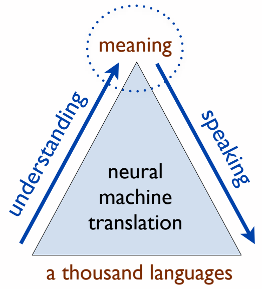
translations as semantic mirrors

"Found in Translation"

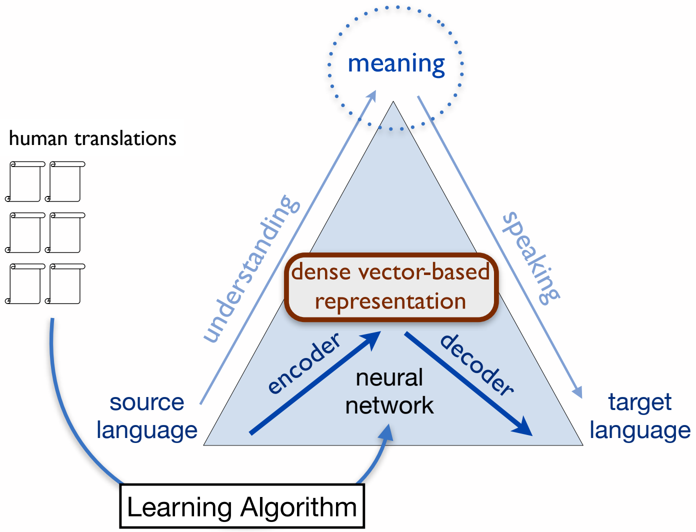
Emerging Language Spaces Learned From Massively Multilingual Corpora

using massive linguistic diversity

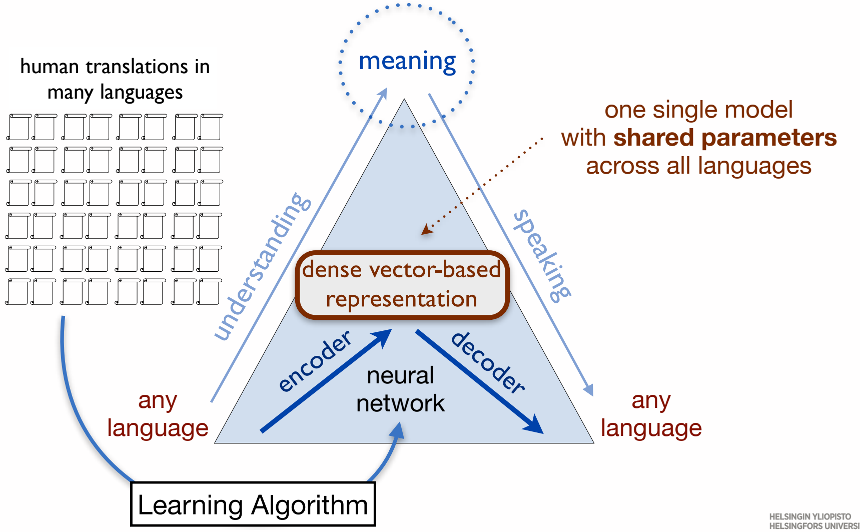


Jörg Tiedemann
University of Helsinki
jorg.tiedemann@helsinki.fi

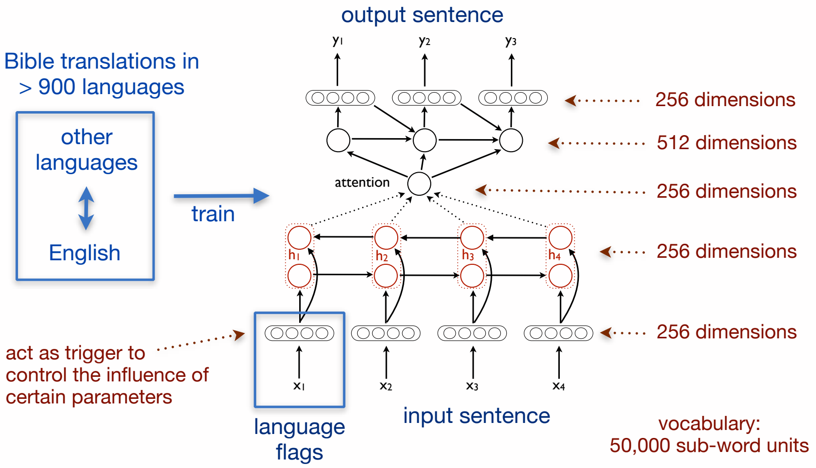
Machine Translation (MT)

