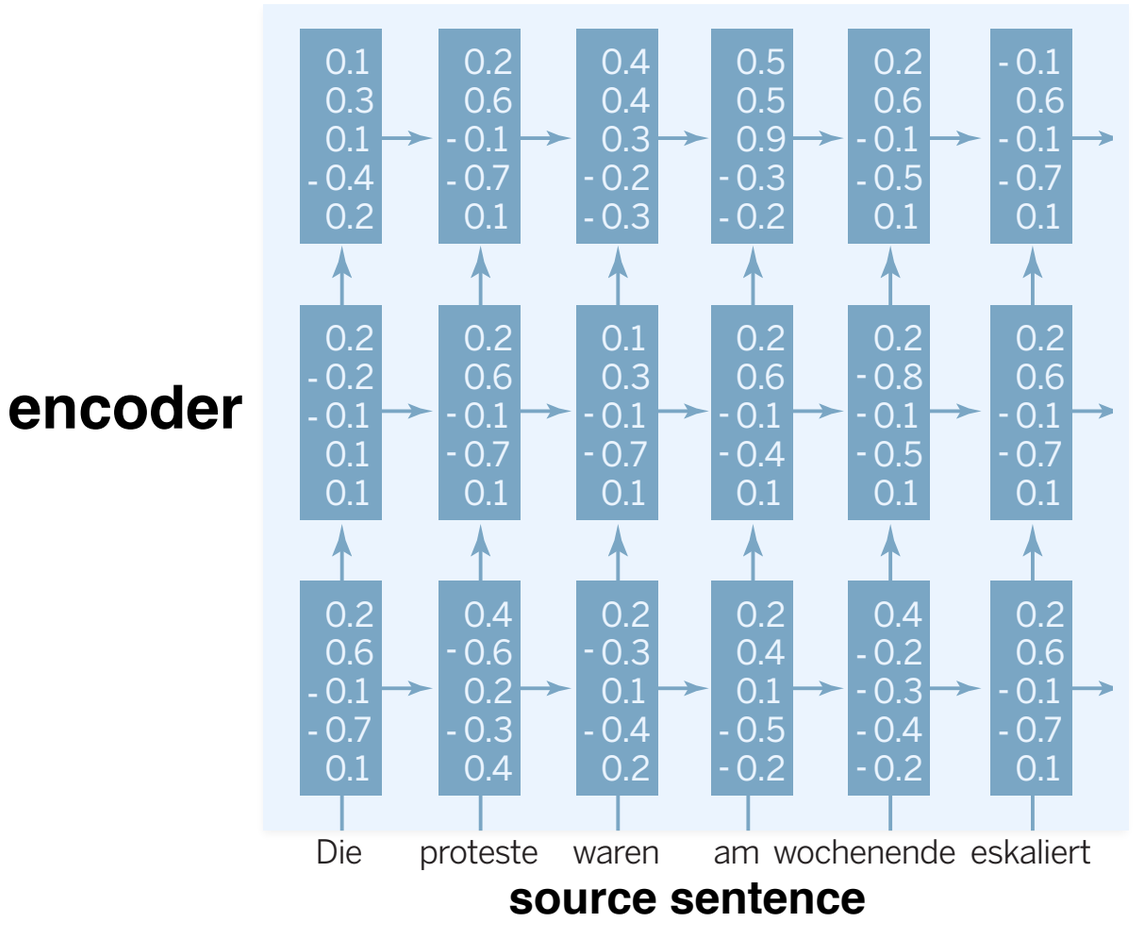
Machine Translation

- Encoder decoder Network



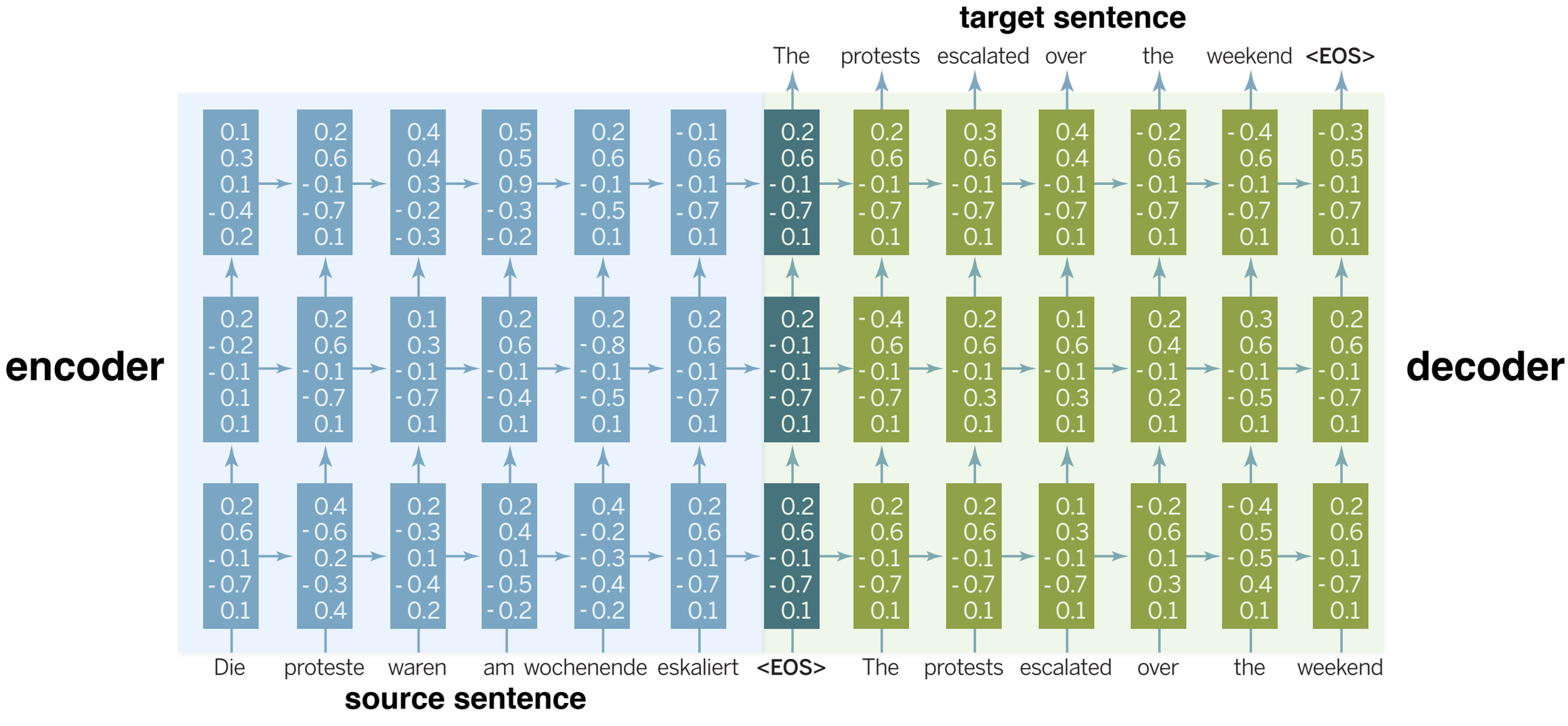
Machine Translation

- Encoder decoder Network
  - maintain contextual information from early until late in a sentence

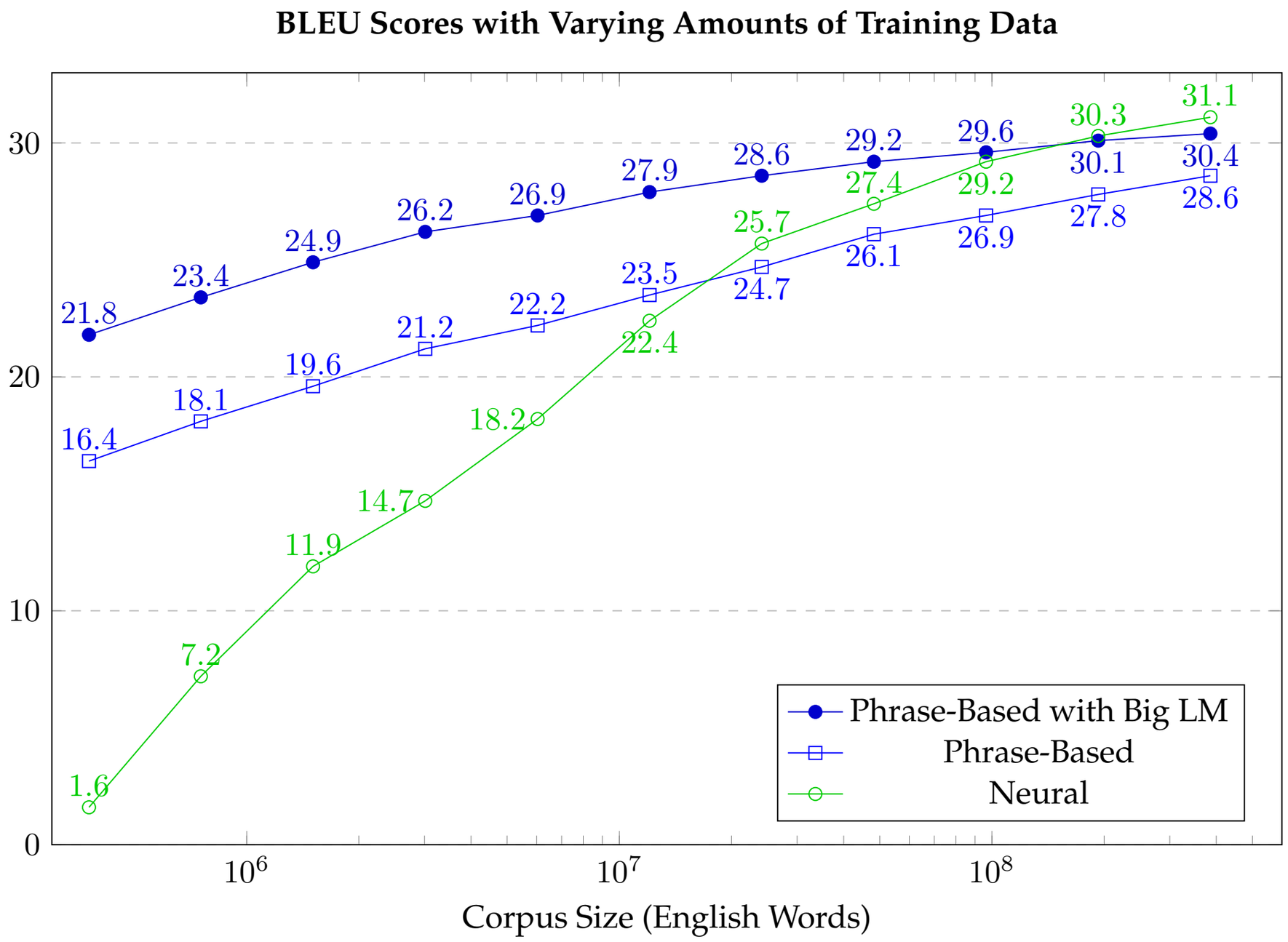


Machine Translation

- Encoder decoder Network
  - maintain contextual information from early until late in a sentence



Machine Translation



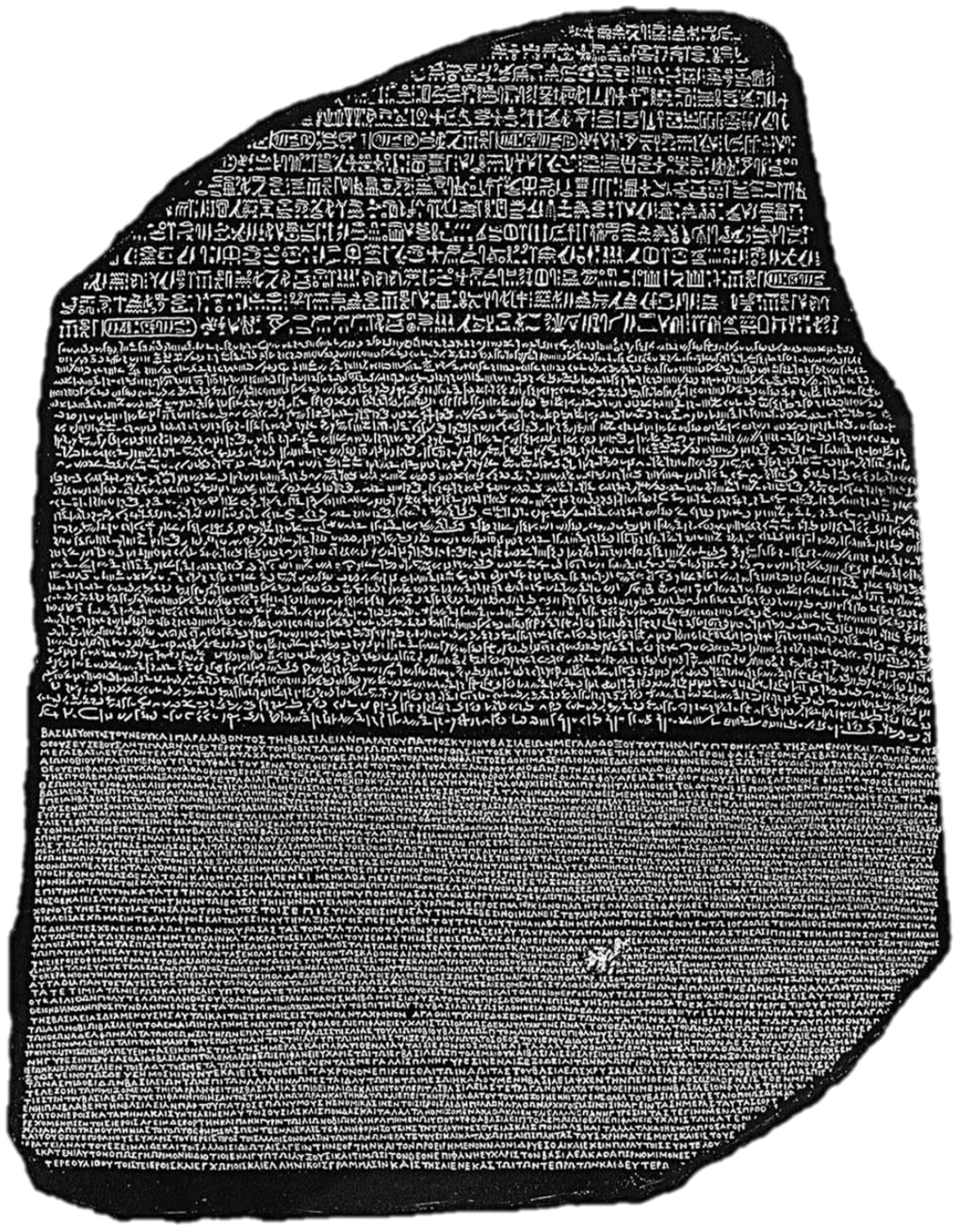
Neural MT, Philipp Koehn

Machine Translation

Ratio	Words	**Source:** *A Republican strategy to counter the re-election of Obama*
$\frac{1}{1024}$	0.4 million	*Un órgano de coordinación para el anuncio de libre determinación*
$\frac{1}{512}$	0.8 million	*Lista de una estrategia para luchar contra la elección de hojas de Ohio*
$\frac{1}{256}$	1.5 million	*Explosión realiza una estrategia divisiva de luchar contra las elecciones de autor*
$\frac{1}{128}$	3.0 million	*Una estrategia republicana para la eliminación de la reelección de Obama*
$\frac{1}{64}$	6.0 million	*Estrategia siria para contrarrestar la reelección del Obama .*
$\frac{1}{32}+$	12.0 million	*Una estrategia republicana para contrarrestar la reelección de Obama*

Figure 13.49: Translations of the first sentence of the test set using neural machine translation system trained on varying amounts of training data. Under low resource conditions, neural machine translation produces fluent output unrelated to the input.

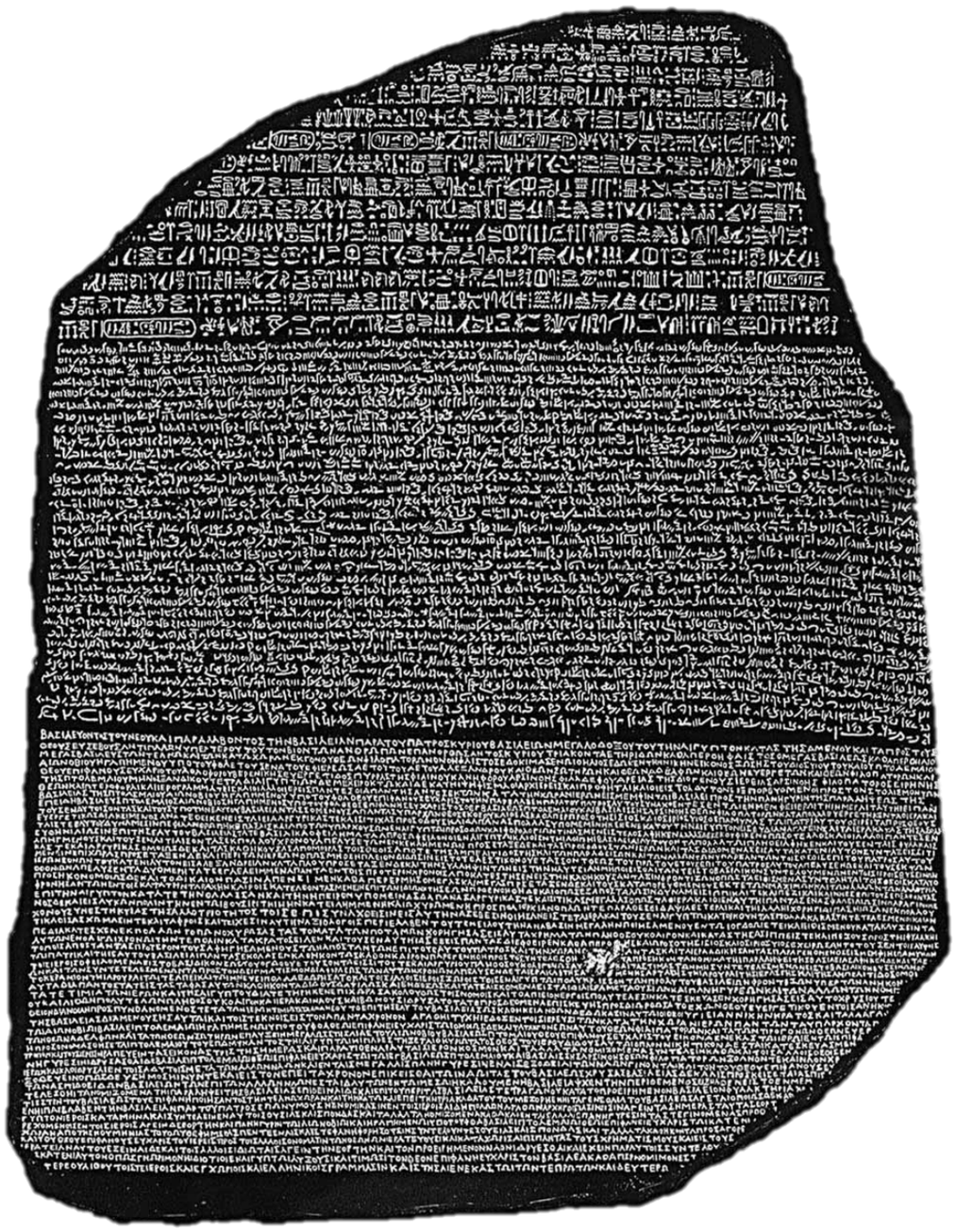
Machine Translation



Machine translation (MT) models typically require large, **sentence-aligned bilingual texts** to learn good translation models

