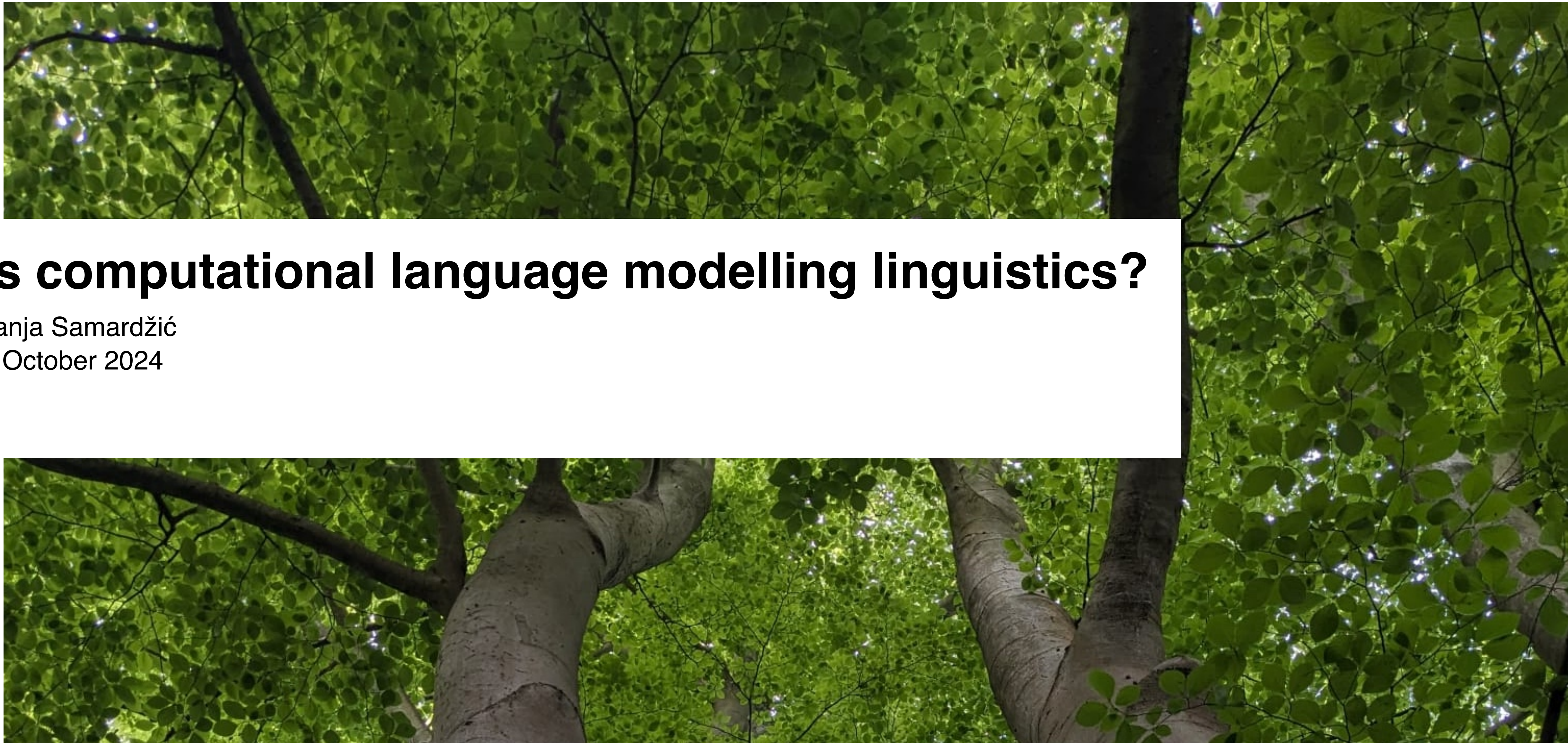


Is computational language modelling linguistics?

Tanja Samardžić

7 October 2024

Machine learning and language

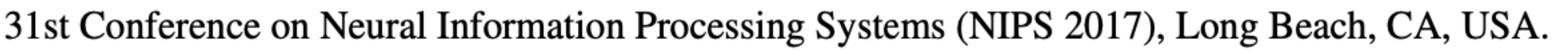
Transformers were invented for **natural language processing (NLP)**

processing = transforming

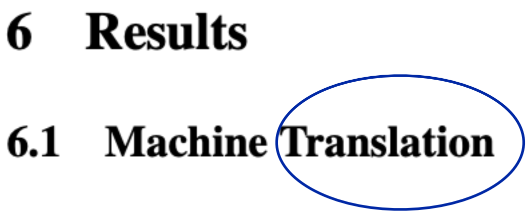
German sentence → English

any sentence → LABEL

Attention Is All You Need



31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.



6 Results

6.1 Machine Translation

Is this linguistics?

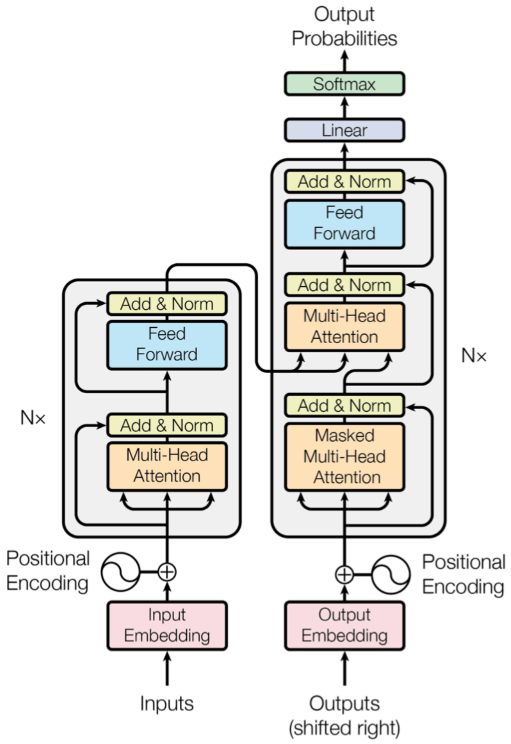


Figure 1: The Transformer - model architecture.

Language and linguistics

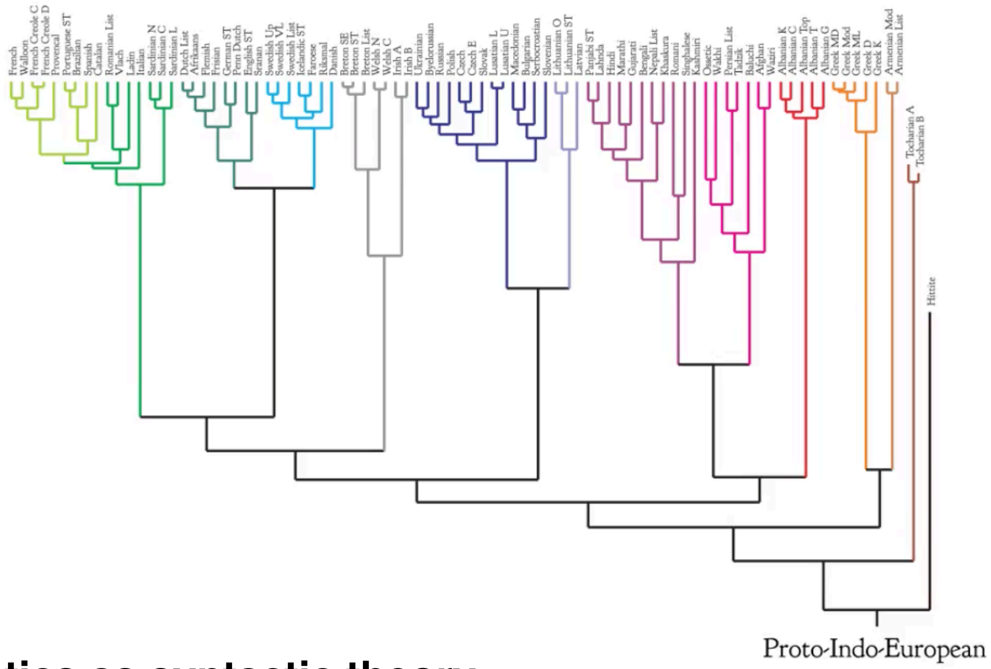
Language as art
Speaking or writing "correctly"
Speaking a foreign language
Being able to translate
Being able to teach a language
Write dictionaries

Philology
Interpret the meaning of complex texts
Knowing the particularities of a (single) language
Knowing the associated culture and history

Language as a cognitive capacity
Studied mostly in psychology

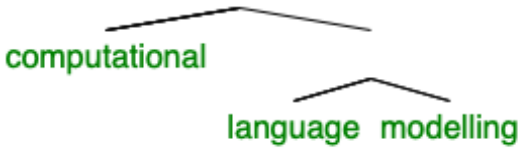
Language as social phenomenon
Studied mostly in psychology

General linguistics as typology

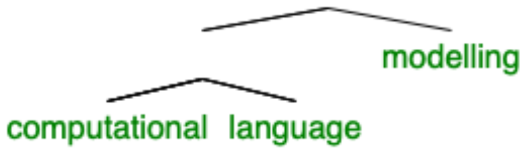


General linguistics as syntactic theory

This?


