Automatic syntactic analysis

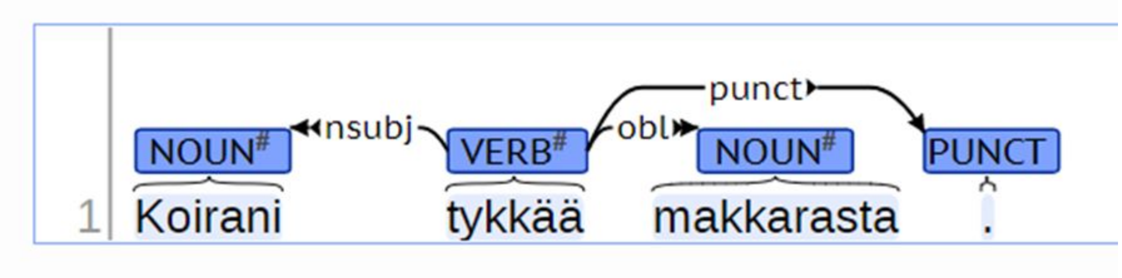
Veronika Laippala
veronika.laippala@utu.fi

TURKUNLP
.ORG
UNIVERSITY OF TURKU

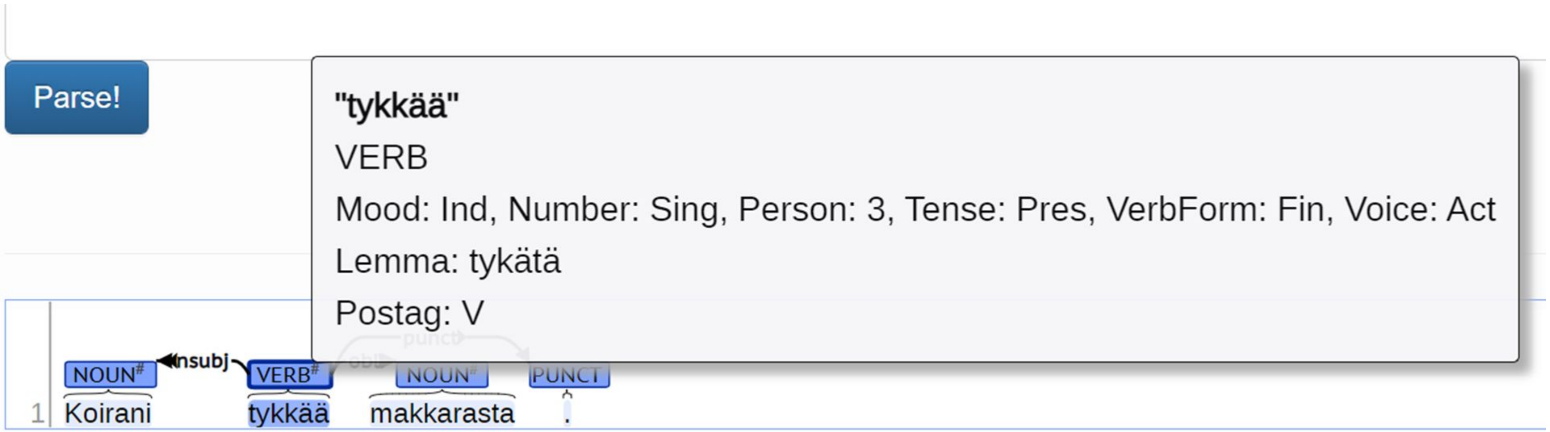
Topic of the day

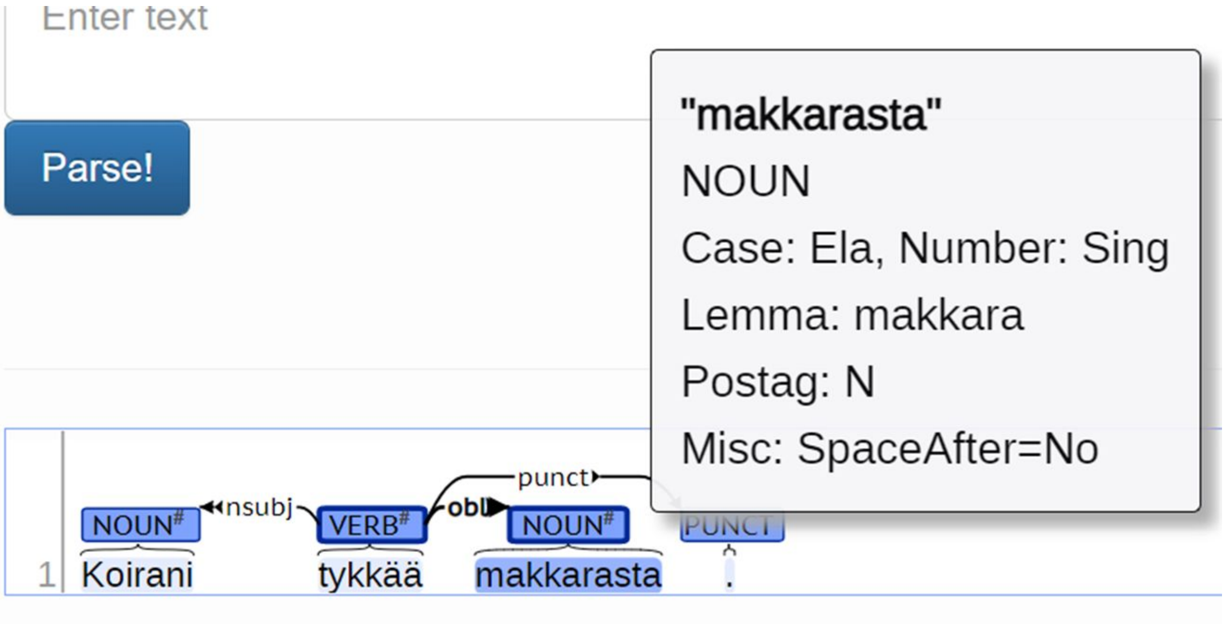
- Automatic syntactic analysis
- Theory
  - What it is?
  - What can it be used for?
  - How does it work?
- Practice
  - Running syntax analysers
  - Analysing output
  - Done partially on Google colab - you need Google credentials for that