Parallel Data

and the lack of it



sufficient quantities available for only a **few** language pairs

we need

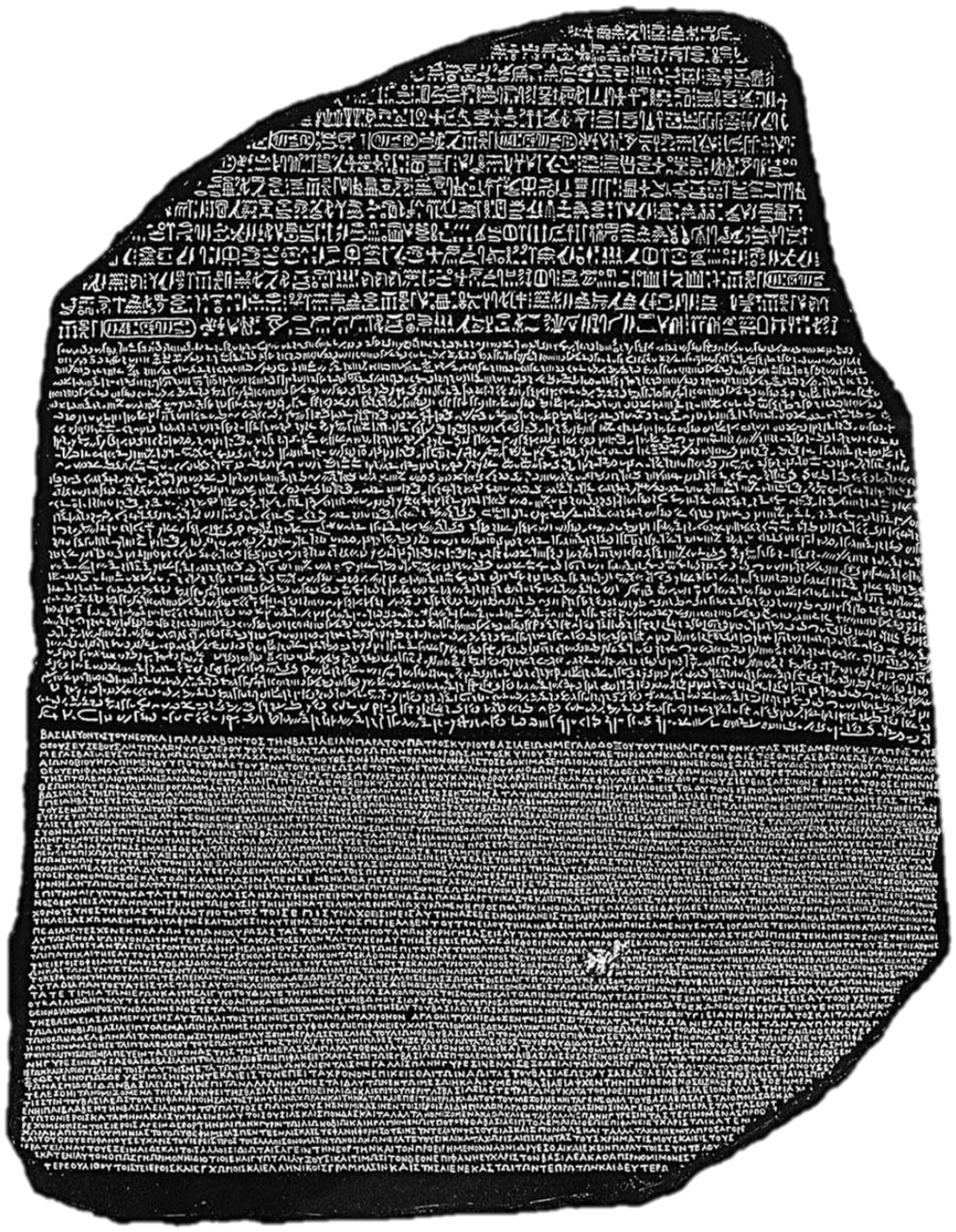
>10^7 word tokens

worth of parallel sentences

for effective neural MT

Parallel Data

and the lack of it

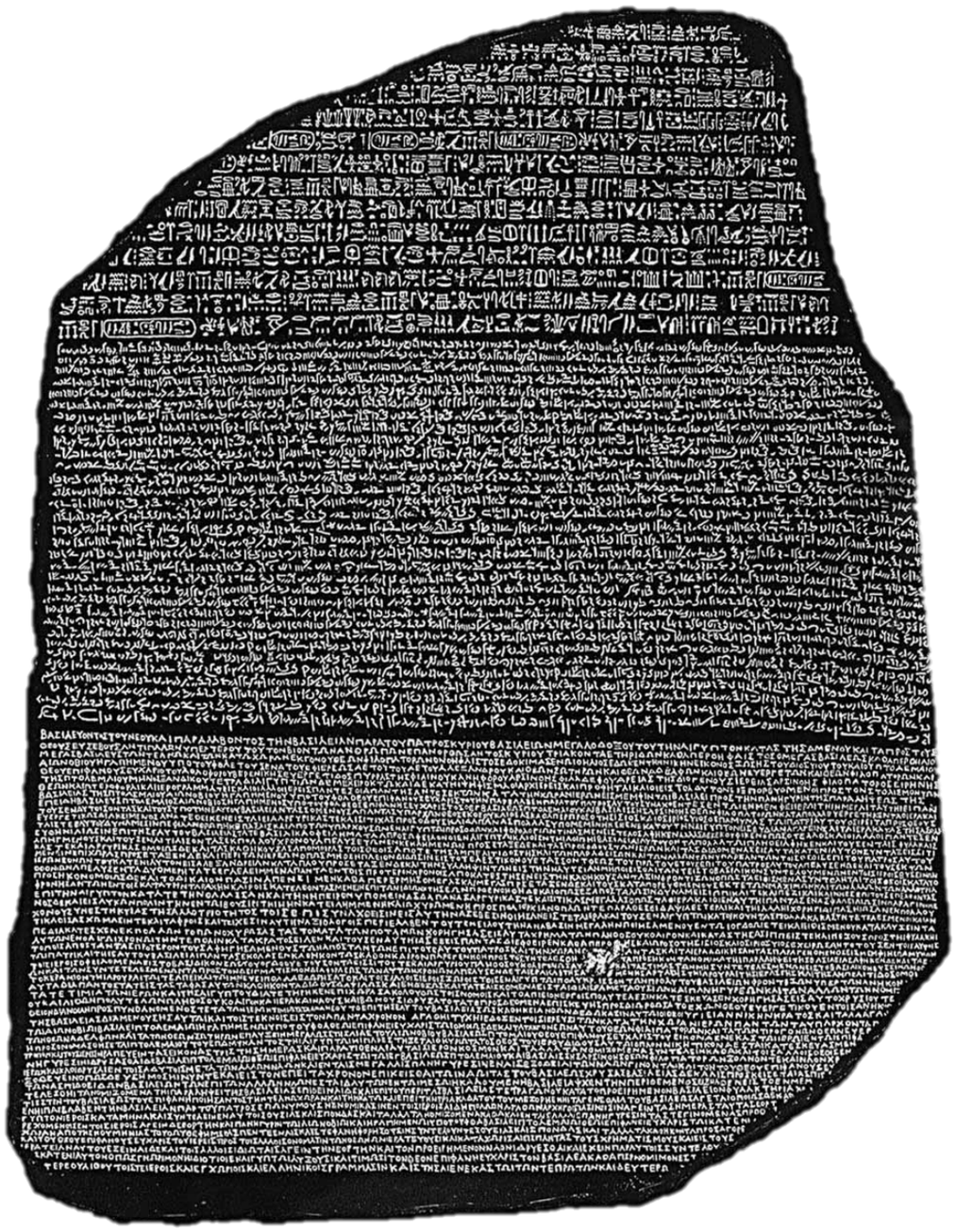


sufficient quantities available for only a **few** language pairs

out-of-vocabulary words

Parallel Data

and the lack of it



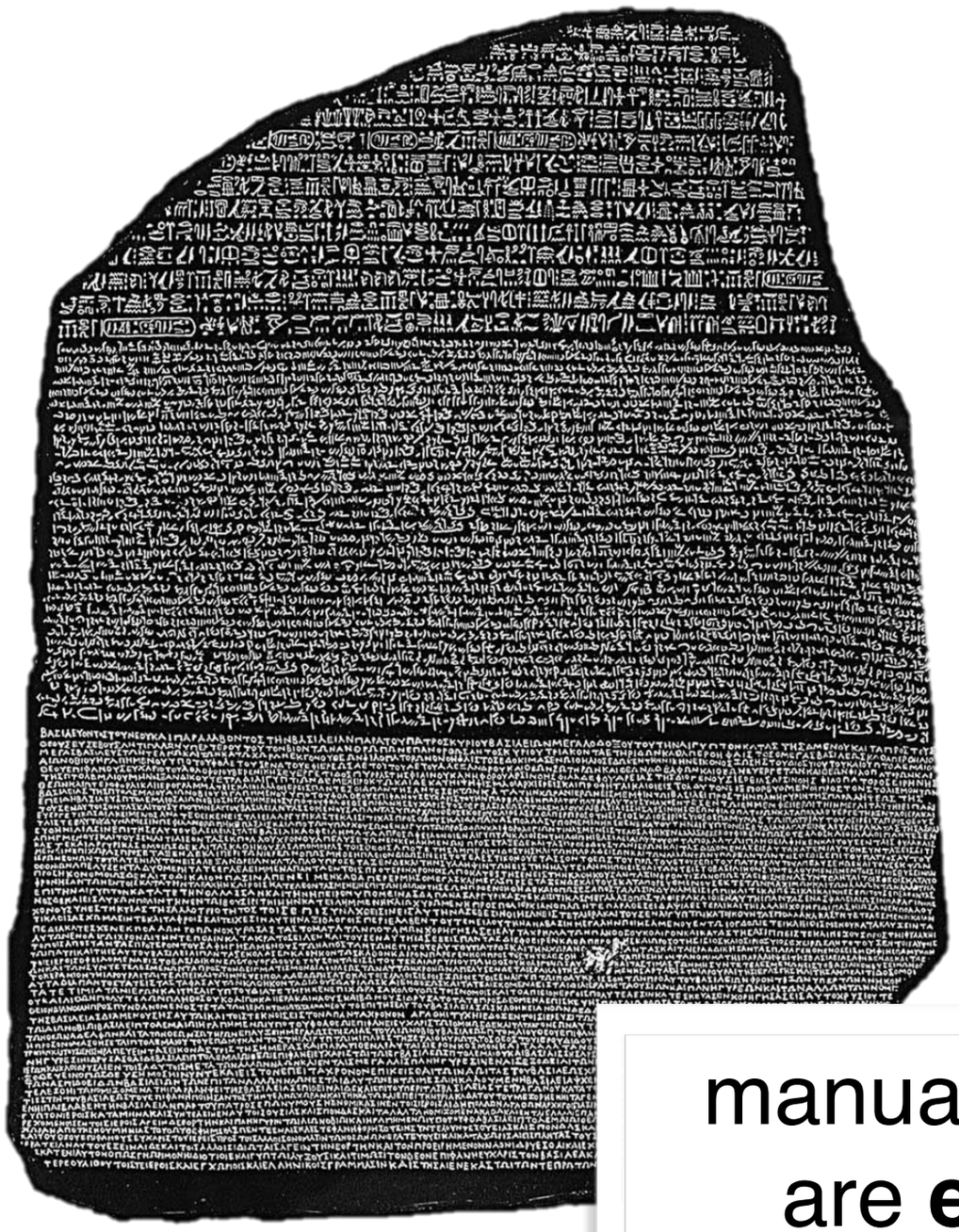
sufficient quantities available for only a **few** language pairs

out-of-vocabulary words

alignments are **inaccurate** with limited parallel texts

Parallel Data

and the lack of it



sufficient quantities available for only a **few** language pairs

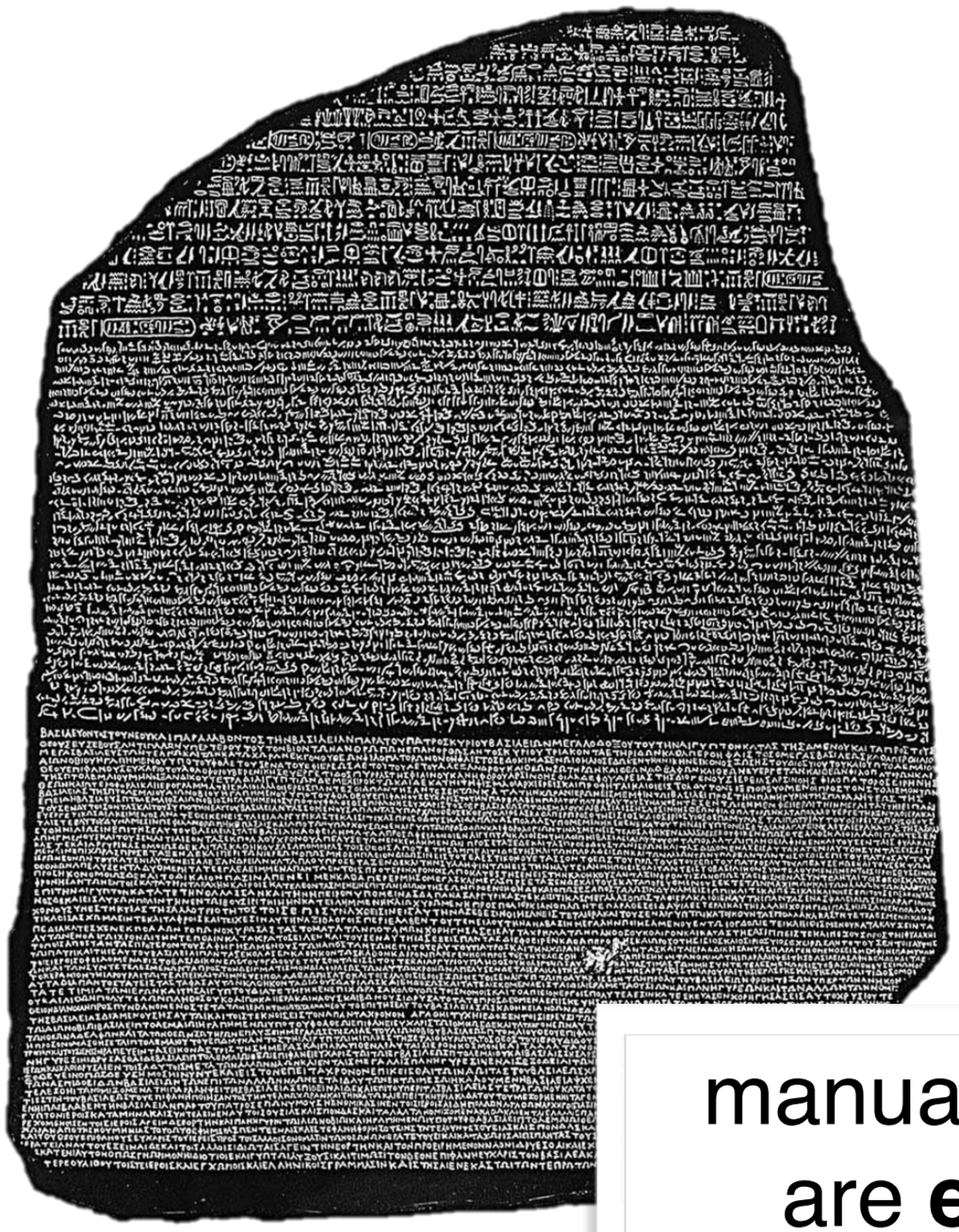
out-of-vocabulary words

alignments are **inaccurate** with limited texts

manual translations are **expensive**!

Parallel Data

and the lack of it



sufficient quantities available for only a **few** language pairs

out-of-vocabulary words

alignments are **inaccurate** with limited texts

manual translation are **expensive**!

dictionaries are **incomplete**

Dictionaries

appaq
1. adj. Snow-white; pure white: -- *su köpükliri* snow-white spray/ -- *qar* pure white snowflake/ ¶ *qağa balam* --, *kirpä balam yumşaq* the crow sees its young snow-white, the hedgehog sees its young soft (the crow loves its young best, the hedgehog loves its young best). [K. *ap-pak*/Ta. *apak*].
2. adj. Beloved; lovable: -- *qiz* lovable girl/ *appiqim* my beloved. Cf. **dilräba, söyümlük.** [U. *oppoq* (#1-2)].

appay
n. Polite address for older woman.

aptap
n. Sunshine: -- *ötüş* med. insolatio; sunstroke/ -- *sun-* to sunbathe, bask in the sun/ -- *yä-* to be exposed to sunshine/ -- *yigän qoğun* sun-drenched melon/ *kiyim-keçäkni --qa sal-* to air out clothes. [U. *oftob*].

aptappäräs
1. bot. n. Sunflower *(Helianthus annuus).* [U. *oftobparast*].

aptappäräs

limited in size

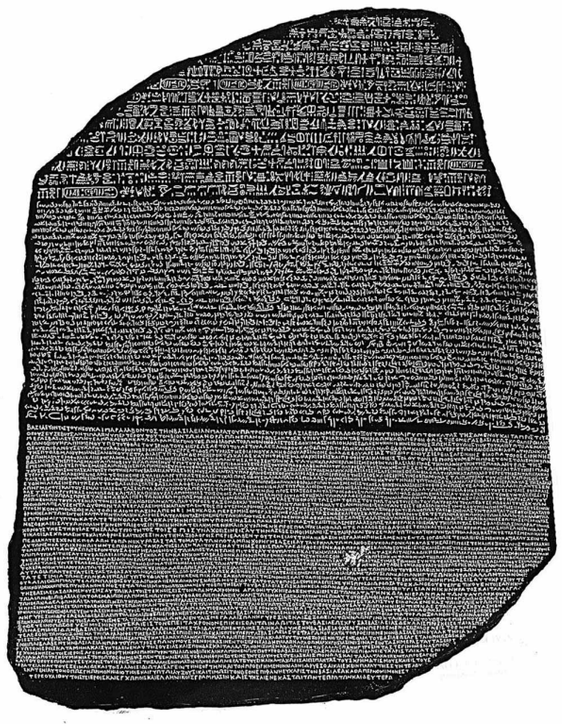
not machine **readable**

more **definition** than translation

not straightforward to use!

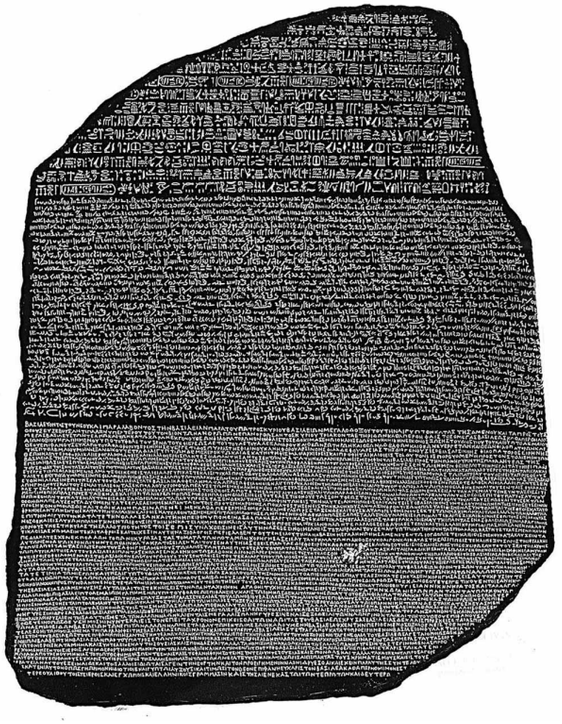
especially by machines

Can We Learn with Fewer Labels?



Bilingual Lexicon Induction

few **bilingual** data



+

plentiful **monolingual** data

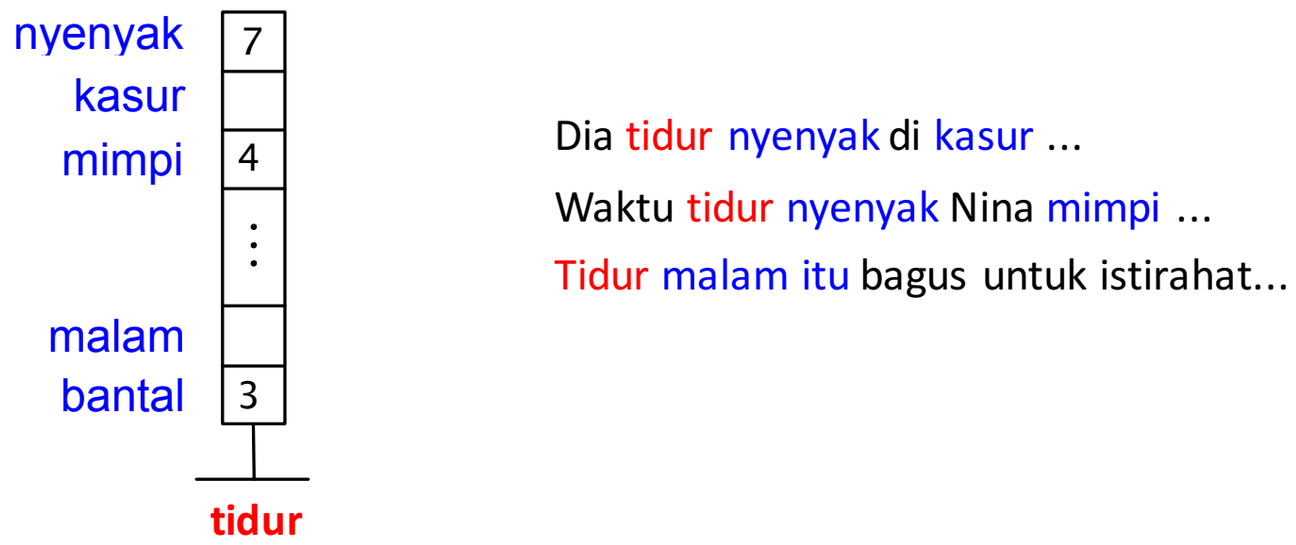
Bilingual Lexicon Induction

using contextual similarity

- Measure contextual similarity (Rapp, 99; Klementiev et al., 2012)

