Research at TurkuNLP

Veronika Laippala

TURKUNLP
.ORG

UNIVERSITY
OF TURKU

TURKUNLP .ORG

TurkuNLP

(NLP = Natural Language Processing)

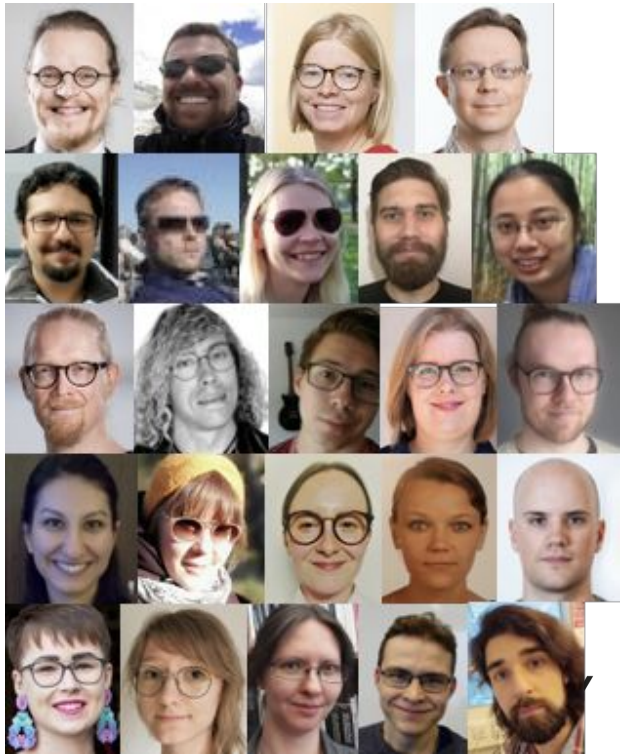
Founded in 2001, now ~25 members

Varied backgrounds, both **CS and languages**

Broad range of projects on **variety of NLP topics and digital linguistics**

Strong focus on **machine learning**-based approaches to language

If you do modern NLP on Finnish, you're probably using some tools and/or resources developed in Turku!



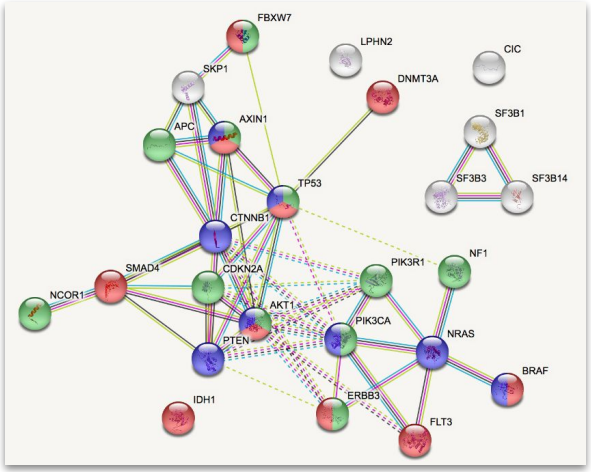
What do we do?

TURKUNLP
.ORG

UNIVERSITY
OF TURKU

TURKUNLP .ORG

Ongoing projects



Part 1: Text processing

- Universal Dependencies for Finnish and Turku Neural Parser
- Turku NER Named Entity Recognition

Part 2: Neural Language Models

- FinBERT and FinGPT
- Very large language data - Finnish Internet Parsebank

Part 3: Understanding texts

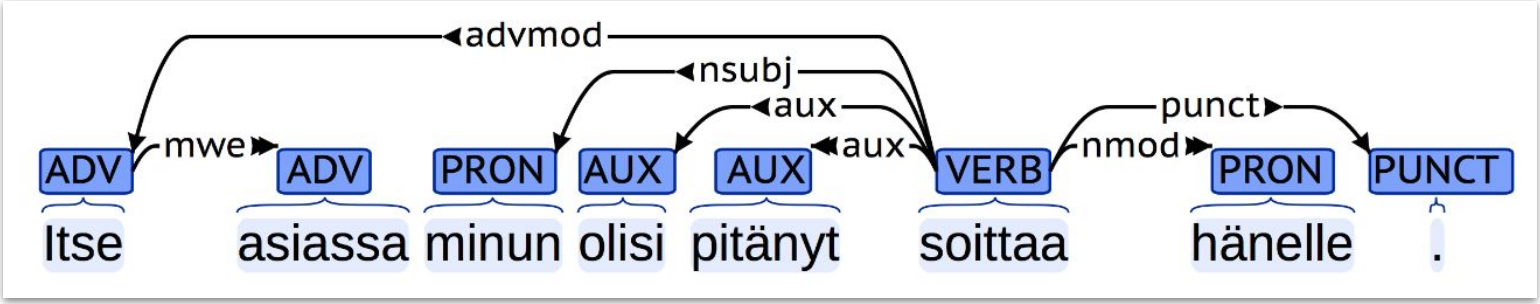
- Toxicity detection
- Paraphrases
- Web registers (genres)
- Parliamentary debates

UNIVERSITY OF TURKU

Syntactic parsing of Finnish

Syntactic analysis (or parsing) of Finnish is a major long-term research focus at TurkuNLP

Task setting: take raw unstructured text, produce representation of words, their base forms, morphological features, and syntactic dependencies



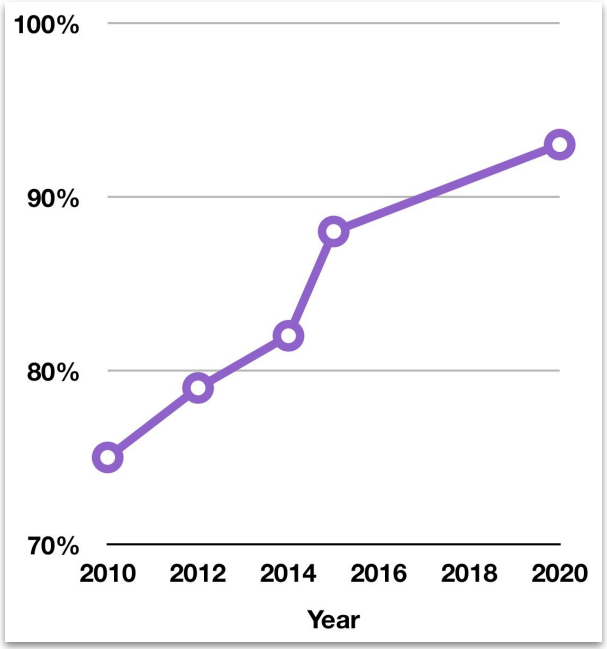
Syntactic parsing of Finnish

Manually annotate training data from scratch (several person-years of effort)

Train ever-better parsers on this data with improving ML methodology

Methodological progress brought parsing accuracy **from useless to practically human level** on “clean” text

Most recent improvements from introduction of **deep learning** models



Kanerva et al. (2018) Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task

Syntactic parsing of Finnish

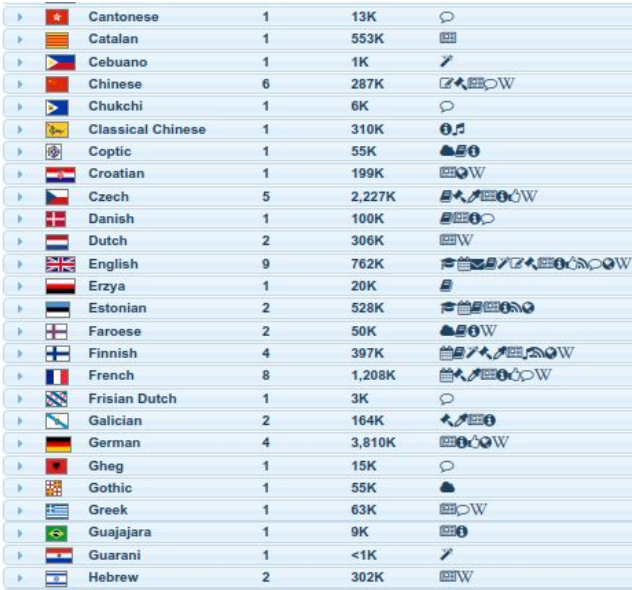
Data initially introduced as **Turku Dependency Treebank**

Now included as **UD-Finnish-TDT** in Universal Dependencies, the largest uniform, multilingual collection of syntactically annotated corpora

(Modern parsers primarily trained on UD)

Building and maintaining datasets is an important task in NLP, often very undervalued

→ High-quality datasets can persist much longer than specific machine learning approaches



Cantonese	1	13K	
Catalan	1	553K	
Cebuano	1	1K	
Chinese	6	287K	W
Chukchi	1	6K	
Classical Chinese	1	310K	
Coptic	1	55K	
Croatian	1	199K	W
Czech	5	2,227K	W
Danish	1	100K	
Dutch	2	306K	W
English	9	762K	W
Erzya	1	20K	
Estonian	2	528K	
Faroese	2	50K	W
Finnish	4	397K	W
French	8	1,208K	W
Frisian Dutch	1	3K	
Galician	2	164K	
German	4	3,810K	W
Gheg	1	15K	
Gothic	1	55K	
Greek	1	63K	W
Guajajara	1	9K	
Guarani	1	<1K	
Hebrew	2	302K	W

universaldependencies.org

Haverinen et al. (2014) Building the essential resources for Finnish: the Turku Dependency Treebank
Pyysalo et al. (2015) Universal Dependencies for Finnish

Universal Dependencies (UD)

Open community effort to create cross-linguistically consistent treebank annotation for many languages

https://universaldependencies.org/

Motivation of UD

Increasing interest in multilingual NLP

- Studies involving several languages
- Multilingual evaluation

Cross-lingual learning - learn from English data, adapt the knowledge for Finnish data

- All these studies rely on annotated data
- Traditionally annotation schemes differ between corpora annotated for the same task

TURKUNLP .ORG

Included steps

Segmentation

Morphological analysis

Lemmatization

Dependency structure

UNIVERSITY OF TURKU

Included steps

Segmentation

- Tokenization, sentence splitting
- White space + punctuation
- However, note "20 000", emoticons ":)", abbreviations "e. g."