Automatic syntactic analysis

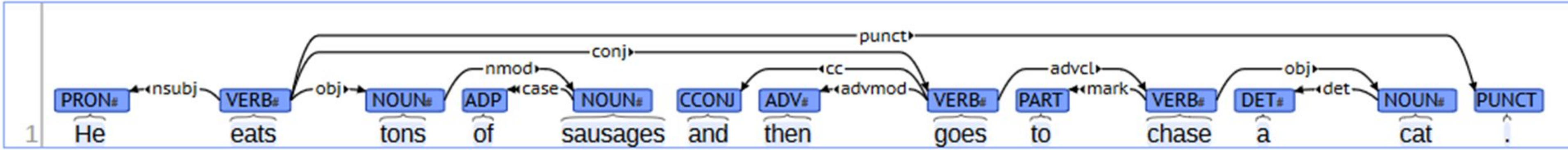


- Typed dependency relations between words


Parse!
"tykkää"
VERB
Mood: Ind, Number: Sing, Person: 3, Tense: Pres, VerbForm: Fin, Voice: Act
Lemma: tykätä
Postag: V
NOUN
nsubj
VERB
NOUN
PUNCT
1
Koirani
tykkää
makkarasta
.


Enter text
Parse!
"makkarasta"
NOUN
Case: Ela, Number: Sing
Lemma: makkara
Postag: N
Misc: SpaceAfter=No
nsubj
obl
punct
NOUN
VERB
NOUN
PUNCT
1
Koirani
tykkää
makkarasta
.

- 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



What could this be used for?

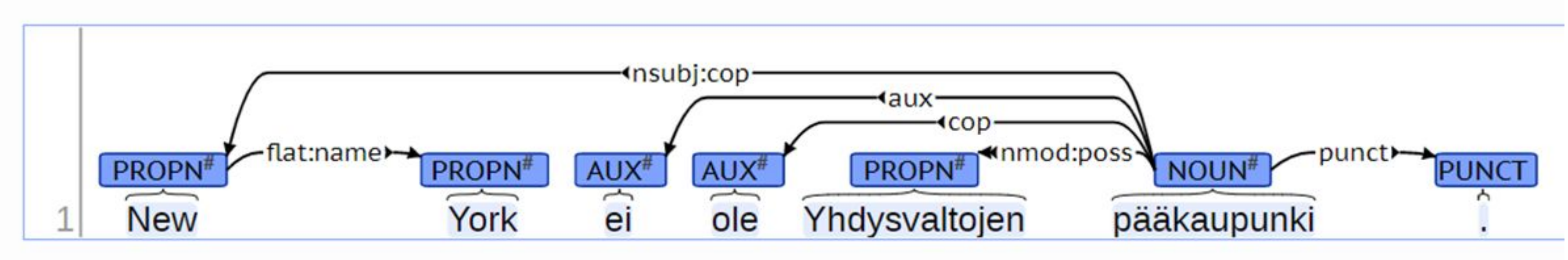
- Information extraction
  - Lemmas
  - Filtering (e.g., away auxiliaries, conjunctions, etc.)
  - Focusing (e.g., keep only adjectives)
- Linguistics
  - Fetch any information you would want!
  - Earlier also features for machine learning, but less so now

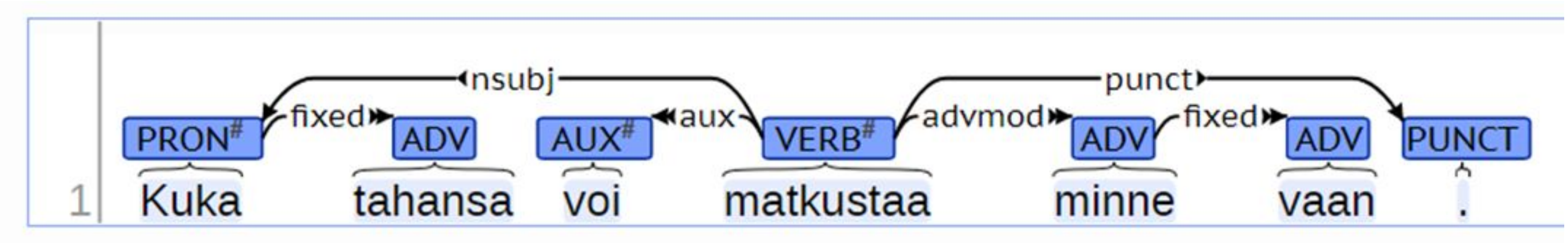
Syntax analysis preceded by a number of tasks

- Tokenization: Identification of words
- Sentence splitting: Identification of sentences