  — Words appearing in similar context are probably related

  - First collect context



Bilingual Lexicon Induction

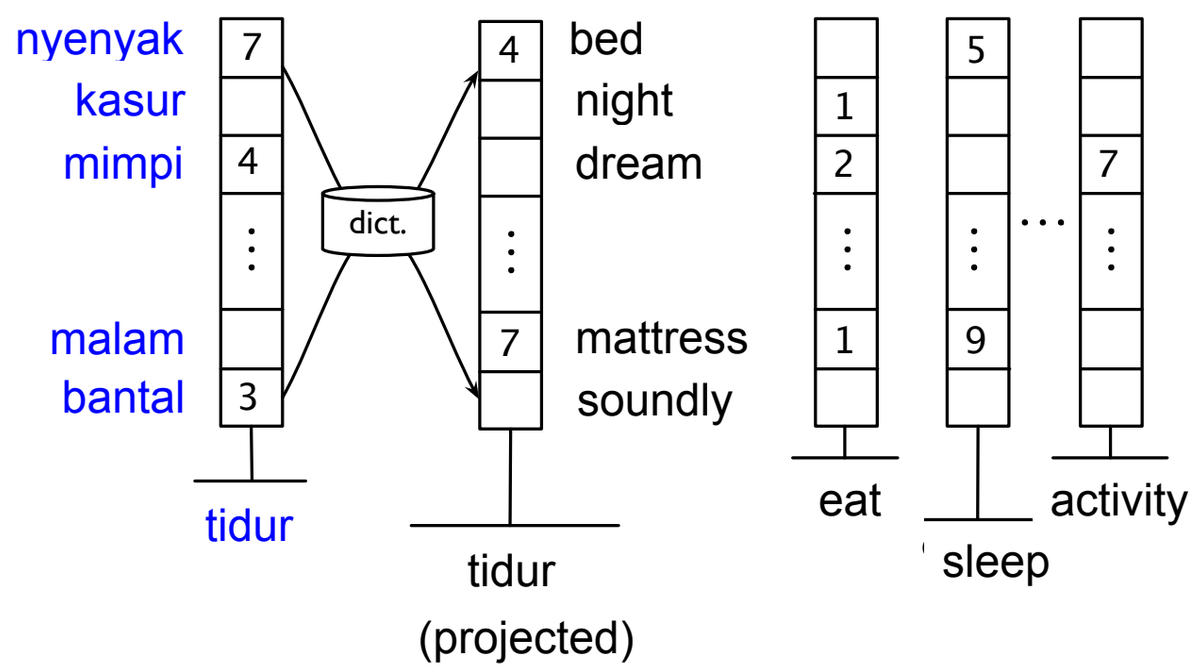
using contextual similarity

- Measure contextual similarity (Rapp, 99; Klementiev et al., 2012)

  — Words appearing in similar context are probably related

  - First collect context

- Then, project through a seed dictionary and compare vectors



Bilingual Lexicon Induction

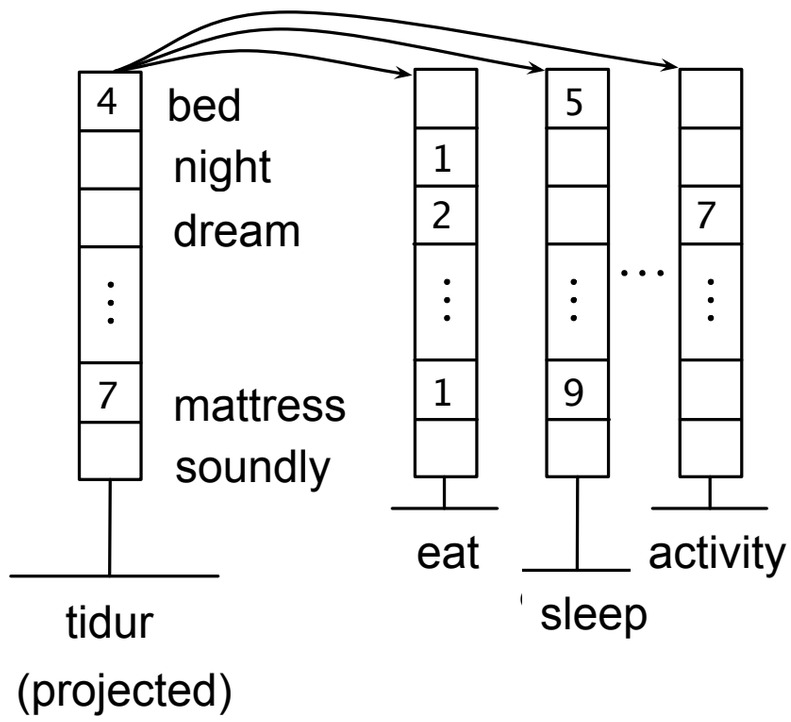
using contextual similarity

- Measure contextual similarity (Rapp, 99; Klementiev et al., 2012)

  — Words appearing in similar context are probably related

  - First collect context

- Then, project through a seed dictionary and compare vectors



Bilingual Lexicon Induction

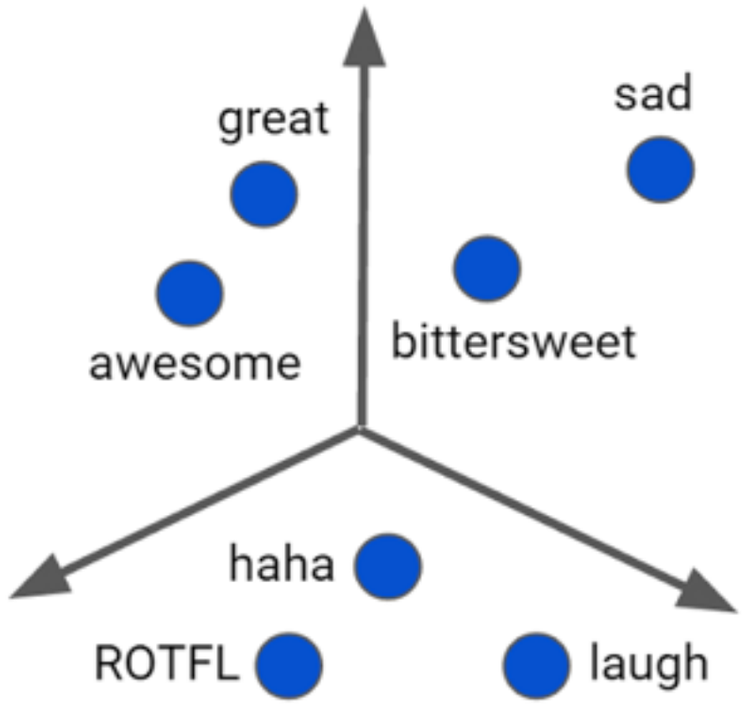
using contextual similarity

use **word embeddings** instead of the sparse vector

dense vectors

vectors of real numbers
that represent words
s.t. similar words have
similar representations

great:	0.5	-0.3	1.5	…
awesome:	0.6	-0.3	1.4	…
sad:	-0.7	0.5	-0.1	…



Vector Space Models

- e.g., vector space models to capture word meanings "**you shall know a word by the company it keeps (Firth, J. R. 1957:11)**"

pasta, lamb,
cheese, mushroom

citrus, apple,
orange, lime

aromatic, nose,
scent, perfume

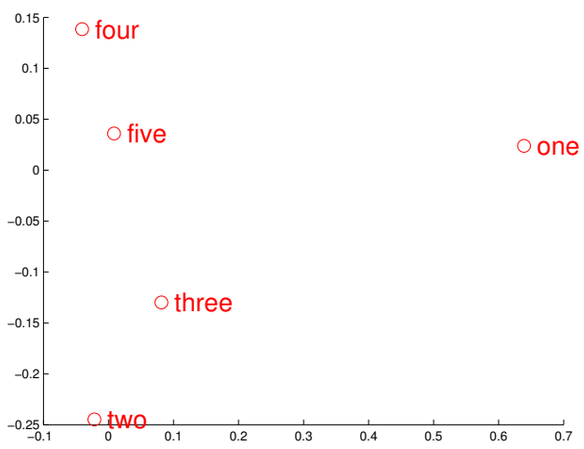
http://methodmatters.blogspot.com/2017/11/using-word2vec-to-analyze-word.html

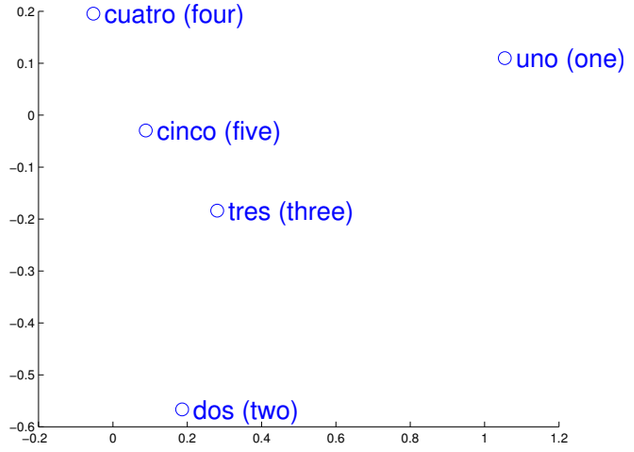
Bilingual Lexicon Induction

using contextual similarity

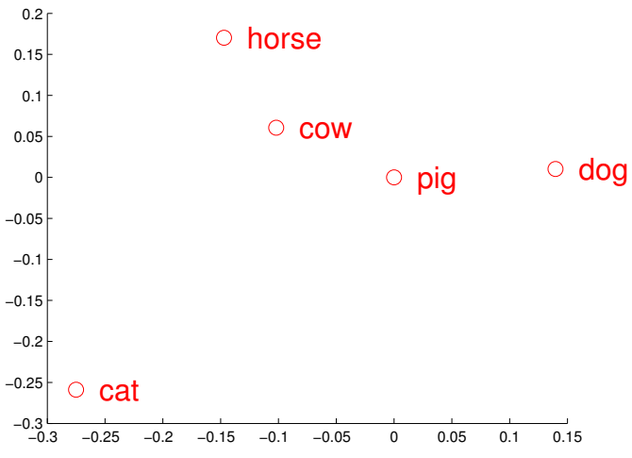
- Word vector representations in different languages might have similar geometric arrangements (Mikolov, T., Le, Q.V. and Sutskever, I., 2013)

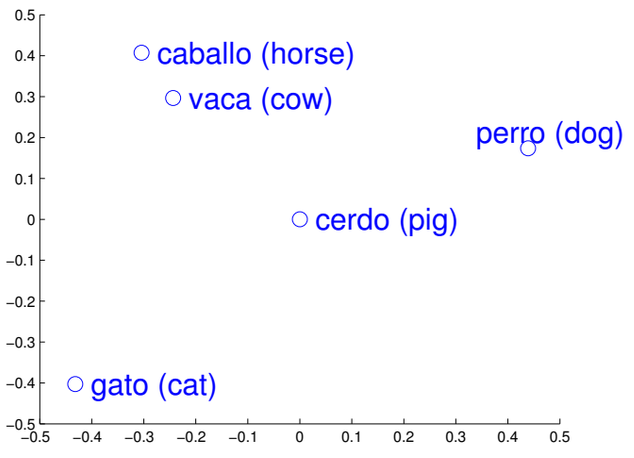
english





spanish

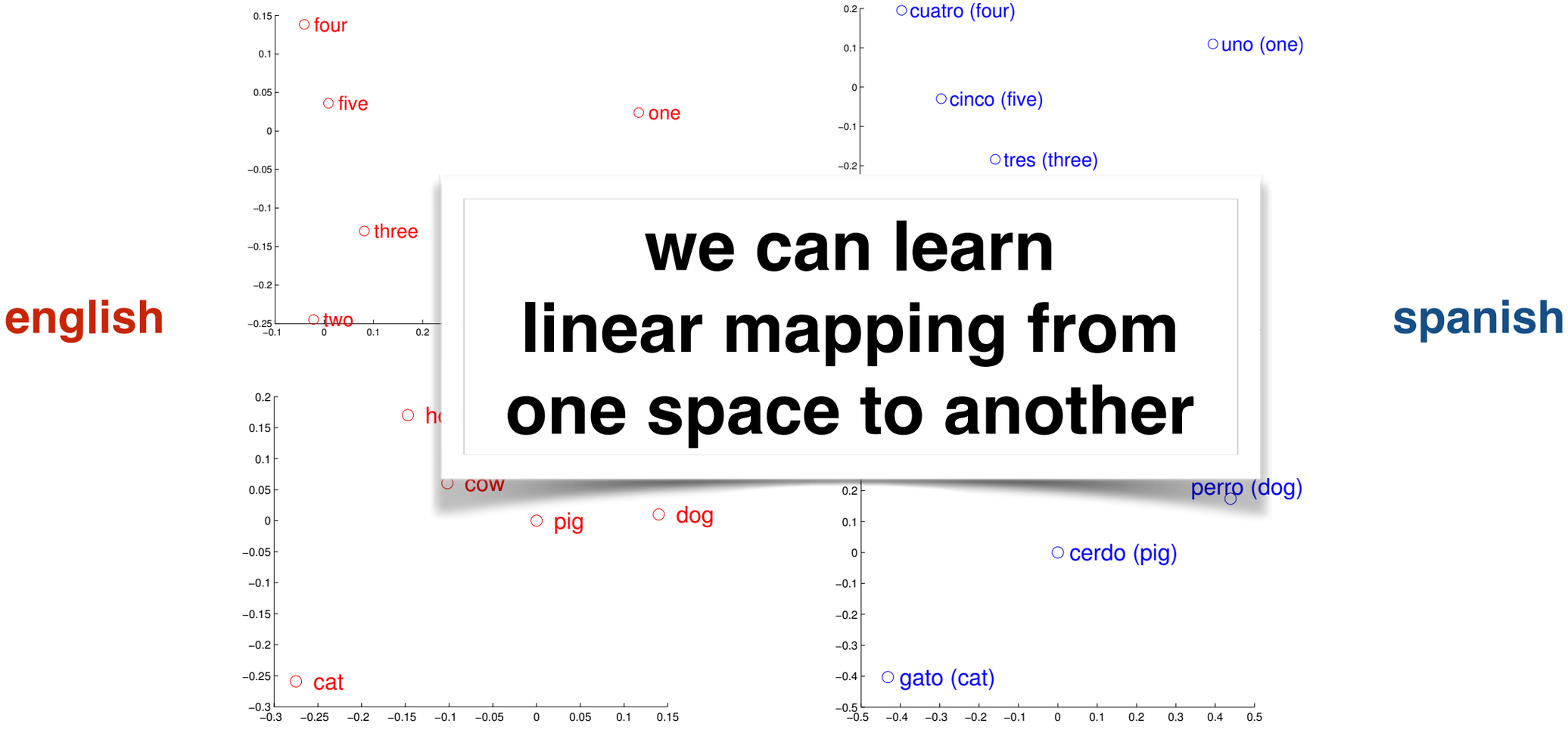




Bilingual Lexicon Induction

using contextual similarity

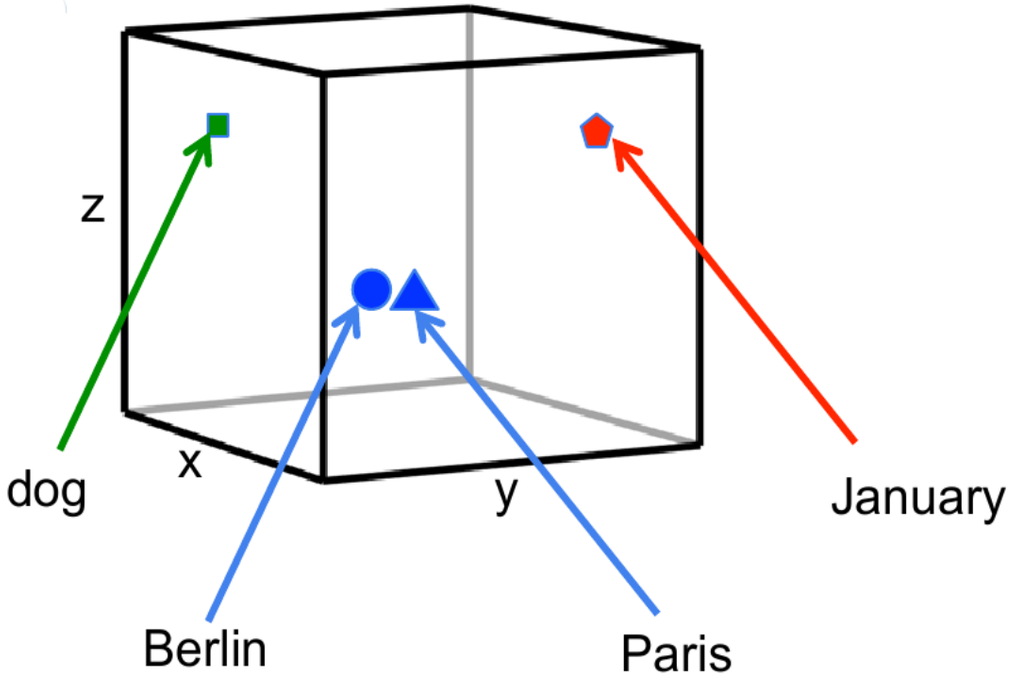
- Word vector representations in different languages might have similar geometric arrangements (Mikolov, T., Le, Q.V. and Sutskever, I., 2013)



Bilingual Word Embeddings

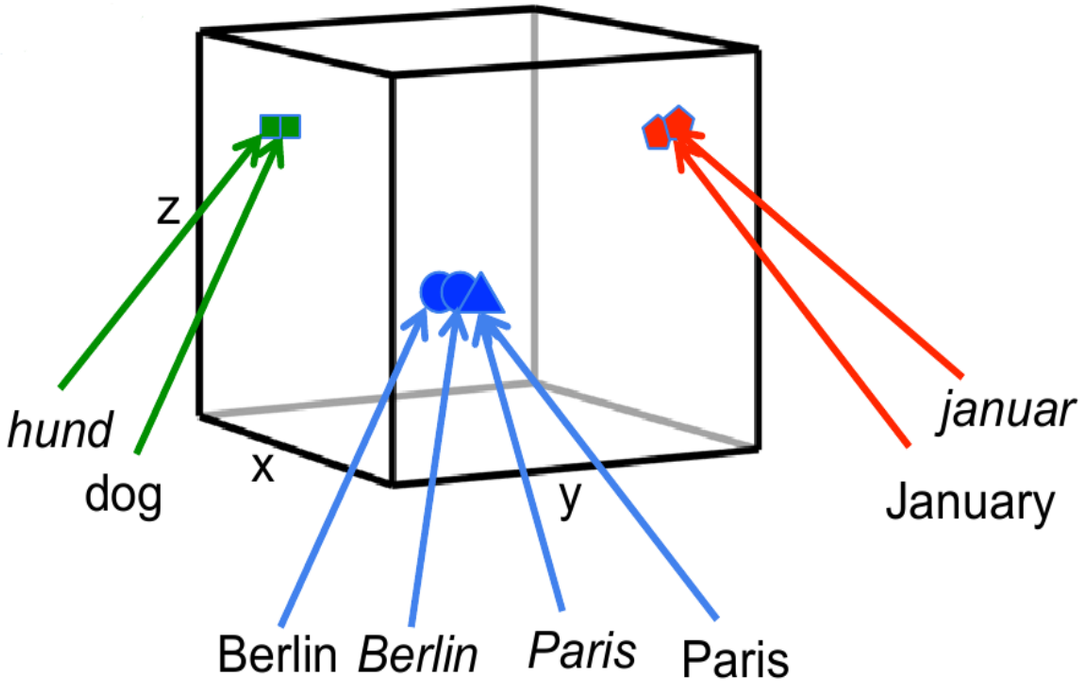
using contextual similarity

such that similar words are close together
in the **bilingual space**



Monolingual

vs



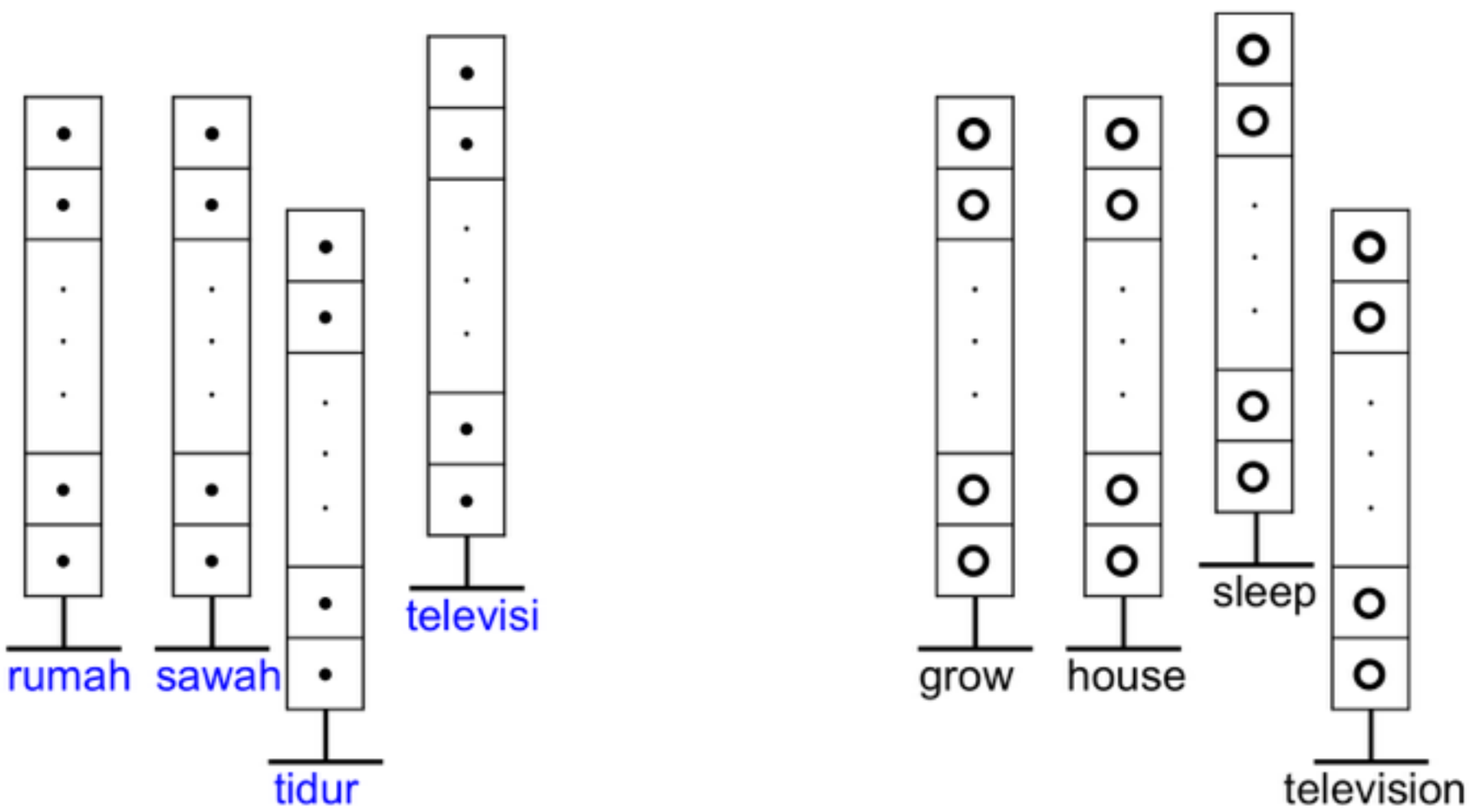
Bilingual

[Gouws, 2015]