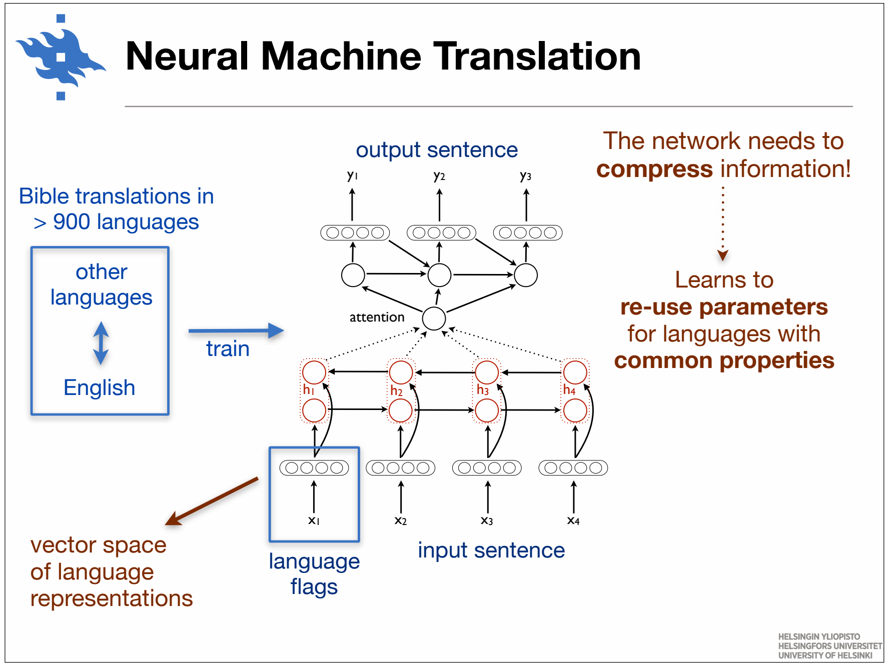


Multilingual Translation Models


meaning
human translations in many languages
one single model with **shared parameters** across all languages
understanding
speaking
dense vector-based representation
encoder
decoder
neural network
any language
any language
Learning Algorithm

Neural Machine Translation





Neural Machine Translation

Bible translations in > 900 languages

other languages ↕ English

train

output sentence

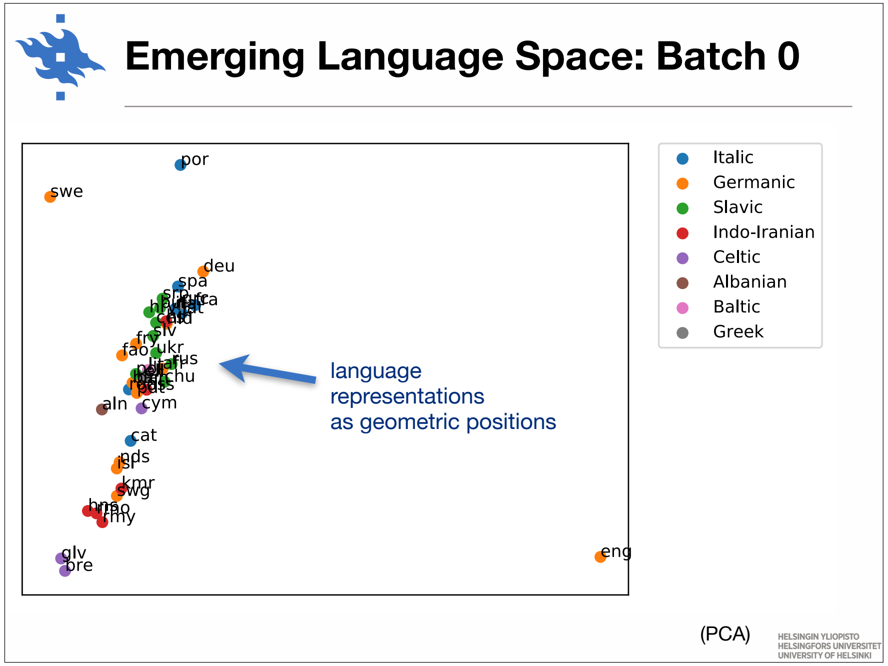
The network needs to **compress** information!