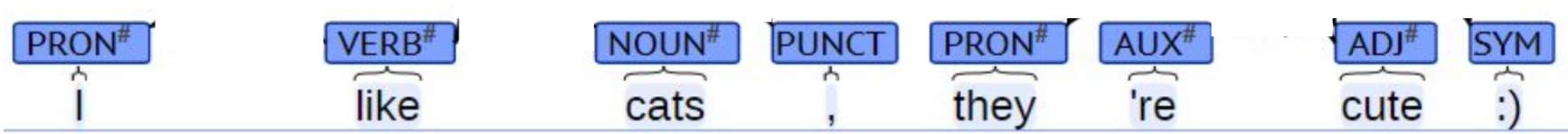
Morphological analysis

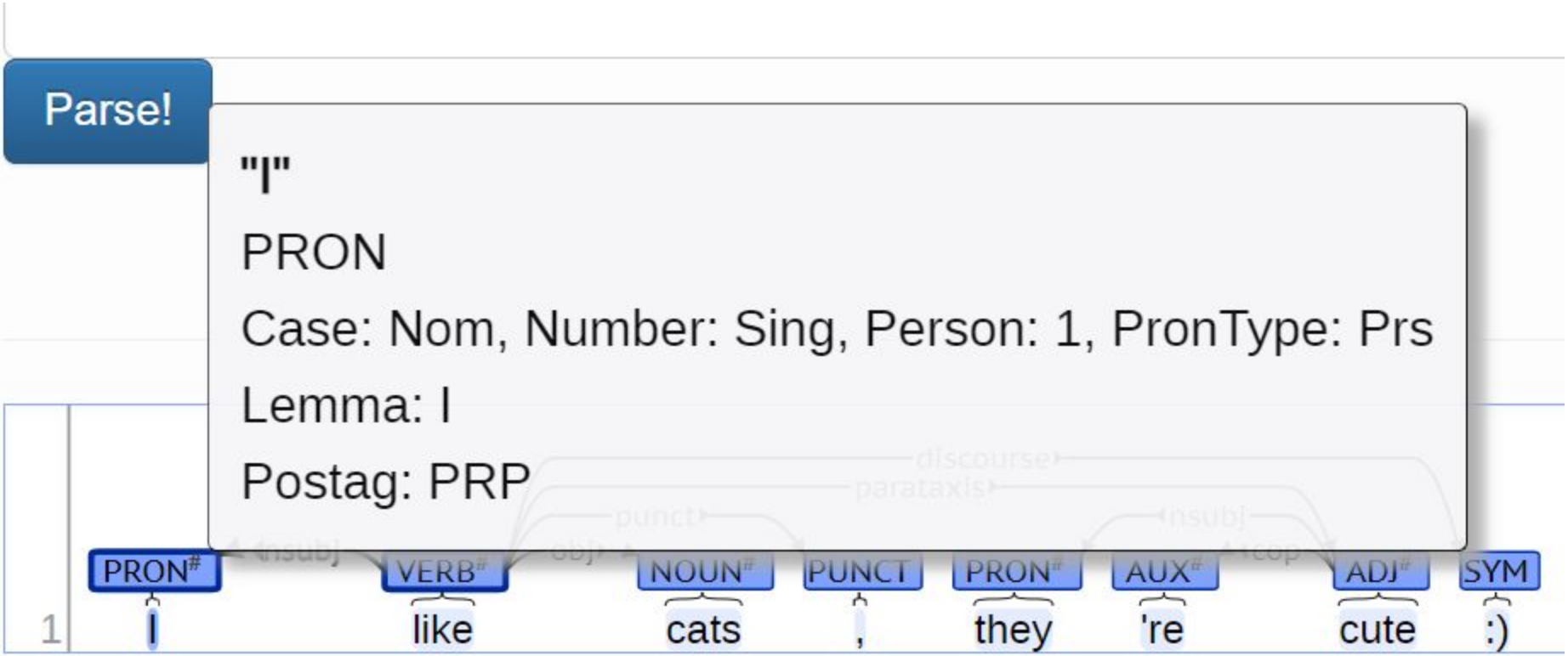
- Part-of-speech (15 classes in Finnish)
- Morphological features (many many features, co-occur “Case=Nom|Number=Sing”

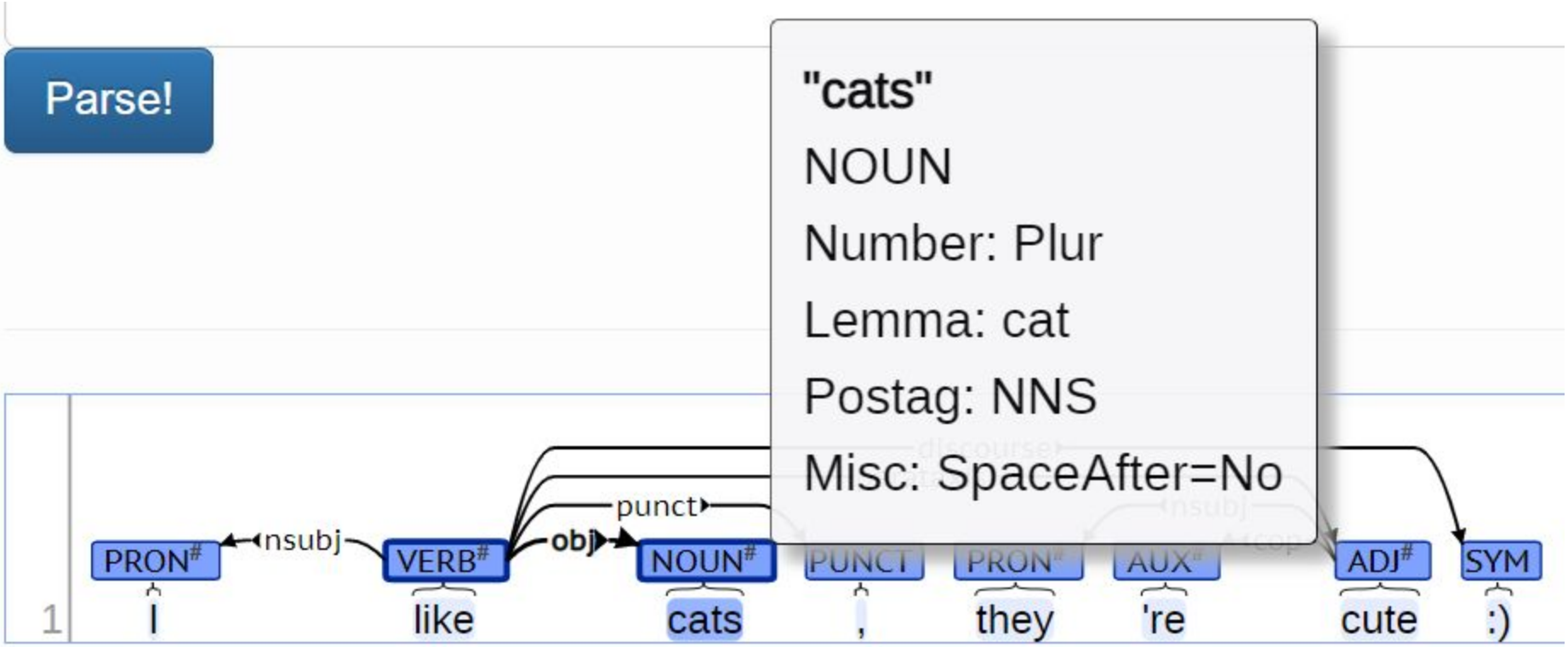
Lemmatization

- Canonical baseform - what would be in a dictionary?

I like cats , they 're cute :)

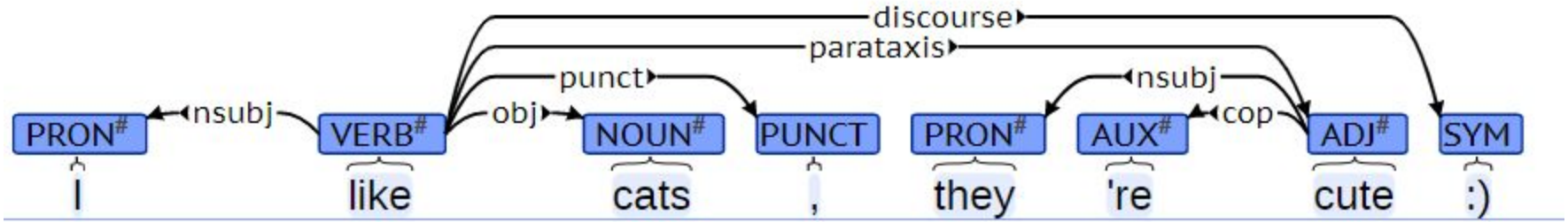

PRON#
VERB#
NOUN#
PUNCT
PRON#
AUX#
ADJ#
SYM
I
like
cats
,
they
're
cute
:)


Parse!
"I"
PRON
Case: Nom, Number: Sing, Person: 1, PronType: Prs
Lemma: I
Postag: PRP
PRON#
VERB#
NOUN#
PUNCT
PRON#
AUX#
ADJ#
SYM
1
I
like
cats
,
they
're
cute
:)


Parse!
"cats"
NOUN
Number: Plur
Lemma: cat
Postag: NNS
Misc: SpaceAfter=No
nsubj
punct
obj
PRON#
VERB#
NOUN#
PUNCT
PRON#
AUX#
ADJ#
SYM
1
I
like
cats
,
they
're
cute
:)

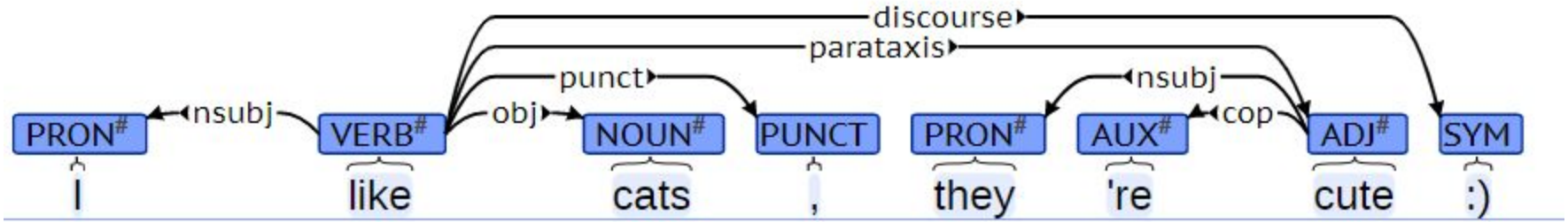
Dependency syntax

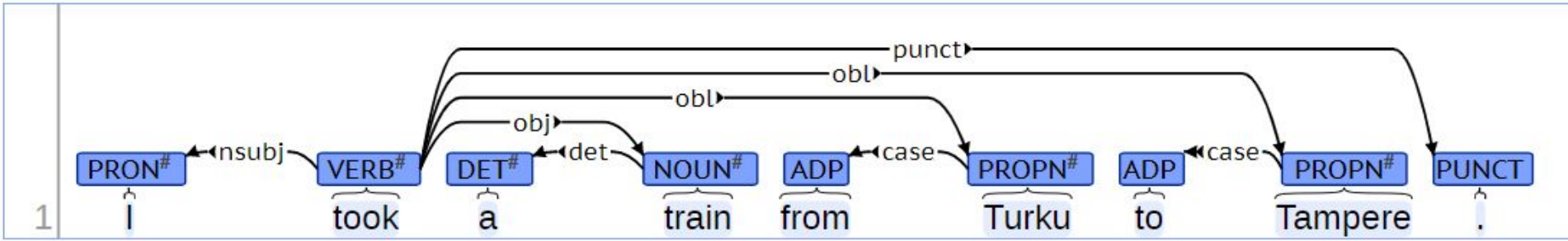
- Typed dependency relations between words
- Basic dependency representation forms a tree
  - Exactly one word is the head of the sentence
  - All other words are dependent on another word in the sentence