Bilingual Lexicon Induction

using contextual similarity

start from the **monolingual** word vectors

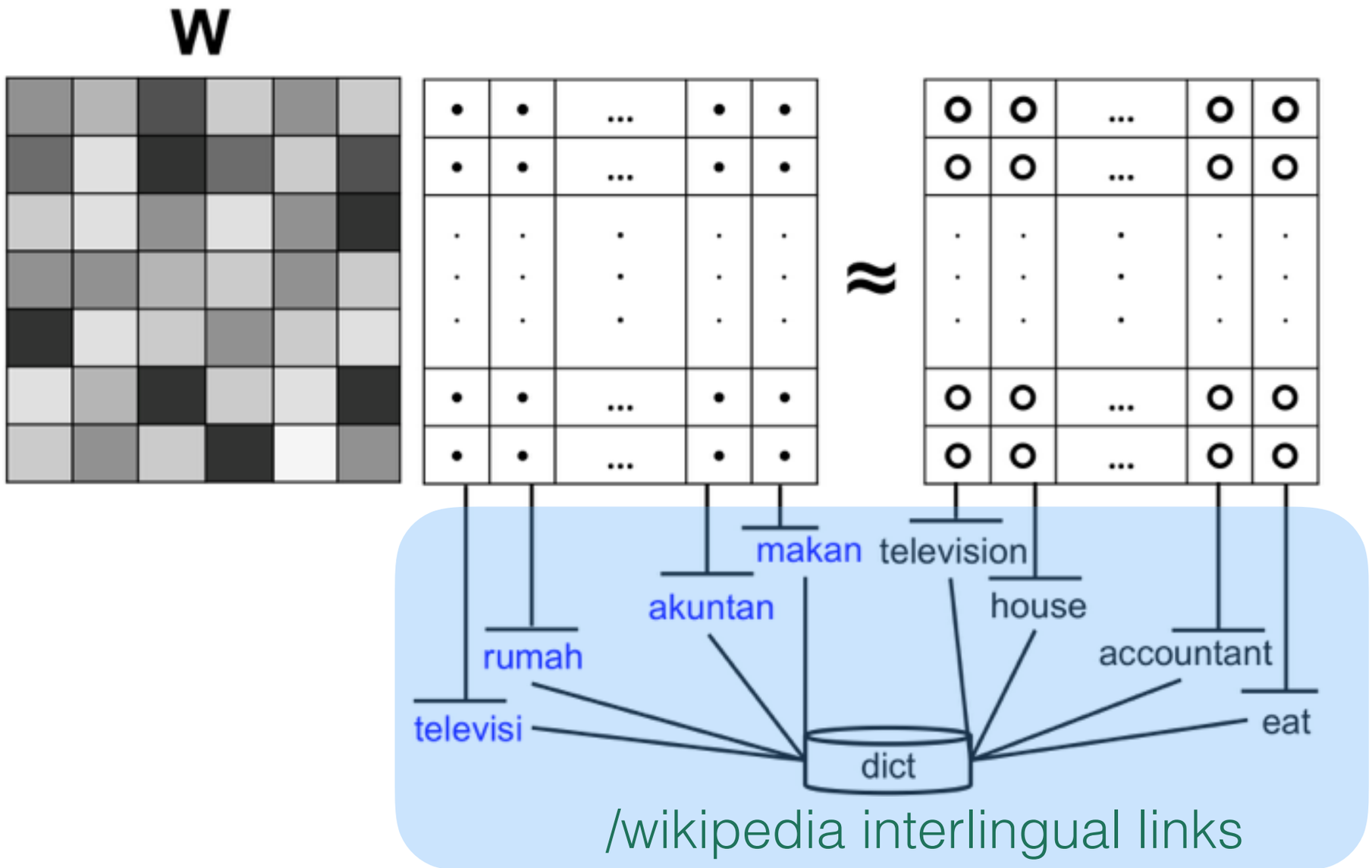


Bilingual Lexicon Induction

using contextual similarity

learn the mapping W between vector spaces

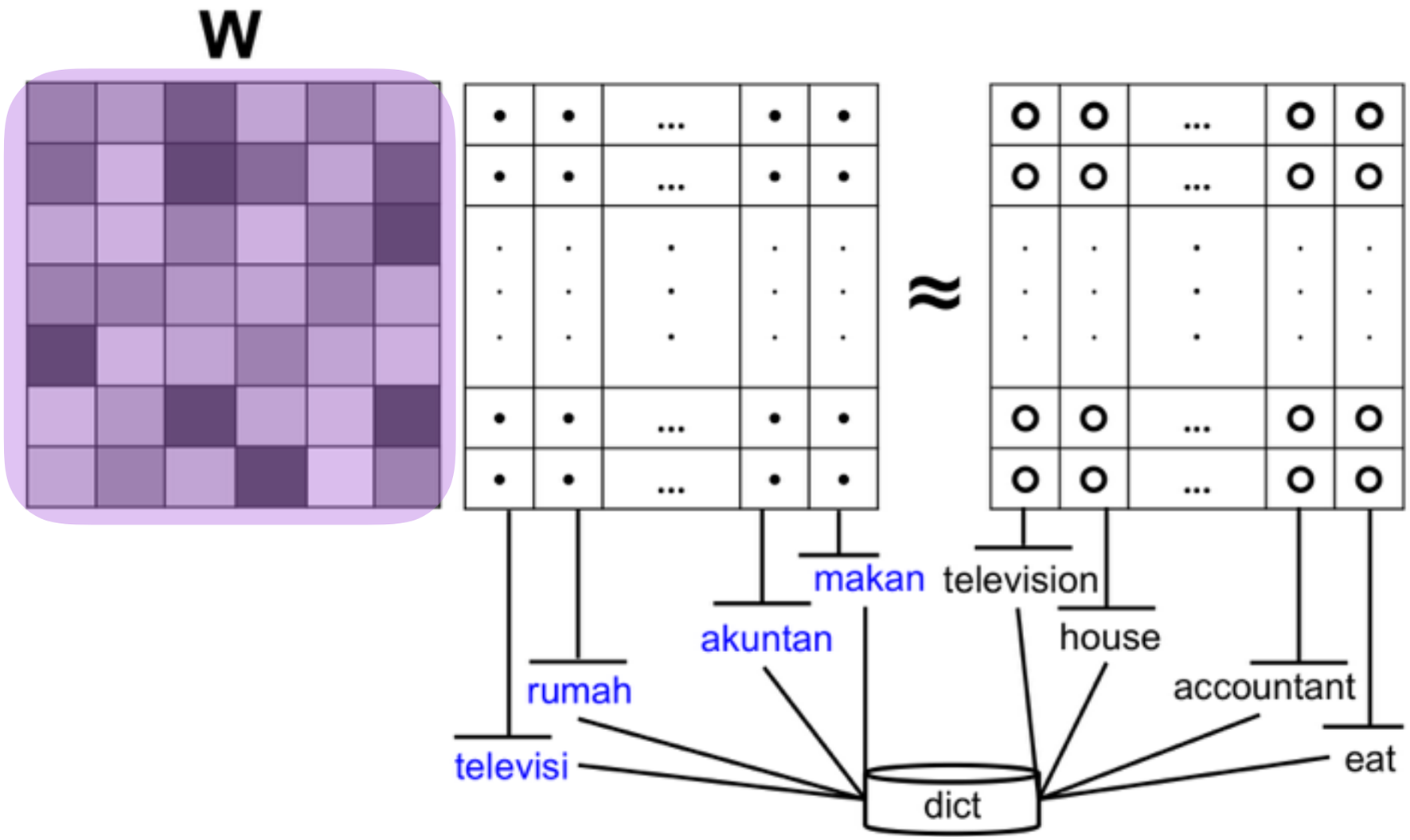
using seed dictionary



Bilingual Lexicon Induction

using contextual similarity

learn the **mapping W** between vector spaces



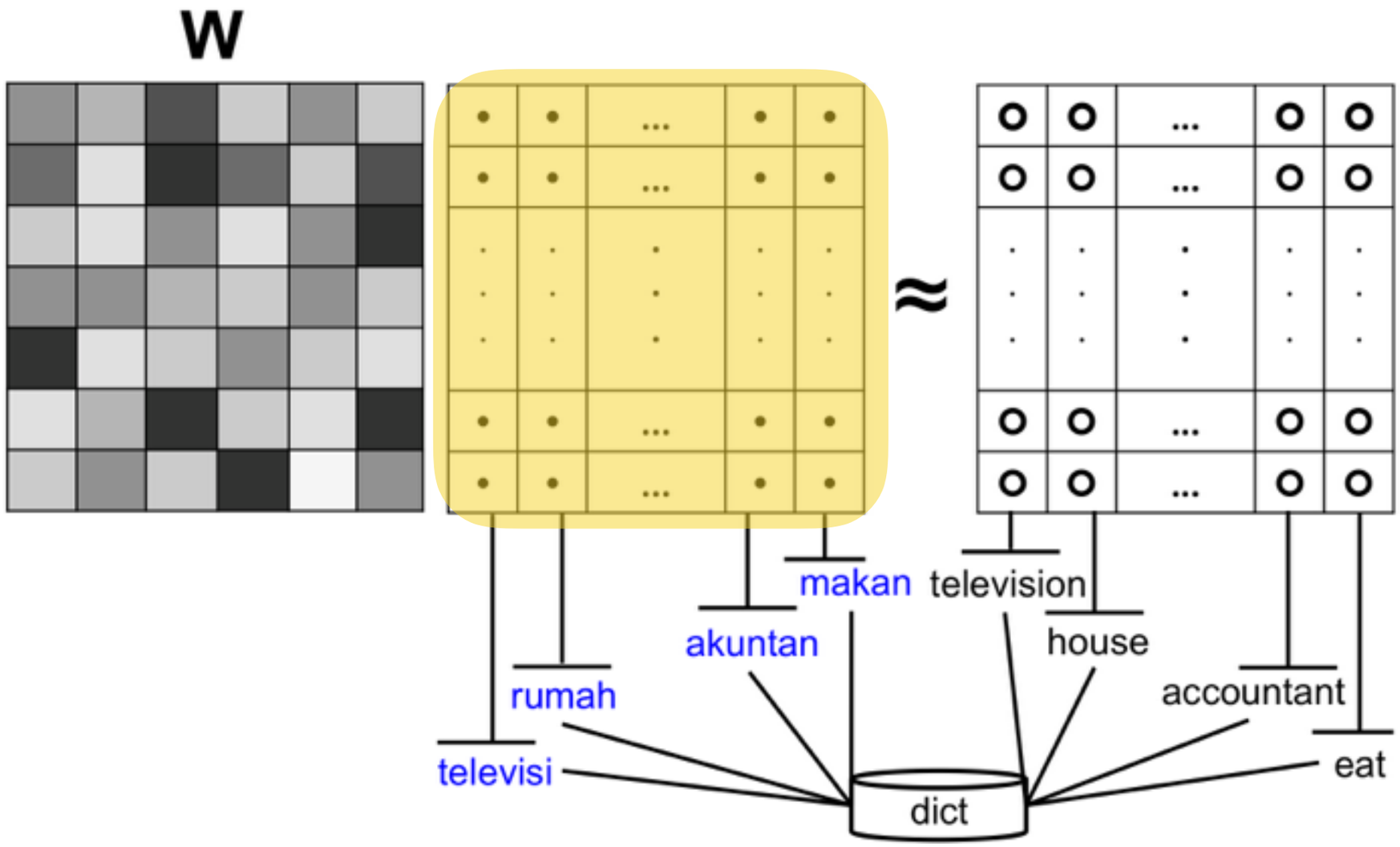
wikipedia interlingual links

Bilingual Lexicon Induction

using contextual similarity

learn the mapping W between vector spaces

when applied to the vectors of the source language words



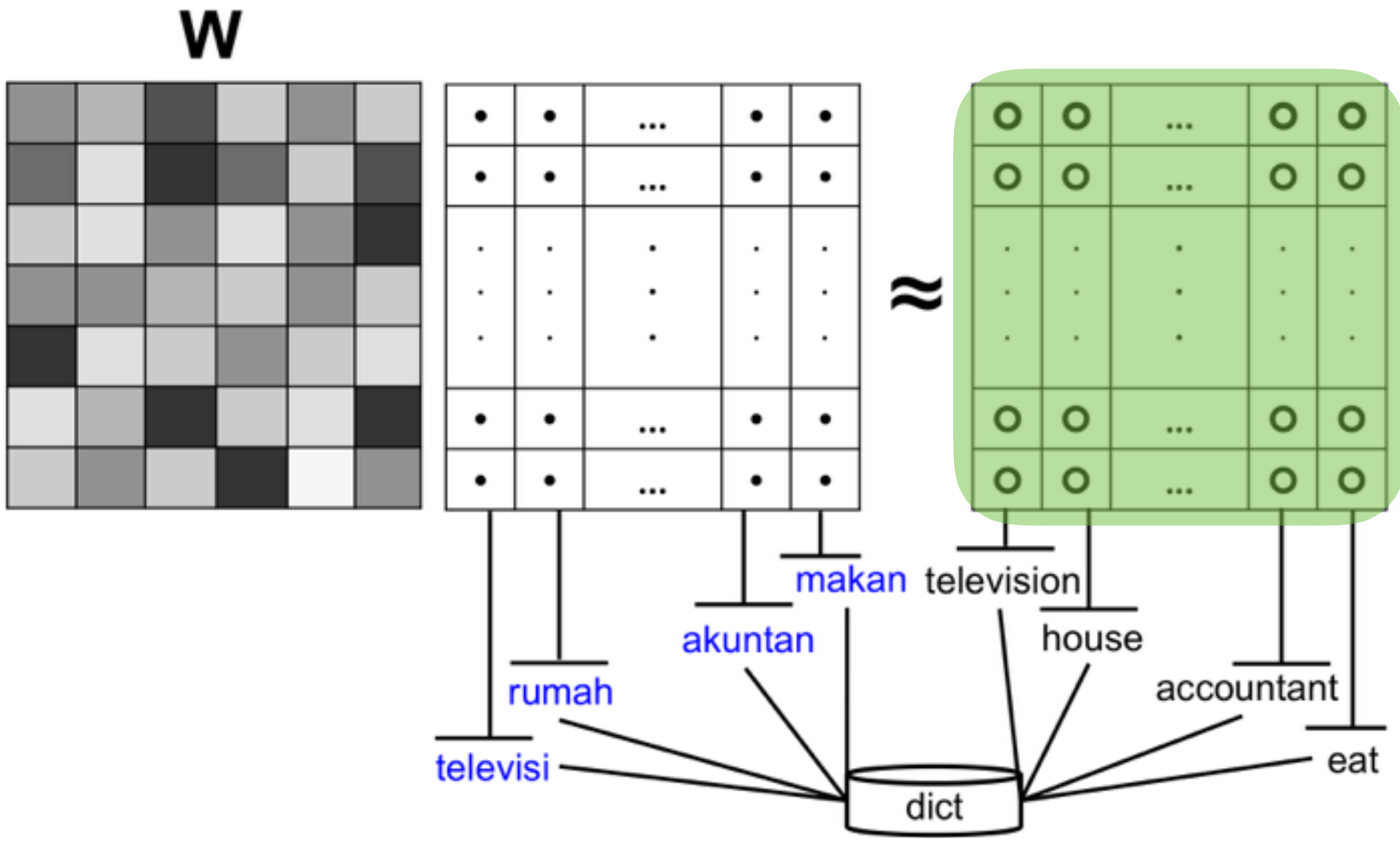
wikipedia interlingual links

Bilingual Lexicon Induction

using contextual similarity

learn the mapping W between vector spaces

will output the vectors of their translations in the target language

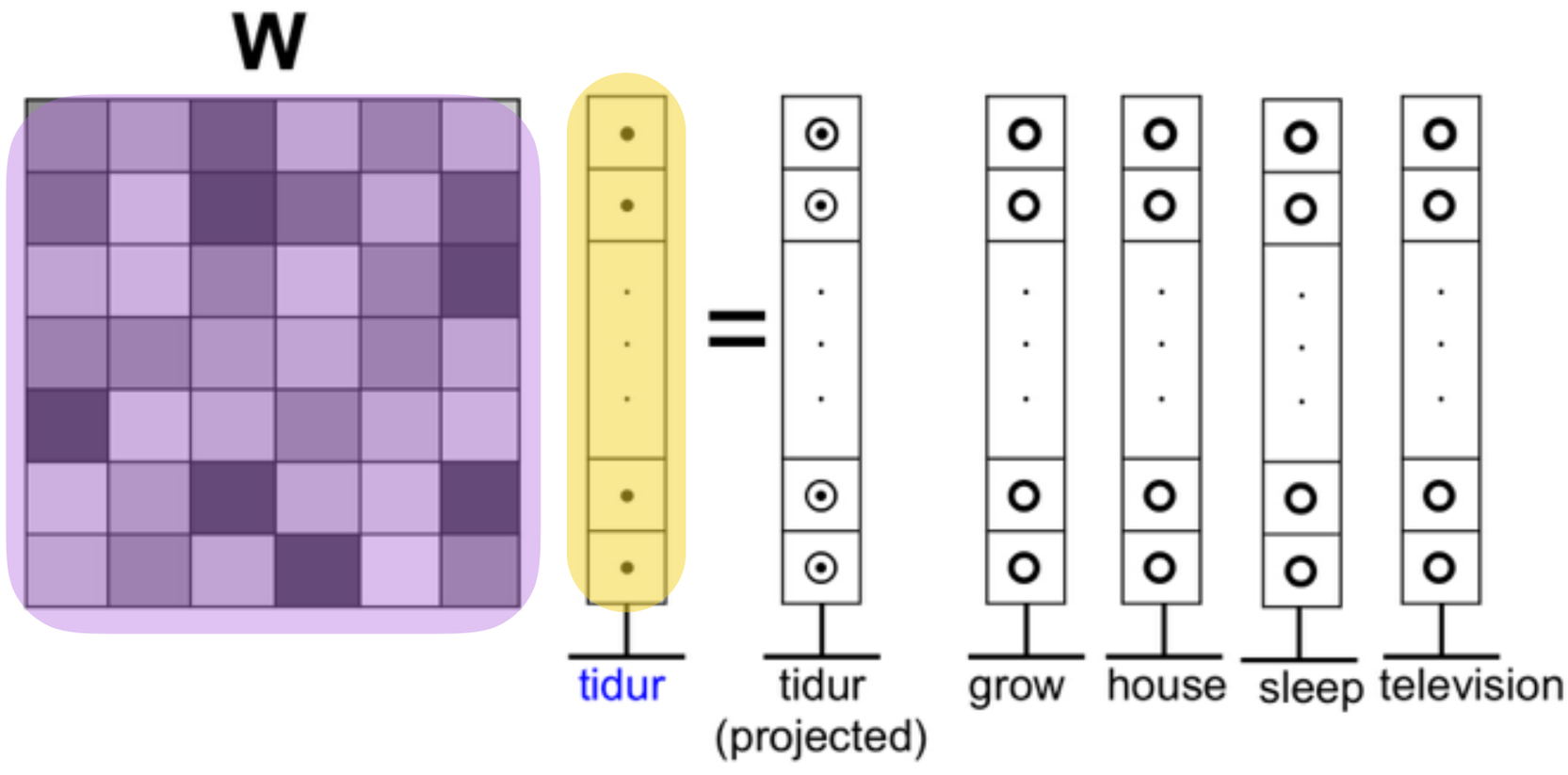


wikipedia interlingual links

Bilingual Lexicon Induction

using contextual similarity

once we learn the **mapping W**,
given the **vector of any source word**

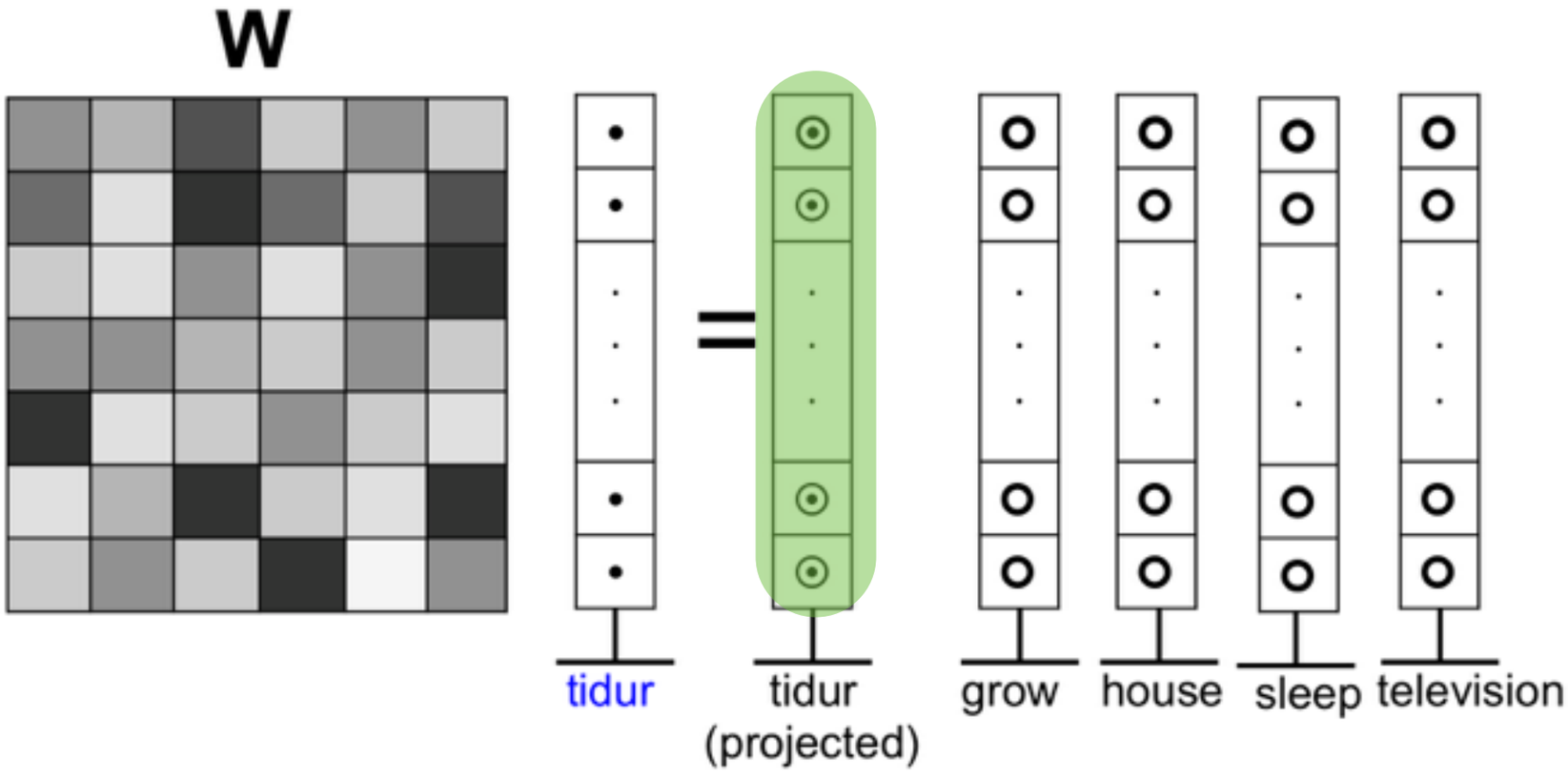


Bilingual Lexicon Induction

using contextual similarity

we can project its vector to

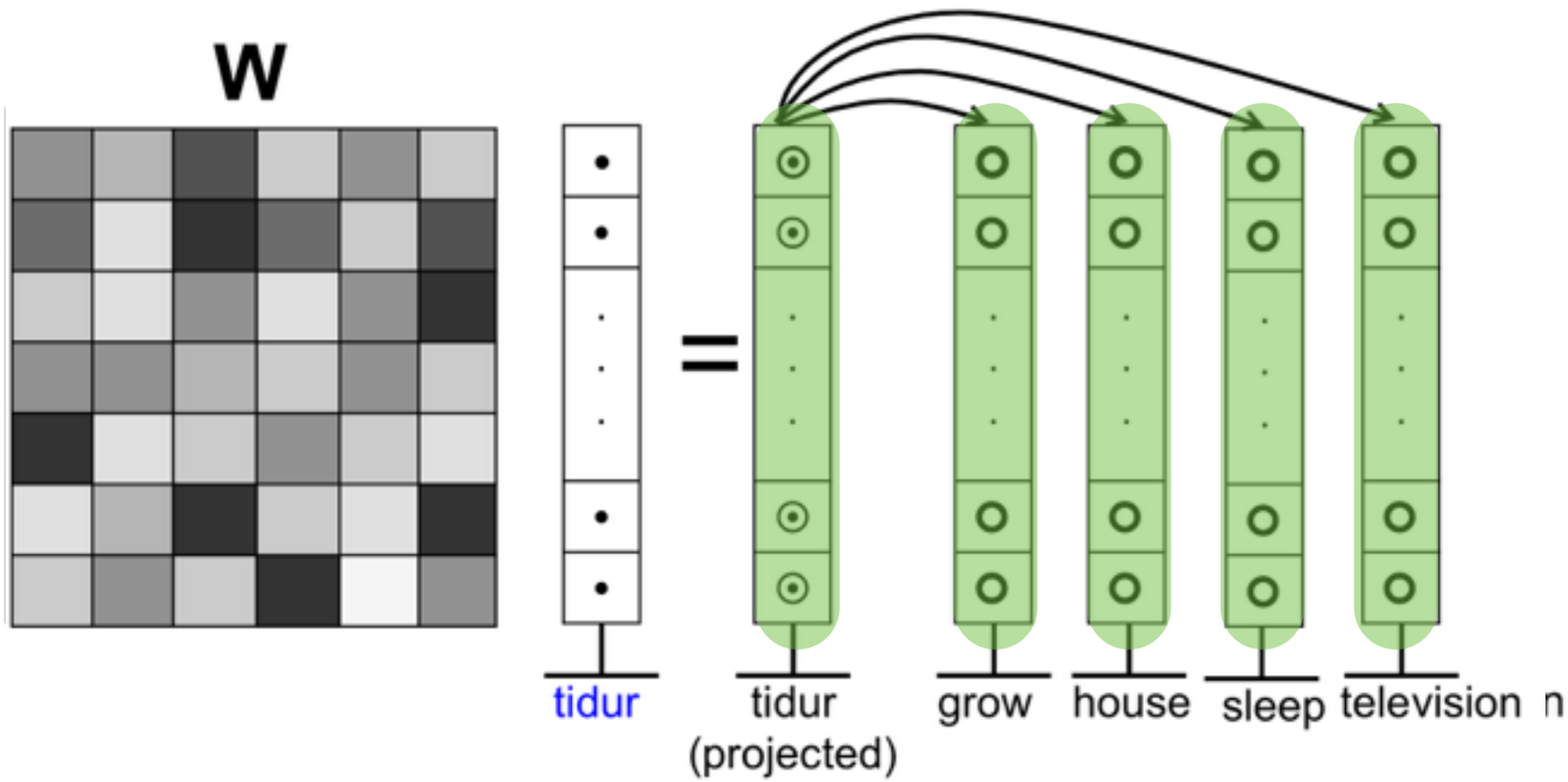
the vector space of the target language



Bilingual Lexicon Induction

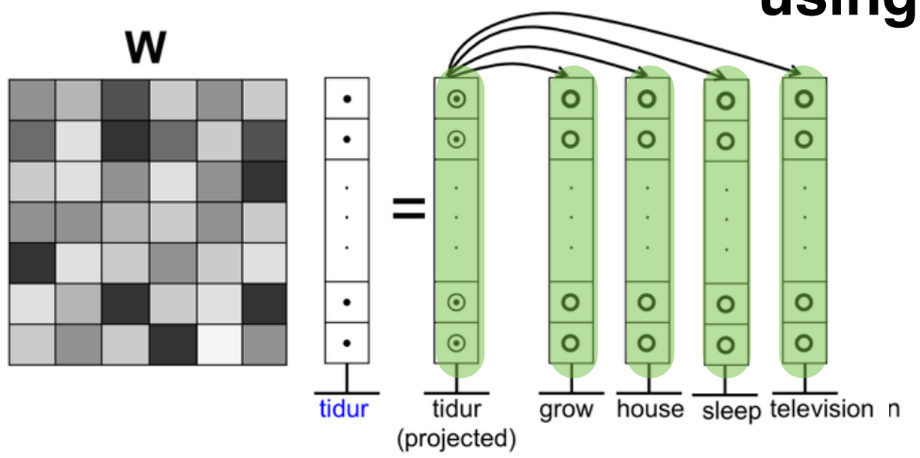
using contextual similarity

and compare the **projected vectors** to get translations



Bilingual Lexicon Induction

using contextual similarity

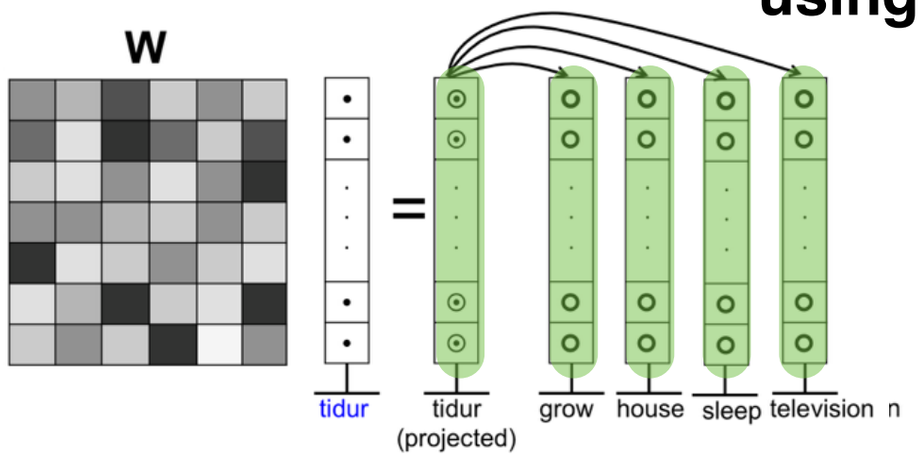


- **linear** mapping does better on closely related languages (English-Spanish)

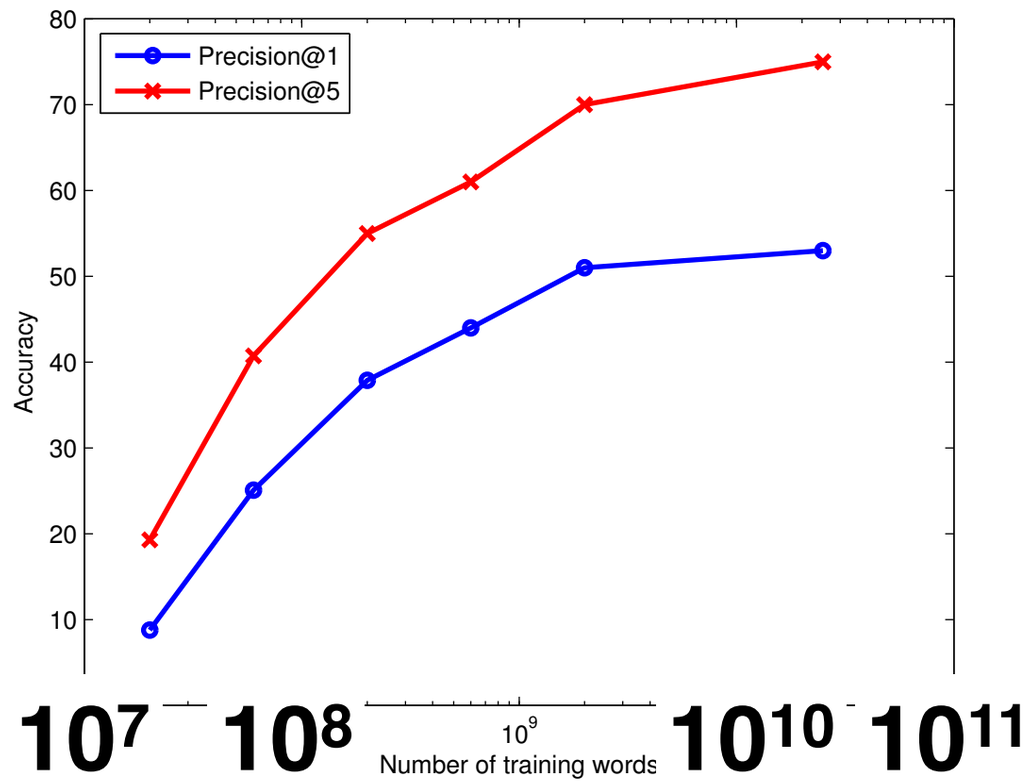
Translations	P@1	P@5
Spanish to English	44%	62%