computational
language modelling

This?


modelling
computational language

http://mshang.ca/syntree/:

Computational modelling as a means of integrating philological knowledge

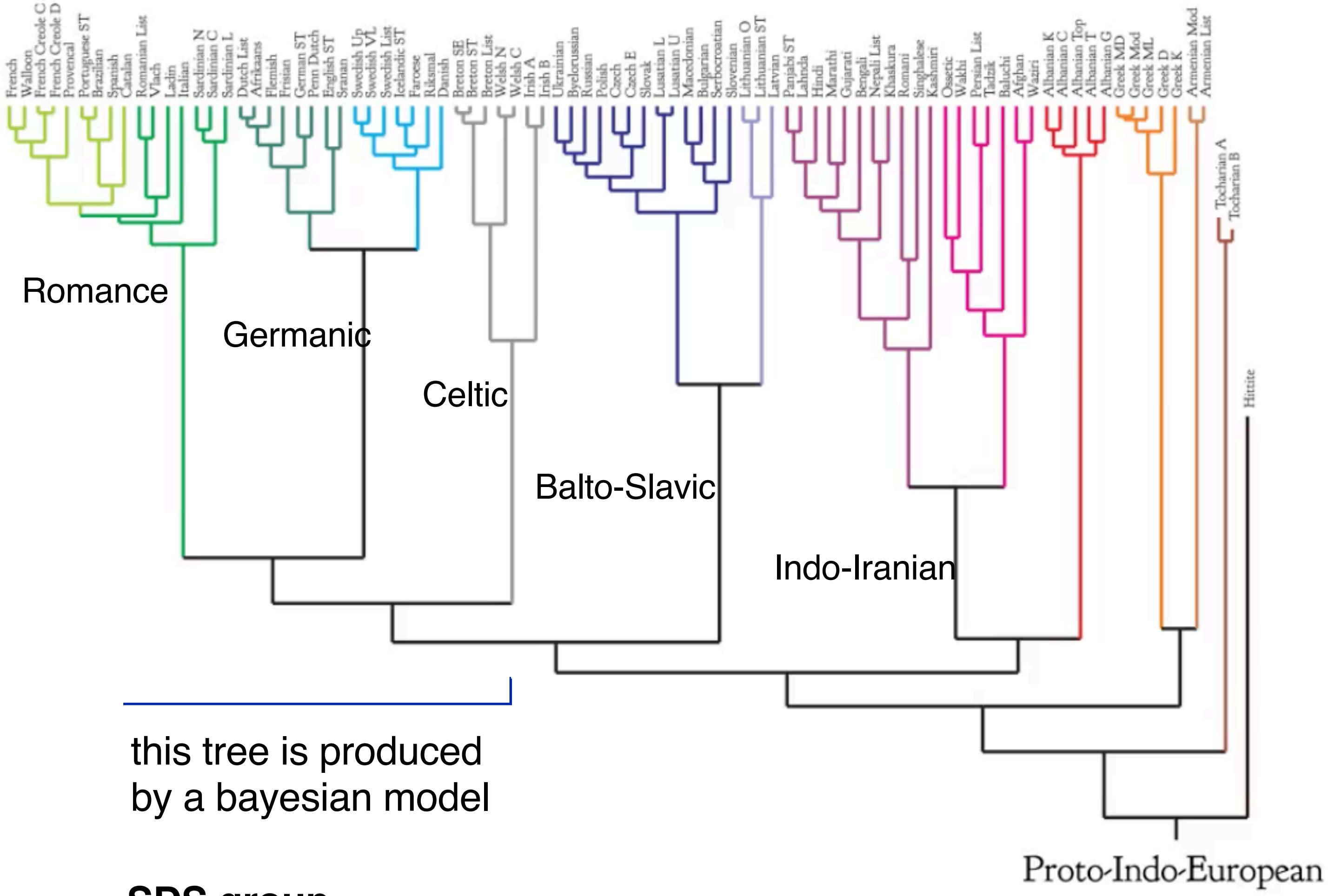
Welcome to Glottolog 5.0

Comprehensive reference information for the world's languages, especially the lesser known languages.

Information about the different languages, dialects, and families of the world ('languoids') is available in the Languages and Families sections. The References section contains bibliographical information. You can query the bibliographical database by filtering the table view or using a complex query involving genealogical affiliation, document type, and macro-area.

Catalogue of languages and families

Glottolog provides a comprehensive catalogue of the world's languages, language families and dialects. It assigns a unique and stable identifier (the Glottocode) to (in principle) all languoids, i.e. all families, languages, and dialects. Any variety that a linguist works on should eventually get its own entry. The languoids are organized via a genealogical classification (the Glottolog tree) that is based on available historical-comparative research (see also the Languoids information section).



this tree is produced
by a bayesian model

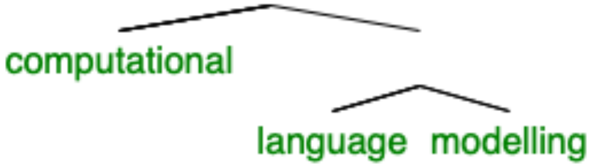
SDS group,
Language and Space Lab, UZH

Computational modelling as a means of interpreting the meaning of texts

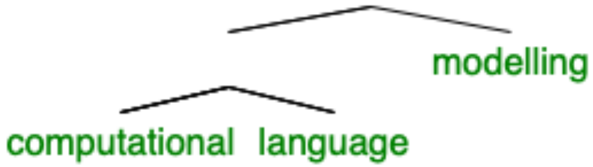
Natural language processing (NLP)

$$\Psi(y, x)$$

$$p(y, x)$$



$x =$ the string "computational language modelling"
(represented as a vector of features)



Various predictions (NLP tasks)

$y =$ a syntactic tree showing the relations between the words

$y =$ a synonym or a paraphrase

$y =$ sentiment (is this something good or bad?)

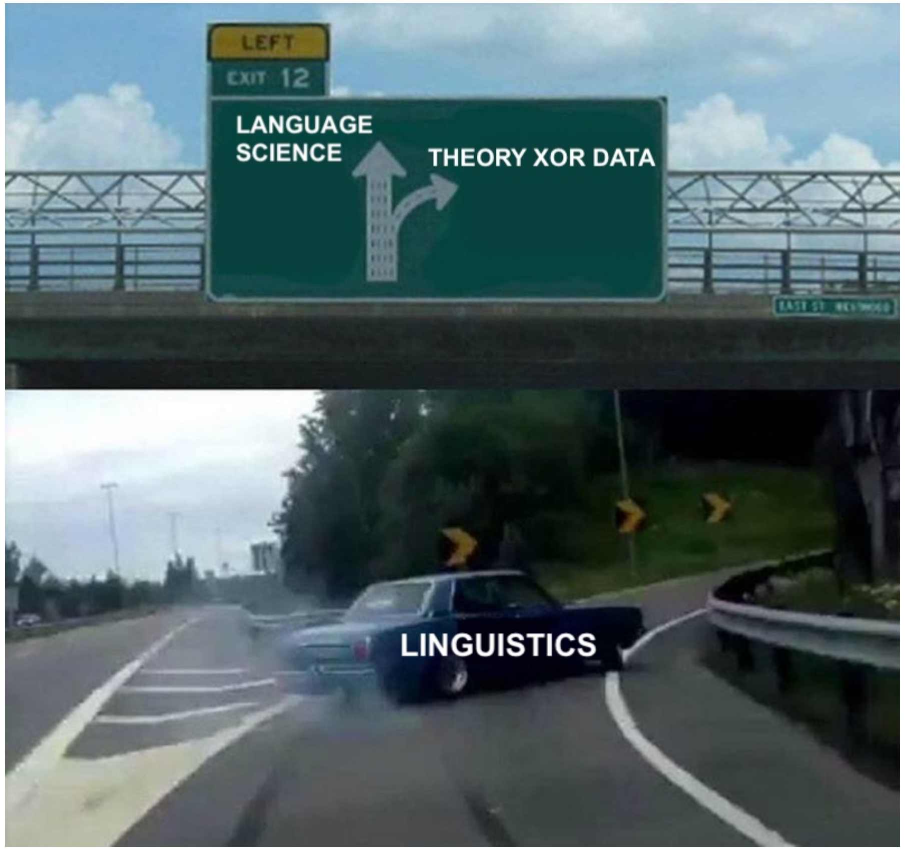
$y =$ translation to another language

many more ...