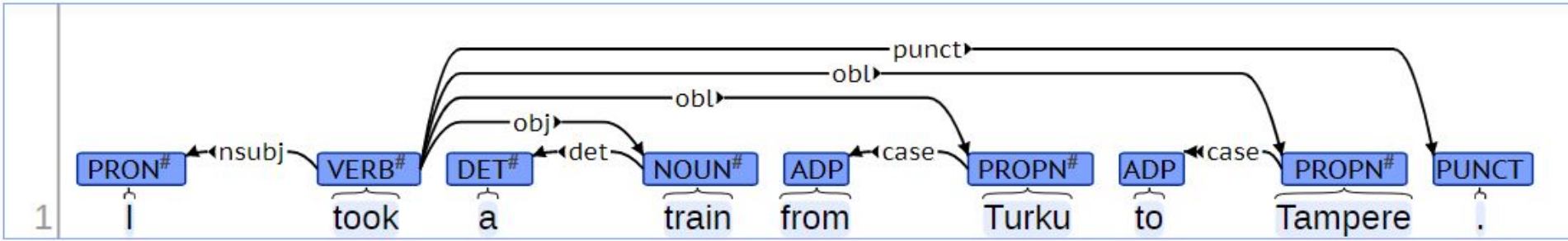


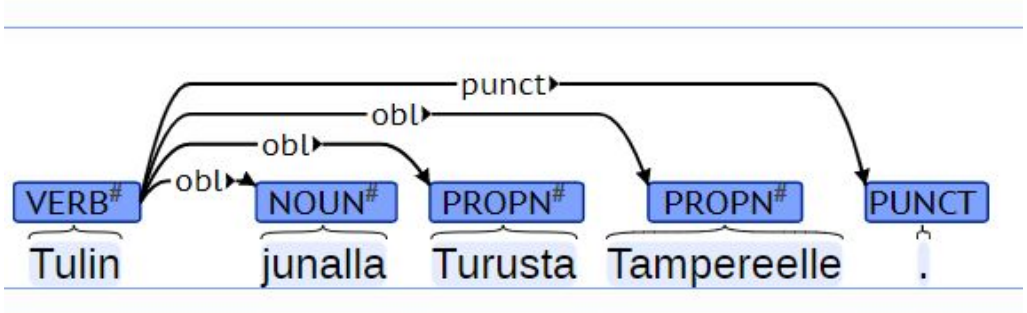
Dependency syntax

- Dependency relations primarily between content words
  - Function words attach as direct dependents of the most closely related content word
  - Content words as heads maximizes parallelism between languages because content words vary less than function words between languages
  - Punctuation attaches to the head of the clause or phrase to which they belong
- 37 universal relations + language-specific subtypes
- https://universaldependencies.org/u/dep/index.html






punct
obl
obl
obj
PRON# nsubj VERB# DET# det NOUN# ADP case PROPN# ADP case PROPN# PUNCT
1 I took a train from Turku to Tampere .


punct
obl
obl
obl
VERB# NOUN# PROPN# PROPN# PUNCT
Tulin junalla Turusta Tampereelle .

Named Entity Recognition

Named Entity Recognition (NER) is a key step for extracting structured information from text

Task setting: take raw text, identify spans mentioning proper names (and often e.g. dates) and assign each a type (`PERSON`, `LOCATION`, …)

Two manually annotated corpora for NER introduced in Turku, one initially starting as an annotation project on NLP course

Luoma et al. (2020) A broad-coverage corpus for Finnish named entity recognition
Luoma et al. (2021) Fine-grained Named Entity Annotation for Finnish

Named Entity Recognition

NER remains an active topic of research at TurkuNLP

(Model performance continues to improve!)

Luoma et al. (2020) A broad-coverage corpus for Finnish named entity recognition
Luoma et al. (2021) Fine-grained Named Entity Annotation for Finnish

Try these out!

Demo at http://epsilon-it.utu.fi/parser_demo/

Output explained at:

https://github.com/TurkuNLP/gf_summerschool/blob/main/gf_parser_output_explained.ipynb

Try a parser trained on TDT:

https://github.com/TurkuNLP/gf_summerschool/blob/main/trankit.ipynb

NER:

https://github.com/TurkuNLP/gf_summerschool/blob/main/NER.ipynb

Part 2: Language Models

UNIVERSITY OF TURKU

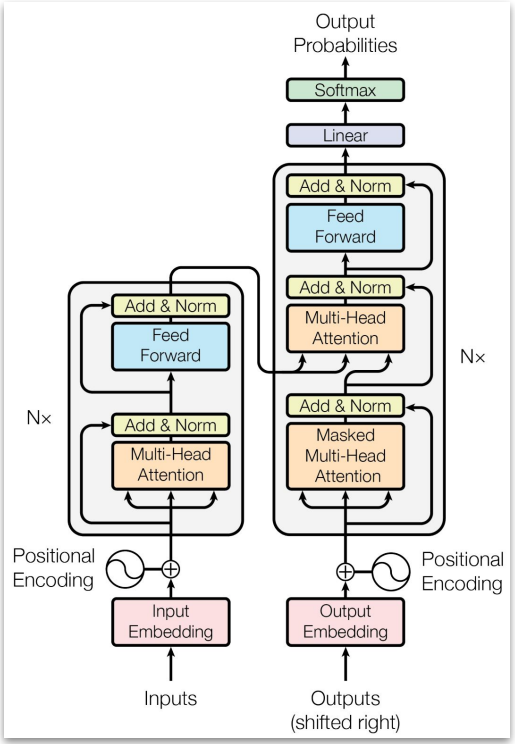
Deep neural language models

Deep neural language models are the basis of much of modern NLP

Transfer learning approaches using the **Transformer architecture** particularly effective

Dedicated models mostly introduced for **English, Chinese** and other large languages

To have state-of-the-art models for small languages like Finnish, we need to make them ourselves

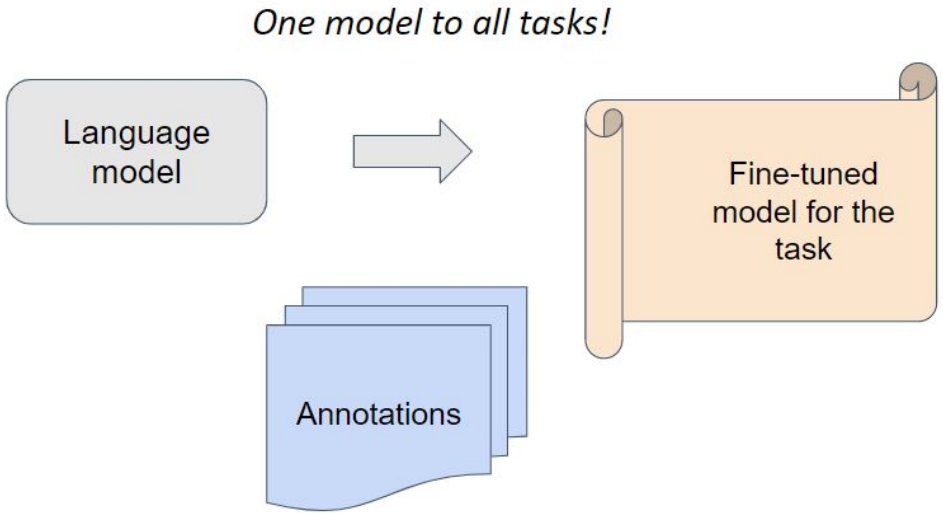


Vaswani et al. (2017) Attention Is All You Need

Language models

Trained on very large amounts of data

Can be fine-tuned to downstream tasks



TURKUNLP .ORG

Finnish Internet Parsebank

- Finnish Internet Parsebank the most important big data for Finnish
- Crawl, clean, and analyze as much of Internet Finnish as possible
- Approx. 8 billion (8,000,000,000) words of reasonably clean text
- Available soon via Kielipankki!

UNIVERSITY OF TURKU

Two types of language models

- Masked LLMs
- Generative models

FinBERT

A version of Google's BERT **deep transfer learning model** for Finnish

Trained using masked language modeling

- Predict a given masked token in the input
- *I have watched this [MASK] and it was awesome.*
- *What is [MASK] name?*

Model with **110B parameters**, trained on **3B words** for 1M steps on CSC supercomputer puhti

http://turkunlp.org/finbert
https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1
Virtanen et al. (2019) Multilingual is not enough: BERT for Finnish

TURKUNLP .ORG

UNIVERSITY OF TURKU

TURKUNLP .ORG

FinBERT

Available at https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1

UNIVERSITY OF TURKU

FinGPT3

GPT-3 -like generative language models for Finnish

Models ranging from **110M - 176B parameters** trained for **300B tokens** on LUMI supercomputer

Data: news, crawls, social media, national library collections

Released **openly** earlier this year

https://turkunlp.org/gpt3-finnish

EXAMPLE GENERATION:
Suomenkielisen tekoälyteknologian tulevaisuudelle on keskeisen tärkeää, että suomalainen koulujärjestelmä tarjoaa riittävän perusosaamisen tekoälyn hyödyntämiseen myös lapsille ja nuorille. Koulutusjärjestelmämme tulee taata lapsille jo varhaisessa vaiheessa valmiudet, tiedot ja taidot, joilla he pystyvät luomaan ja jakamaan itse tietoa tekoälyyn liittyen. Tämän lisäksi tulee kiinnittää huomiota tekoälyn opettamiseen ja siihen, miten tekoäly linkittyy eri oppiaineisiin. […]

FinGPT3

LUMI: Fastest supercomputer in Europe (3rd in world)

2560 GPU nodes w/AMD MI250X (20480 GCDs)

Peak performance 550 Pflops (pre-exascale)

Large allocations (1M+ GPUh) via EUHPC and national orgs

TURKUNLP .ORG

FinGPT-3

We have trained generative GPT-3-like models for Finnish. We are currently in the process of evaluating and documenting the models and finalizing their release.

Prior to the full release, we are offering access to some of the models via the Hugging Face model repository:

- Finnish GPT-3 small
- Finnish GPT-3 medium
- Finnish GPT-3 large
- Finnish GPT-3 xl
- Finnish GPT-3 3B
- Finnish GPT-3 8B
- Finnish GPT-3 13B
- BLOOM + Finnish 176B *coming soon!*

UNIVERSITY OF TURKU

TURKUNLP .ORG

Try this out!

Available at https://huggingface.co/TurkuNLP/gpt3-finnish-13B

Try the notebook at

https://github.com/TurkuNLP/gf_summerschool/blob/main/text_generation_example.ipynb

UNIVERSITY OF TURKU

Part 3: Understanding what's in the data

UNIVERSITY OF TURKU

Many projects towards understanding language use

Toxicity detection

Paraphrase detection

Detecting web registers (genres) to make sense of web data

Exploring political ideologies through parliamentary speech

TURKUNLP .ORG

Toxicity

Automatic identification of **toxic language** is crucial for moderating social media channels and for cleaning datasets used to train LLMs

Task setting: given text, predict all applicable toxicity labels

Problem: no data for Finnish! Manual annotation is time-consuming, expensive and difficult.

Suggested solution: let's machine translate! And test on native Finnish!