Czech to English	25%	45%
Vietnamese to English	24%	40%

Bilingual Lexicon Induction

using contextual similarity



- also depends on size of monolingual corpus



Bilingual Lexicon Induction

for distant languages:

- Non-linear mapping between the language spaces

mapping W

$$\sum_{\mathbf{x}_f \in \mathbf{X_F}} \sum_{\mathbf{x}_e \in \mathbf{X_E}} ||\mathbf{x}_f - \phi^{(4)} s(\phi^{(3)} s(\phi^{(2)} s(\phi^{(1)} \mathbf{x}_e)))||^2$$

s = tanh

for lack of monolingual texts:

- Add other signals of similarities
  - related languages
  - images

Bilingual Lexicon Induction

using related languages

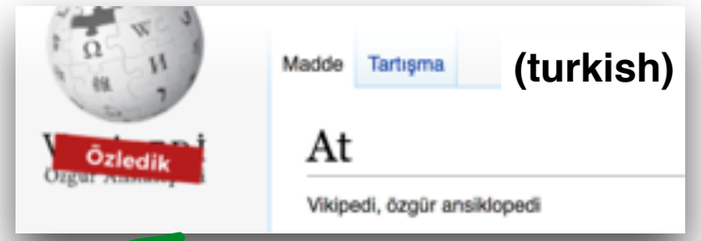
ئات

(uyghur)

source language

(low resource)

e.g. **uyghur**



horse

(english)

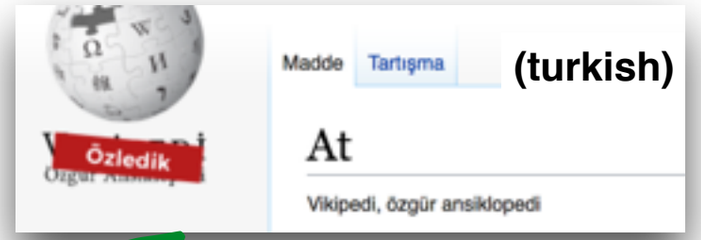
...

Bilingual Lexicon Induction

using related languages

ئات

(uyghur)



horse

(english)

source language

(low resource)

e.g. **uyghur**

intermediate language

(higher resource)

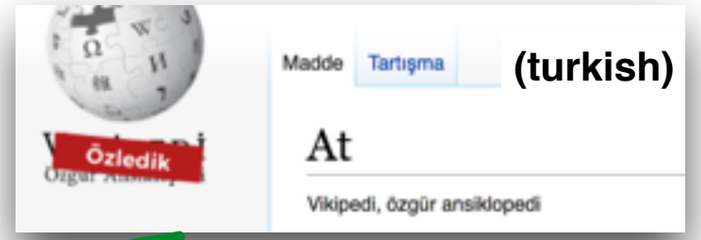
e.g. **uzbek**

Bilingual Lexicon Induction

using related languages

ئات

(uyghur)



horse

(english)

source language

(low resource)

e.g. **uyghur**

intermediate language

(higher resource)

e.g. **uzbek**

target language

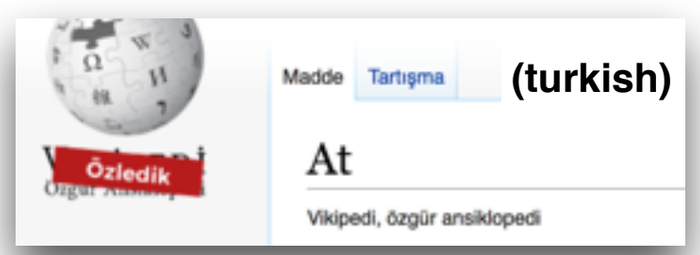
e.g. **english**

Bilingual Lexicon Induction

using images

ئات

(uyghur)



(turkish)

horse

(english)

...

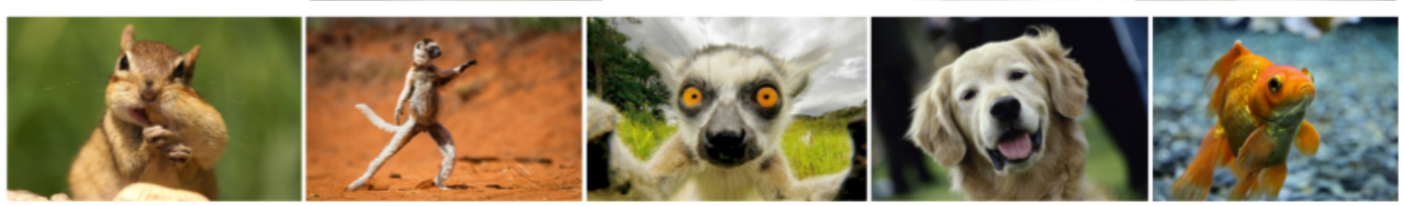
Bilingual Lexicon Induction

using images

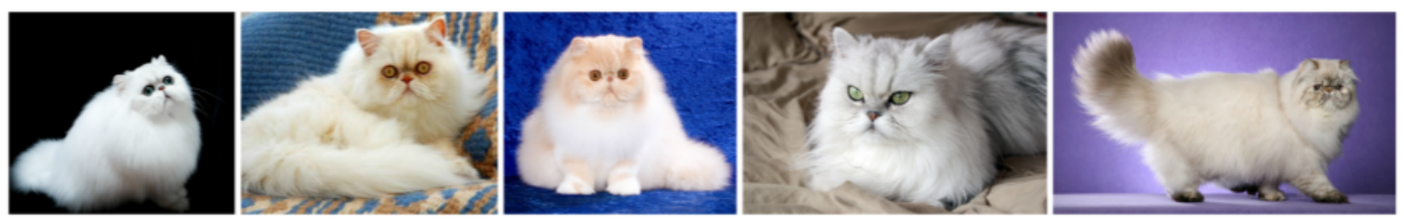
kucing ⇨

cat ⇨

animal ⇨



persian ⇨



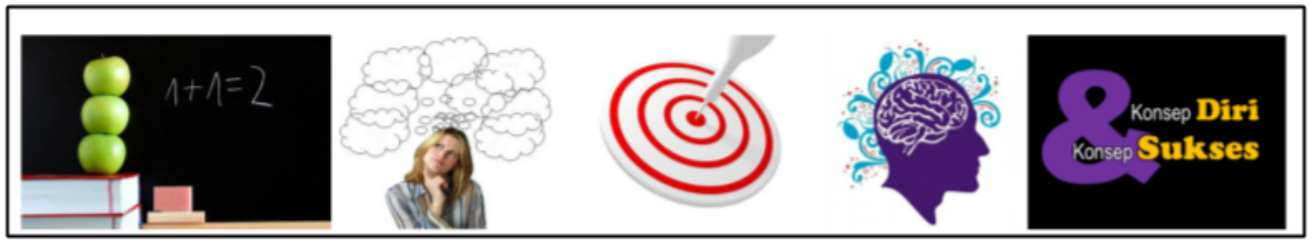
pet ⇨

Identify **translations via images** associated with words in different languages that have a **high degree of visual similarity**

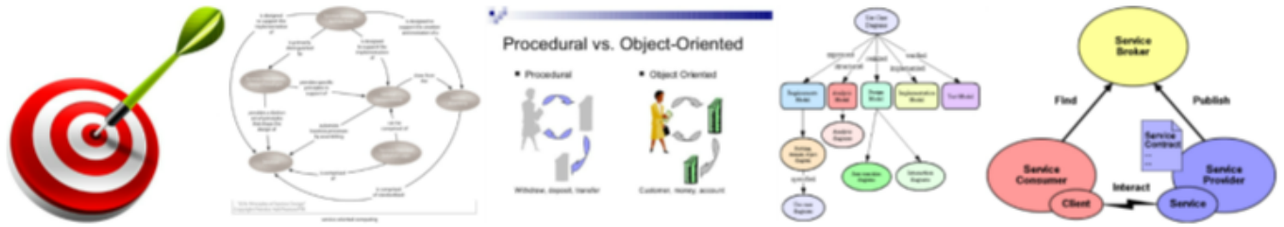
Bilingual Lexicon Induction

using images

konsep ⇨



oriented ⇨



department ⇨

gifted ⇨

top-level ⇨

Bilingual Lexicon Induction

using images

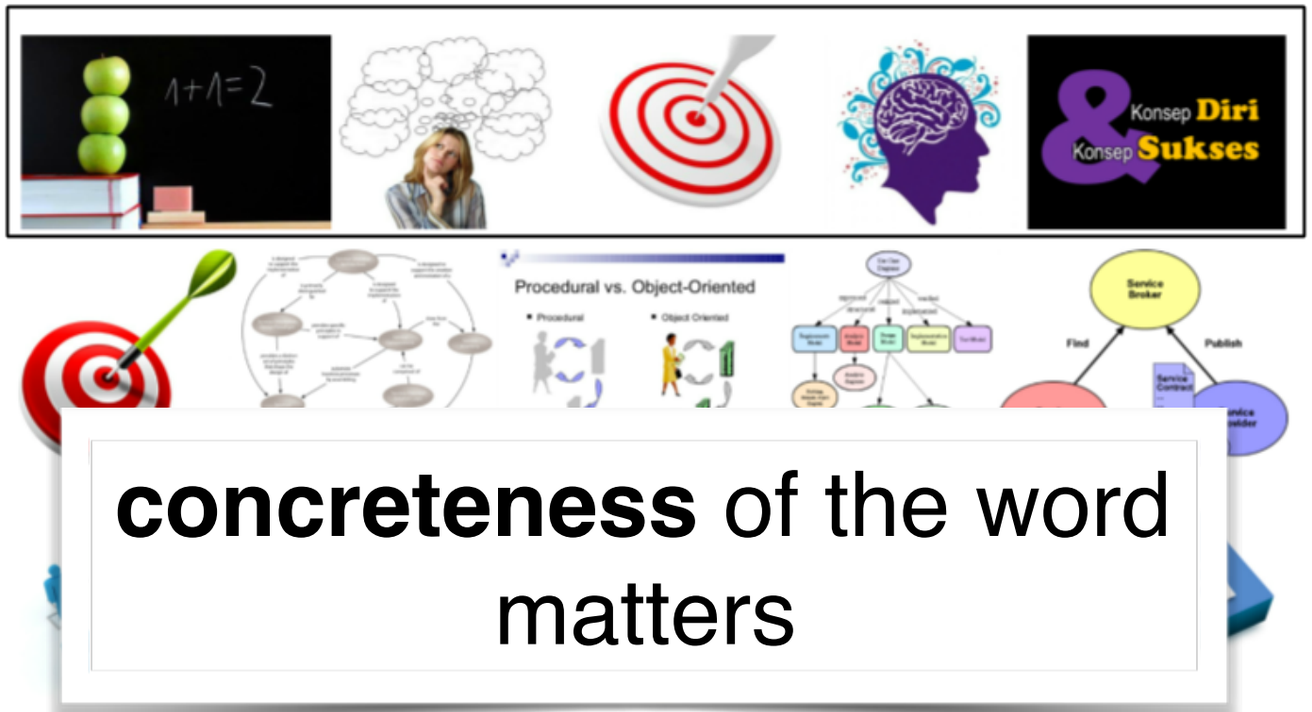
konsep ⇨

oriented ⇨

department ⇨

gifted ⇨

top-level ⇨



Bilingual Lexicon Induction

using images

sepatu

concreteness

Bilingual Lexicon Induction

using images

sepatu

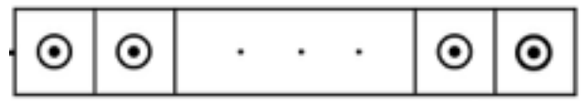
concreteness

image similarity

Bilingual Lexicon Induction

using images

sepatu **sapato zapato**



concreteness

image similarity

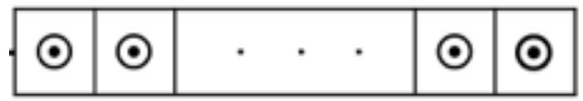
context similarity

+ related languages

Bilingual Lexicon Induction

using images

sepatu **sapato zapato**



concreteness

image similarity

context similarity

+ related languages

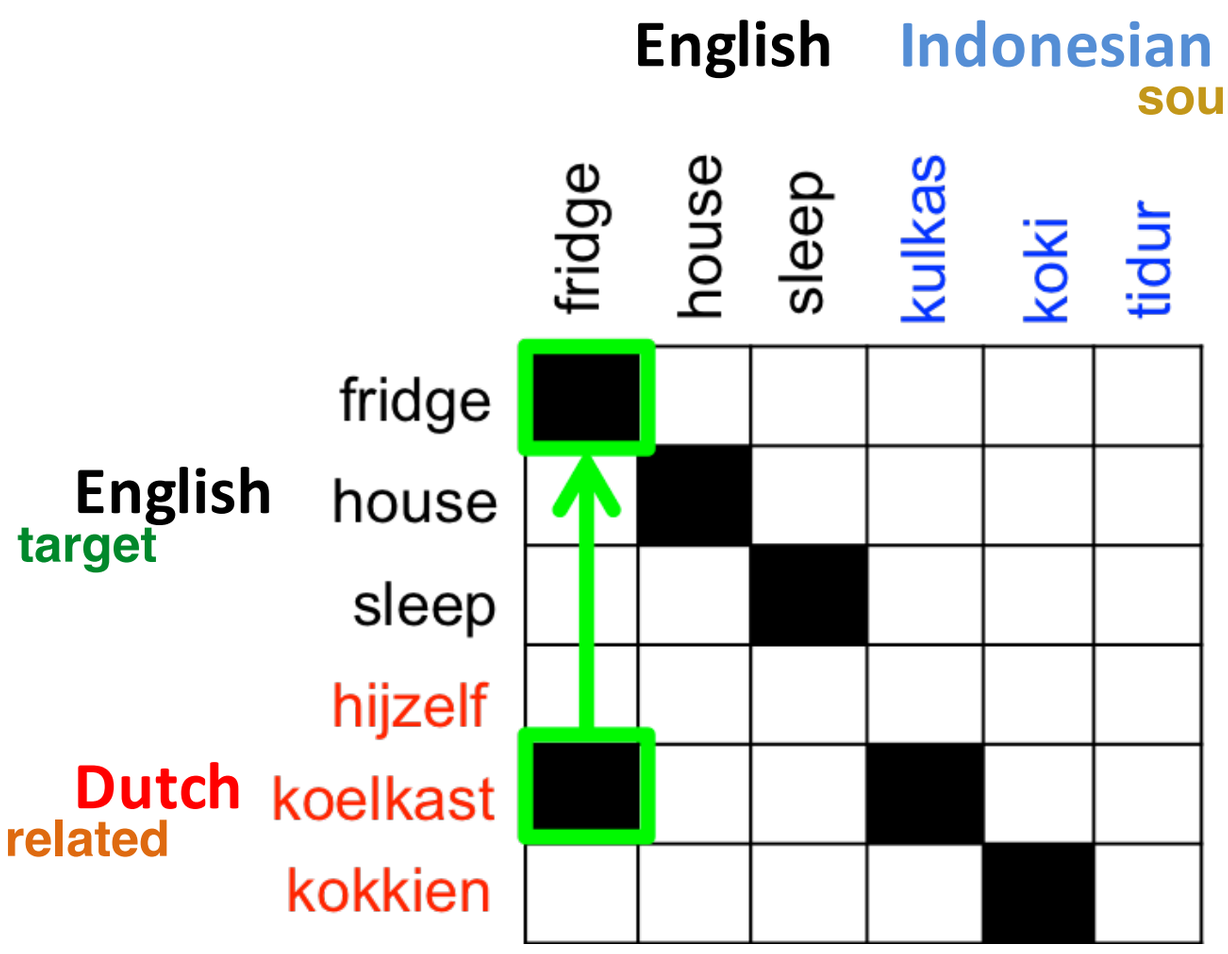
candidate translation: shoes

Bilingual Lexicon Induction

- Non-linear mapping between the language spaces
- Add other signals of similarities
  - related languages
  - images

Translation as Matrix Completion

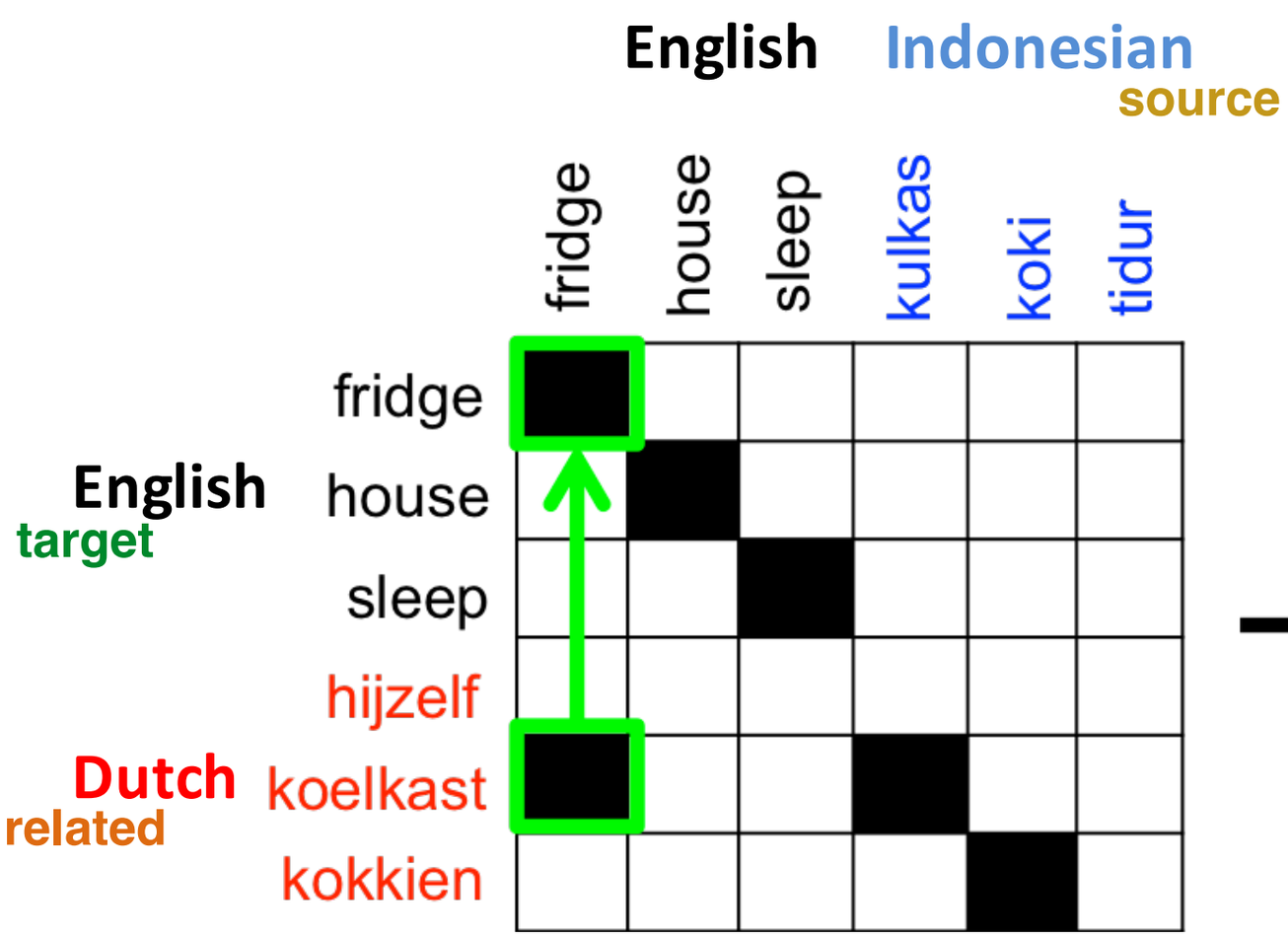
using related languages

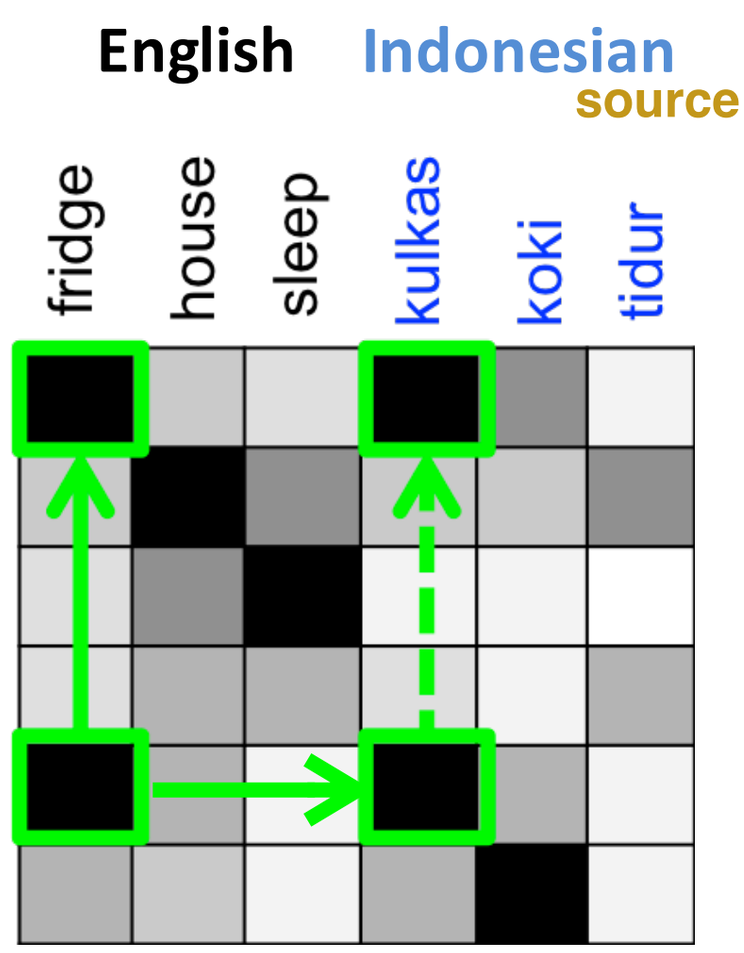


wikipedia interlingual links as **observed** translations

Translation as Matrix Completion

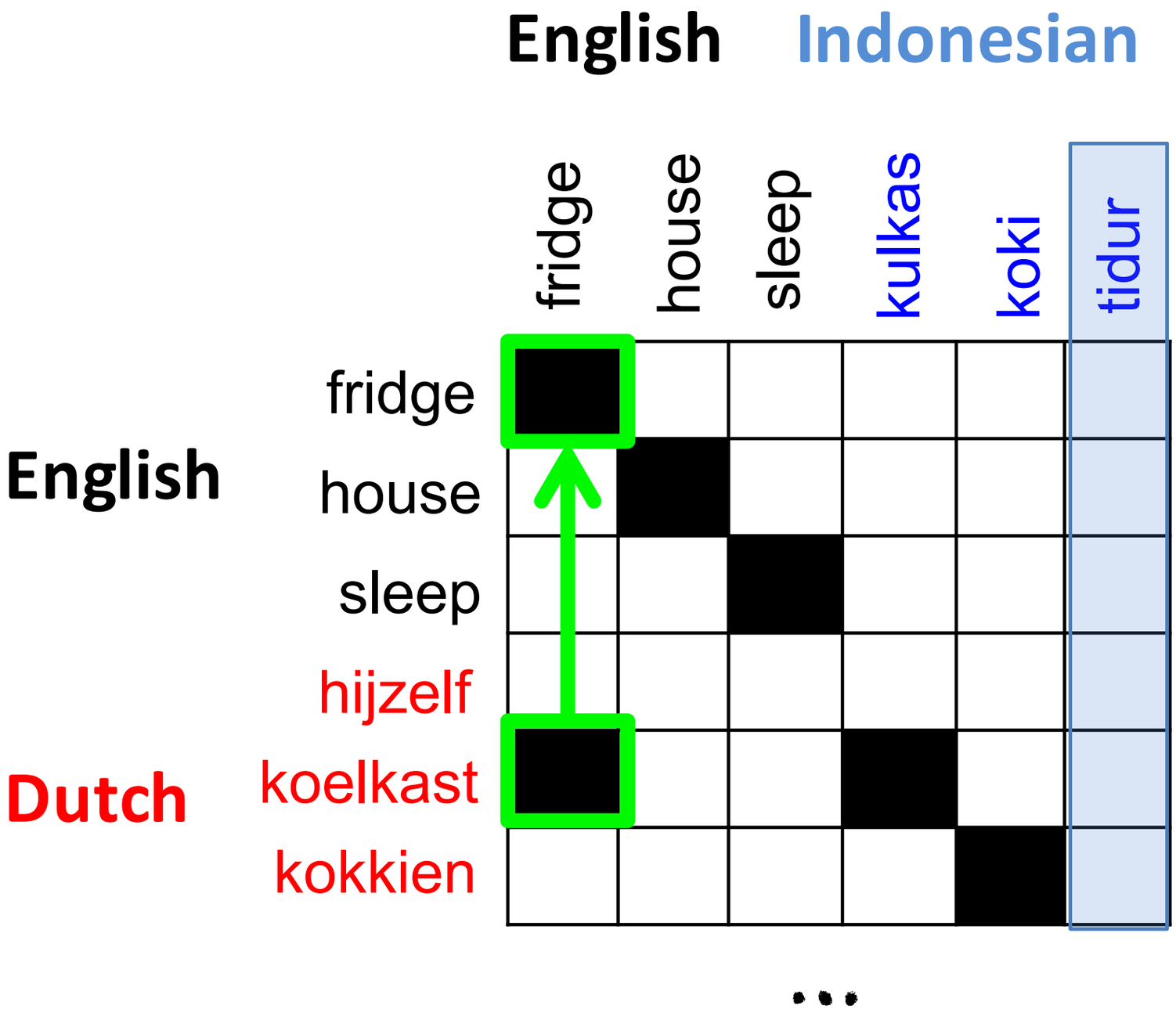
using related languages



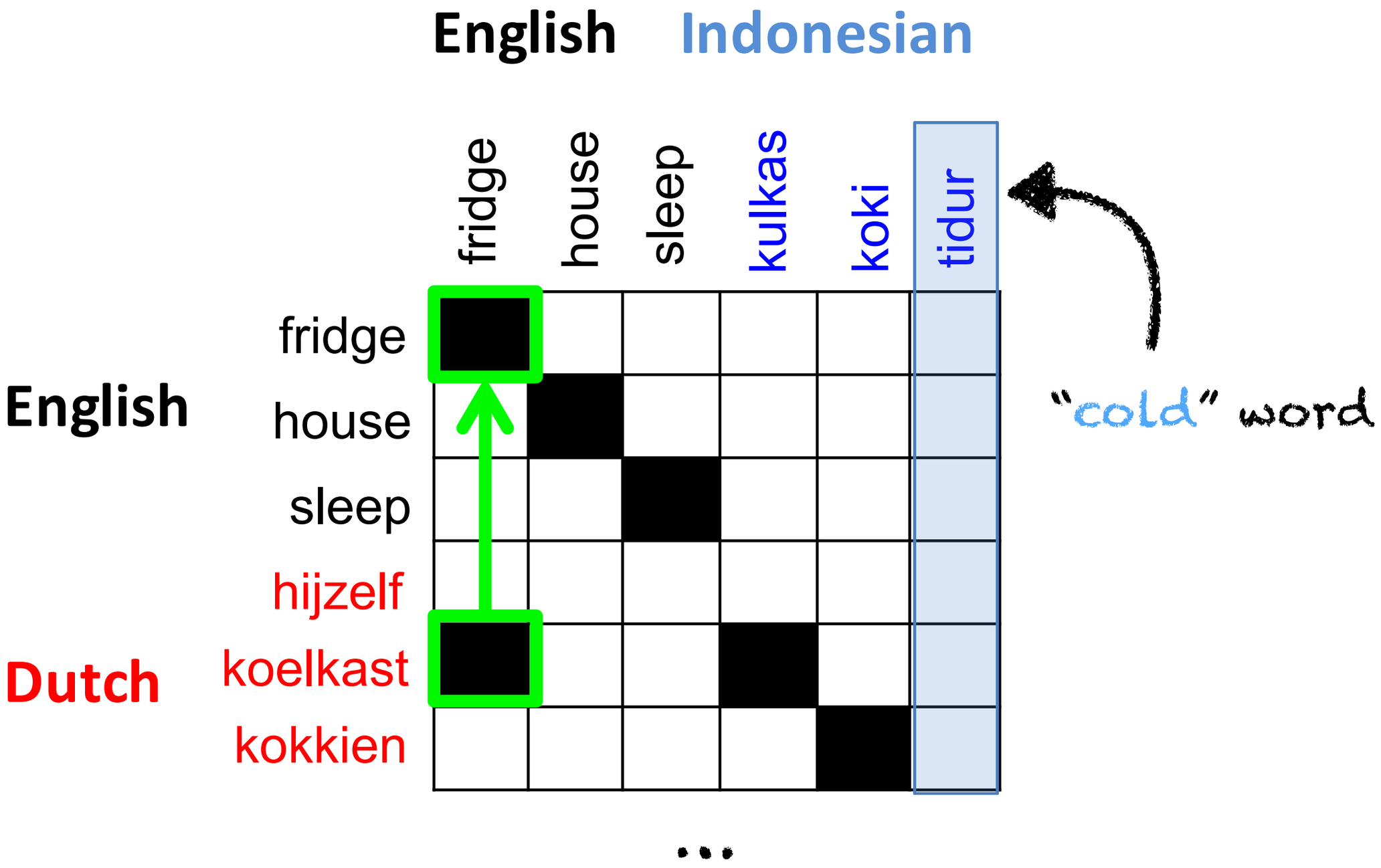


project the **observed** translations

Cold Start Issues



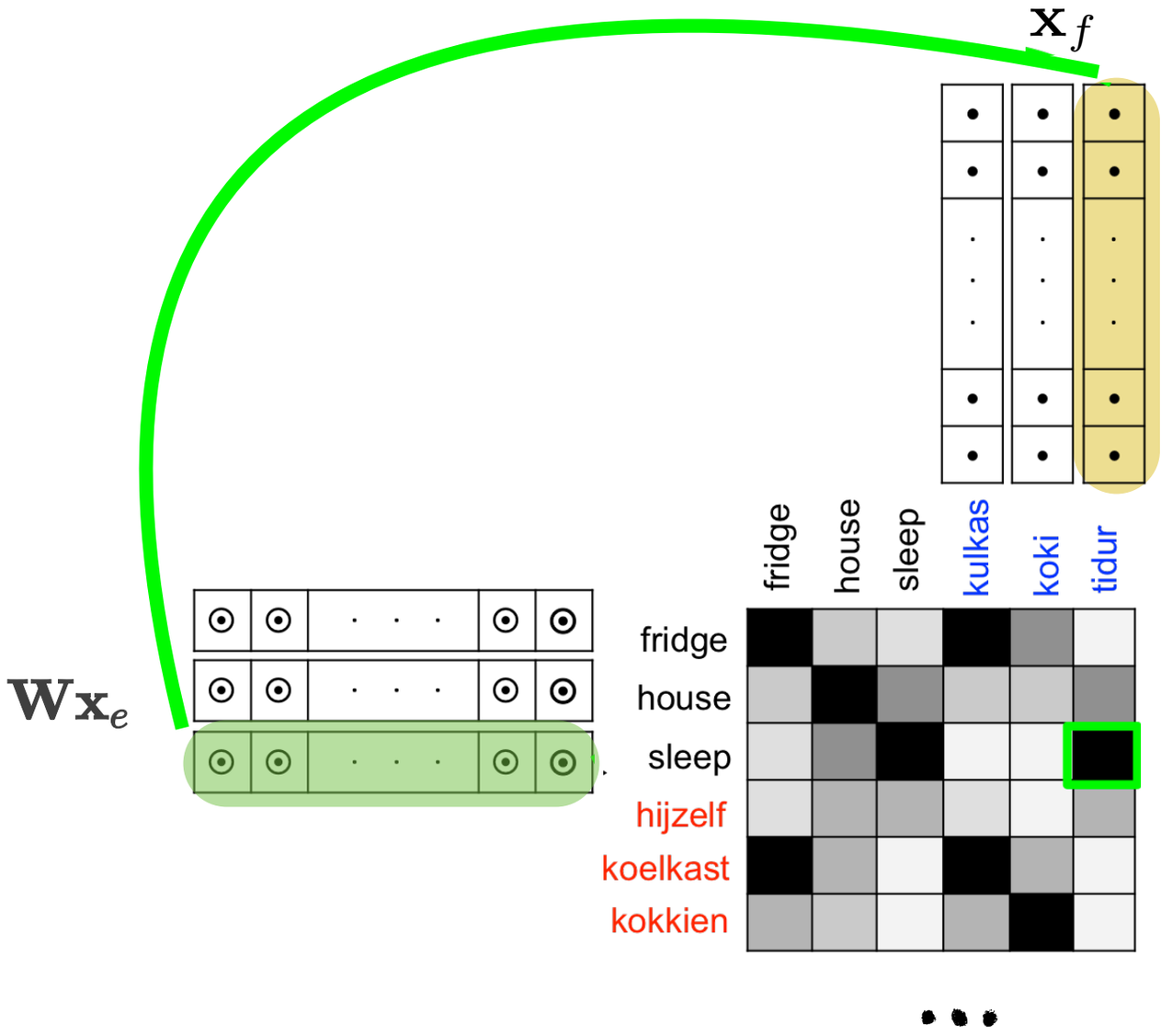
Cold Start Issues



Translation as Matrix Completion

using context similarities

use **bilingual word vectors** as **additional signals** for inferring translations



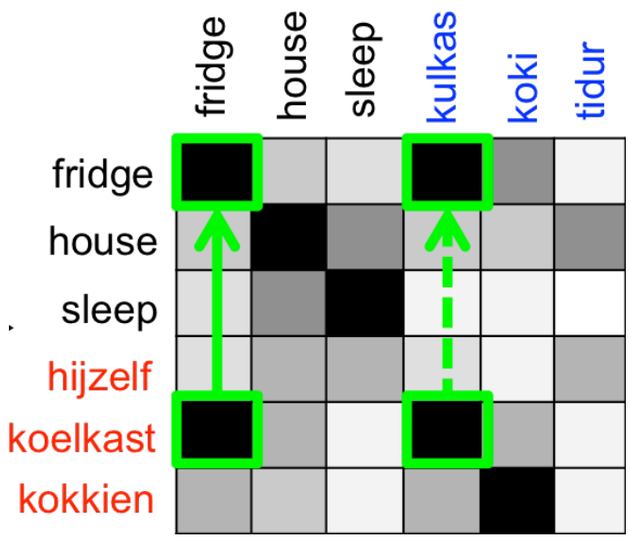
Experiments

back-off to “noisier” signals of similarities
when encountering **cold** words

Experiments

back-off to “noisier” signals of similarities
when encountering **cold** words

use **related** languages

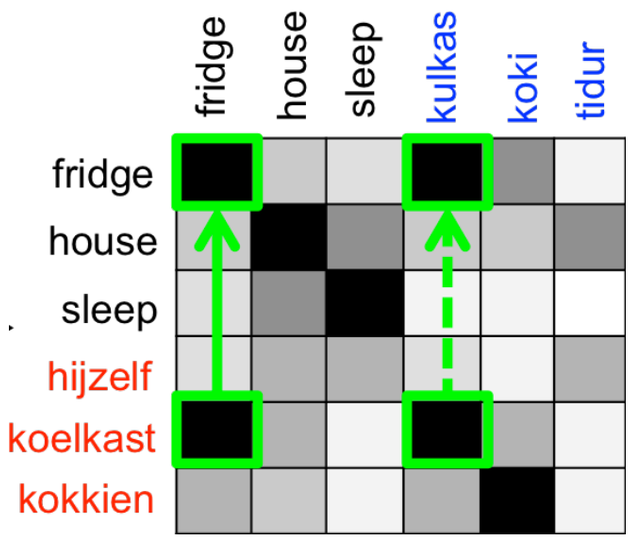


• • •

Experiments

back-off to “noisier” signals of similarities
when encountering **cold** words

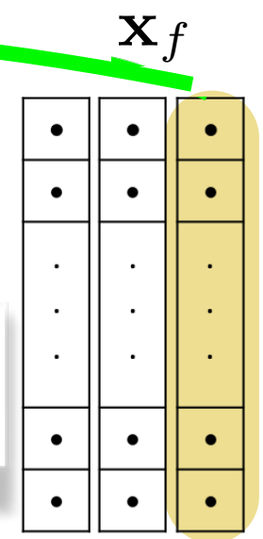
use **related** languages



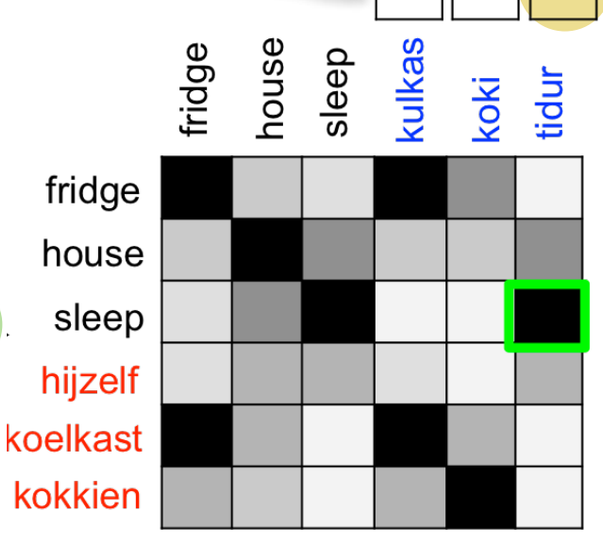
back-off

$\mathbf{W}\mathbf{x}_e$

use **context** similarity



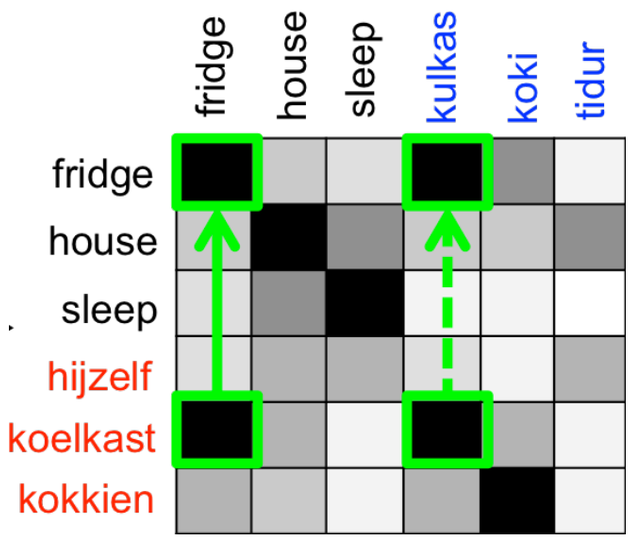
$\mathbf{x}_f$



Experiments

back-off to “noisier” signals of similarities
when encountering **cold** words

use **related** languages



back-off

$\mathbf{W}\mathbf{x}_e$

$\mathbf{x}_f$

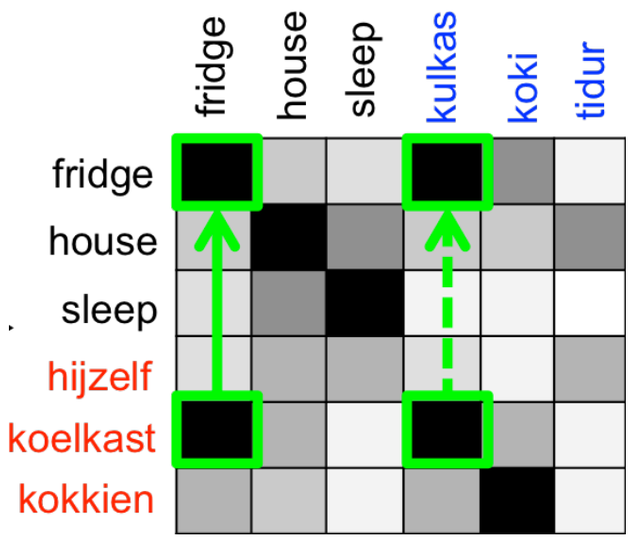
use **context** similarity

use **image** similarity

Experiments

back-off to “noisier” signals of similarities
when encountering **cold** words

use **related** languages



back-off

$\mathbf{W}\mathbf{x}_e$

$\mathbf{x}_f$

use **context** similarity

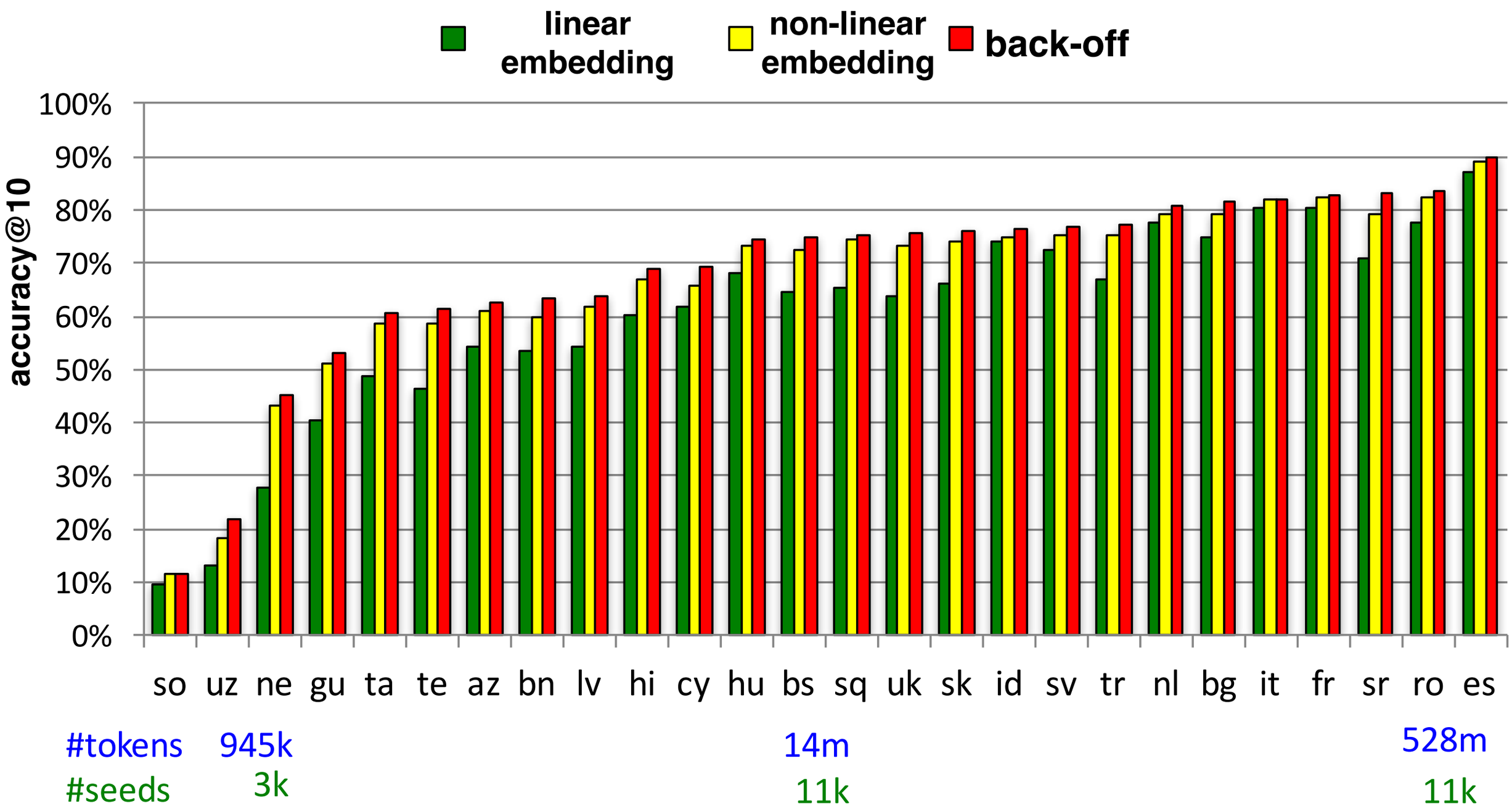
use **image** similarity

use **transliteration**

obama ⟷ ობამა

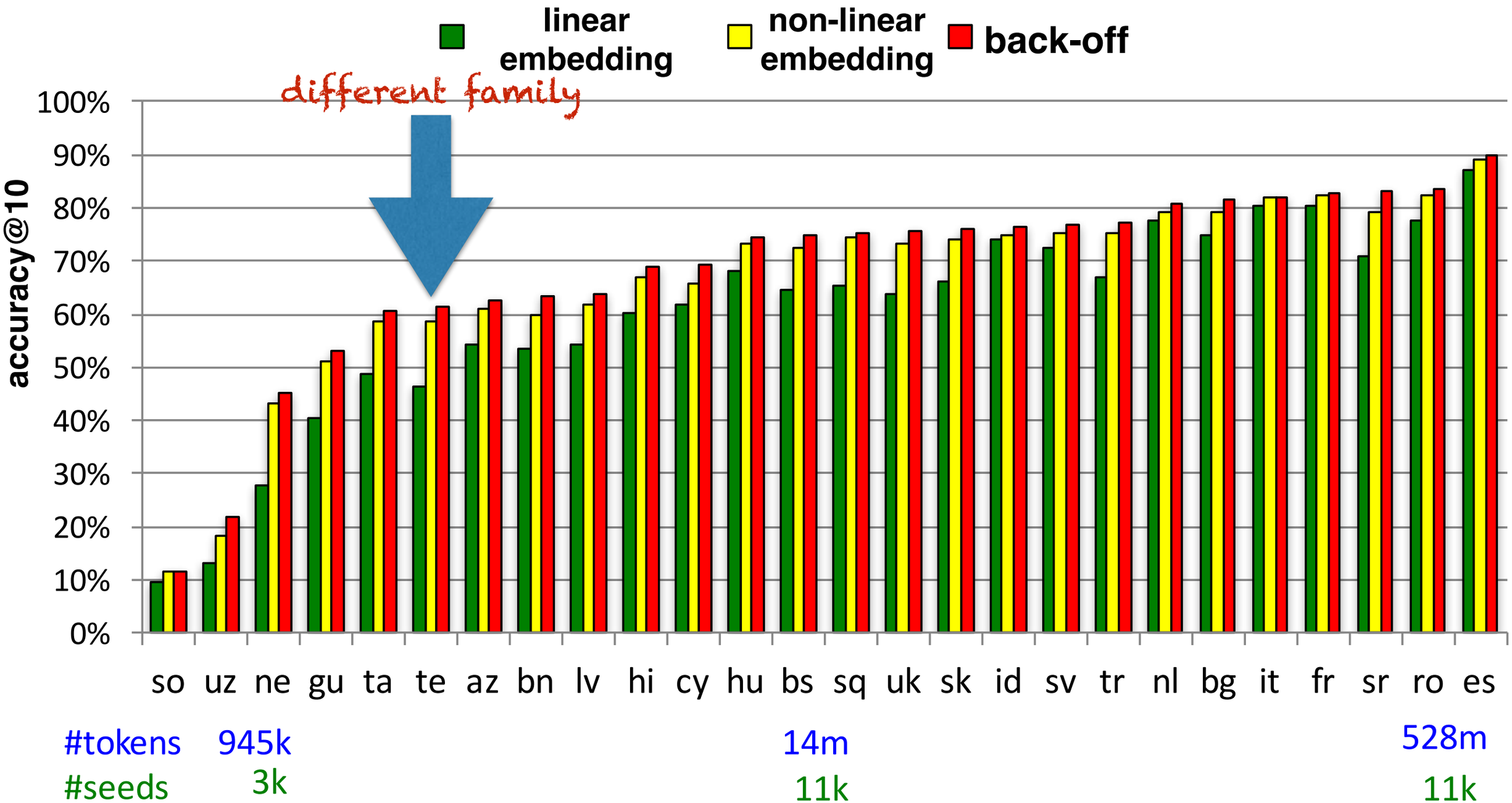
Experiments

using related languages + context similarities



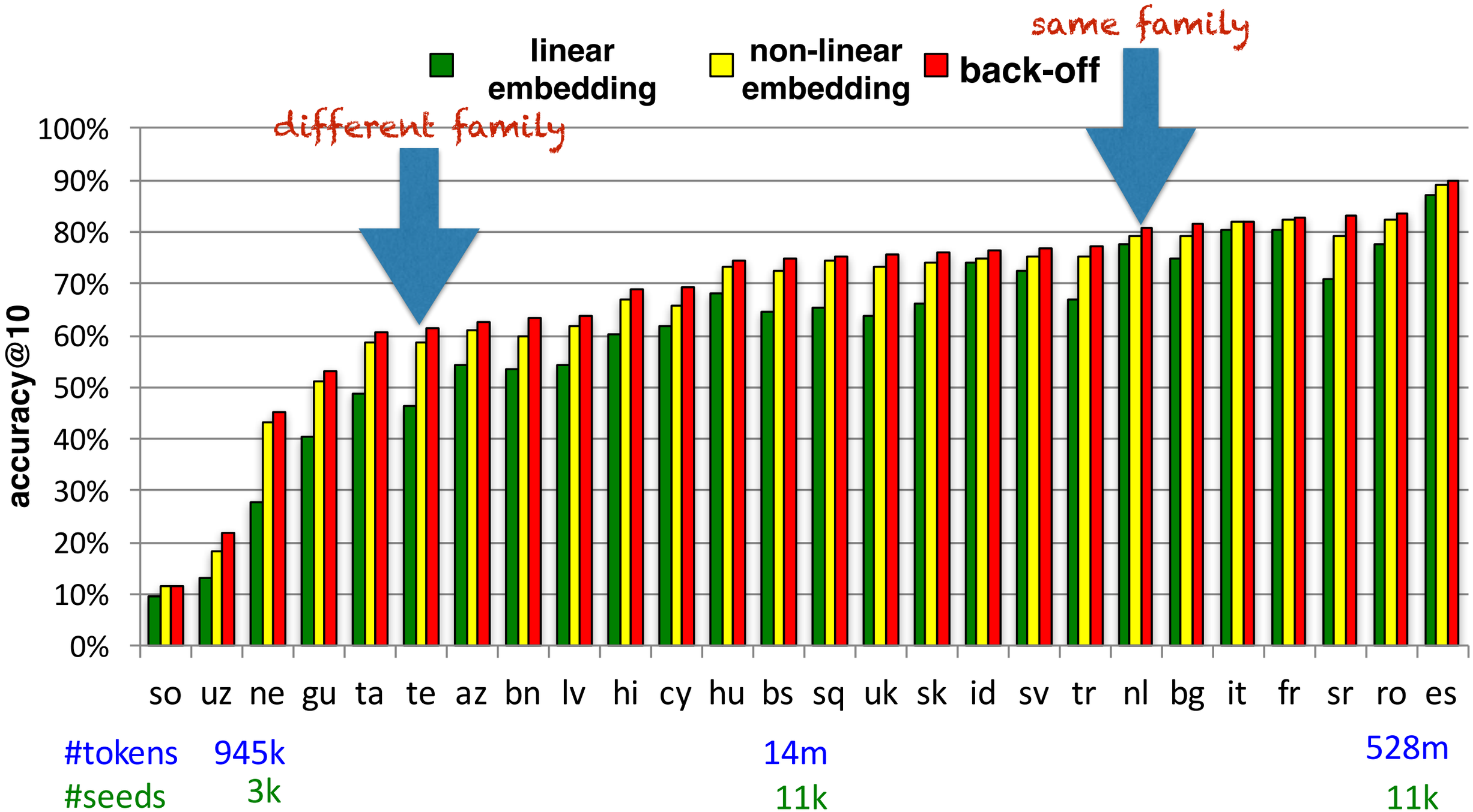
Experiments

using related languages + context similarities



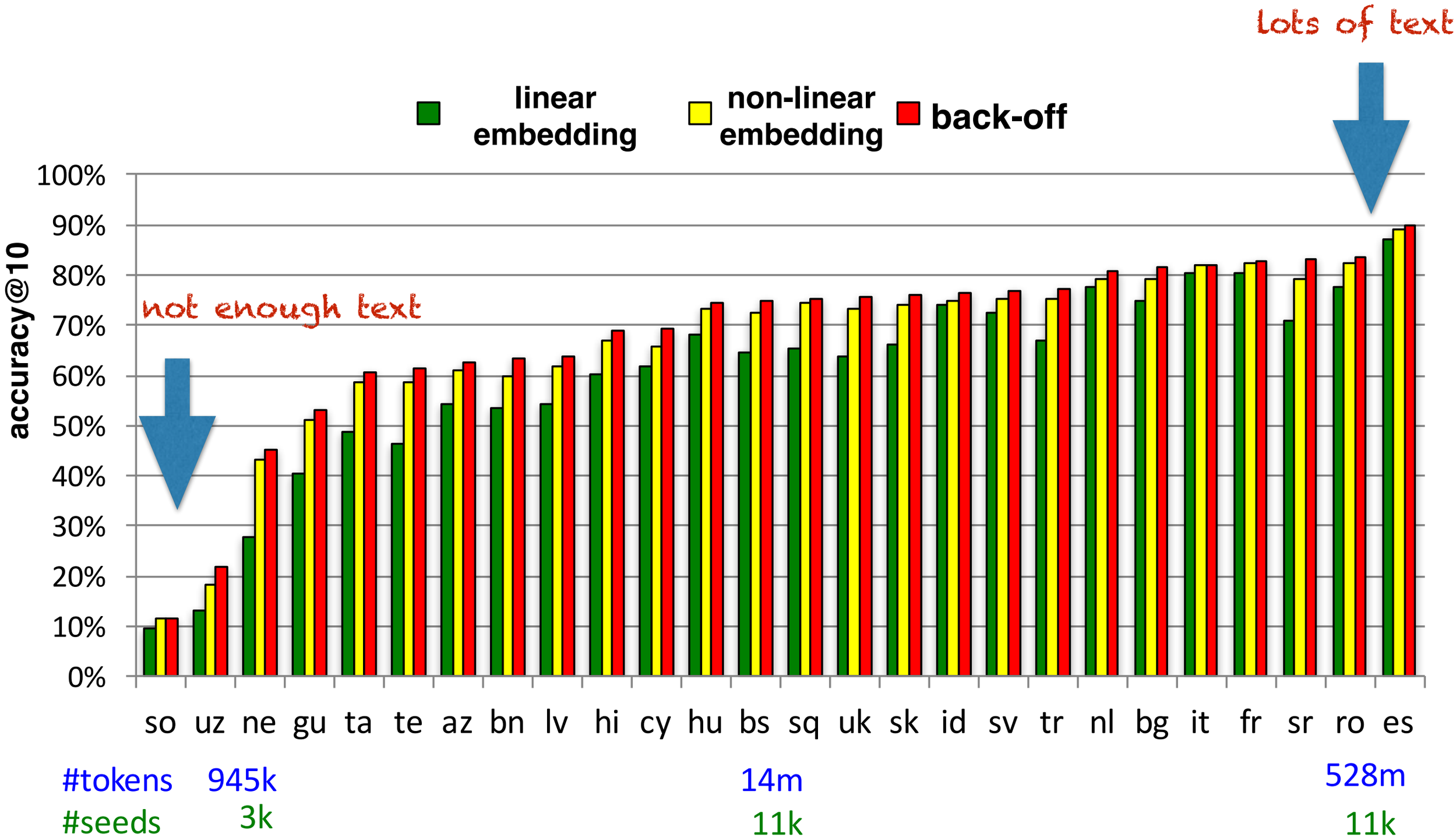
Experiments

using related languages + context similarities



Experiments

using related languages + context similarities



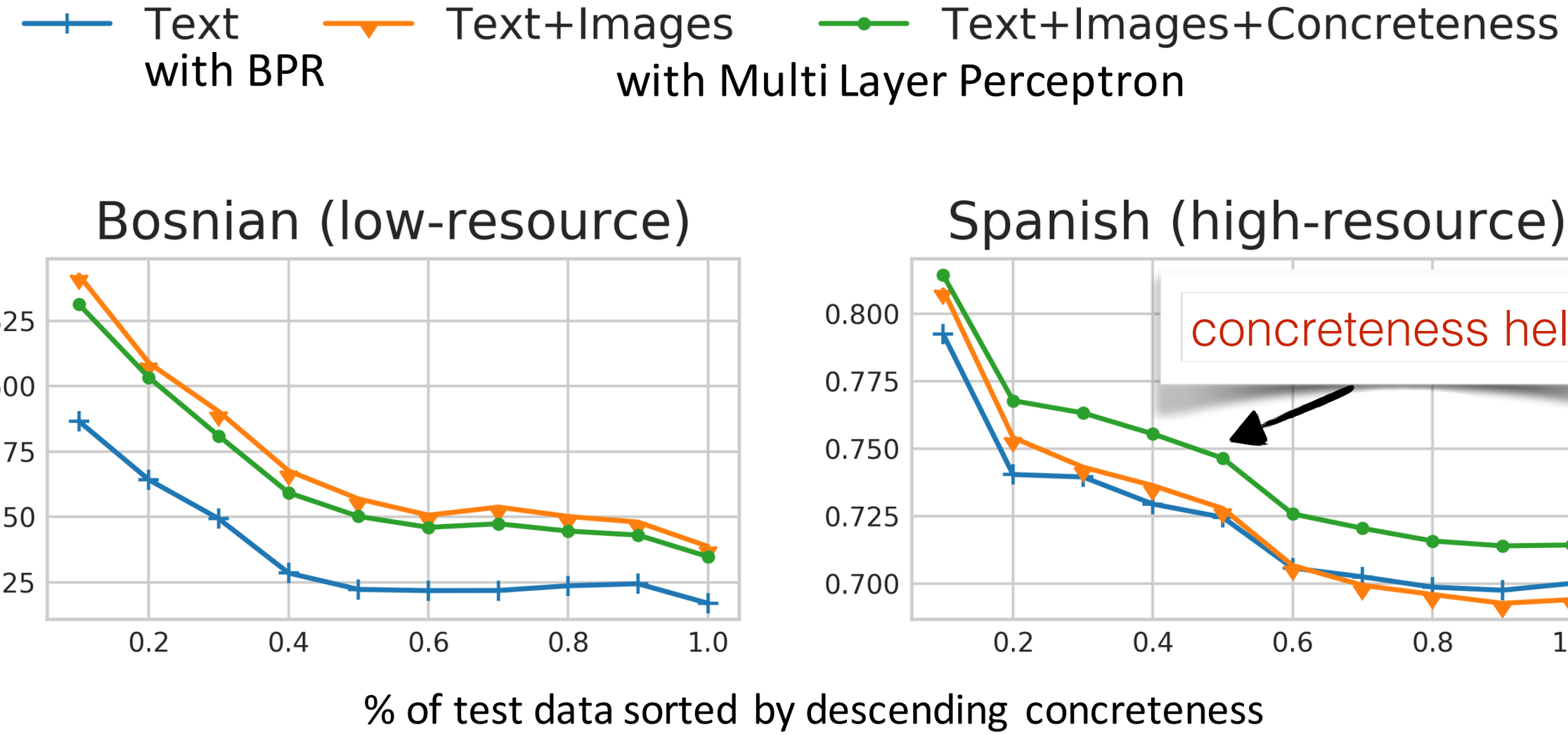
Experiments

Text + Image

when not enough text, image can help

Experiments

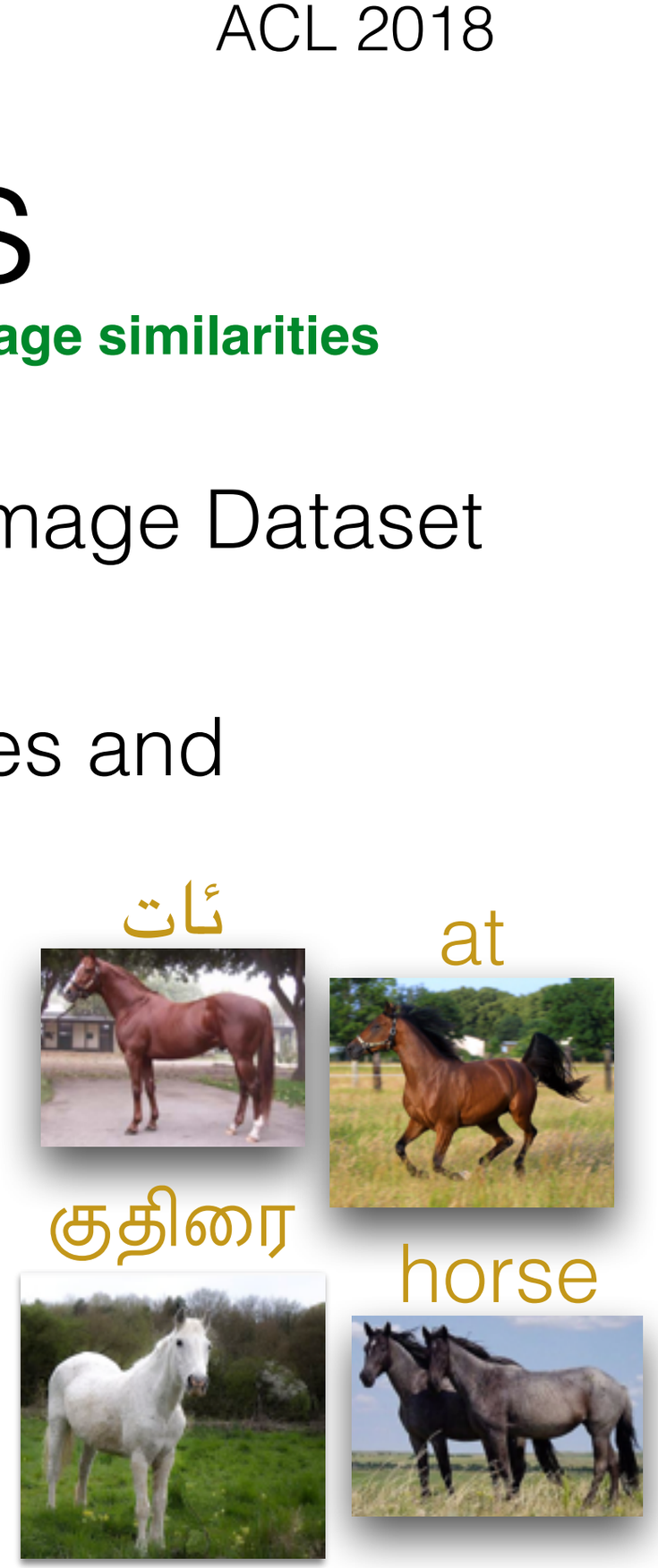
using related languages + context similarities + image similarities



Experiments

using related languages + context similarities + image similarities

- Released the Massively Multilingual Image Dataset (MMID)
  - images of words from 100 languages and images of their English translations
  - up to 10,000 words per language
  - up to 100 images per word
  - http://multilingual-images.org/



In Submission

Experiments

using related languages + context similarities + transliteration

names are often just written in

different scripts in different languages

transliterated

obama ⟷ ობამა ⟷ Օբաման

In Submission

Experiments

using related languages + context similarities + transliteration

transliteration is useful for translating
names and other *rare words*

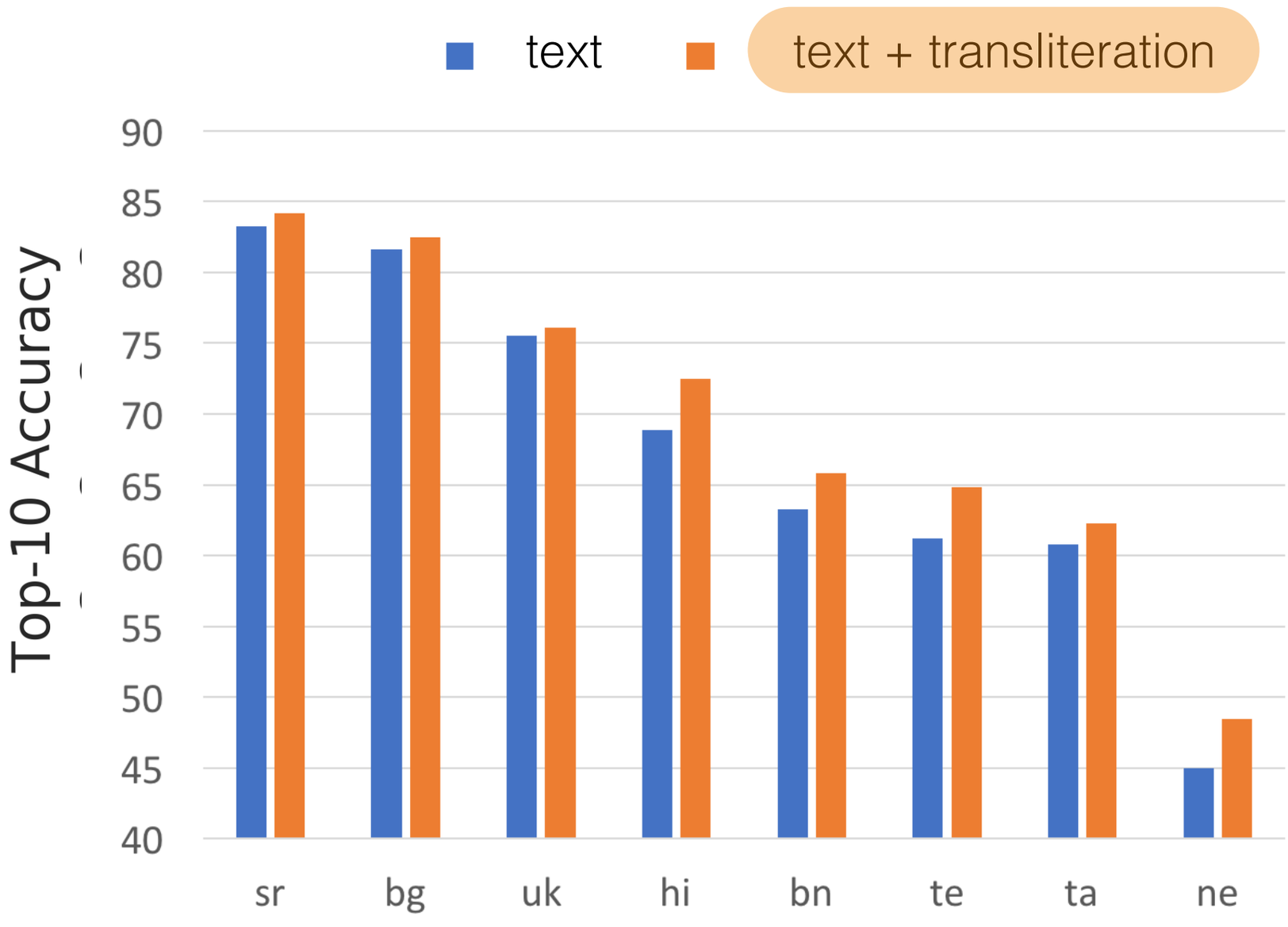
Text	+Translit	Text	+Translit	Text	+Translit
चित्रा	चित्रा	हलि	हलि	मधुमती	मधुमती
savita	chithra	brockley	★ hill	xan ★	madhumati
rohini	★ chitra	chatsworth	brockley	neela	xan
rashmi	savita	woodhouse	chatsworth	madhumati	neela
ashwini	rohini	riverside	woodhouse	zeenat	zeenat
rane	rashmi	clifton	riverside	seema	seema

related, but incorrect names

In Submission

Experiments

using related languages + context similarities + transliteration



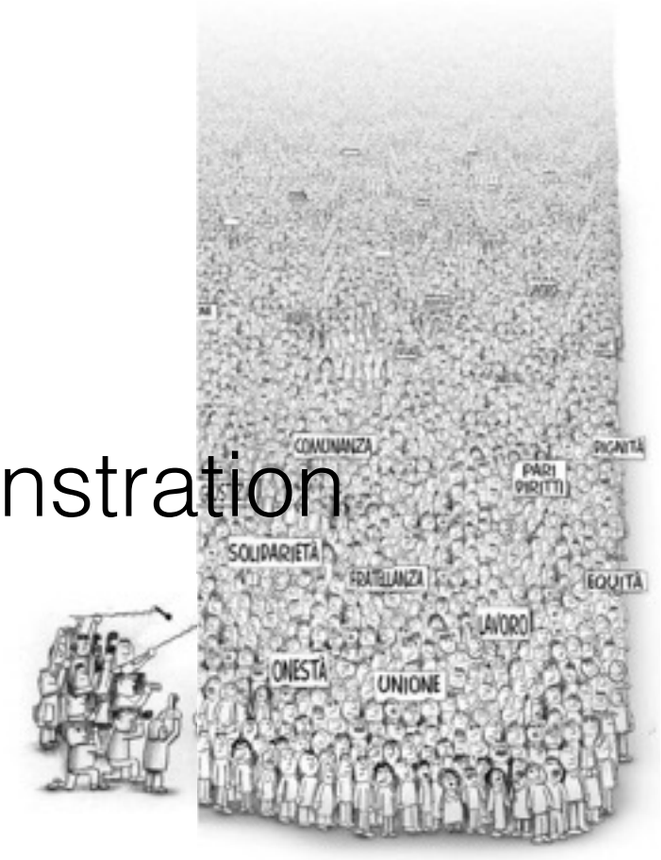
Current Work

- Learn better multimodal multilingual representation
  - extend our multilingual image datasets to cover images of more words
  - using weak supervision (noisy captions) of many images, learn images of new objects

Current Work

- Learn better multimodal multilingual representation
  - in developing multilingual and multimodal tools for analyzing public communication

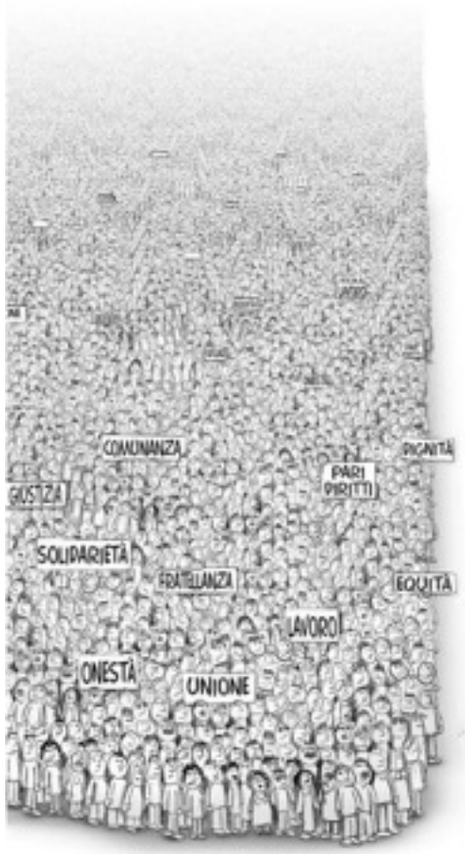
peaceful demonstration



https://criticalmediareview.wordpress.com/2015/10/19/what-is-media-framing/

Current Work

- Learn better multimodal multilingual representation
  - to develop multilingual and multimodal tools for analyzing public communication



angry protesters

https://criticalmediareview.wordpress.com/2015/10/19/what-is-media-framing/

Thank You!