Tokenization

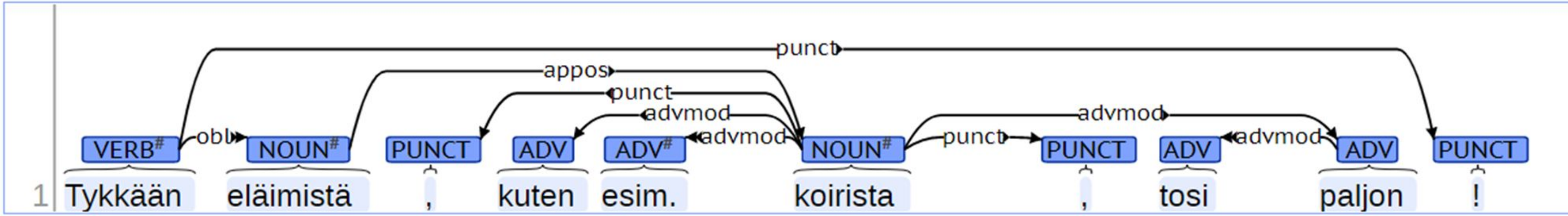
- Difficulty depends on the language



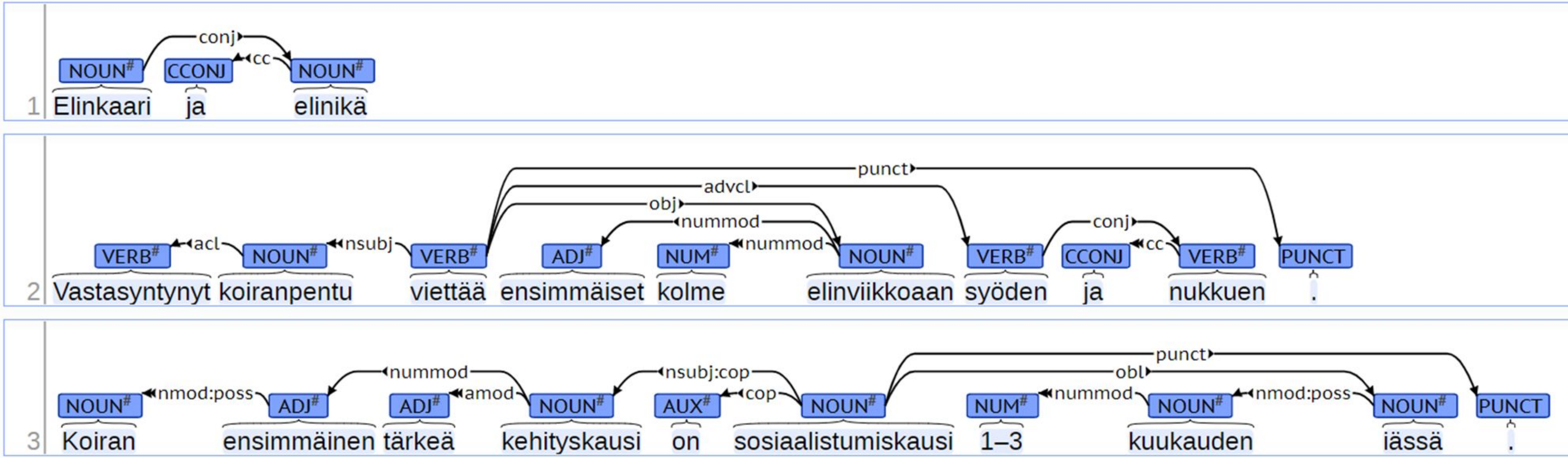


Sentence splitting

- Need to separate abbreviations from other punctuation

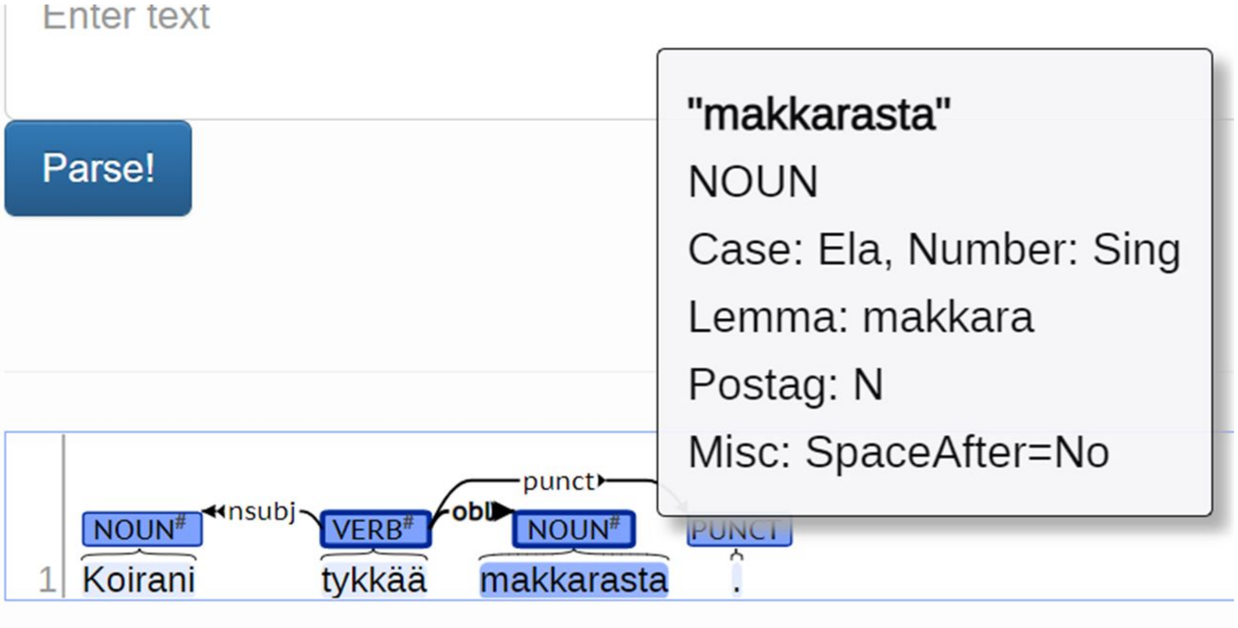


Titles etc. can also make the sentence splitting harder



After sentence splitting + tokenization

- Part-of-speech classes: Adjective, Noun, Auxiliary, etc.
- Dependency relations: subject, object, attributive adjective, etc.
- Morphology: plural, singular, cases, person, etc.


Enter text
Parse!
"makkarasta"
NOUN
Case: Ela, Number: Sing
Lemma: makkara
Postag: N
Misc: SpaceAfter=No
punct
nsubj
obl
NOUN
VERB
NOUN
PUNCT
1
Koirani
tykkää
makkarasta
.

How does automatic syntactic analysis work then?