Multilingual language model vs. DeepL vs. OpusMT

Eskelinen et al. (2023) Toxicity Detection in Finnish Using Machine Translation

UNIVERSITY OF TURKU

Data 1/2

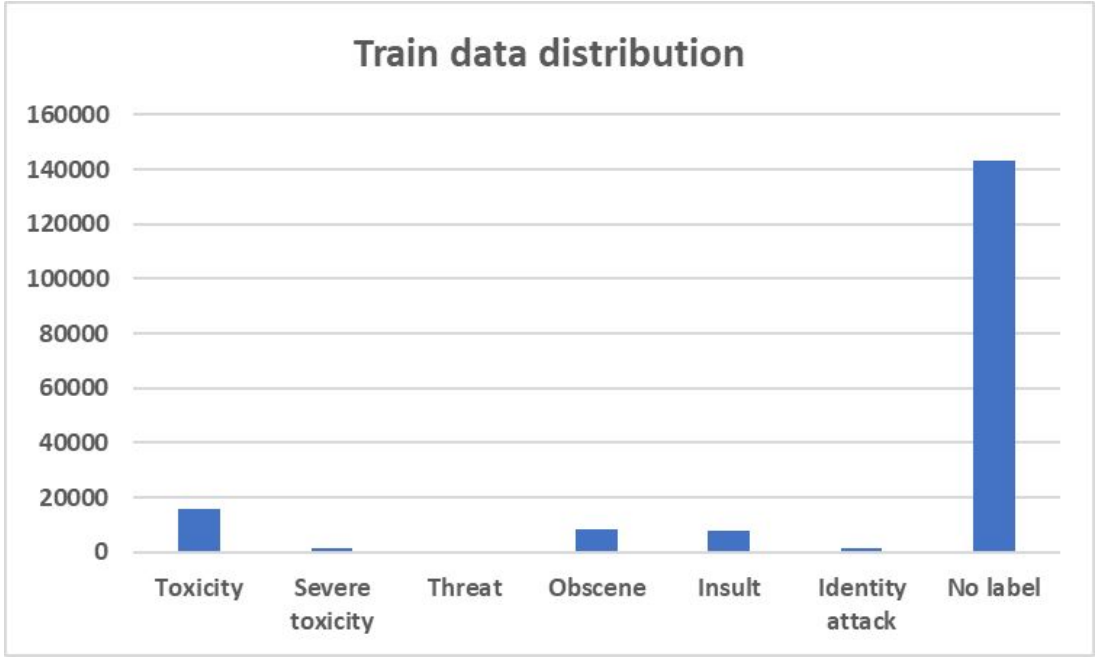
Jigsaw Toxicity dataset

Train & test size:

159,571 & 63,978

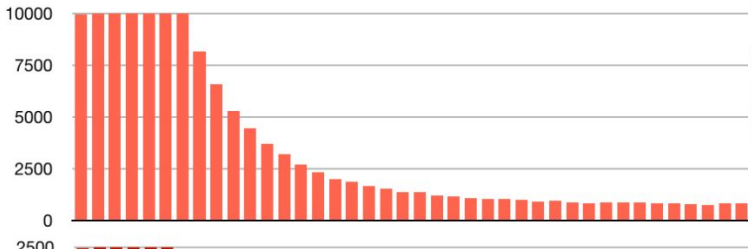
Labels:

toxicity, severe toxicity, identity attack, insult, obscene and threat



Data 2/2

Suomi24 annotated test

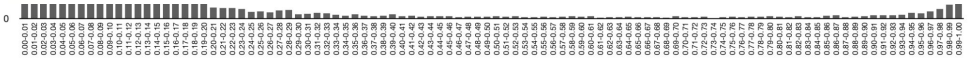


Annotation process
-> Guidelines

Agreement

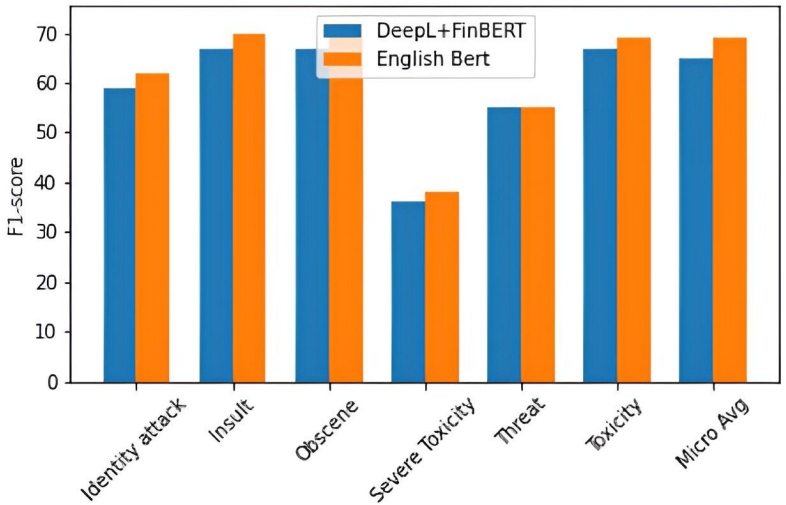
	Initial	After discussion
Toxicity	58%	54%
Severe toxicity	63%	66%
Threat	82%	80.3%
Obscene	69%	62%
Insult	47.5%	49.6%
Identity attack	54.5%	66.6%
Mean	62.3%	63%

Table 3: Unanimous inter-annotator agreement (IAA) for the native Finnish toxicity dataset



Results 1/2

Model	Train	Test	F1-micro
BERT	En	En	0.69
FinBERT	Fi-DeepL	Fi-DeepL	0.66
FinBERT	Fi-Opus-MT	Fi-Opus-MT	0.65
XLM-R	Fi-DeepL	Fi-DeepL	0.65
XLM-R	En	Fi-DeepL	0.57
XLM-R	Fi-DeepL+En	Fi-DeepL	0.65
BERT	Backtr-En	En	0.67

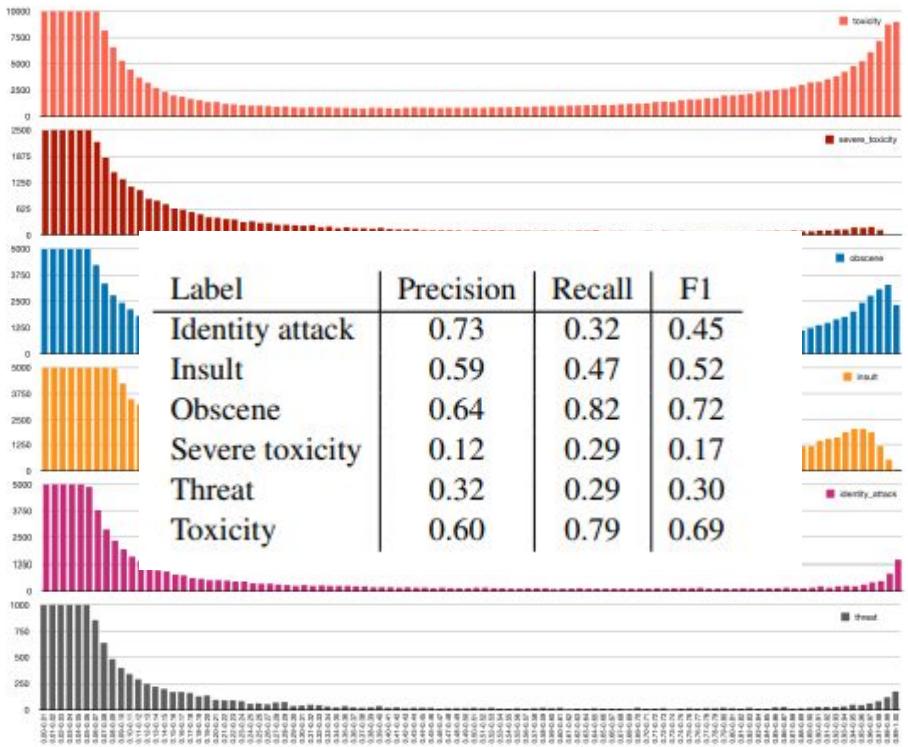


Results 2/2

Annotated results

	Prec	Rec	F1
FinBERT-DeepL	0.57	0.59	0.58
FinBERT-DeepL Weighted	0.61	0.74	0.67
XLMR-En	0.50	0.40	0.45
XLMR-En Weighted	0.50	0.40	0.45

Table 6: Micro evaluation results for the native Finnish dataset using threshold 0.5.



Label	Precision	Recall	F1
Identity attack	0.73	0.32	0.45
Insult	0.59	0.47	0.52
Obscene	0.64	0.82	0.72
Severe toxicity	0.12	0.29	0.17
Threat	0.32	0.29	0.30
Toxicity	0.60	0.79	0.69

Let's try it out!

https://github.com/TurkuNLP/gf_summerschool/blob/main/Toxicity_Demo.ipynb

TURKUNLP
.ORG

Turku paraphrase corpus

- 100,000+ paraphrase pairs
- All manually selected and classified
- Data:
  - Movie subtitles
  - Aligned news from different sources
  - Student exam answers to the same question
  - Different translations of the same texts by students

https://turkunlp.org/paraphrase.html

TURKUNLP
.ORG

Paraphrase

Paraphrase pairs with minimal lexical overlap

‘*Pohdi haastattelujen positiivisia ja negatiivisia puolia.*’:

...Haastatteluja voi olla vaikea vertailla keskenään tai laittaa järjestykseen, jos ne ovat kovin erilaisia toisistaan (esimerkiksi strukturoimattomissa haastatteluissa). **Ryhmähaastattelussa vaarana on, että osa puhuu liikaa ja osa ei saa ollenkaan suunvuoroa.** Haastateltavalla pitäisi aina olla luottavainen ja turvallinen olo, jotta haastattelusta saadaan kaikki irti…

...Jos on kyseessä ryhmähaastattelu, niin osallistujat saattavat puhua toistensa päälle, jolloin on hankala saada selvää. Henkilön elekielen oikea tulkinta voi olla haastavaa. **Ryhmähaastattelussa ujoimmat saattavat jäädä ilman ääntä ja vahvimmat henkilöt jyrätä omalla mielipiteellään.**

...Ääripäät jäävät taustalle ja lopputulos voi olla tylsä konsensus. **Ei päästä esiin hiljaisempia ja ujompia yksilöitä vaan äänekkäimmät nousevat esille.** Keskustelu voi karata liikaa jos moderaattori on on täysin ulkona keskustelun ohjaamisessa, vaikka toki roolin tulee olla lähtökohtaisesti suhteellisen näkymätön…

TURKUNLP
.ORG

Try this out!

http://epsilon-it.utu.fi/sbert400m

https://app-kaiku.rahtiapp.fi/semantic

Multilingual modeling of Web registers

UNIVERSITY OF TURKU

WIKIPEDIA

BBC NEWS

SUOMI 24

Google books

NGRAM VIEWER

Τι κάνετε;

WHAT DO THESE DOCUMENTS REPRESENT?

Article Talk

Borgio Verezzi

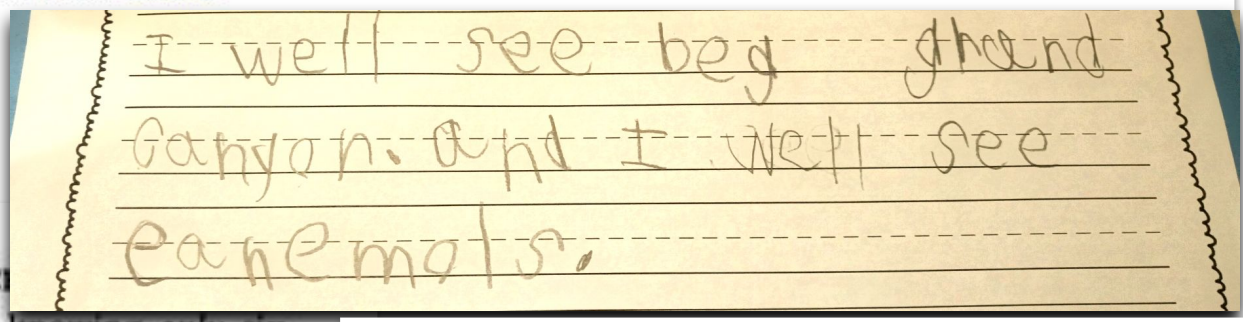
From Wikipedia, the free encyclopedia

Borgio Verezzi (Ligurian: *Bòrzi Veresso*) is a *comune* (municipality) in the Province of Savona in the Italian region Liguria, located about [illegible] southwest of Genoa and about 20 kilometres (12 mi) southwest of Savona.