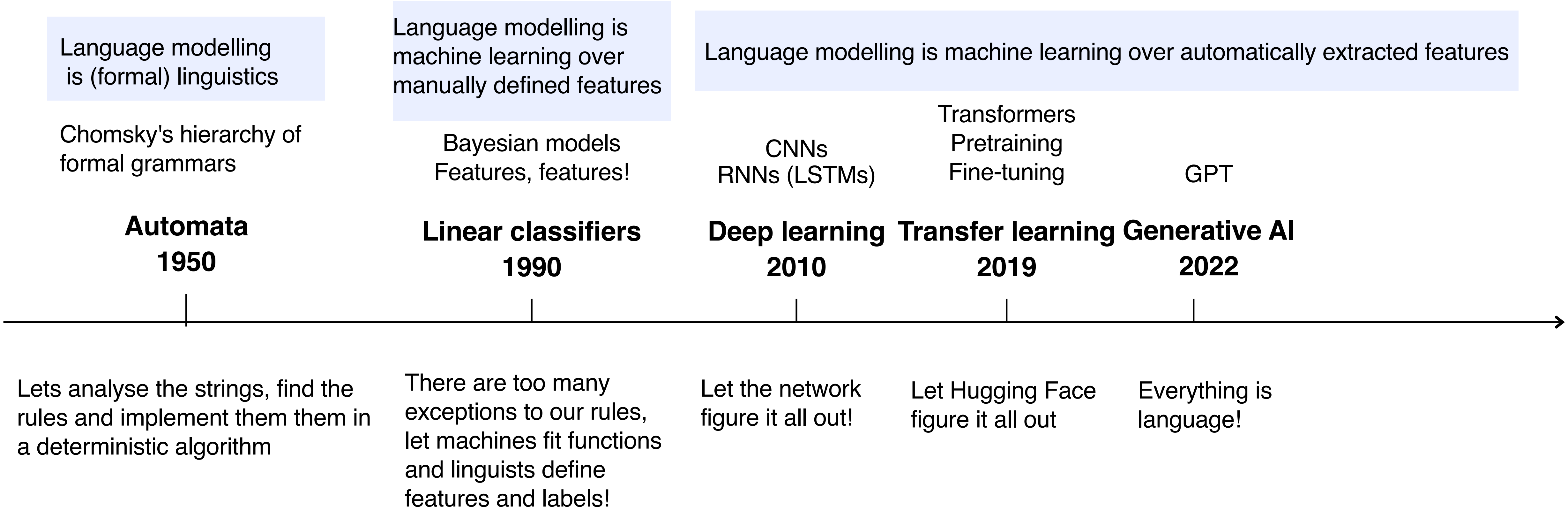
Machine learning as a replacement for linguistics in NLP

"Statistical revolution" in AI in 1990s makes a split in NLP

LEFT
EXIT 12
LANGUAGE SCIENCE
HOCUS-POCUS AI
NLP



From linguistics to language



Language modelling is (formal) linguistics

Chomsky's hierarchy of formal grammars

Automata
1950

Lets analyse the strings, find the rules and implement them them in a deterministic algorithm

Language modelling is machine learning over manually defined features

Bayesian models
Features, features!

Linear classifiers
1990

There are too many exceptions to our rules, let machines fit functions and linguists define features and labels!

Language modelling is machine learning over automatically extracted features

CNNs
RNNs (LSTMs)

Deep learning
2010

Let the network figure it all out!

Transformers
Pretraining
Fine-tuning

Transfer learning
2019

Let Hugging Face figure it all out

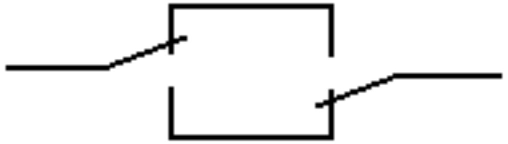
GPT

Generative AI
2022

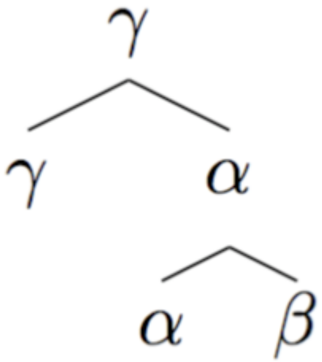
Everything is language!

Linguistics as theory XOR data

Theory (Generative Grammar)
Counting observations is irrelevant to understanding language capacity!
Especially since **1959**



Typology
Construction Grammar
Corpus Linguistics
Let the data speak for themselves!



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Wikimedia Common

Scientific language modelling is still possible as linguistics

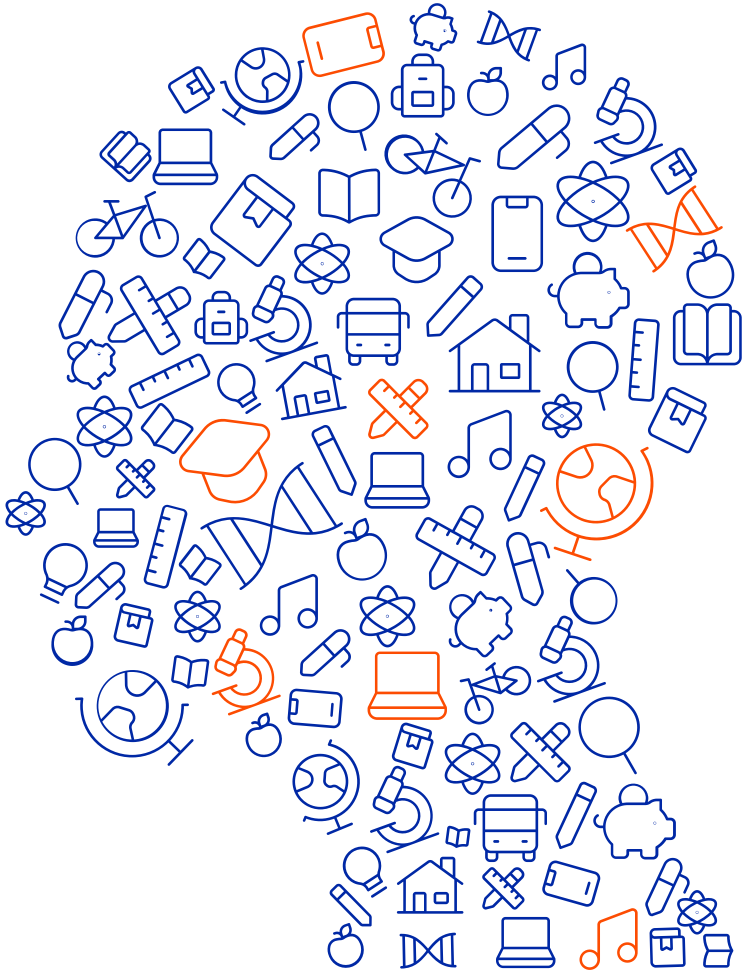
Theory AND data
For now, *information theory*, but likely relevant to syntax

A lot of data
Text samples from many languages

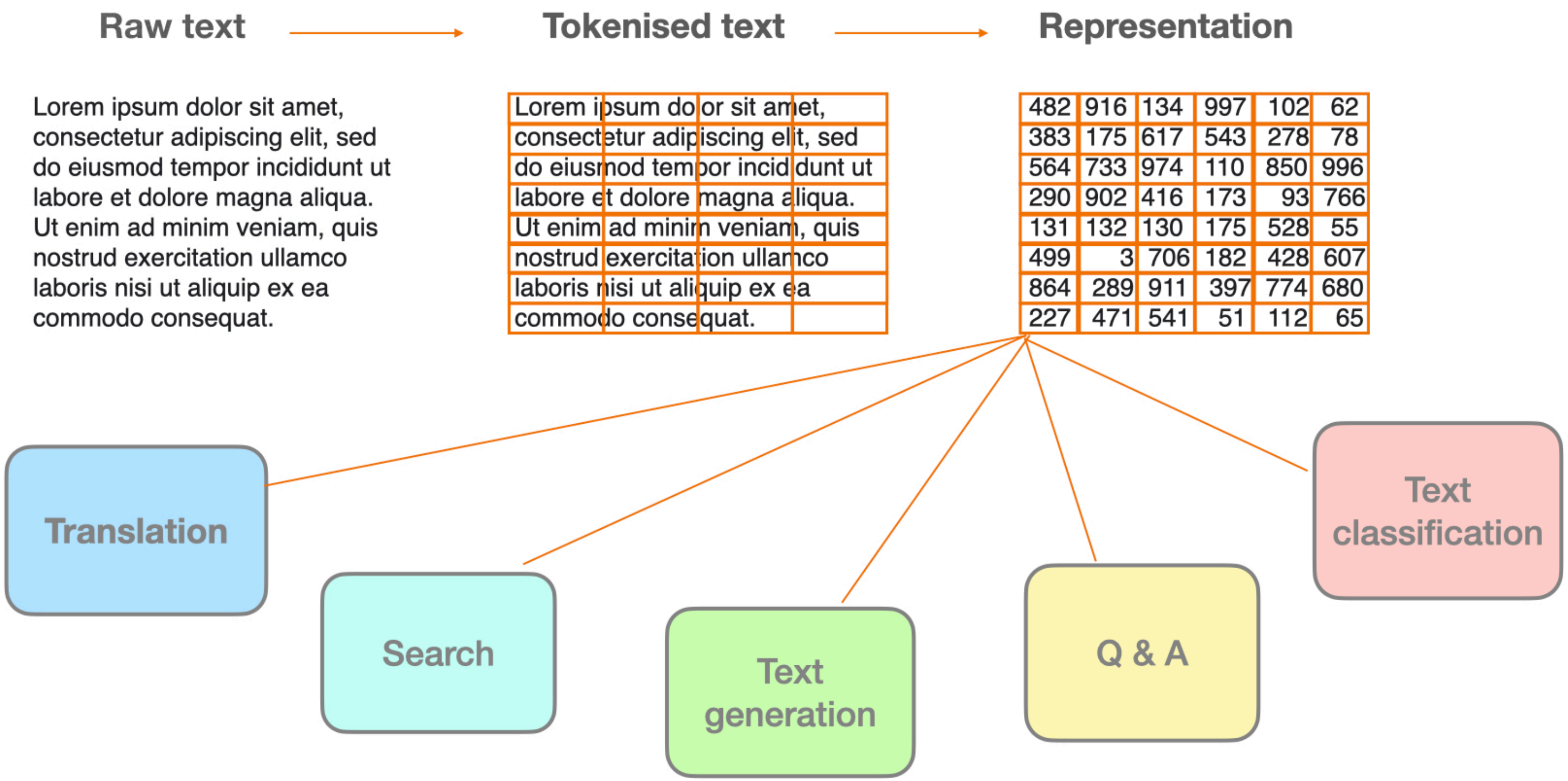
Computing
Simple "cheap" methods at the core of the research, transfer learning for tests

Problem selection
An interesting theoretical problem, but also of an immediate use in practice

1. Text tokenisation
2. Cross-lingual transfer



Text tokenisation as an interesting fundamental problem



Raw text → Tokenised text → Representation

Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed
do eiusmod tempor incididunt ut
labore et dolore magna aliqua.
Ut enim ad minim veniam, quis
nostrud exercitation ullamco
laboris nisi ut aliquip ex ea
commodo consequat.

Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed
do eiusmod tempor incididunt ut
labore et dolore magna aliqua.
Ut enim ad minim veniam, quis
nostrud exercitation ullamco
laboris nisi ut aliquip ex ea
commodo consequat.

482	916	134	997	102	62
383	175	617	543	278	78
564	733	974	110	850	996
290	902	416	173	93	766
131	132	130	175	528	55
499	3	706	182	428	607
864	289	911	397	774	680
227	471	541	51	112	65

Translation

Search

Text generation

Q & A

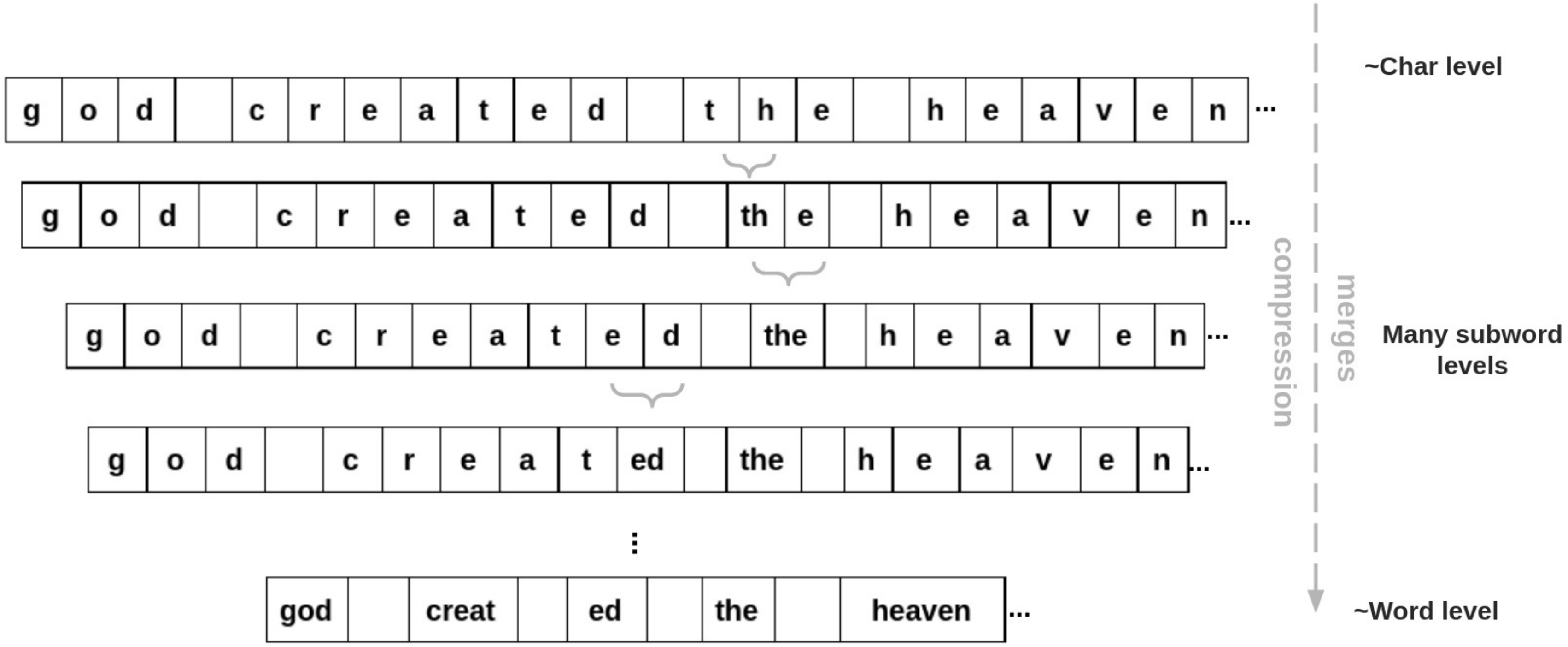
Text classification

How to segment input text?

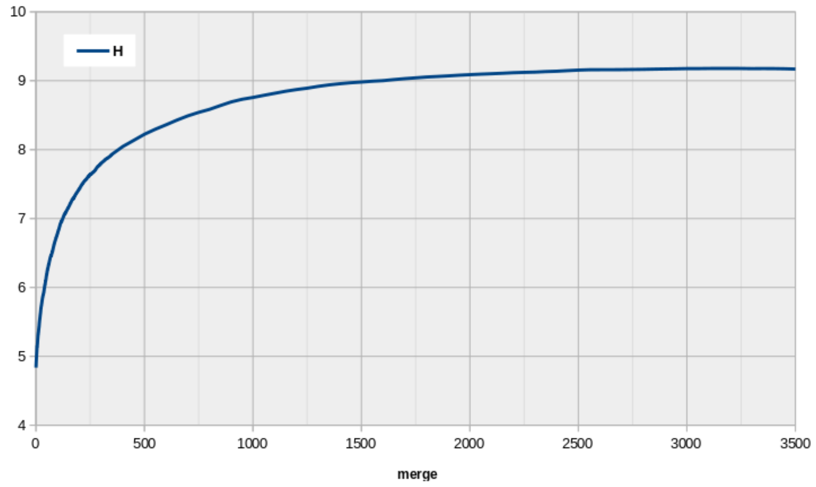
Words are tokens
Traditional view strongly influenced by English

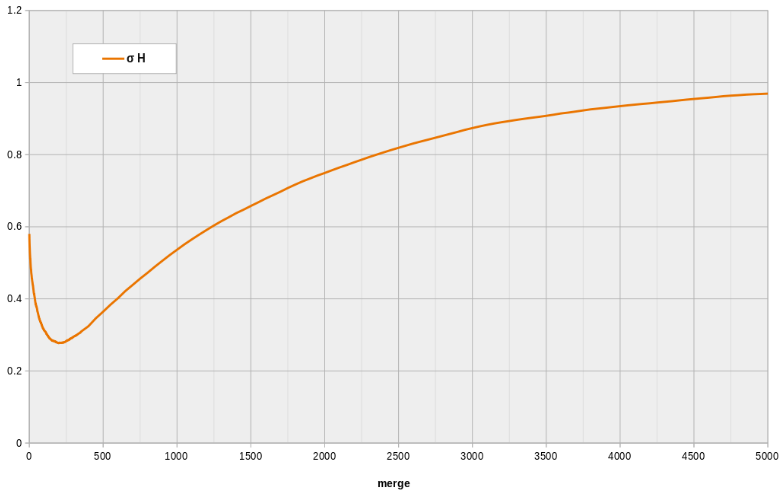
Subword tokenisation
BPE Introduced in **2016**
For some reason works with NNs

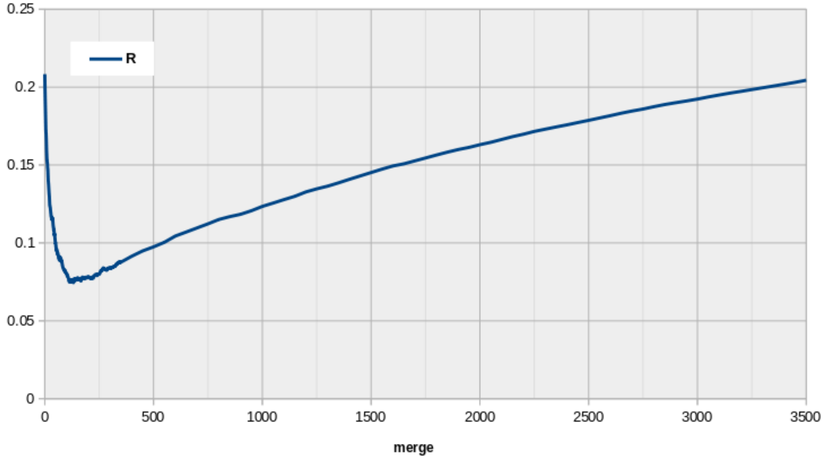
Problems
-- What subwords?
-- Pre-trained models come with a selected (arbitrary) tokenisation
-- Discrimination against languages other than English


~Char level
g o d c r e a t e d t h e h e a v e n ...
g o d c r e a t e d th e h e a v e n ...
g o d c r e a t e d the h e a v e n ...
g o d c r e a t ed the h e a v e n ...
⋮
god creat ed the heaven ...
compression
merges
Many subword levels
~Word level

How about stopping BPE at minimum redundancy? (BPE-MR)





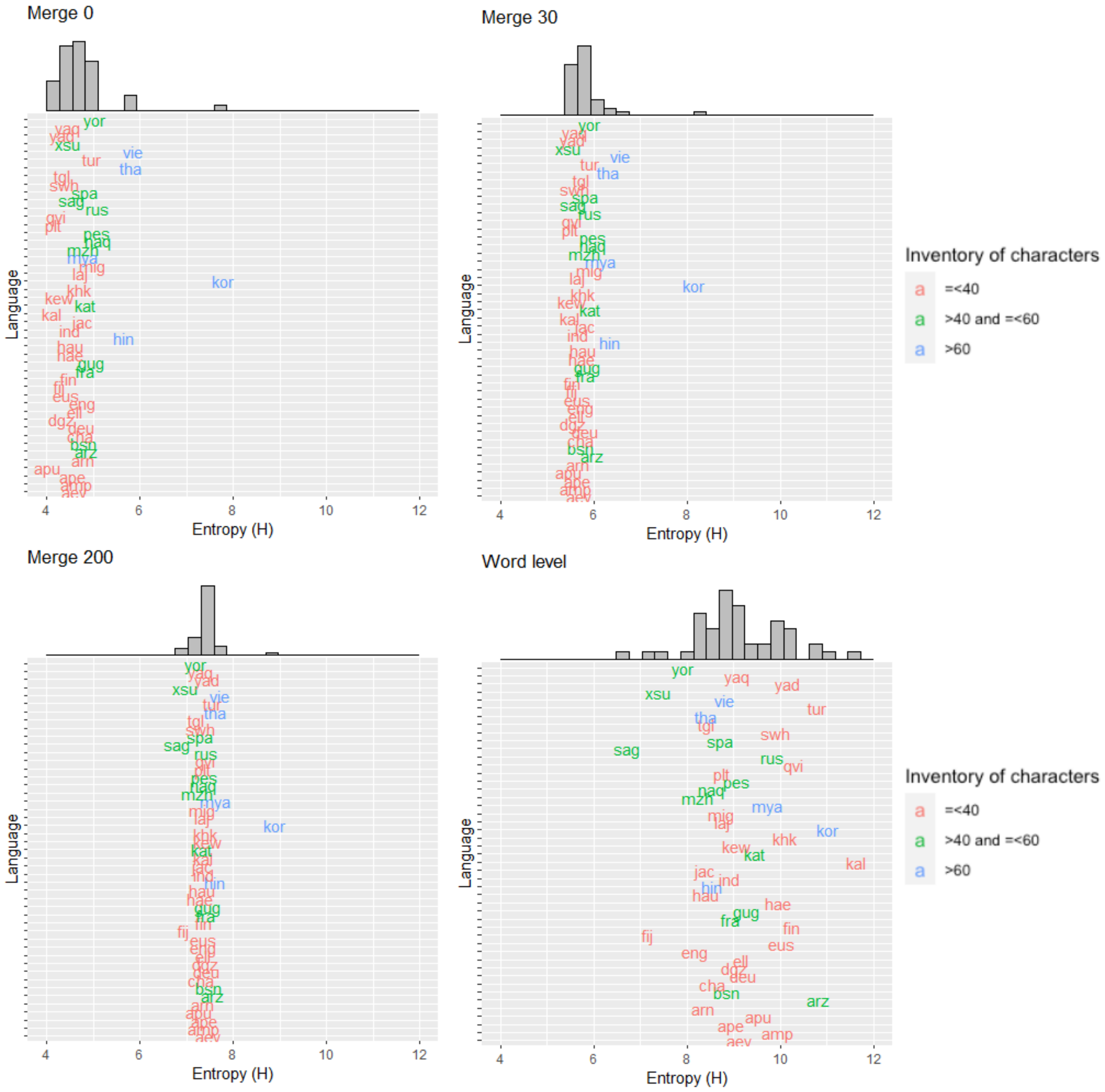


Minimum redundancy, converging text entropy
There is an area of subword tokenisations where text redundancy is minimised. In the same area, text entropy growth slows down and the values across languages start converging

From characters to words: the turning point of BPE merges
Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, Tanja Samardžić
EACL2021

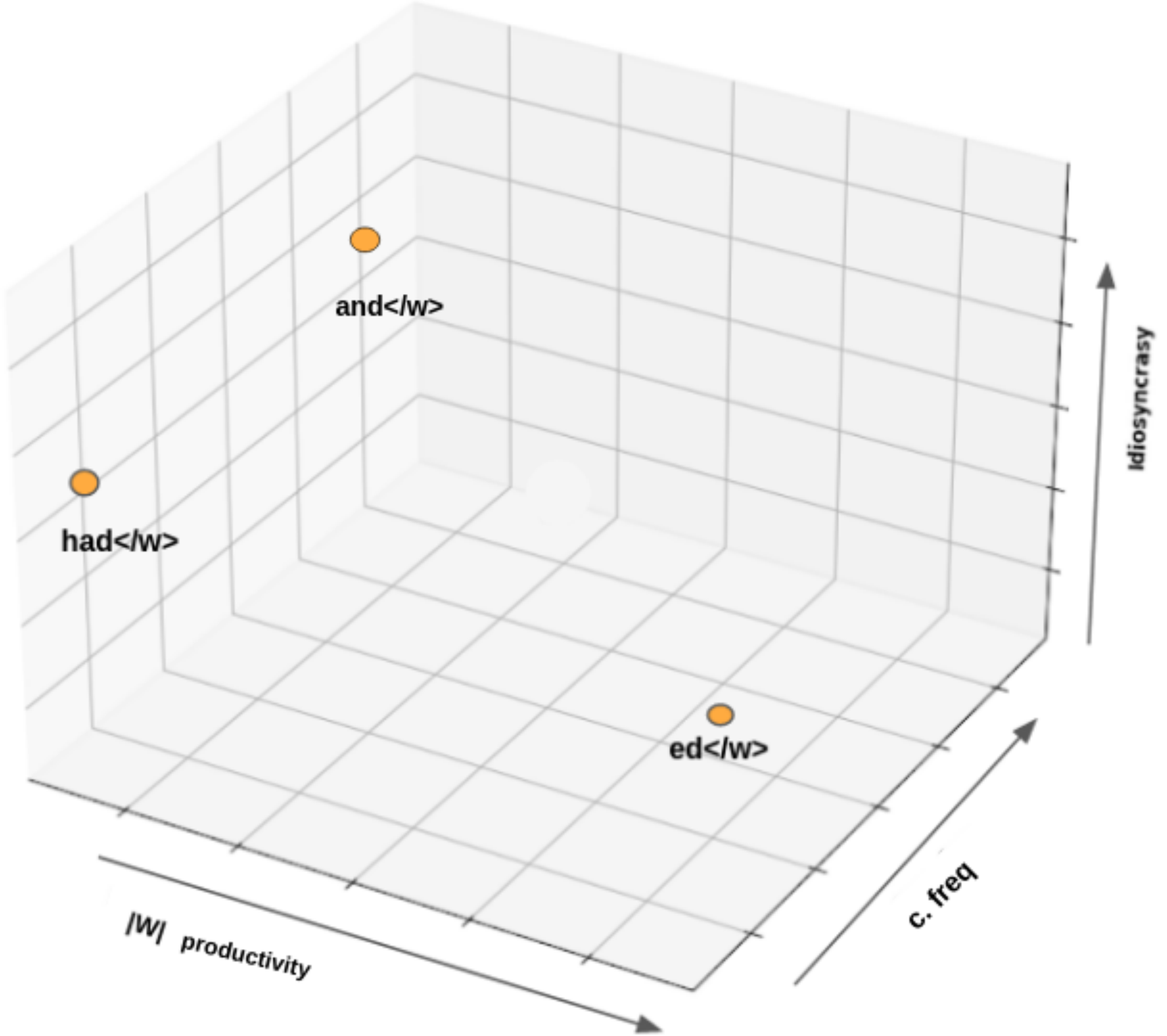
Text entropy across languages

Early merges (200-350)
Text entropy almost the same across 47 languages in the parallel Bible corpus



From characters to words: the turning point of BPE merges
Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, Tanja Samardžić
EACL2021

What are the BPE units at minimum redundancy?



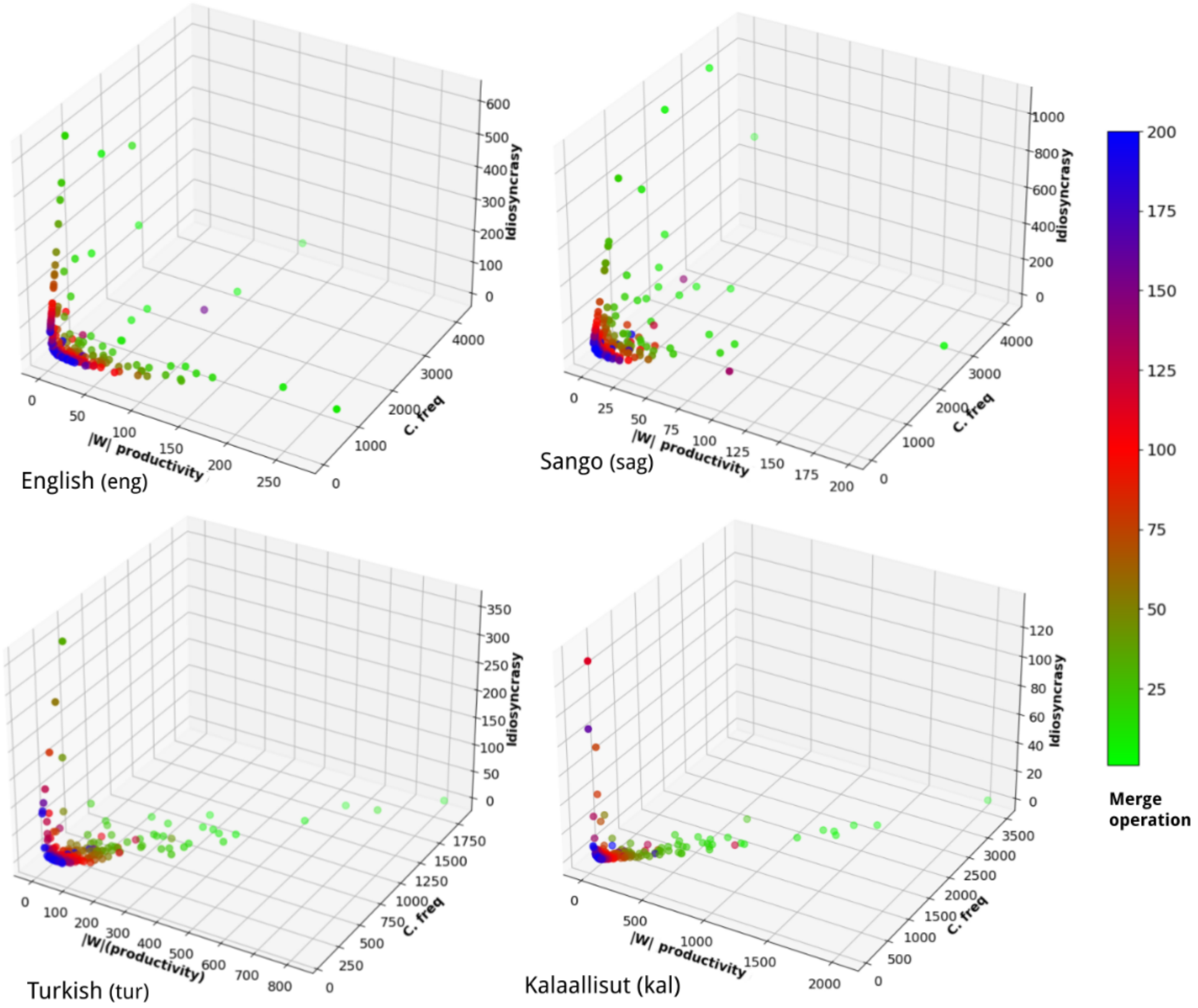
Observations in a 3D space
It looks like we have productive affixes on the floor, and function words on the left-hand side wall

Languages through the Looking Glass of BPE Compression
Ximena Gutierrez-Vasques, Christian Bentz, Tanja Samardžić
Computational Linguistics 2023

How do languages look like in this 3D space?

Languages have different shapes
Those with longer, more complex words tend to have most items on the floor, those with short words tend to have most items on the wall

This BPE units that are merged first are the most discriminative

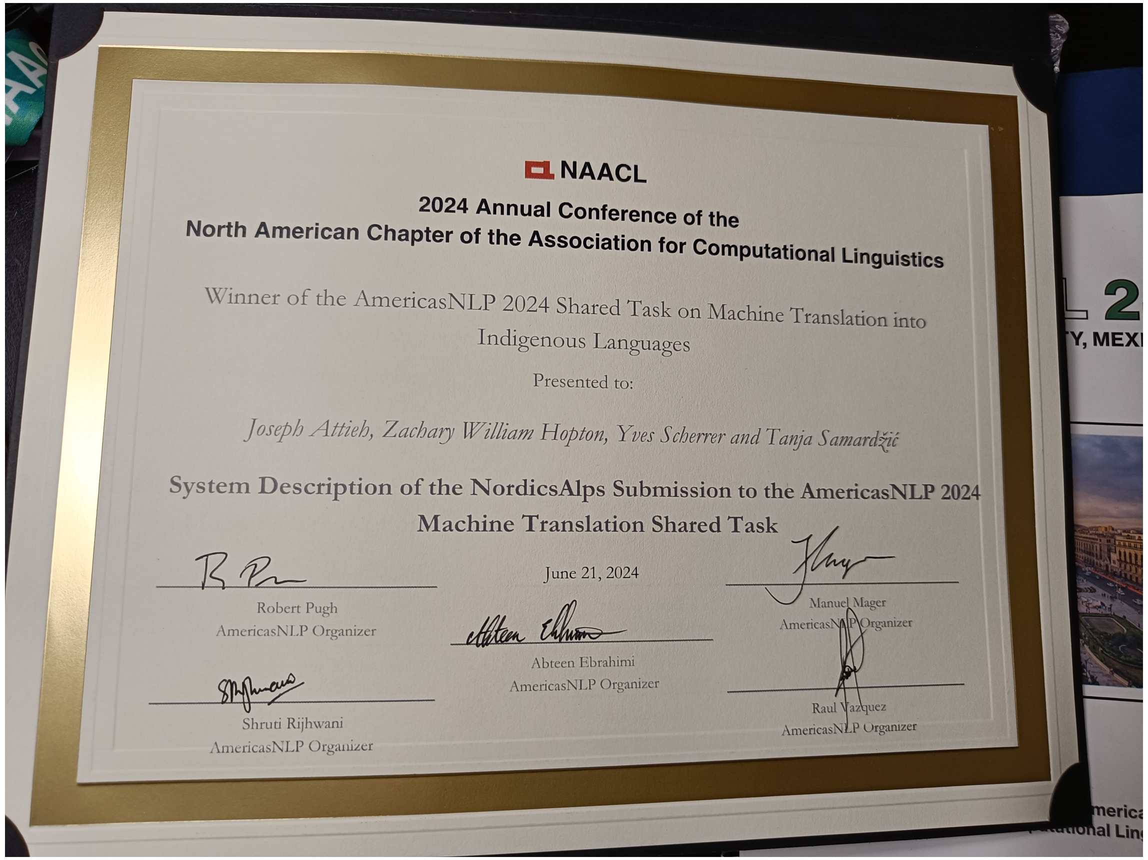


Languages through the Looking Glass of BPE Compression
Ximena Gutierrez-Vasques, Christian Bentz, Tanja Samardžić
Computational Linguistics 2023

Implications for machine translation

Translation from Spanish into 11 American indigenous languages
Hñähñu, Wixarika, Nahuatl, Guaraní, Bribri, Rarámuri, Quechua, Aymara, Shipibo-Konibo, Asháninka, Chatino

NordicAlps
The only system that outperformed last year's winner



NAACL

2024 Annual Conference of the
North American Chapter of the Association for Computational Linguistics

Winner of the AmericasNLP 2024 Shared Task on Machine Translation into Indigenous Languages

Presented to:

Joseph Attieh, Zachary William Hopton, Yves Scherrer and Tanja Samardžić

System Description of the NordicsAlps Submission to the AmericasNLP 2024 Machine Translation Shared Task

June 21, 2024

Robert Pugh
AmericasNLP Organizer

Abteen Ebrahimi
AmericasNLP Organizer

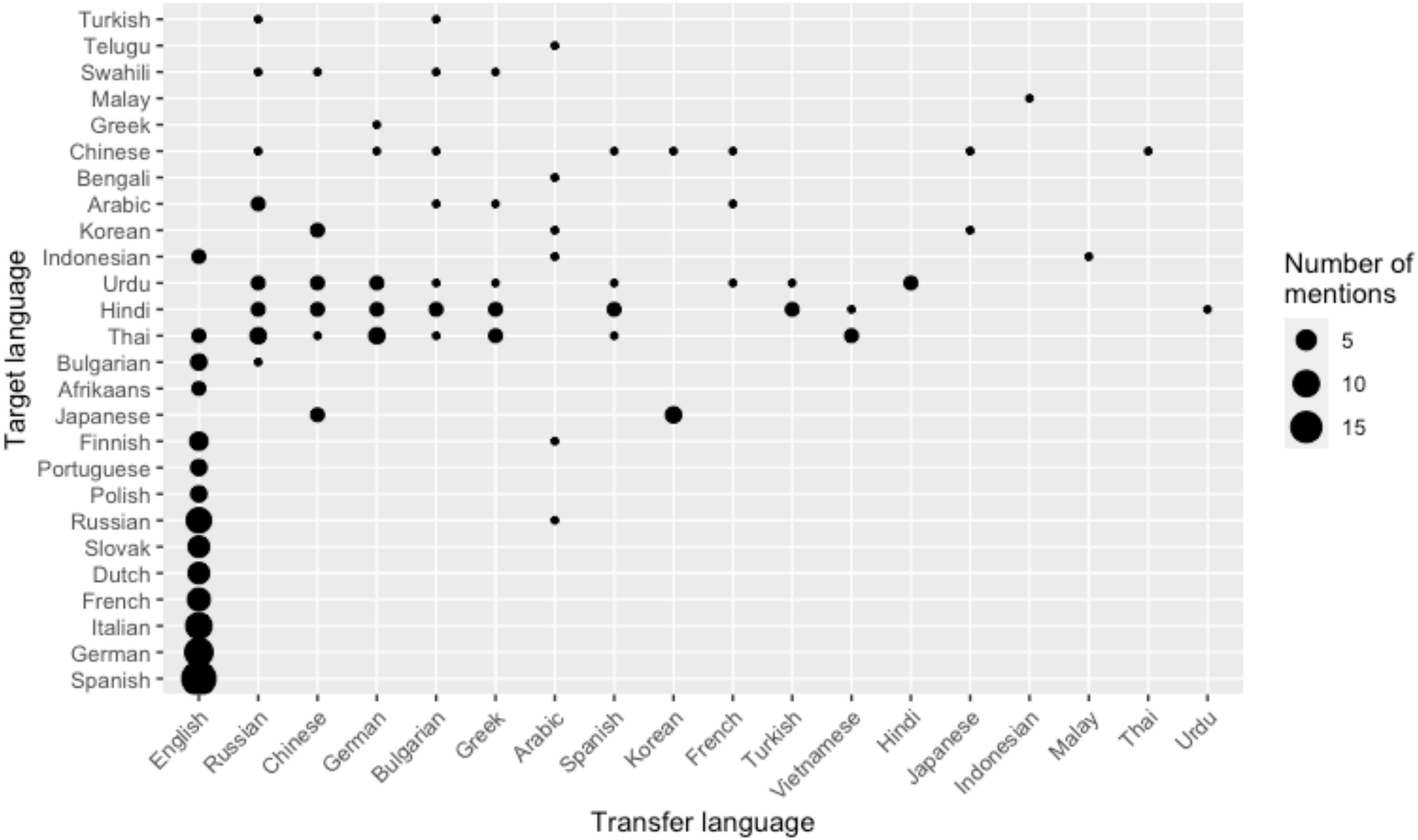
Manuel Mager
AmericasNLP Organizer

Shruti Rijhwani
AmericasNLP Organizer

Raul Vazquez
AmericasNLP Organizer

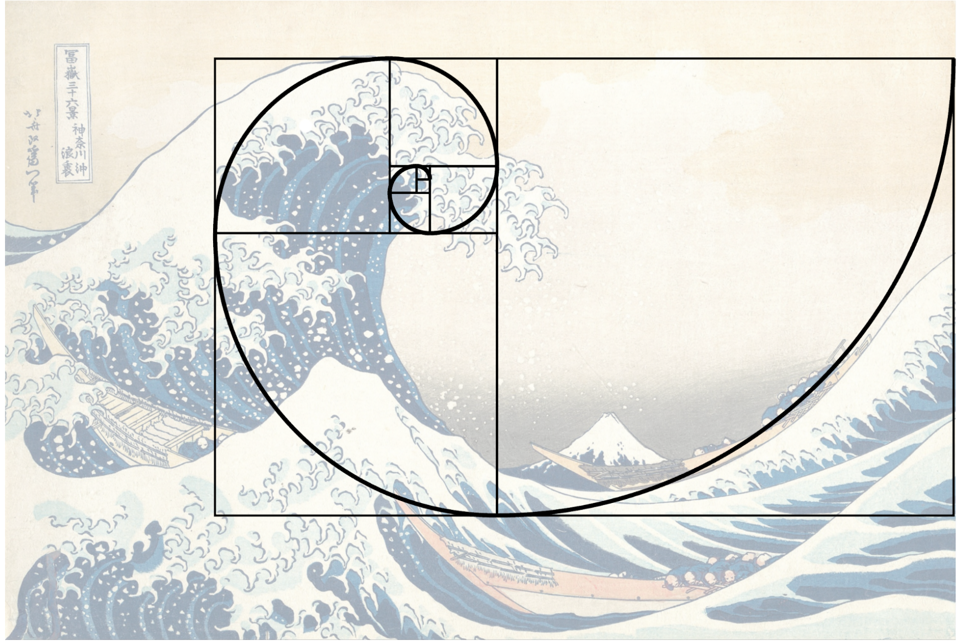
System Description of the NordicsAlps Submission
Joseph Attieh, Zachary Hopton, Yves Scherrer, Tanja Samardžić
AmericasNLP 2024

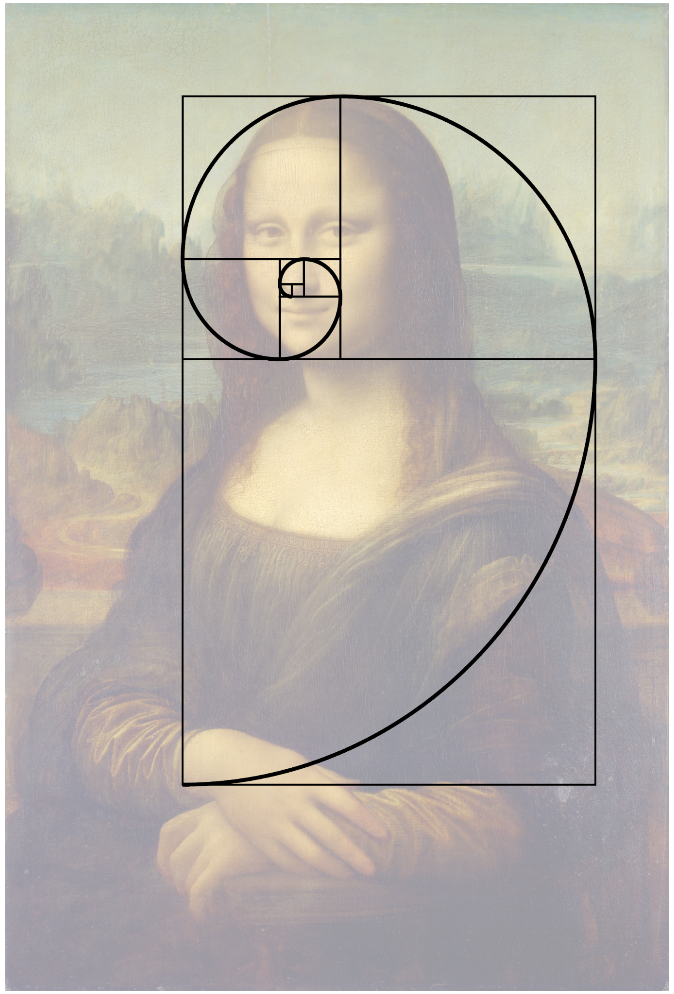
Cross-lingual transfer of language models is unpredictable



Subword evenness (SuE) as a predictor of cross-lingual transfer to low-resource languages
Olga Pelloni, Anastassia Shaitarova, Tanja Samardžić
EMNLP 2022

Subword geometry as a predictor of cross-lingual transfer





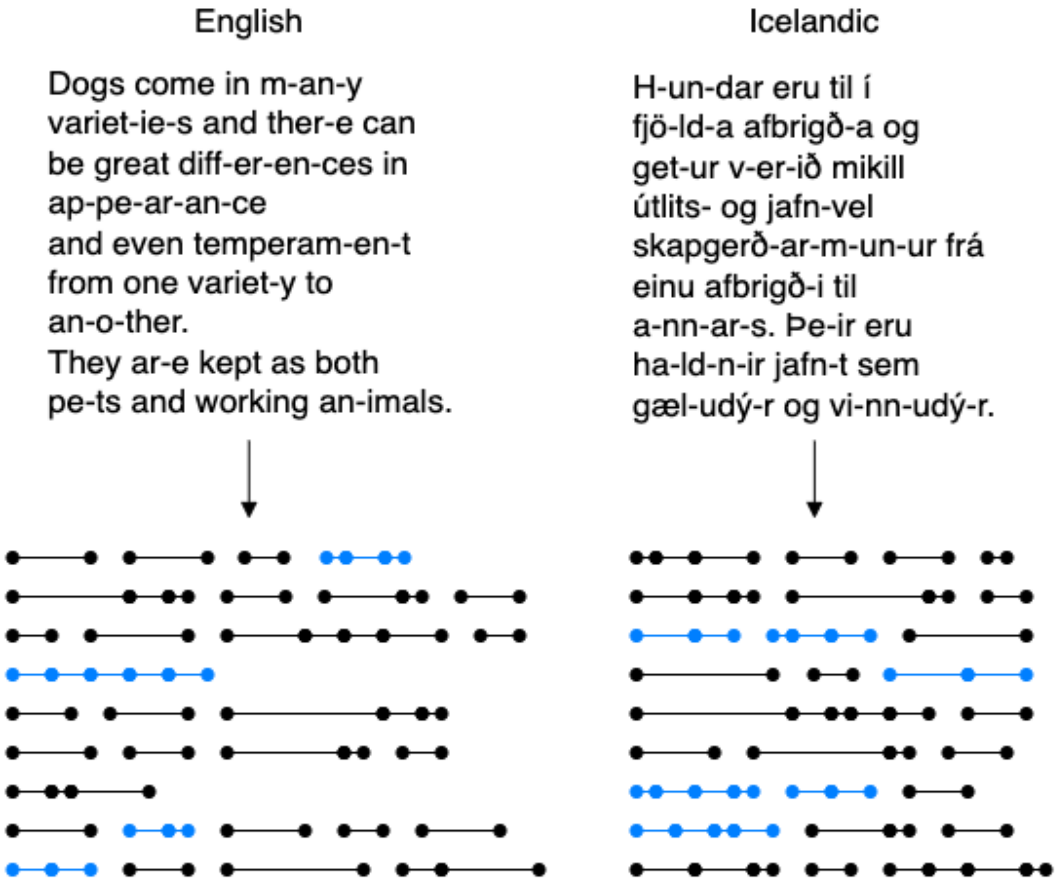
Subword evenness (SuE) as a predictor of cross-lingual transfer to low-resource languages
Olga Pelloni, Anastassia Shaitarova, Tanja Samardžić
EMNLP 2022

Subwords as lines

BPE-MR tokenisation
We stop BPE at minimum redundancy and look at what we get as geometric patterns

Evenness
Some patterns are more even than others

Subword evenness (SuE) as a predictor of cross-lingual transfer to low-resource languages
Olga Pelloni, Anastassia Shaitarova, Tanja Samardžić
EMNLP 2022



English

Dogs come in m-an-y
variet-ie-s and ther-e can
be great diff-er-en-ces in
ap-pe-ar-an-ce
and even temperam-en-t
from one variet-y to
an-o-ther.
They ar-e kept as both
pe-ts and working an-imals.

Icelandic

H-un-dar eru til í
fjö-ld-a afbrigð-a og
get-ur v-er-ið mikill
útlits- og jafn-vel
skapgerð-ar-m-un-ur frá
einu afbrigð-i til
a-nn-ar-s. Þe-ir eru
ha-ld-n-ir jafn-t sem
gæl-udý-r og vi-nn-udý-r.

Finnish

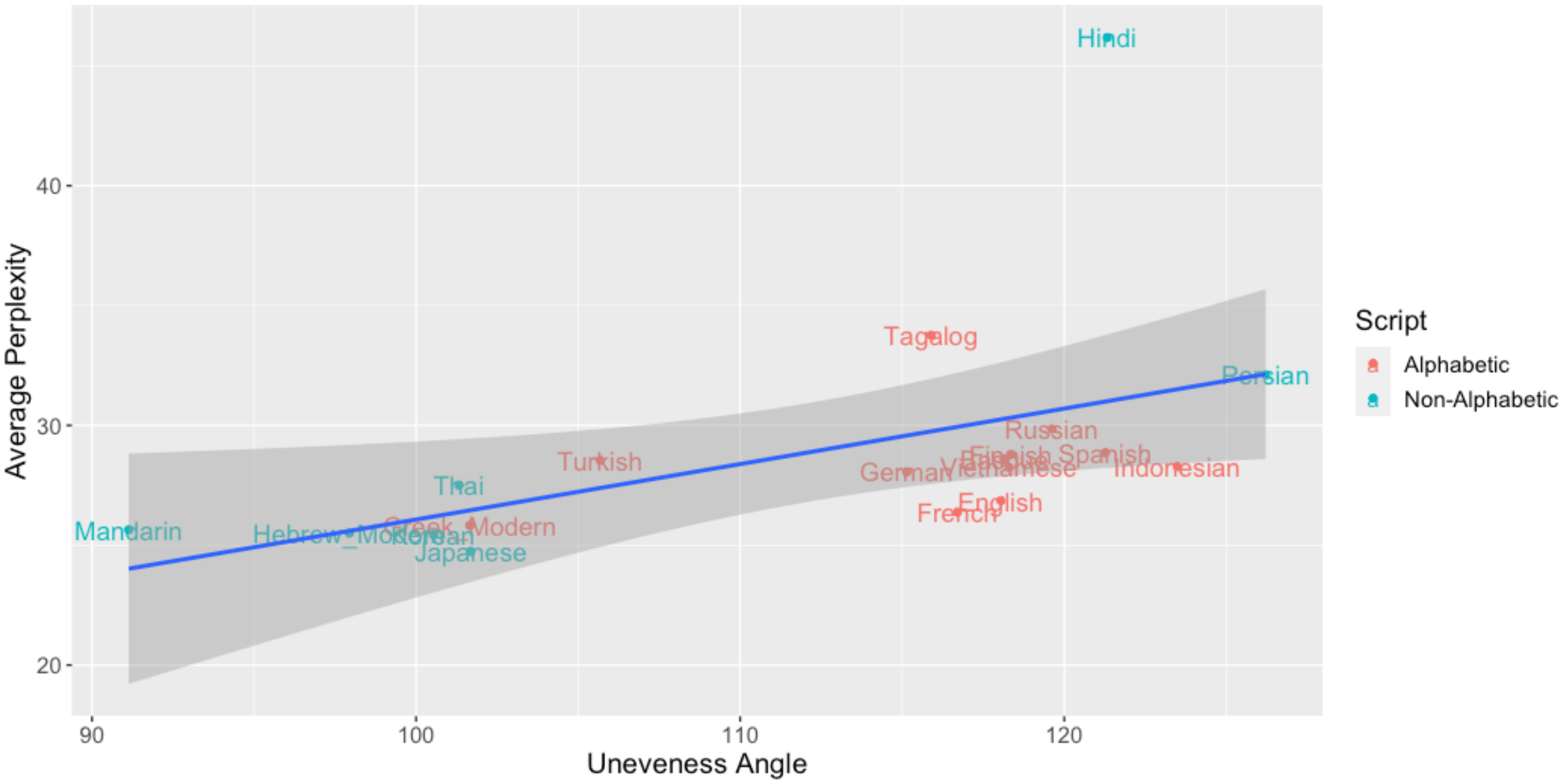
K-oi-ria on m-on-ia
lajikke-ita, ja
ulk-on-äö-ss-ä ja jopa
lu-on-tee-ss-a voi
olla suu-ria ero-ja
lajikke-esta t-oi-se-en.
Nii-tä pid-et-ään
se-kä lemm-ikke-in-ä
et-tä työeläim-in-ä.

Predicting the perplexity of a transferred language model

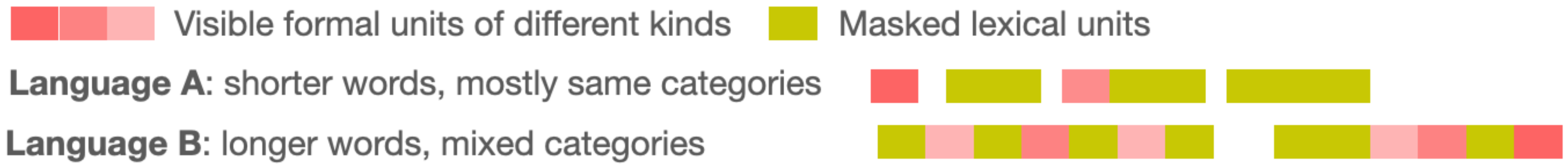
Languages
49 languages from the TeDDi sample

TeDDi Sample: Text Data Diversity Sample for Language Comparison and Multilingual NLP
Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Pelloni, and Tanja Samardžić
LREC 2022

Subword evenness (SuE) as a predictor of cross-lingual transfer to low-resource languages
Olga Pelloni, Anastassia Shaitarova, Tanja Samardžić
EMNLP 2022



Future work: separate form and content in any language



Visible formal units of different kinds

Masked lexical units

Language A: shorter words, mostly same categories

Language B: longer words, mixed categories

Represent formal units
Can we learn the syntax of formal units by ignoring what is not formal?

Efficient language embeddings
Can we learn language representations from sequences of formal units?

General data efficiency
Can we model the meaning of texts better if we know at least something about their structure?

Conclusion

Linguistics currently not part of computational language modelling
This situation is due to complex history of the study of natural language, not a necessity

Linguistics might require computational language modelling
Given the complexity of natural language, its scientific study might depend on our ability to deal with a lot of data and computational modelling. Computational modelling is a way to make explicit predictions and test them. It can be the way towards reuniting data and theory.

Partial overlap between NLP and linguistics
Not all NLP has to serve linguistics, and not all linguistics has to rely on NLP, but there can and should be an overlap. Example topics are text tokenisation and cross-lingual transfer of language models.

SMASH
An opportunity for a more scientific computational language modelling!