Treebanks and parsers

Treebank = a collection of manually syntax annotated sentences

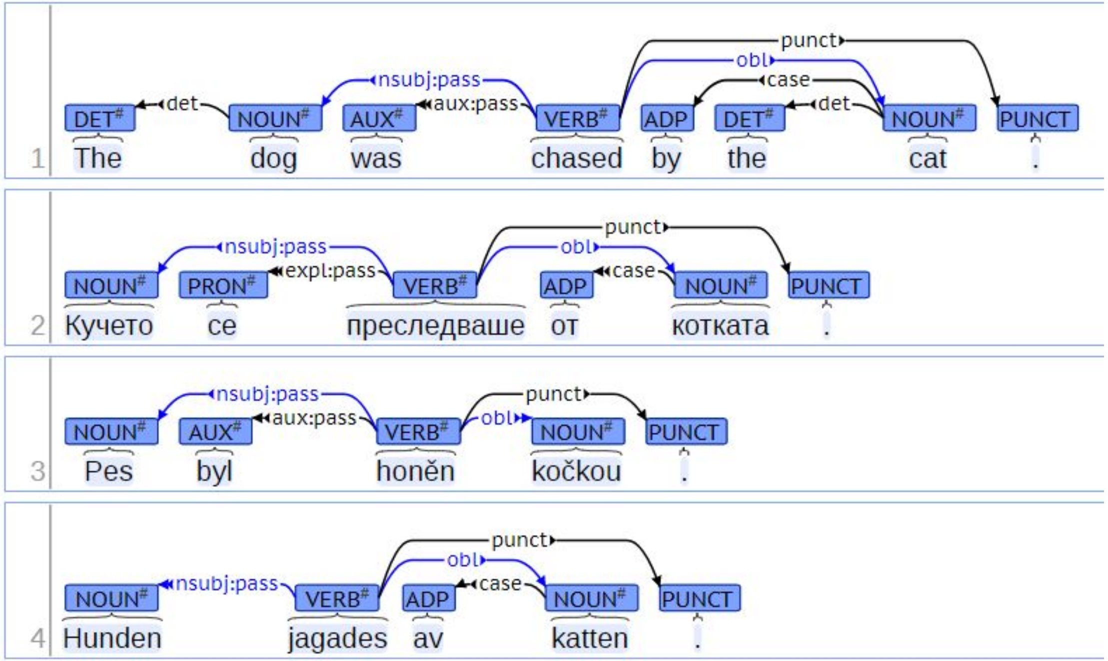
A syntax parser = a computer program that performs automatic syntactic analysis

Treebanks and parsers

- Modern syntactic analysis dependency-based
- Modern parsers based on machine learning
- Need for treebanks
- For Finnish: Turku Dependency Treebank
    - https://github.com/UniversalDependencies/UD_Finnish-TDT/tree/master

Universal dependencies

- **Cross-linguistically valid**



Current UD Languages

Information about language families (and

- Afrikaans
- Akkadian
- Akuntsu
- Albanian
- Amharic
- Ancient Greek
- Ancient Hebrew
- Apurina
- Arabic
- Armenian
- Assyrian
- Bambara
- Basque
- Beja
- Belarusian
- Bengali
- Bhojpuri
- Breton
- Bulgarian
- Buryat
- Cantonese
- Catalan
- Cebuano
- Chinese
- Chukchi
- Classical Chinese
- Coptic
- Croatian
- Czech
- Danish
- Dutch
- English
- Erzya
- Estonian
- Faroese
- Finnish
- French
- Frisian Dutch
- Galician
- German
- Gothic
- Greek
- Guajajara
- Guarani
- Hebrew
- Hindi
- Hindi English
- Hittite
- Hungarian
- Icelandic
- Indonesian
- Irish
- Italian
- Japanese
- Javanese
- Kaapor
- Kangri
- Karelian
- Karo
- Kazakh
- Khunsari
- Kiche
- Komi Permyak
- Komi Zyrian
- Korean
- Kurmanji
- Latin
- Ligurian
- Lithuanian
- Livvi
- Low Saxon
- Madi
- Makurap
- Maltese
- Manx
- Marathi
- Mbya Guarani
- Moksha
- Munduruku
- Naija
- Nayini
- Neapolitan
- North Sami
- Norwegian
- Old Church Slavonic
- Old East Slavic
- Old French
- Old Turkish
- Persian
- Polish
- Pomak
- Portuguese
- Romanian
- Russian
- Sanskrit
- Scottish Gaelic
- Serbian
- Skolt Sami
- Slovak
- Slovenian
- Soi
- South Le
- Spanish
- Soi
- South Levantine Arabic
- Spanish
- Swedish
- Swedish Sign Language
- Swiss German
- Tagalog
- Tamil
- Tatar
- Teko
- Telugu
- Thai
- Tupinamba
- Turkish
- Turkish German
- Ukrainian
- Umbrian
- Upper Sorbian
- Urdu
- Uyghur
- Vietnamese
- Warlpiri
- Welsh
- Western Armenian
- Wolof
- Xibe
- Yakut
- Yoruba

https://universaldependencies.org/

Design principles of the treebanks

- Needs to have a solid linguistic foundation
- Be transparent and accessible to non-specialists
- Support well downstream language understanding tasks

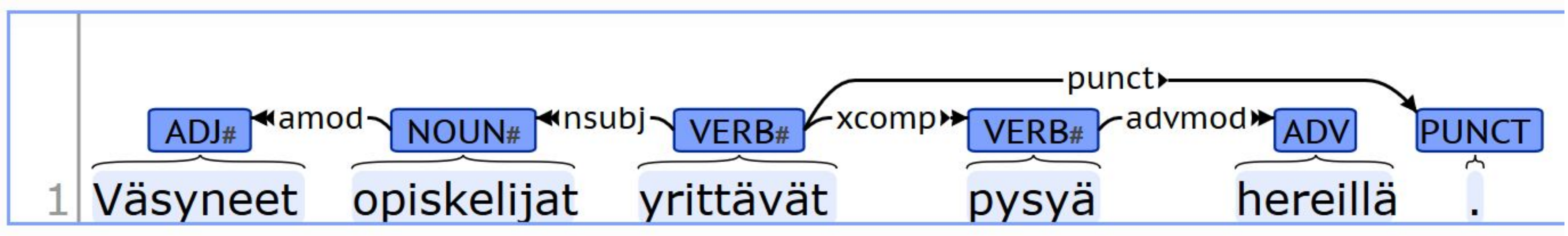
Disclaimer

- Despite of the common guidelines, adaptation can vary between different treebanks
  - Language-specific additions
  - Language-specific omissions
  - Treebank level annotation practices

Turku-neural-parser-pipeline

http://turkunlp.org/Turku-neural-parser-pipeline/

A neural parsing pipeline for segmentation, morphological tagging, dependency parsing and lemmatization with pre-trained models for more than 50 languages. Top ranker in the CoNLL-18 Shared Task.