Contents [hide]

1 Geography
2 Main sights
3 References

WIKIPEDIA
The Free Encyclopedia



"I am no longer surprised at your knowing *only* six accomplished women. I rather wonder now at your knowing *any*."

"Are you so severe upon your own sex as to doubt the possibility of all this?"

"*I* never saw such a woman. *I* never saw such capacity, and taste, and application, and elegance, as you describe, united."

Mrs. Hurst and Miss Bingley both cried out against the injustice of her implied doubt, and were both protesting that they knew many women who answered this description, when Mr. Hurst called them to order, with

Ingredients

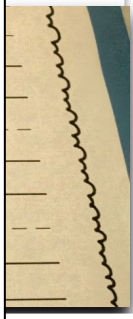
- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 cup butter, softened
- 1 teaspoon vanilla
- 1 egg
- 2 1/4 cups Gold Medal™ all-purpose flour

Article Talk

WHAT DO THESE DOCUMENTS REPRESENT?

Project goals

1. Understand language use on the web
2. Develop automatic register identification in order to be able to add document metadata to web-crawled datasets (not just in English, but also in other languages)



Are you so severe upon your own sex as to doubt the possibility of all this?"

"I never saw such a woman. I never saw such capacity, and taste, and application, and elegance, as you describe, united."

Mrs. Hurst and Miss Bingley both cried out against the injustice of her implied doubt, and were both protesting that they knew many women who answered this description, when Mr. Hurst called them to order, with

3/4	cup granulated sugar
3/4	cup packed brown sugar
1	cup butter, softened
1	teaspoon vanilla
1	egg
2 1/4	cups Gold Medal™ all-purpose flour

Why is the modeling of web registers difficult?

- In restricted corpora
  - A pre-determined set of registers
  - Documents selected to represent specific registers
- On the unrestricted web
  - The set of registers unknown
  - No gatekeepers to control language use
  - → Hybrid documents display characteristics of several registers, such as news+opinion
  - → Study of registers much more difficult than it would be for any restricted corpora
  - → Lack of representative corpora featuring the unrestricted web

The Web Library of Babel: evaluating genre collections

Serge Sharoff,[†] Zhili Wu,[†] Katja Markert[‡]

4. Conclusions

The results are relatively negative. The collections are not comparable to each other: even when categories in a collection are described in a very similar way, e.g., `FAQs` in SANTINIS and `help` in KI-04, their actual content is considerably different. When the similarity between genre collections is tested using cross-classification, the accuracy is also quite low. This shows the limits of the existing web-genre collections: if each of them is so different from any other, neither of them can be treated as a good representative for the entire web. The experiments also show that humans disagree on genre annotation of randomly selected webpages, throwing doubt on their reliability as well as on their representativeness.

The jury is still out on the best set of features useful for AGI. Character n-grams can capture many relevant generalisations not possible for other feature types, such as genre-specific prefixes and suffixes (unlike word forms), subcategories within general POS classes (unlike POS tags), but their efficiency is often related to the ability to identify *topics* exemplifying particular genres in available collections. This is the reason why the accuracy often drops when we go beyond the training set. In addition, as the datasets used

Corpus of online registers of English (CORE)
(Egbert, Biber and Davies 2015)

- Unrestricted sample of the searchable web
- ~50,000 documents, > 50 million words
- Manually annotated for registers
- Covers **the full range of registers and documents on the open searchable English Web!**
- **Register scheme developed in a data-driven manner**

TURKUNLP
.ORG

Narrative NA
news report / news blog, sports report, personal blog, travel blog, historical article, short story

Opinion OP
review, opinion blog, advice

Informational Description IN
description of a thing, Description of a person, research article, information blog

Interactive Discussion ID

How-to HI
how-to instructions, recipe

Informational Persuasion IP
description with intent to sell, editorial

Lyrical LY
poem, song lyrics

Spoken SP
formal speech, interview

Machine Translated MT

Multilingual register corpora

TURKUNLP .ORG

- Finnish ~10,000 documents
- Swedish ~4,000 documents
- French ~4,000 documents
- Ongoing: Turkish ~3,000 documents

Evaluation datasets

Arabic	92
Catalan	111
Spanish	100
Hindi	161
Indonesian	1190
Portuguese	334
Urdu	160
Chinese	317
Japanese	100
Farsi	106
Norwegian	150

Web register annotation guidelines

The annotation task consists of two steps: deciding whether to accept or reject a document, and giving a register label / labels to the accepted documents.

When to accept or reject a document

When to give a document several labels

Short list of register labels and their abbreviations

Video instructions to the annotation on Prodigy

Please note that

- You can have a look at how the document website looks like by following the document url on the annotator
- The annotation decision should, however, base on the text on the annotator

prodigy

PROJECT INFO

DATASET	tr-batch-02
RECIPE	registers
VIEW ID	choice

PROGRESS

THIS SESSION	8
TOTAL	3,648

97%

ACCEPT	7
REJECT	1
IGNORE	0

HISTORY

- İlim ve Geleceğin İnşasın... ✓
- Beylikdüzü Viessmann S... ✓
- Mobil Sitelerde En Çok Y... ✓
- Develi'de ilçe protokolü t... ✓
- EPSON T0348 Muadil M... ✓

MUTLU EVLILIĞIN SIRRI:H2O

https://zehirliok.com/klm/mutlu-evliligin-sirri-h2o.html

"Evlendikten sonra aşk öldü diyenler, aşk ateşini beslemiyorlar" diyen Prof. Tarhan'a göre aşkı ölümsüz kılmanın formülü: H2O. Peki bu formülü nasıl hayata geçirmek gerekiyor?

'Kadın Psikolojisi' adını taşıyan kılavuz kitabında evlilik ilişkisini H2O (suyun kimyasal formülü) simgesiyle tanımlayan Prof. Dr. Nevzat Tarhan, 'Hidrojen ve oksijen, atmosferde ayrı ayrı dolaşıyor, birleşince suyu oluşturuyor. Eğer evliliğinizde sevdiğinizle uyum içindeyseniz siz de H2O formülünü uygulamışsınızdır' diyor.

Kadın erkek ilişkisini H2O (su) simgesiyle tanımlayan Prof. Tarhan, 'Hidrojen ve oksijen, atmosferde ayrı ayrı dolaşıyor, birleşince suyu oluşturuyor. Eğer ilişkinizde sevdiğinizle uyum içindeyseniz ve 'biz' olmanın güzelliğini yaşıyorsanız H2O formülünü uygulamışsınız demektir' diye konuştu. İşte Prof. Tarhan'ın kadın-erkek ilişkisi üzerine söyledikleri:

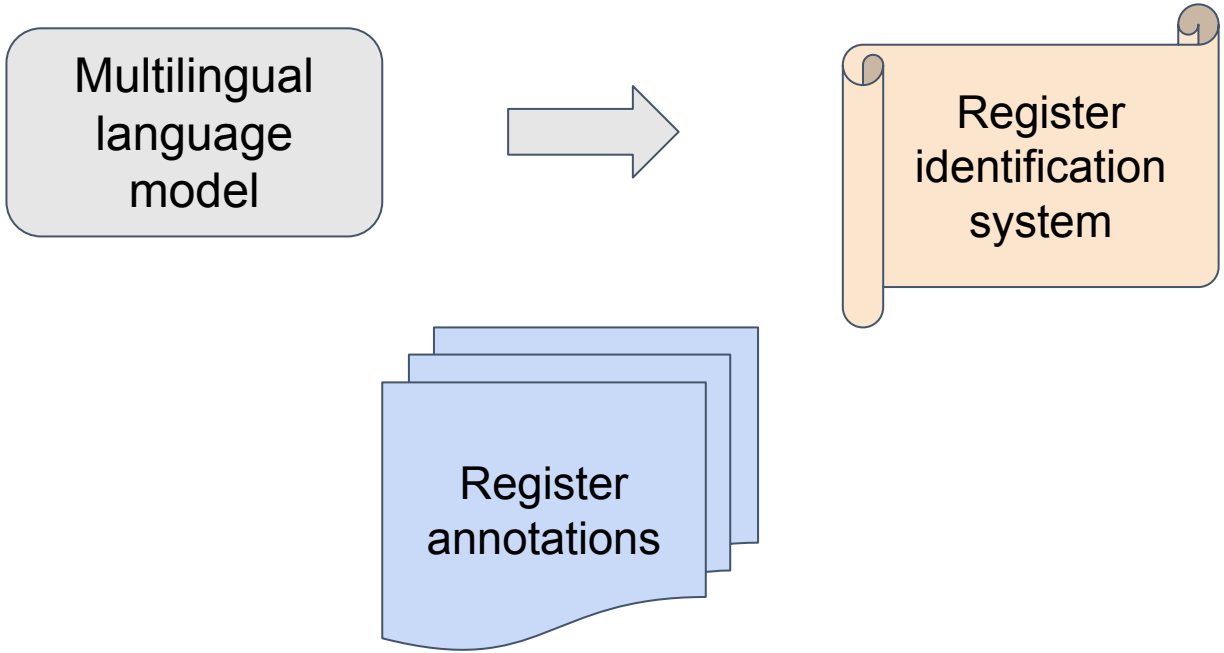
* Taraflar arasında, güven, saygı ve sevgi sarsılmışsa mutlu olmak mümkün değildir. Çok zıt kişilikli insanlar, iyi iletişim kurarak birbirini anlıyor. Kişilik yapıları benzeyenler ise kötü iletişim kurdukları için anlaşamıyor.
* İnsan 100 kapılı bir saraya benziyor. İyi iletişimin temelinde de, hep kapalı kapıları zorlamak yerine, açık kapıları bulup iletişim kurmak yatıyor.
* Aşk bir ateş gibidir, bakılırsa büyüyor. Bakılmazsa sönüp gidiyor. Evlendikten sonra aşk öldü diyenler, aşk ateşini beslemiyorlar.
* Erkek ve kadın birbirlerini eleştirmekten ilişkiyi yürütmeye zaman kalmıyor.
* Uzlaşmada 'altın orta nokta kuralı' var. Tartışma çıktığında erkek bir adım, kadın bir adım atıyor, orta bir noktada buluşup anlaşmaya çalışıyorlar.