Learns to **re-use parameters** for languages with **common properties**

vector space of language representations

language flags

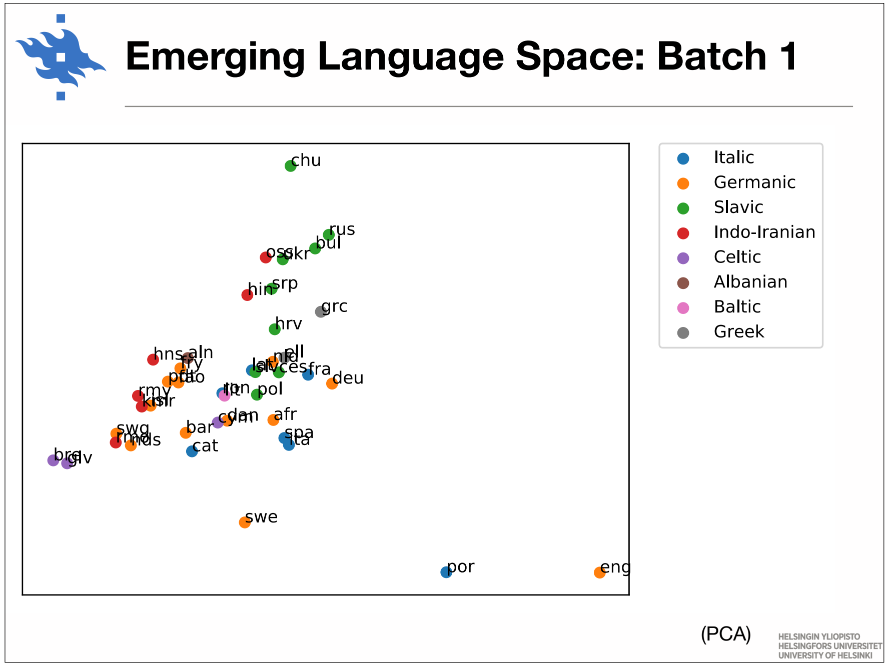
input sentence



Emerging Language Space: Batch 0

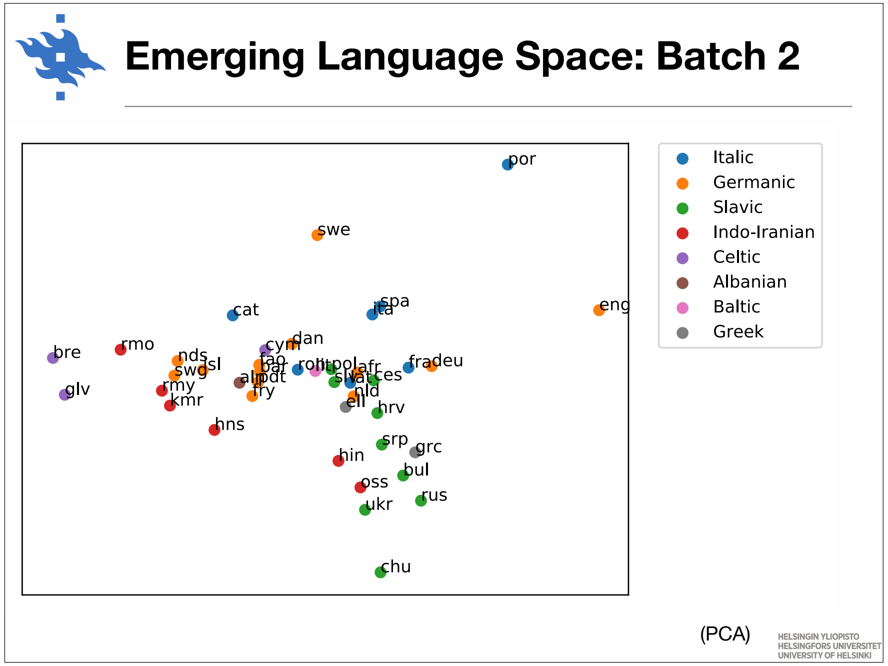
language representations as geometric positions

(PCA)



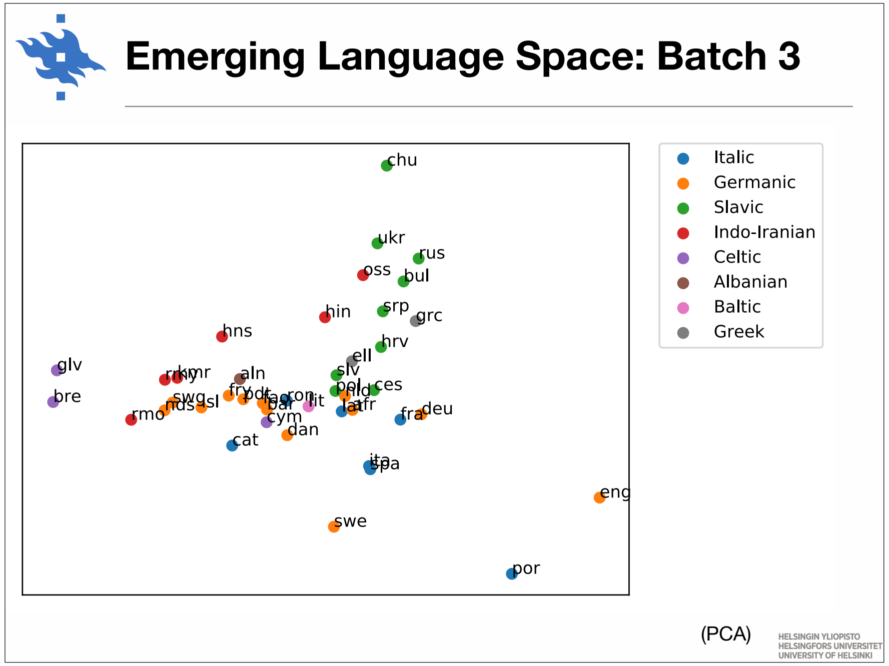
Emerging Language Space: Batch 1

(PCA)



Emerging Language Space: Batch 2

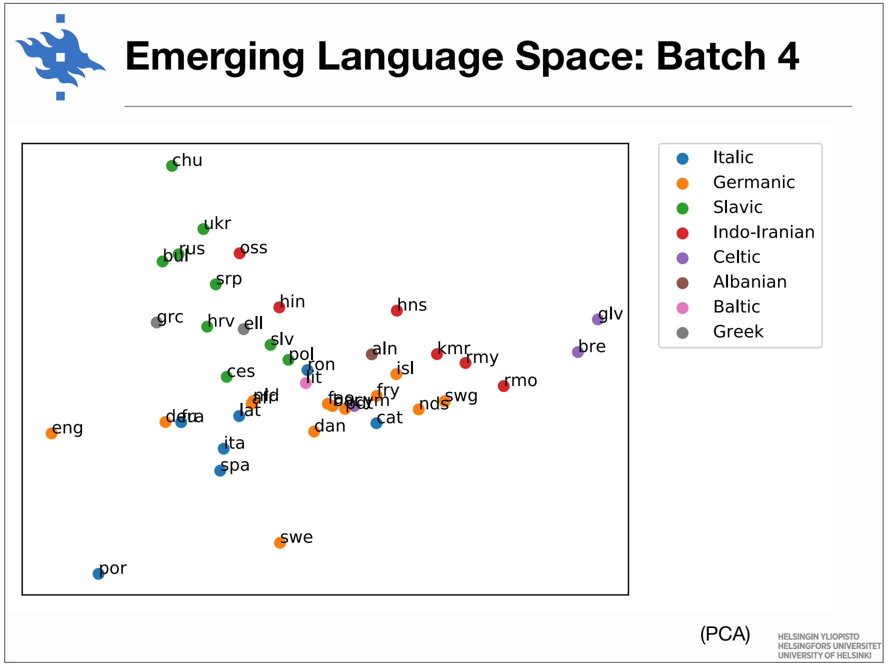
(PCA)