"Väsyneet"

ADJ

Case: Nom, Degree: Pos, Number: Plur

Lemma: väsynyt

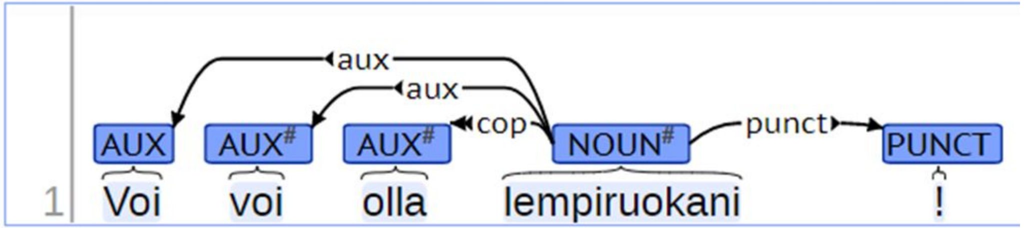
Postag: A

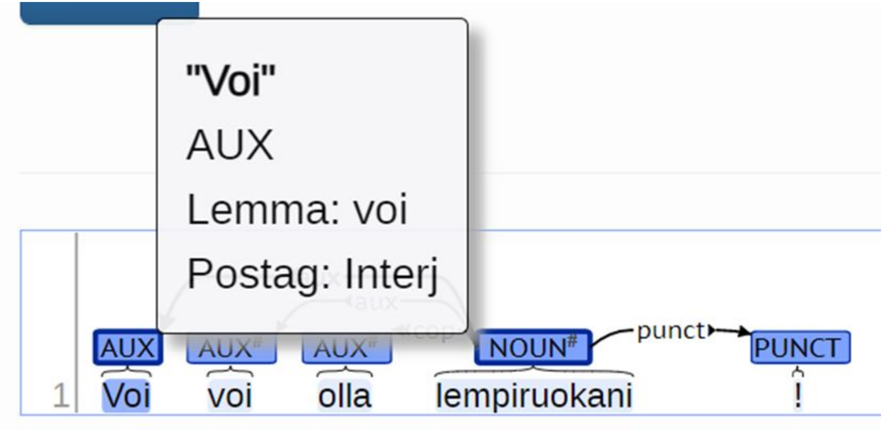
Parser performance on TDT

- Haverinen et al. 2014
  - LAS 81%, POS 94.4, LEMMA 91.8%
- Kanerva et al. 2018
  - LAS 86.6%, POS 96.66%, LEMMA 95.32%
- Latest:
  - LAS 91%, POS 97.8%, LEMMA 96.13%

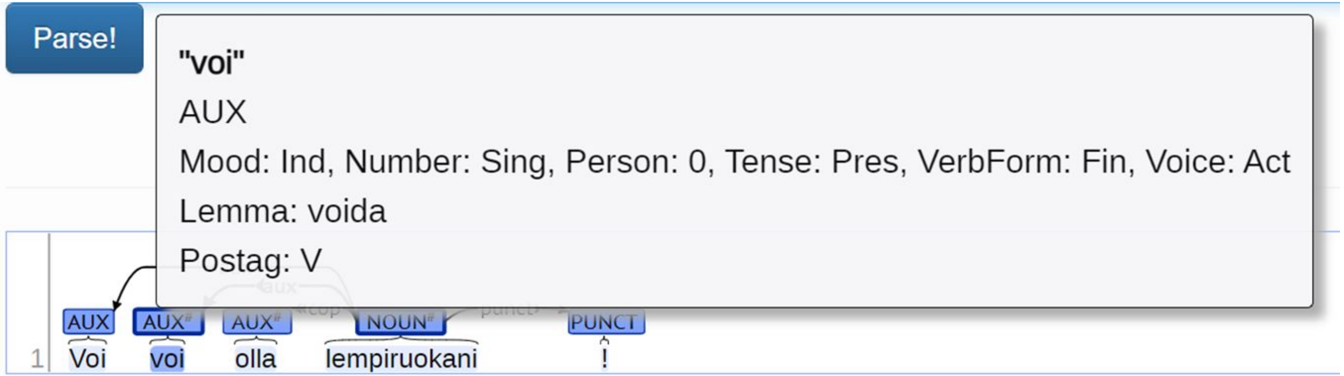
What then can cause trouble?

Ambiguity as a disturbing property of human language…

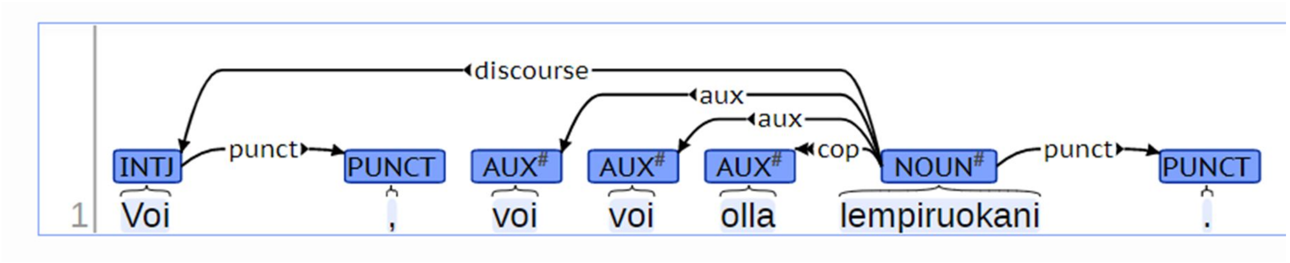

aux
aux
cop
punct
AUX
AUX#
AUX#
NOUN#
PUNCT
1
Voi
voi
olla
lempiruokani
!



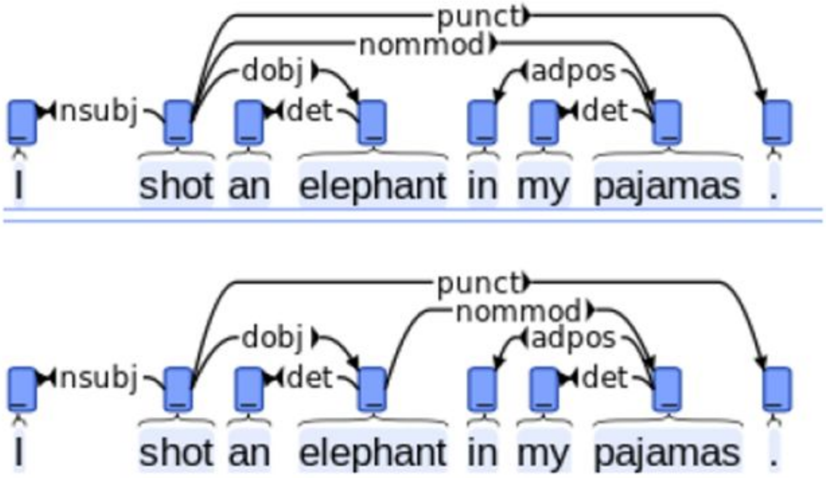
"Voi"
AUX
Lemma: voi
Postag: Interj
punct
AUX
AUX
AUX
NOUN
PUNCT
1
Voi
voi
olla
lempiruokani
!

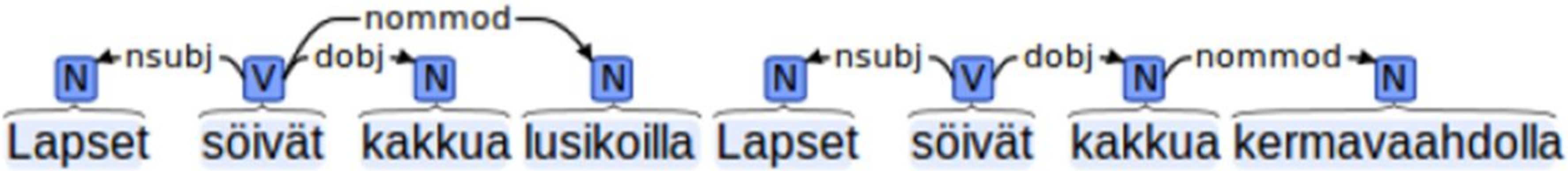


Parse!
"voi"
AUX
Mood: Ind, Number: Sing, Person: 0, Tense: Pres, VerbForm: Fin, Voice: Act
Lemma: voida
Postag: V
AUX
AUX
AUX
NOUN
PUNCT
1
Voi
voi
olla
lempiruokani
!


discourse
aux
aux
punct
cop
punct
INTJ
PUNCT
AUX#
AUX#
AUX#
NOUN#
PUNCT
1
Voi
,
voi
voi
olla
lempiruokani
.

Ambiguity in syntax




nommod
nsubj
dobj
N
V
N
N
Lapset
söivät
kakkua
lusikoilla
nsubj
dobj
nommod
N
V
N
N
Lapset
söivät
kakkua
kermavaahdolla

Word order

Subject and object can be either before or after the verb

- Lippuja voivat ostaa aikuiset. vs Aikuiset voivat ostaa lippuja.
- Aikuisia katselivat lapset. vs Lapset katselivat aikuisia.
- Or even: Sotilaita seurasi ihmisiä.

Lisäksi osmat eli objektin sijamuodossa olevat määrän adverbiaalit ovat luonnollisesti hankaliä jäsentimelle, sillä pintamuodosta huolimatta niitä pitäisi merkitä nmod-dependenssillä

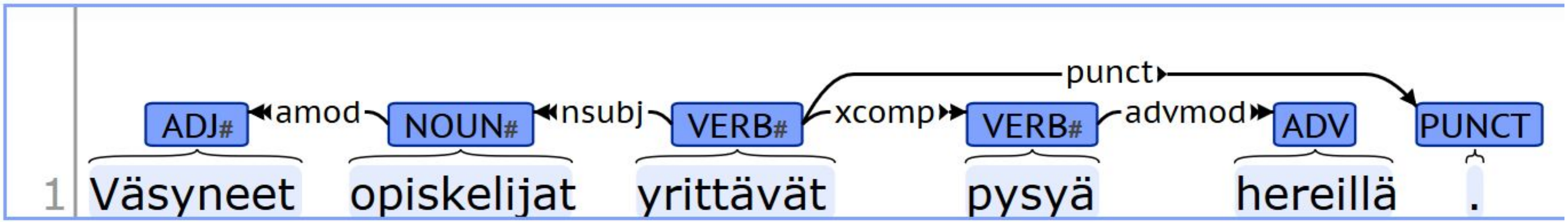
- Koira juoksi koko vuoden.
- Munkki paastosi kuukauden.

Towards practice

Turku-neural-parser-pipeline

http://turkunlp.org/Turku-neural-parser-pipeline/

A neural parsing pipeline for segmentation, morphological tagging, dependency parsing and lemmatization with pre-trained models for more than 50 languages. Top ranker in the CoNLL-18 Shared Task.



"Väsyneet"

ADJ

Case: Nom, Degree: Pos, Number: Plur

Lemma: väsynyt

Postag: A

UNIVERSITY OF OREGON
Natural Language Processing group

PEOPLE PUBLICATIONS SOFTWARE

Trankit is a light-weight Transformer-based Toolkit for multilingual Natural Language Processing (NLP). It pro
pipeline for fundamental NLP tasks over 100 languages, and 90 pretrained pipelines for 56 languages.

Trankit can be easily installed via pip: **pip install trankit**

For more information, please check out our github repo, documentation, and technical paper.

Usage

```
from trankit import Pipeline
# initialize a pipeline on English
p = Pipeline(lang='english', gpu=True, cache_dir='./cache')

doc = '''Michael helped shoot the majority of my firm's website and we could not have been happier.'''

# perform all tasks on the input
all = p(doc)

sents = p.ssplit(doc) # sentence segmentation
tokens = p.tokenize(doc) # tokenization
```

http://nlp.uoregon.edu/trankit

TURKUNLP
.ORG

id wordform lemma upos pos morpho head deptype

1	They	they	PRON	PRP	Case=Nom|Number=Plur	2	nsubj
2	buy	buy	VERB	VBP	Number=Plur|Person=3|Tense=Pres	0	root
3	and	and	CONJ	CC	_	4	cc
4	sell	sell	VERB	VBP	Number=Plur|Person=3|Tense=Pres	2	conj
5	books	book	NOUN	NNS	Number=Plur	2	obj
6	.	.	PUNCT	.	_	2	punct

https://universaldependencies.org/format.html

More practice

1. Let’s see what Turku Neural Parser demo gives us
2. We can also try out Trankit on Google Colab

Not secure | epsilon-it.utu.fi/parser_demo/

Online Parser Demo

[Turku NLP Group]

This is a demo of the Finnish dependency parsing pipeline. Pipeline includes text segmentation, morphological tagging, dependency parsing, and lemmatization. This demo is meant for testing only, if you need to parse a lot of text, please download the parser and models from here and run it locally. In case of problems, contact TurkuNLP Group (contact information here).

Easiest way to run the parser locally on Linux/Mac/Windows is Docker image. See instructions here.

Finnish-neural

Parser: http://turkunlp.github.io/Turku-neural-parser-pipeline/

Treebank documentation: http://universaldependencies.github.io/docs/

What to try?

1:

- Try to parse simple sentences you made up and analyze how the system works.
- What happens to ambiguous words / constructions?
- (Esimerkiksi suomen voi, kuusi, parka (takki tai ressukka), englannin virke Time flies like an arrow.

2:

- The parsers are trained on treebanks that usually do not include a lot of informal language or dialects. Let’s evaluate this a bit!
- Parse some ready sentences collected from somewhere, representing different genres and / or dialects.
- What kinds of mistakes do you get? Which mistakes are the most common? Can you note differences between genres?

What to do with conllu data then?

Topic of the day

- Using syntax-analysed data as the basis of linguistic analysis
  - A lot of potential for linguistics
- Graphical interface
  - Easy to use for small files
- Command line
  - Much more flexible than corpus software
  - Allows for more advanced and efficient processing
  - Standard solutions eventually make their way to software, but new and more complex methods are not (necessarily) available there

Two research designs in corpus linguistics
(see Biber & Jones 2009, Biber, Conrad & Reppen 1998)

Table 1.1 *Association patterns in language use*

A. Investigating the use of a linguistic feature (lexical or grammatical)
   (i) Linguistic associations of the feature
      – lexical associations (associations with particular words)
      – grammatical associations (associations with particular grammatical constructions)
   (ii) Non-linguistic associations of the feature
      – distribution across registers
      – distribution across dialects
      – distribution across time periods

B. Investigating varieties or texts (e.g., registers, dialects, historical periods)
   (i) Linguistic association patterns
      – individual linguistic features or classes of features
      – co-occurrence patterns of linguistic features

What to do with conllu data then?

- Type A
  - *Zero or pronominal subject? (Helasvuo & Kyröläinen 2016)*
  - *Near-synonyms (Biber, Conrad and Reppen 1998: 93)*
- Type B
  - *Filtering -- filter away function words etc.*
  - *Normalisation with lemmatisation*
  - *Lexico-grammatical characteristics of texts (Biber 1988)*

Today focus on type B

- Characteristics of entire texts
- Focus on *keyness*
  - A loose theoretical framework for analyzing important characteristics of a set of texts (Scott 1997, Scott & Tribble 2006)
  - Keywords ~ a group of words that function as a key to the text
  - Tell about the style and aboutness of the texts

Transportation/ lodging	Tourism	Narrative/ description	Physical features	Places/attractions	Food/drink
airport	adventure	afternoon	beach	city	beer
biking	arrived	amazing	beaches	gardens	delicious
boat	attractions	around	cliffs	museum	dinner
booked	destination	beautiful	hills	park	lunch
bus	explore	day	island	places	restaurant
ferry	exploring	enjoyed	islands	shops	restaurants
flight	guide	famous	mountain	town	
flights	holiday	hour	mountains	village	
headed	locals	located	river	villages	
hike	photo	lovely	rocks		
hiking	photos	nearby	sea		
hostel	sights	night	trees		
hostels	tour	north	water		
hotel	tourist	scenery			
hotels	tourists	scenic			
journey	tours	south			
ride	travel	spectacular			
road	travellers	steep			
streets	travelling	stunning			
trail	trip	sun			
trails	visit	sunny			
walk	visited	sunset			
walked	visiting	swimming			
walking	visitors	weather			

Biber & Egbert 2018

In practice

- Two corpora: target + reference
- Comparison of the words in the target corpus and in the reference corpus
- Keywords the ones that are over- or underrepresented in the target corpus
- Result a list of keywords
- Filtering + lemmatisation useful in the preprocessing

Methods used to extract keywords

“The standard”:

Frequency list of target corpus words

Frequency list of reference corpus words

Challenges with keyword analysis

- Text should be the unit of observation!
- (Can you think of why?)
- How to evaluate the keywords?

Text dispersion (Egbert & Biber 2019)

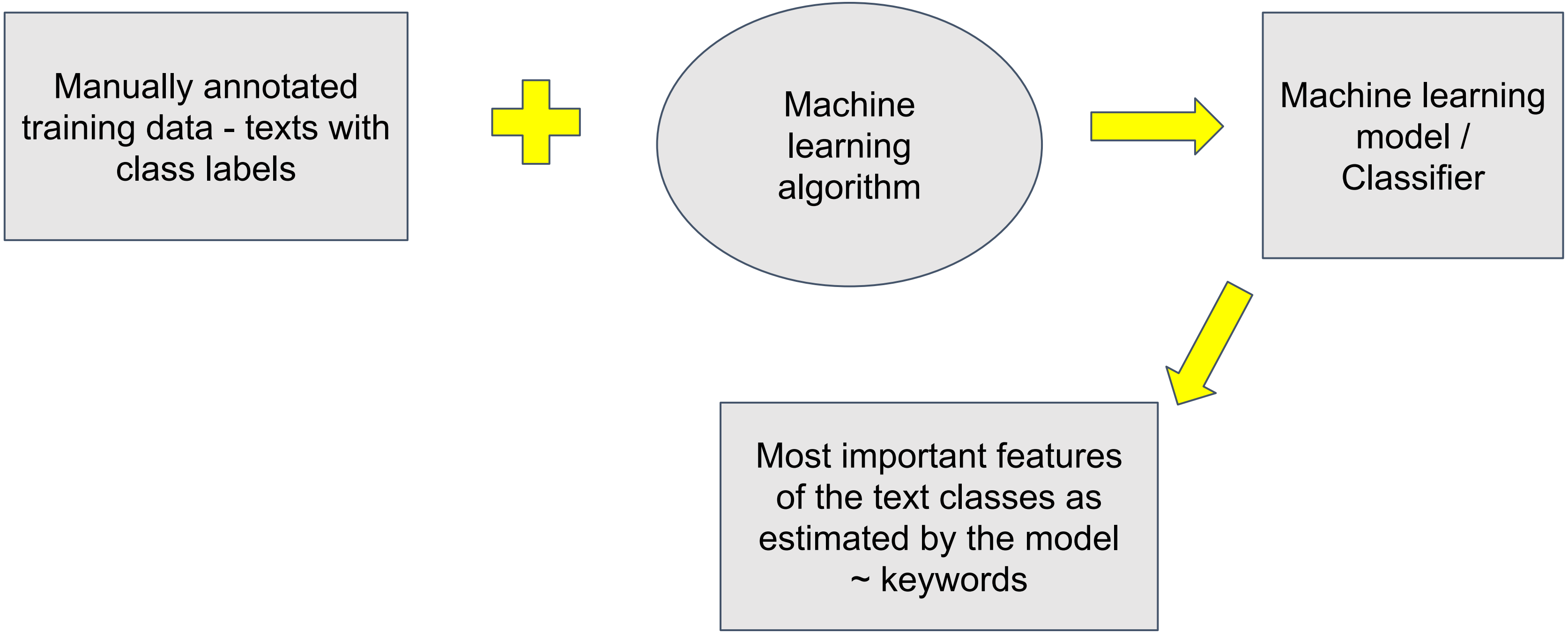
Document frequencies of the target corpus words

Document frequencies of the reference corpus words

Challenges with keyword analysis

- How to evaluate globally how well the keywords reflect **the target and reference corpora?**
- How to evaluate locally how well the keywords reflect **individual documents?**
- One possible solution to these challenges is to set up the task as text classification between the target and reference corpora

Text classification

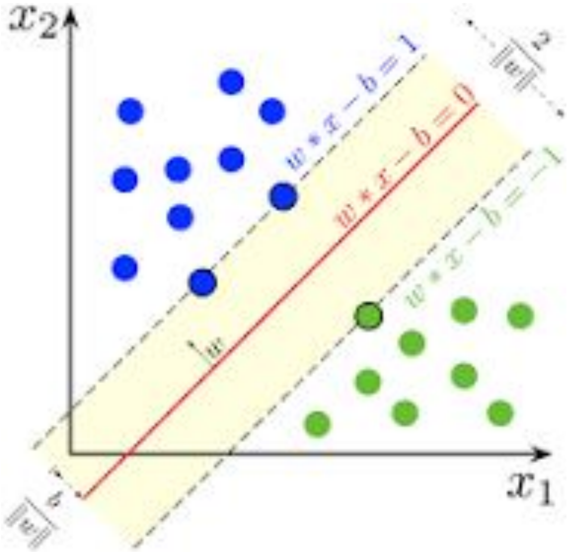


Text classification in keyword analysis

- Answers many of the challenges listed previously
  - Text as a unit of observation
  - Global evaluation: model performance
  - Local evaluation: how confident is the classifier about the text?
- Machine learning not new in linguistics
  - Random forests are often applied to study e.g. the geographical preference of -ing or to complements (e.g. Dehors & Gries 2016)

Support vector machines

- Good with sparse data (such as texts)!
- Feature importance often used to explain predictions in NLP (e.g. Sharoff et al. 2010)



Text classification

- Split data to train and test sets
  - ML focuses on *predicting* new instances
  - → Evaluation done on the test set that has not been seen during the training
- Typical evaluation measures precision, recall, f1-score (balanced and harmonized mean of precision+recall)
- Featurization - how are the data represented to the classifier?
  - Words, but also lemmas, syntactic information, etc.
  - → Comparison of the importance of words / other features for the classes in the data (see e.g. Laippala et al. 2021)

References

Biber, D., Conrad, S. and Reppen, R (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge.

Biber, D., & Jones, J. K. (2009). Quantitative methods in corpus linguistics. In Corpus Linguistics: An International Handbook (Vol. 2, pp. 1286-1304). De Gruyter Mouton.

Biber, D., Egbert, J. (2018) Register Variation Online. Cambridge.

Deshors, Sandra & Gries, Stefan. (2016). 2016. Deshors, Sandra C. & Stefan Th. Gries. Profiling verb complementation constructions across New Englishes: A two-step random forests analysis of -ing vs. to- complements. International Journal of Corpus Linguistics 21(2): 192-218.. International Journal of Corpus Linguistics. 21. 10.1075/ijcl.21.2.03des.

Egbert, Jesse & Biber, Doug. (2019). Incorporating text dispersion into keyword analyses. Corpora. 14. 77-104. 10.3366/cor.2019.0162.

Helasvuo, Marja-Liisa and Kyröläinen, Aki-Juhani. "Choosing between zero and pronominal subject: modeling subject expression in the 1st person singular in Finnish conversation " Corpus Linguistics and Linguistic Theory, vol. 12, no. 2, 2016, pp. 263-299.

Laippala, Veronika, Jesse Egbert, Douglas Biber, Aki-Juhani Kyröläinen (2021) Exploring the roleof lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. Language resources and evaluation.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC. 2016.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman. Universal Dependencies. Computational Linguistics. 2021.

Mike Scott, Chris Tribble (2006): Textual Patterns: Key Words and Corpus Analysis in Language Education

Sharoff, S., Z. Wo, and K. Markert. 2010. The web library of babel: evaluating genre collections. Proceedings of the Seventh Conference on International Language Resources and Evaluation, 3063–3070.