Yorumlar

Yeni yorum gönder

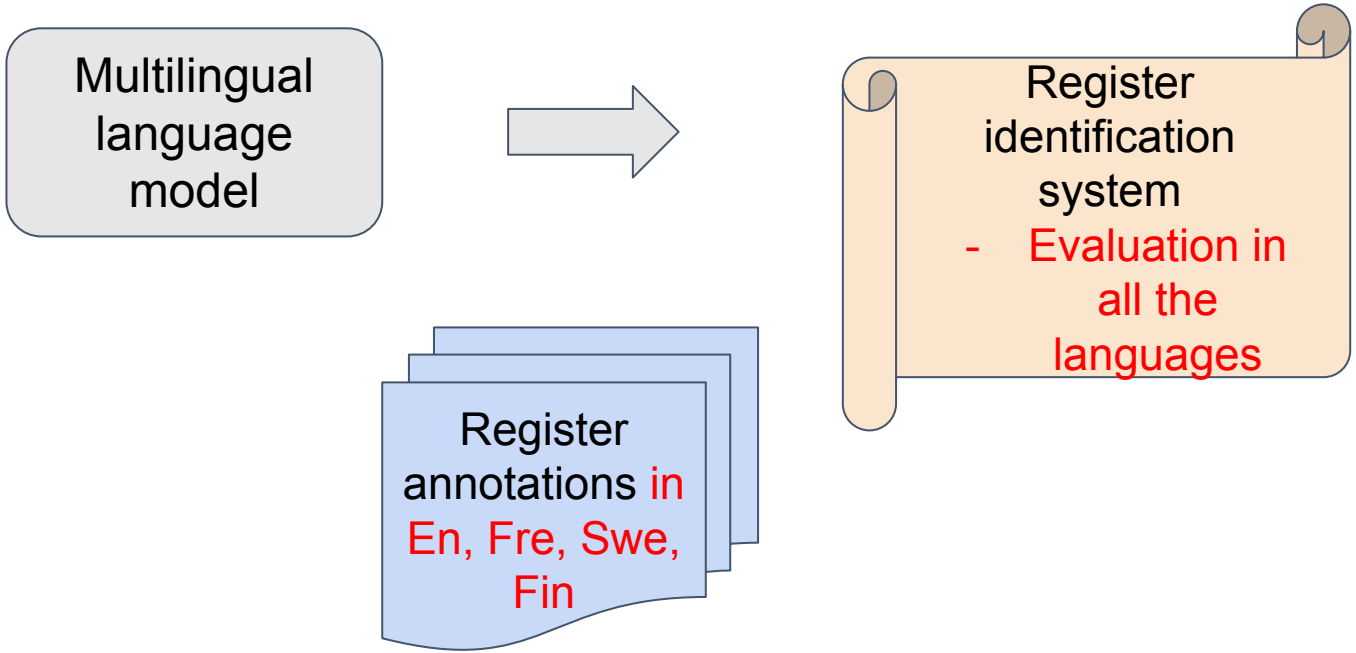
Training multilingual language models

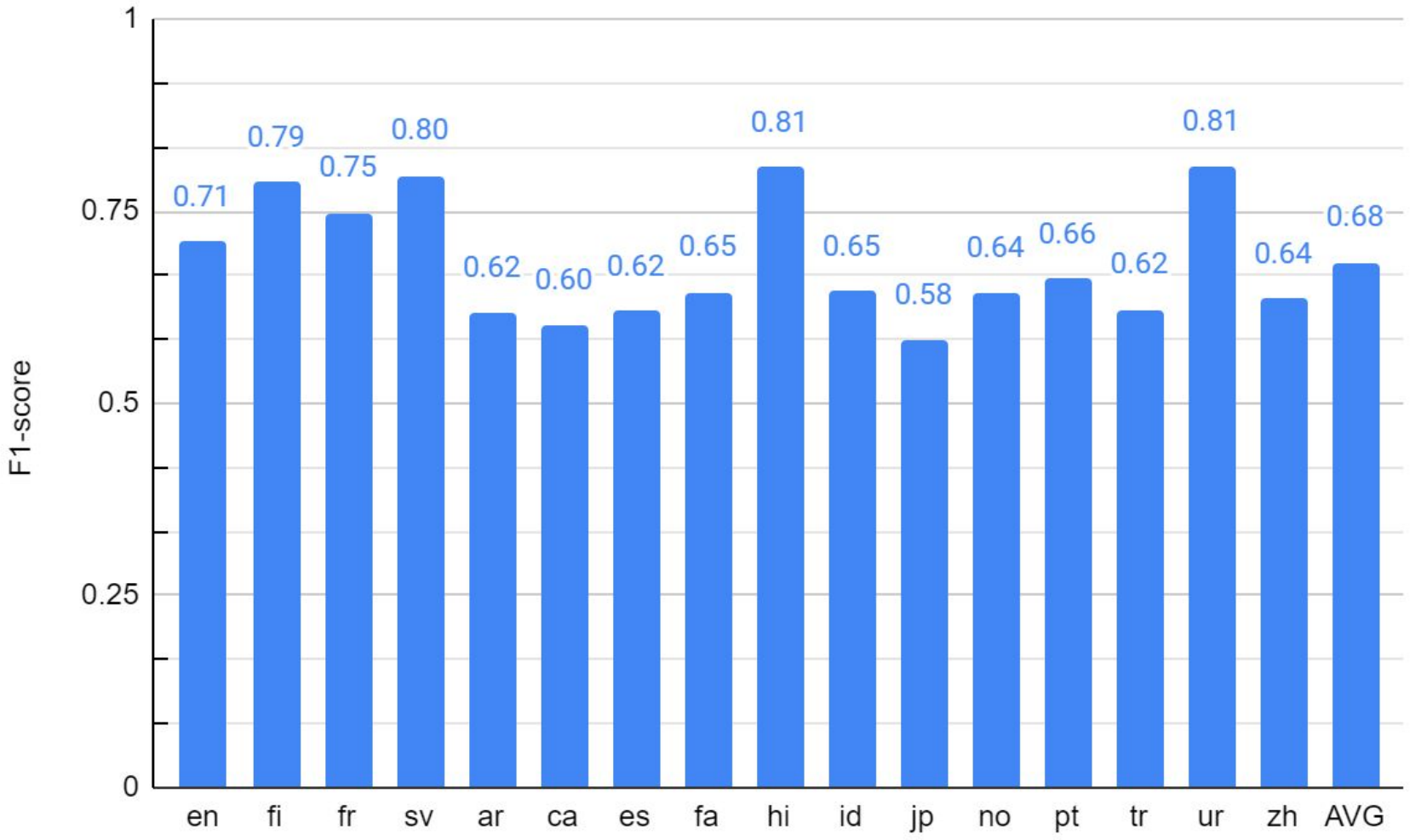
One model to all tasks!



To what extent does multilingual register identification work?

One model to all tasks!




F1-score
1
0.75
0.5
0.25
0
0.71
0.79
0.75
0.80
0.62
0.60
0.62
0.65
0.81
0.65
0.58
0.64
0.66
0.62
0.81
0.64
0.68
en
fi
fr
sv
ar
ca
es
fa
hi
id
jp
no
pt
tr
ur
zh
AVG

Conclusions

- Register identification achieves decent results even in cross-lingual settings
- Zero-shot performance decreases but remains useful

→ The model has clearly learnt something cross-lingual about registers
→ What has the model learnt?

Exploring cross-linguistic representations of Web registers with a deep multilingual model

TURKUNLP
.ORG

UNIVERSITY
OF TURKU

Data

Register Oscar (Laippala et al. 2022)

- A sub-corpus of Oscar (Ortiz Suárez et al., 2020), a massive Web-crawled corpus
- 14 languages, 351M documents

Sample used in the study

- 72,000 documents
- Three languages: English, French, Finnish

Model

Model

- Fine-tuned XLM-R *base* to register classification using register annotations in Finnish, French and English
- Model performance: F1-score of 0.80

Narrative, Informational Description, Informational Persuasion, Opinion, How-to, Interactive Discussion

Document representations in the model

- Embeddings, i.e., vectors in a multilingual high-dimensional space (768 dims)
- Final layer of the fine-tuned XLM-R model

Questions

Option 1:

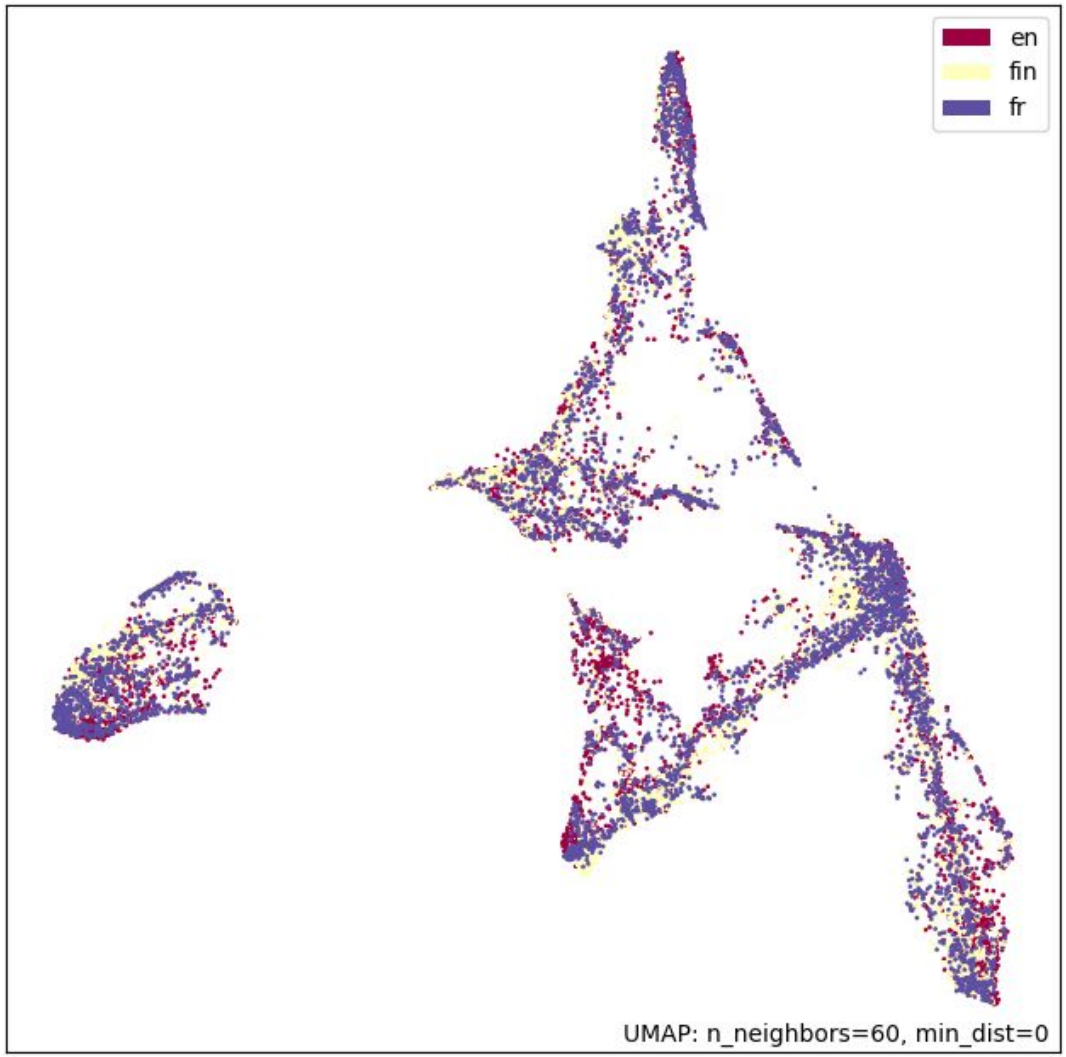
- Learned representations **do not display structuring** with respect to language/register/a combination of languages and registers

Option 2:

- Learned representations display **language-specific and register-specific** structuring

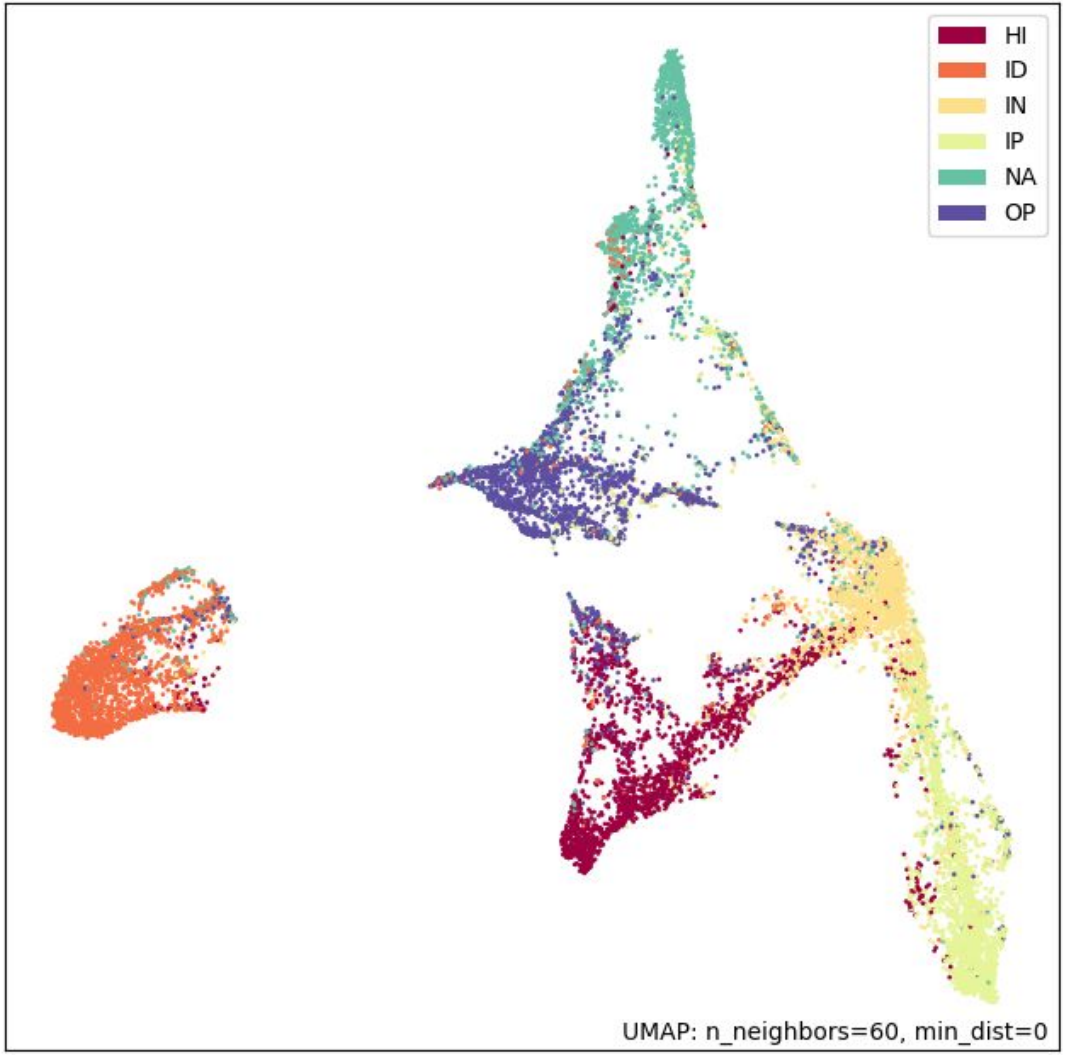
Option 3:

- Learned representations display **register-specific** structuring across languages



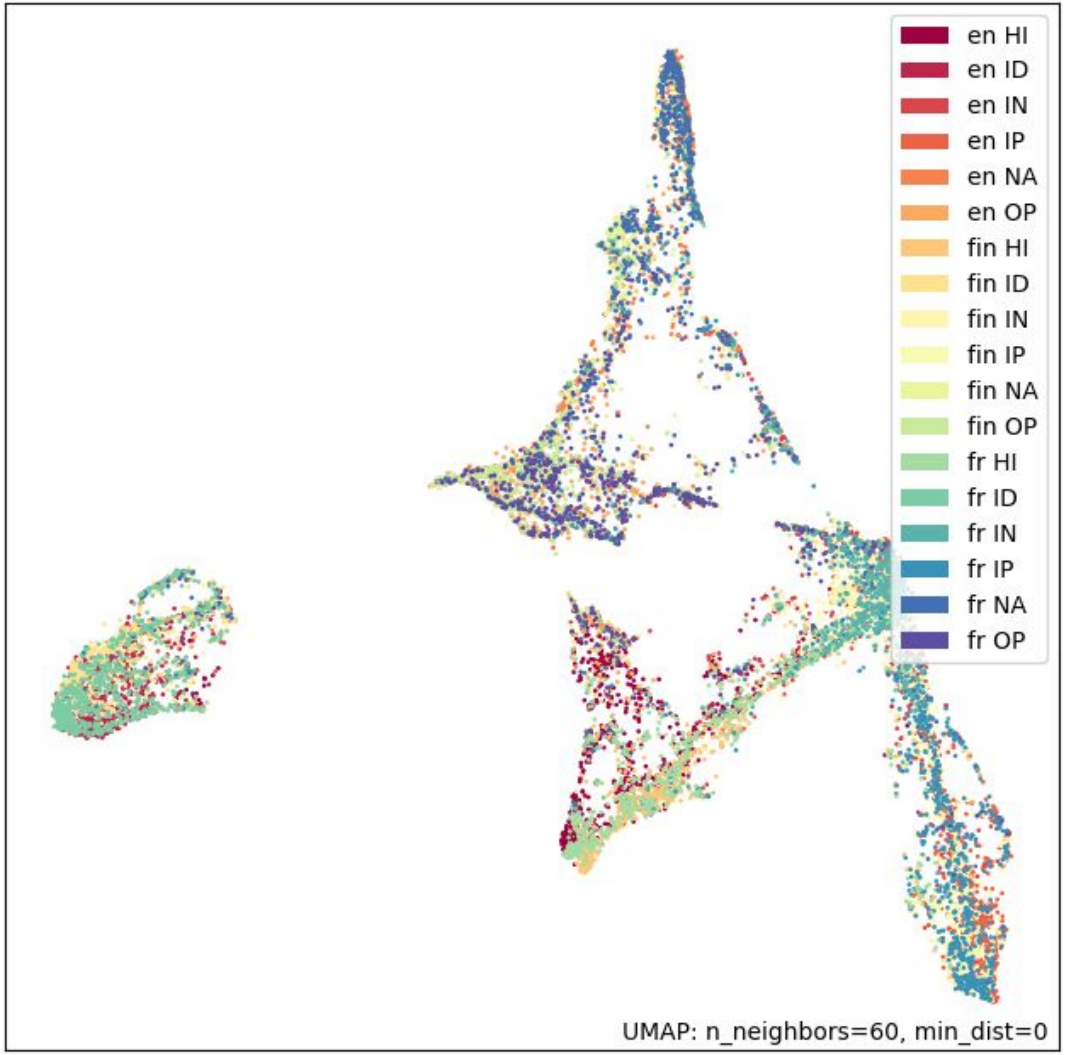
TURKUNLP .ORG

UNIVERSITY OF TURKU



TURKUNLP .ORG

UNIVERSITY OF TURKU

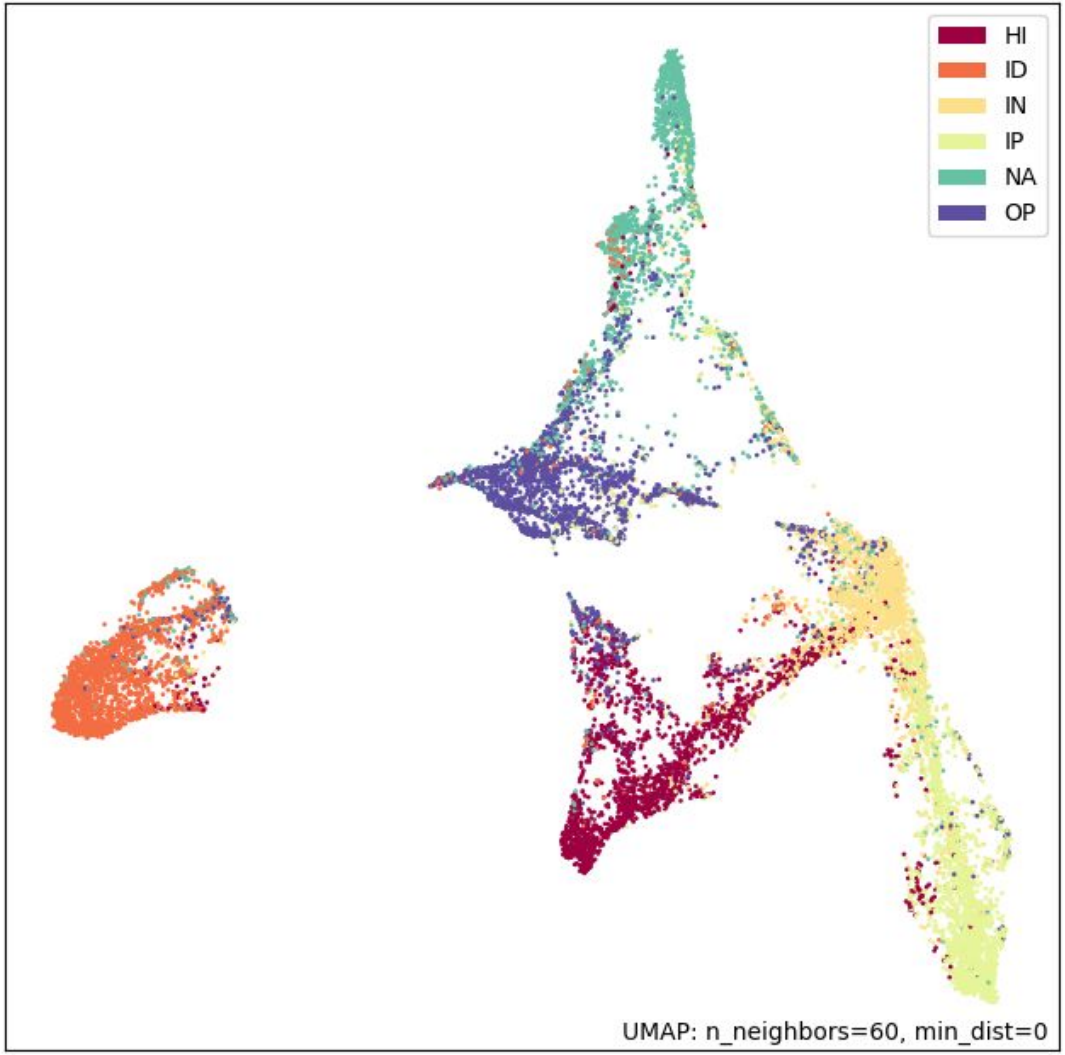


Evaluation

Cluster document embeddings using Kmeans

- 3 clusters corresponding to 3 languages
- 6 clusters corresponding to 6 registers
- 18 clusters corresponding to a combination of languages and registers

Comparison of the cluster solution to the ground truth with Adjusted Rand Index



3 clusters = ARand 0.0

6 clusters = ARand 0.58

18 clusters = ARand 0.23

TURKUNLP .ORG

UNIVERSITY OF TURKU

How to know the basis of the grouping?