Emerging Language Space: Batch 3

(PCA)

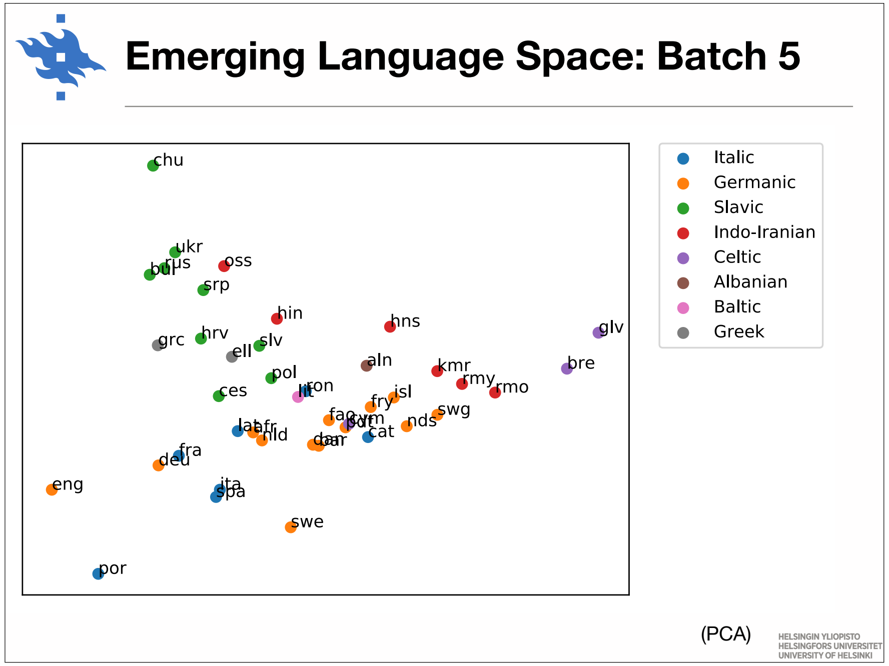
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI



Emerging Language Space: Batch 4

(PCA)

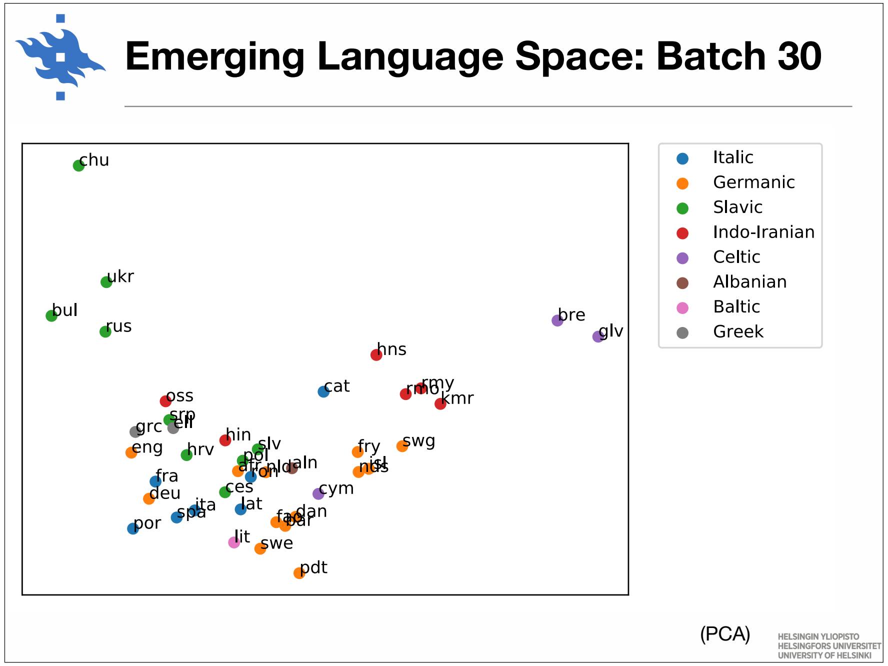
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI



Emerging Language Space: Batch 5

(PCA)

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI



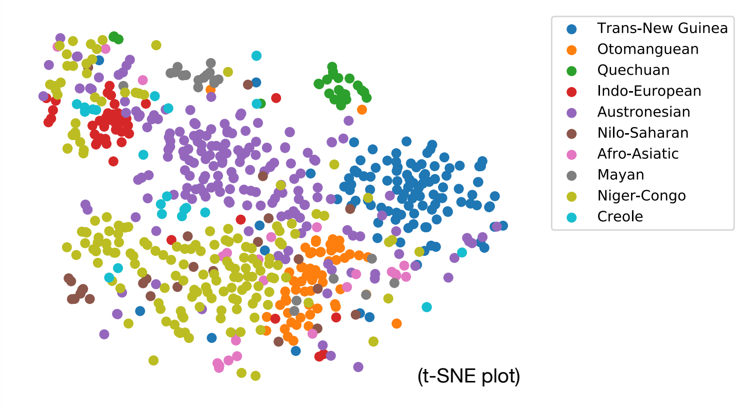
Emerging Language Space: Batch 30

(PCA)

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Language Space of 972 Languages

Rough clusters of language families



(t-SNE plot)

Why Is This Interesting?

Continuous language space

- distances refer to relationship
- languages are not independent discrete units

Completely data-driven approach

- no prior knowledge
- driven by optimizing compression (with translation objective)

Interesting questions for future research

- Can we see specific linguistic properties?
- Combination with other tasks than MT

It Is A Trendy Research Topic

- Academy of Finland project: **Digital language typology: mining from the surface to the core** (Vainio, Toivonen)
- Bjerva, J. and Augenstein, I. (2018) **From Phonology to Syntax: Unsupervised Linguistic Typology at Different Levels with Language Embeddings**, NAACL-HLT 2018.
- Bjerva, J. and Augenstein, I. (2018) **Tracking Typological Traits of Uralic Languages in Distributed Language Representations**, the Fourth International Workshop on Computational Linguistics for Uralic Languages (IWCLUL).
- Chaitanya Malaviya, Graham Neubig, Patrick Littell. **Learning Language Representations for Typology Prediction**, EMNLP 2017.
- Ehsaneddin Asgari and Hinrich Schütze: **Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages**, EMNLP 2017
- ...

What Is Missing - What Is Next?

Many shortcomings

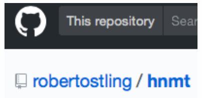
- data sources are limited and of very narrow domains
- the models are simple and generic
- difficult interpretation of results
- very little interaction with general linguistics

Ideas for the future

- emerging linguistic structures (syntax / semantics)
- diachronic models, different registers, ...
- training for specific phenomena with different objectives

CSC - IT CENTER FOR SCIENCE

CSC



This repository Sear

robertostling / hnmt

NVIDIA.

Thank You!

Questions or suggestions?

erc

ACADEMY OF FINLAND

Creating a Massively Parallel Bible Corpus

Thomas Mayer, Michael Cysouw

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI