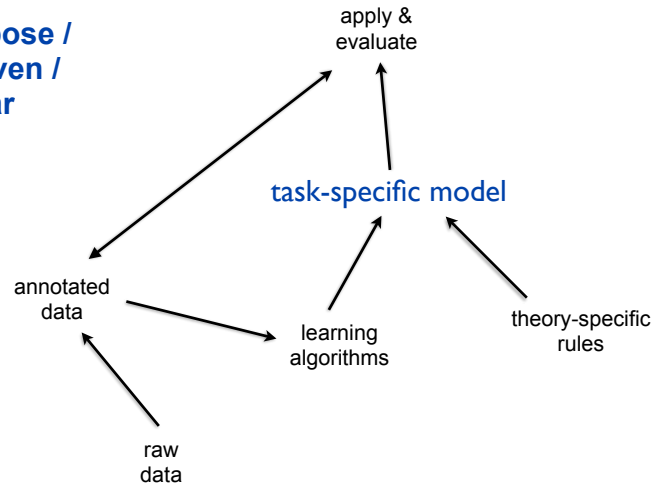
The language continuum and multimodality in LT -

achievements and goals in modern NLP and what we do about them in Helsinki

Jörg Tiedemann
Department of Digital Humanities
University of Helsinki
jorg.tiedemann@helsinki.fi

Traditional approaches

single purpose / theory-driven / modular



Traditional approaches

single purpose / theory-driven / modular

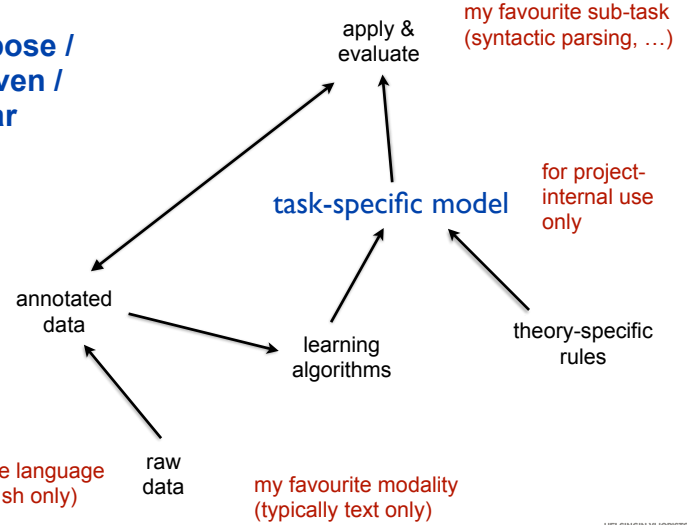
my favourite sub-task (syntactic parsing, …)

my favourite theoretical framework

for project-internal use only

my favourite language (often English only)

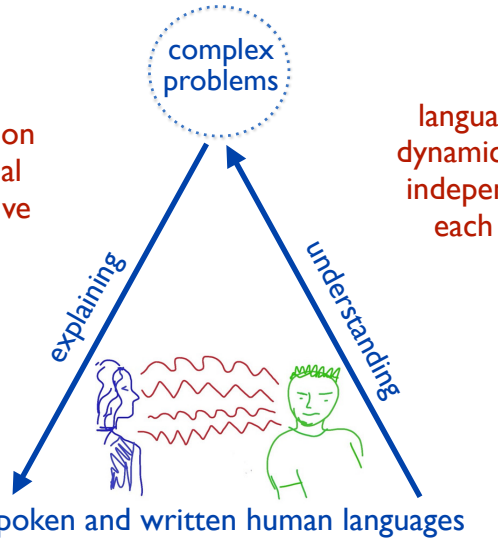
my favourite modality (typically text only)



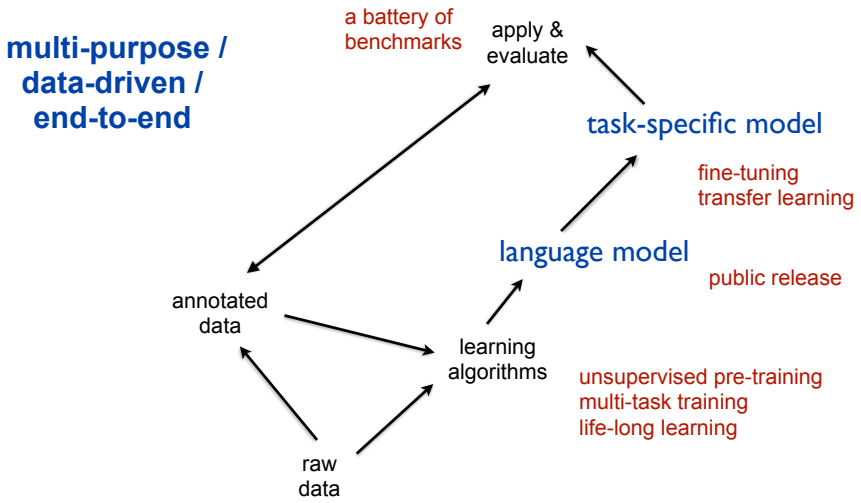
But language is a multi-purpose tool

communication is multimodal and interactive

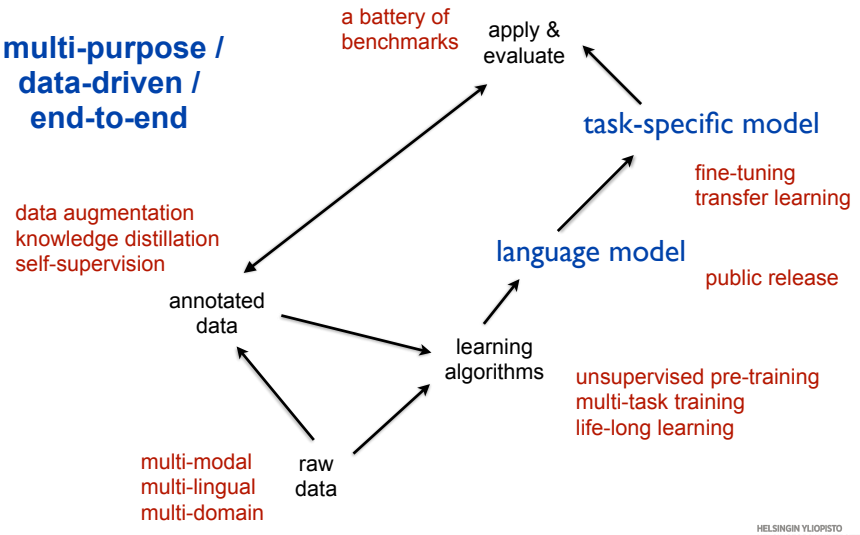
languages are dynamic and not independent of each other



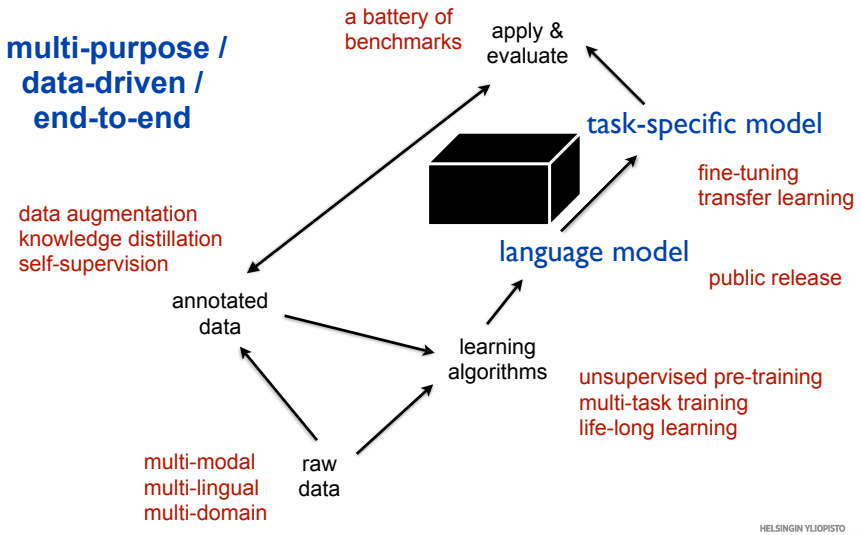
Current trend in language technology



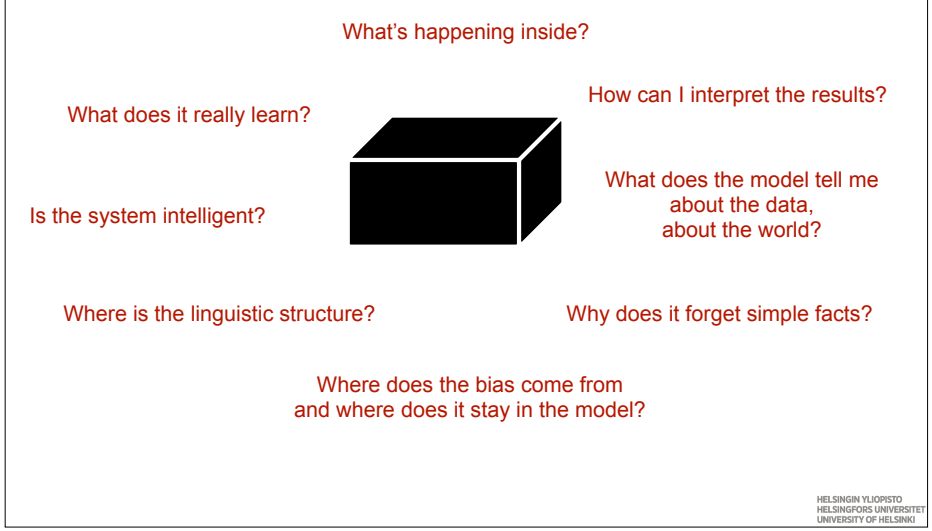
Current trend in language technology



Current trend in language technology



Deep learning and explainability



Research in humanities and explainability

What's happening inside?

How can I interpret the results?

What do humans really learn?

What is intelligence?

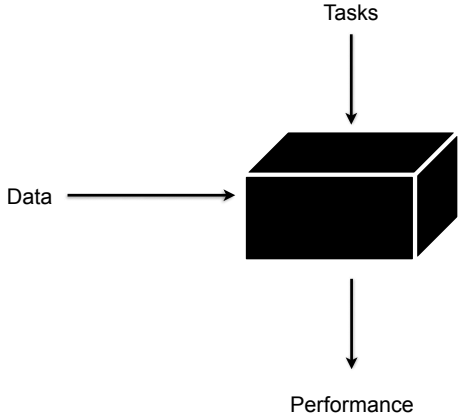
What does this tell me about the things I see and about the world?

Where is the linguistic structure?

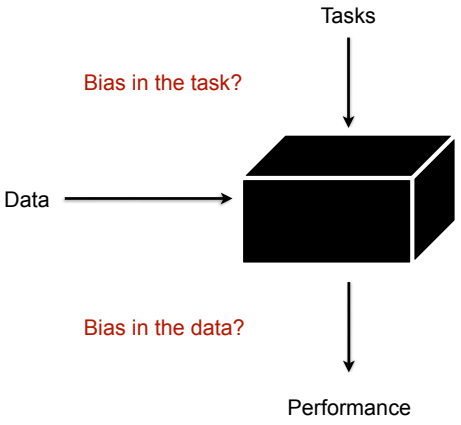
Why do they forget simple facts?

Where does the bias come from and where does it stay in the human brain?

Simulation-based linguistics



Simulation-based linguistics / DH?



Bias in the task?

Emerging structure?

Knowledge acquisition?