UNIVERSITY OF TURKU

SACX — towards keywords explaining multilingual language models

Task: want to understand the linguistic basis of the model

- Preferably on the class-level

Challenge:

- NNs difficult to interpret
- Most methods operate on the level of individual documents

Solution:

- **SACX — towards keywords explaining multilingual language models**

SACX — towards keywords explaining multilingual language models

Keywords as a linguistic tool for characterising corpora

- Typically frequency-based, but can also be predicted (see Kyröläinen & Laippala 2022)

Integrated Gradients (IG) (Sundararajan et al., 2017)

- A general framework for estimating feature importance in deep neural networks at the document level

Our suggestion:

- Class-wise, stable keywords through aggregation of words identified by IG across documents and models

Verbs of communication

Narrative				
English	Frech		Finnish	
announced	communiquées	‘communicated’	kommentoi	‘commented’
reported	annoncée	‘announced’	sanoo	‘says’
commented	explique	‘explains’	uutisoitiin	‘was reported in news’
replied	souligne	‘underline’	kertoi	‘told’
blogged	révèle	‘reveal’	julkisti	‘published’

Directions for making something

How to/Instructional (HI)				
English	French		Finnish	
recipe	recette	‘recipe’	sekoita	‘mix’
tutorial	tuto	‘tutorial’	käyttöohje	‘instructions for use’ (sng)
tutorials	tutoriel	‘tutorial’	reseptillä	‘with a recipe’
recipes	recettes	‘recettes’	viimeistään	‘at the latest’
tips	devez	‘devez’	käyttöohjeet	‘instructions’ (pl)
guide	conseils	‘conseils	asennusohje	‘installation instructions’
threaded	recommandé	‘recommended’	turvallisuusohjeet	‘security instructions’ (pl)
remove	préparer	‘prepare’	turvaohjeet	‘security instructions’ (pl)
steps	guide	‘guide’	ohjeet	‘instructions’ (pl)
accordance	faudra	‘have to’ (fut)	reseptiä	‘of a recipe’

References to writings

Opinion				
English	French		Finnish	
article	article	‘article’	kirjoituksessani	‘in-my-article’
blog	blog	‘blog’	blogissani	‘in-my-blog’
recommend	recommanderais	‘I-would-recommend’	suosittelen	‘I-recommend’
criticized	regrettons	‘we-regret’	ajattelin	‘I-thought’
impressed	satisfaite	‘pleased’	tyytyväinen	‘pleased’
pleasant	géniale	‘great’	loistavalla	‘with-a-great’
disappointed	inutile	‘useless’	absurdi	‘absurd’
complaint	apprécié	‘appreciated’	ironista	‘ironic’
greatful	suffisant	‘sufficient’	hienot	‘nice’

(Stance) verbs

Opinion				
English	French		Finnish	
article	article	‘article’	kirjoituksessani	‘in-my-article’
blog	blog	‘blog’	blogissani	‘in-my-blog’
recommend	recommanderais	‘I-would-recommend’	suosittelen	‘I-recommend’
criticized	regrettons	‘we-regret’	ajattelin	‘I-thought’
impressed	satisfaite	‘pleased’	tyytyväinen	‘pleased’
pleasant	géniale	‘great’	loistavalla	‘with-a-great’
disappointed	inutile	‘useless’	absurdi	‘absurd’
complaint	apprécié	‘appreciated’	ironista	‘ironic’
greatful	suffisant	‘sufficient’	hienot	‘nice’

Evaluation

Opinion				
English	French		Finnish	
article	article	‘article’	kirjoituksessani	‘in-my-article’
blog	blog	‘blog’	blogissani	‘in-my-blog’
recommend	recommanderais	‘I-would-recommend’	suosittelen	‘I-recommend’
criticized	regrettons	‘we-regret’	ajattelin	‘I-thought’
impressed	satisfaite	‘pleased’	tyytyväinen	‘pleased’
pleasant	géniale	‘great’	loistavalla	‘with-a-great’
disappointed	inutile	‘useless’	absurdi	‘absurd’
complaint	apprécié	‘appreciated’	ironista	‘ironic’
greatful	suffisant	‘sufficient’	hienot	‘nice-pl’

Conclusions

Register identification models seem to learn linguistically meaningful representations

More questions

- Language-specific similarities / differences?
- Cultural similarities / differences?

Exploring the stability of political rhetoric in Finnish parliamentary debates using deep learning

Otto Tarkka, Kimmo Elo, Filip Ginter, Veronika Laippala

DHNB 2023 - Political Discourse session, Thursday March 9

UNIVERSITY OF TURKU
Centre for Parliamentary Studies

TURKUNLP .ORG

UNIVERSITY OF TURKU

Motivation and background

- Political rhetoric reflects underlying ideological bias (Finlayson 2013)
- Typically, we encounter political text together with other context cues: familiar logos, colours and political figures, etc.
  - predictive approach
- RQs
  - To what extent can party affiliation be predicted from parliamentary speeches using modern deep learning methods?
  - What linguistic cues does the machine learning model depend on?

UNIVERSITY OF TURKU

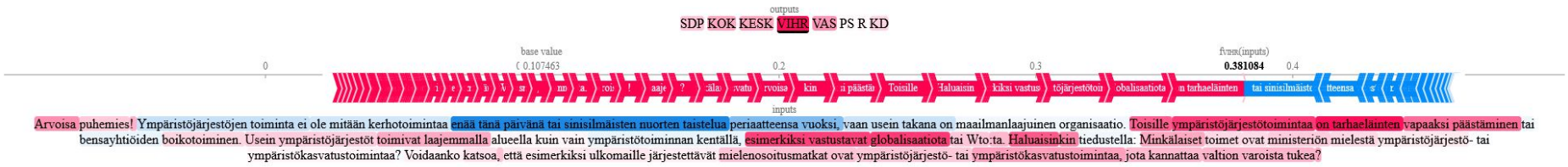
Data

- The FinParl dataset contains all plenary speeches held in the Finnish *eduskunta* (Sinikallio et al. 2012)
- We focus on speeches from 2000 to 2021: over 250,000 speeches, more than 50 million words
  - 80/20 train/test split
- 8 largest parties included
- Light pre-processing

Parliament House in Finland
Photo by Tiina Tuukkanen

BERT model and SHAP

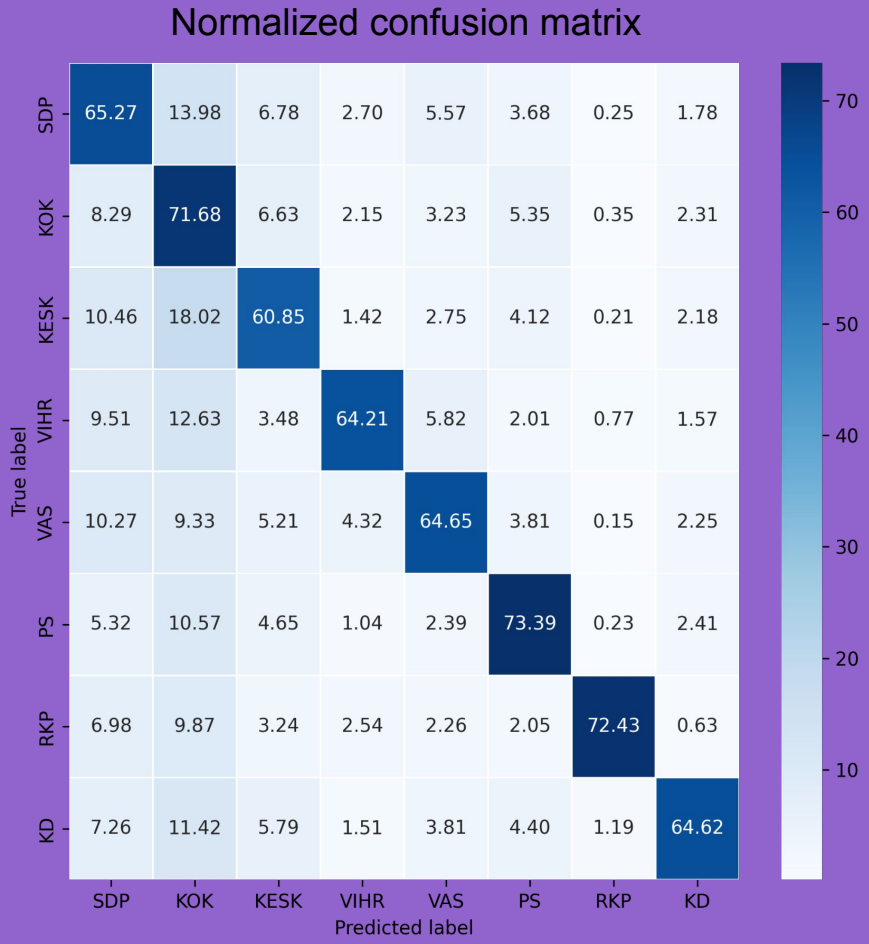
- We train a classifier by fine-tuning the FinBERT deep-learning model (Virtanen et al. 2019)
  - Masked language model: ‘I prefer [MASK] over dogs.’
- Model explainability: SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017)
  - How does removing a word affect the prediction?



UNIVERSITY OF TURKU

Results

UNIVERSITY OF TURKU



Model performance

- Average macro-F1 from 10-fold random subsampling is 0.674
- Most accurately predicted were the Swedish People's Party (RKP) and the Finns Party (PS)

	SDP	KOK	KESK	VIHR	VAS	PS	RKP	KD
F1	0.656	0.638	0.670	0.674	0.656	0.715	0.786	0.634

UNIVERSITY OF TURKU

SHAP keyword examples

- Words with highest mean SHAP values
  - Names of parties and MPs removed
- **Christian Democrats (KD)**: *I think, mister, spiritual, Christian, my spouse, David, amphetamine, Gospel*
- **National Coalition (KOK)**: *blue-green, I do, career, agreement, I held, accused, scold, we too, winning*
- **The Finns (PS)**: *I do, I support, blue, there, socialist, cult, speak ill of, police, website, immigrant*
- **Left Alliance (VAS)**: *madam, on the left, condemn, still, party, workers, supports, Indonesia, customer service, Gaza*

Conclusions

- Even with deep-learning methods, identifying party affiliation from plenary speeches is a difficult machine-learning task and even more difficult for humans
- Model recognises names but also topical and rhetorical keywords
- Party rhetoric in Finland is not strongly polarized but there is a divide into left-wing and right-wing rhetoric
- Model recognises fine statistical differences in word distribution that are missed by humans

TURKUNLP .ORG

Join SIGWAC!

ACL Special Interest Group in Web-as-Corpus

Officers

- Nikola Ljubešić (co-president)
- Benoît Sagot (co-president)
- Veronika Laippala (co-secretary)
- Pedro Ortiz Suarez (co-secretary)

SIGWAC

https://www.sigwac.org.uk/

UNIVERSITY OF TURKU