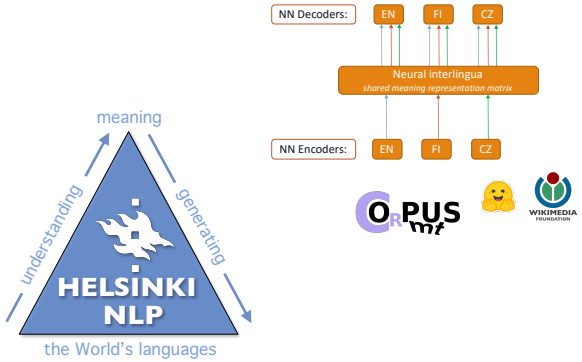
Knowledge transfer?

Bias in the data?

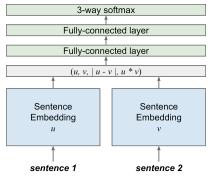
Generalisability across languages and domains?

Bias in the world and in society?

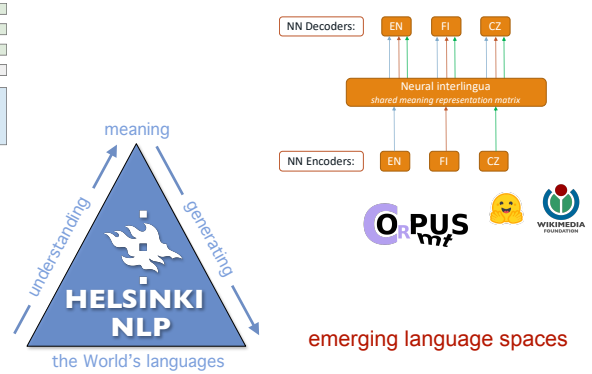
multilingual language & translation models



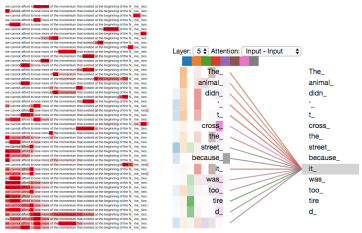
semantic inference tasks



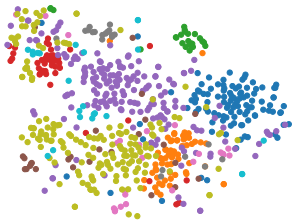
multilingual language & translation models



model interpretation



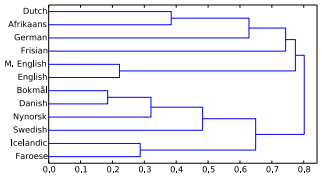
emerging language spaces



Language Model from a 1,000 Languages

Training (actually 990 languages / language variants)

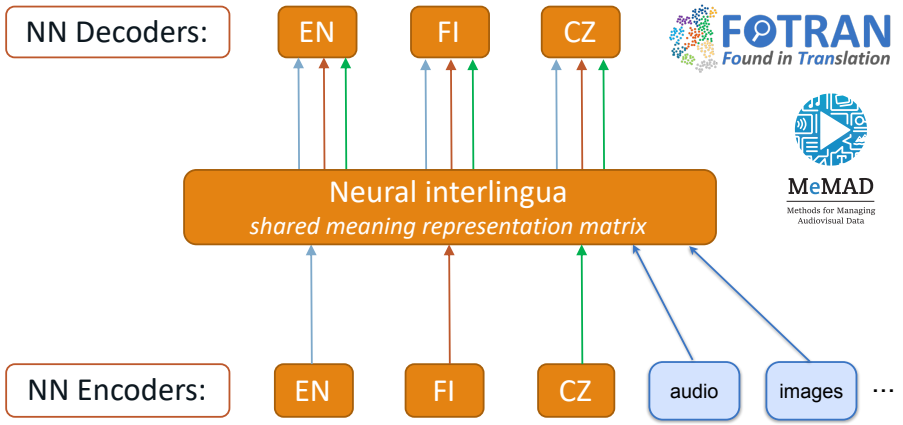
- learn from translations of the Bible
- **emerging language continuum**
- visible language clusters



The model can generate text:

- **turn on Swedish**:
  – *och jehova sade till honom : ” jehova har sagt , och jag skall ...*
- **turn on German**:
  – *und er sprach zu ihnen : siehe , ich bin der herr ...*
- **mix Swedish and German**:
  – *vocken ånner vocken ånnen söhenöckenföcken ...*
- **average of Scandinavian languages**:
  – *og han sa til herrens : " han skal vitnaðus til herrens hjárt*

Multilingual and multimodal translation



Can we reason with neural semantics?

Natural language inference benchmarks:

A black race car starts up in front of a crowd of people.

contradicts

A man is driving down a lonely road.

A soccer game with multiple males playing.

entails

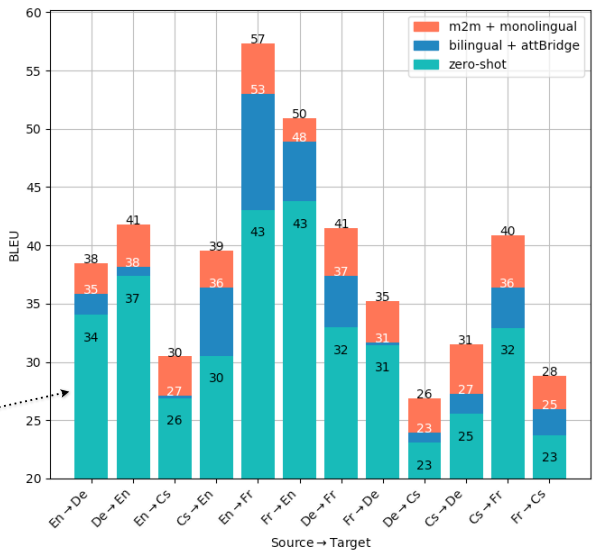
Some men are playing a sport.

Can we reformulate given sentences?

A system that is trained to translate can also paraphrase sentences in one language:

Source	But even as he was on the road going down, his servants met him and reported, saying, Your son lives!
+NLD	And as he was on the road, his servants went down with him, and reported, saying, Thy son lives!
+SPA	But as it was on the road, his servants came to him and told him, "Your own Son lives!"
+ALL	And while he was on the way, his servants came to him, saying, "Your son lives!"

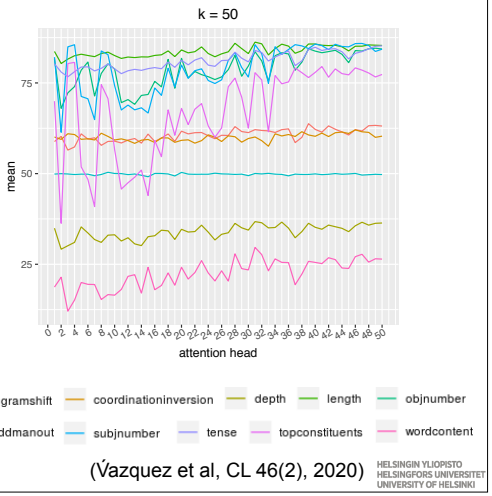
Can we translate between unseen language pairs?



language pair not seen in training data (zero shot)

Where does the model look at and why?

various syntactic and semantic probing tasks:



(Vázquez et al, CL 46(2), 2020)

To sum up

Deep learning and language technology

- black-box models that learn to read/listen and write/speak
- downstream tasks like translation enforce to learn semantics
- continuous language spaces can be learned

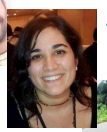
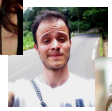
Deep learning and humanities

- black-box models are quite close to traditional humanities
- interpretation of artificial models can be insightful
- we are closer to each other than you may think …

Language Technology in Helsinki

http://blogs.helsinki.fi/language-technology/





fiskmö
finsk-svensk korpus & maskinöversättning

Data collection & MT
2 languages (Finnish/Swedish)
https://blogs.helsinki.fi/fiskmo-project/

FOTRAN
Found in Translation

semantics & MT
> 1,000 languages

https://github.com/Helsinki-NLP

audiovisual data & MT
6 languages





OPUS
http://opus.nlpl.eu
Data collection for MT
> 200 languages

sentimentator

MeMAD
Methods for Managing Audiovisual Data
https://memad.eu


HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI