

(Universal) Dependency Treebanks

A very short introduction

Kaja Dobrovoljc

Faculty of Arts, University of Ljubljana
Jožef Stefan Institute, Ljubljana

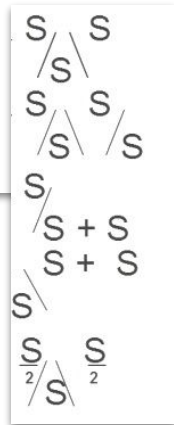
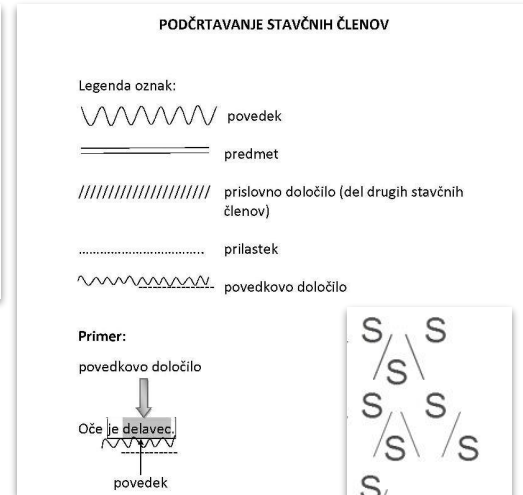
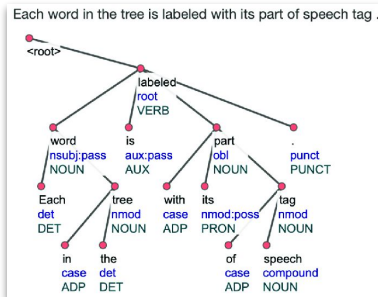
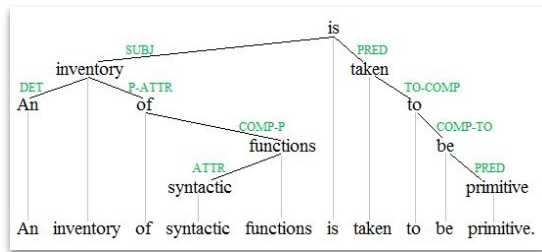
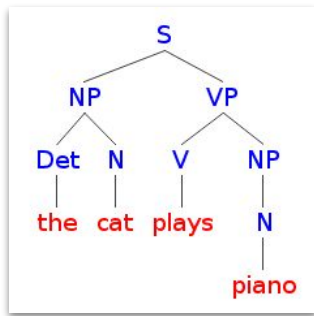
FRI UL, Ljubljana, May 16, 2024

Outline of the talk

- Dependency treebanks
- Universal Dependencies
- Some use cases in NLP
- State-of-the-art for Slovenian
- Conclusion

Treebanks

- Structured text corpora with **syntactically annotated sentences**
 - Sentences as tree-like graphs



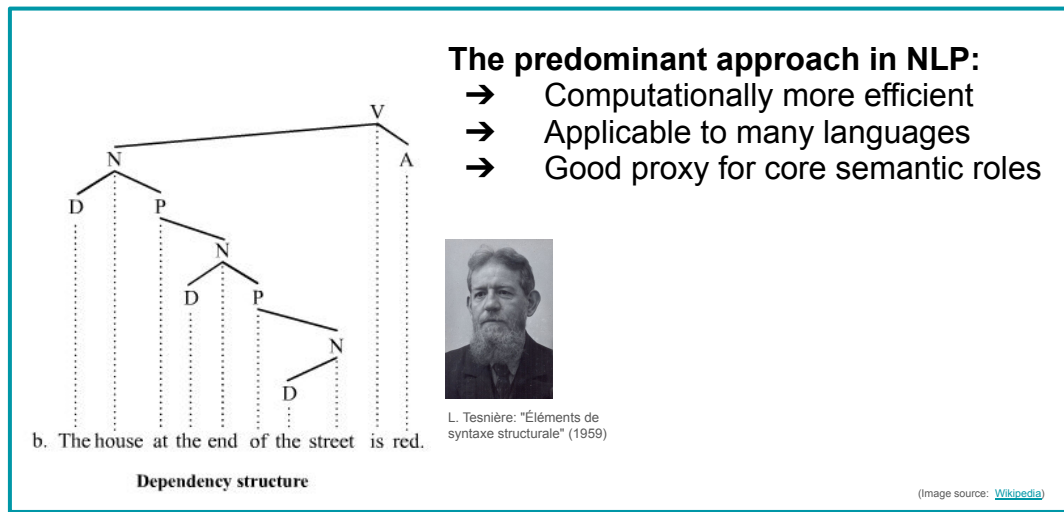
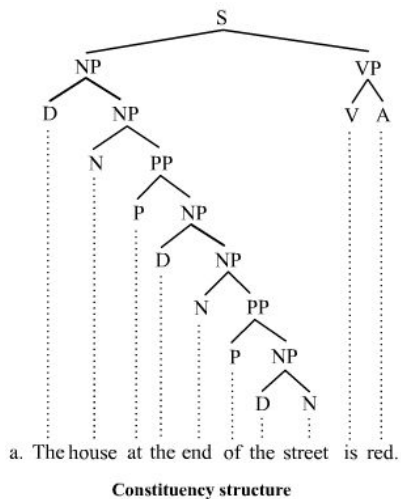
- Useful for linguistics
 - Development of linguistic theories
 - Data-driven grammar description
 - Language documentation and preservation
- ... and NLP

Two main approaches

- **Constituency treebanks:** phrase structure rules
 - e.g. Penn Treebank
- **Dependency treebanks:** syntactic relations between words
 - e.g. Prague Dependency Treebank



N. Chomsky: Syntactic structures (1957)




b. The house at the end of the street is red.

Dependency structure

The diagram shows a dependency structure for the sentence "The house at the end of the street is red." The root node V branches into N and A. The N branches into D and P. The P branches into N and PP. The second N branches into D and P. The third P branches into N. The fourth N branches into D and N. The fifth N branches into D and N. The A branches into V.

The predominant approach in NLP:

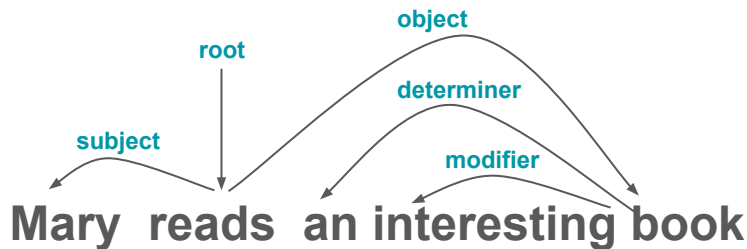
- Computationally more efficient
- Applicable to many languages
- Good proxy for core semantic roles



L. Tesnière: "Éléments de syntaxe structurale" (1959)

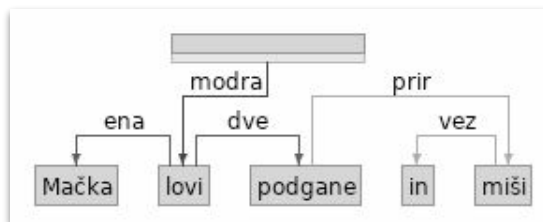
Dependency tree

- Graphical representation of sentence structure with directed edges connecting a **head** (governing word) to a **dependent** (subordinate word).
 - Directed acyclic graph
- Components:
 - **Nodes**: Represent words in the sentence.
 - **Edges**: Define syntactic relationships, showing which word governs another.
 - **Labels**: Specify the type of grammatical function (e.g., subject, object, modifier).
- **Root**: The ultimate head from which all dependencies originate, usually the main verb.

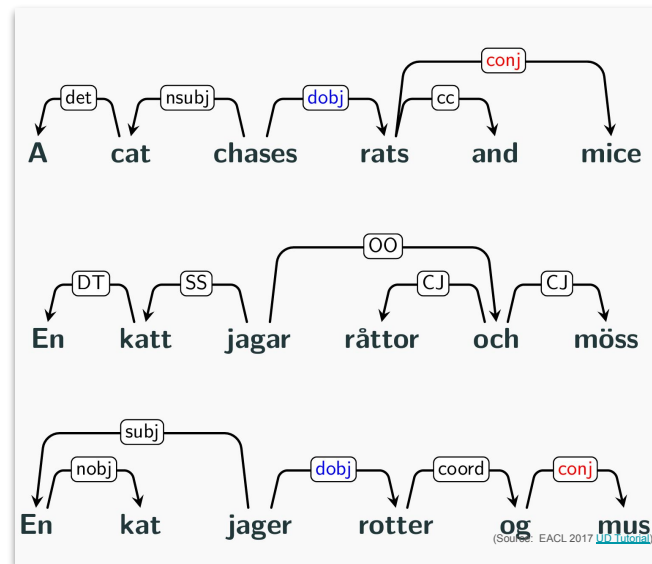


Multitude of annotation schemes

- **Language-, theory- or corpus-specific** approaches to parsing, e.g.
 - PDT-inspired Slovene Dependency Treebank (Džeroski et al. 2006)
 - JOS Dependency Treebank (Ledinek and Erjavec 2009)



- A big problem for:
 - **Comparing empirical results** across languages
 - Doing **cross-lingual transfer and learning**
 - Building and maintaining **multilingual systems**



Universal Dependencies

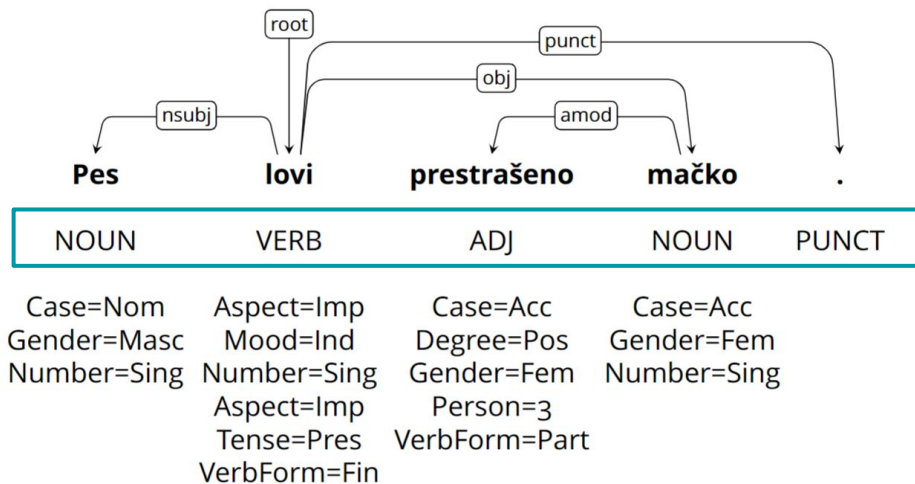
Universal Dependencies (UD)

- Project aimed at developing **cross-linguistically consistent treebank annotation** for many languages, started in **2014**
- Built on common usage and existing de facto standards
 - e.g. Google Universal Part-of-Speech Tagset, Stanford Dependencies
- Aimed at complementing (not replacing) language-specific schemes
 - Application-oriented (not a theory)
 - **Community-driven, open-source**
- Main design principles:
 - Lexicalism: basic annotation units are words
 - Function words modify content words
 - Flexibility: language-specific features and relation subtypes



Annotation scheme

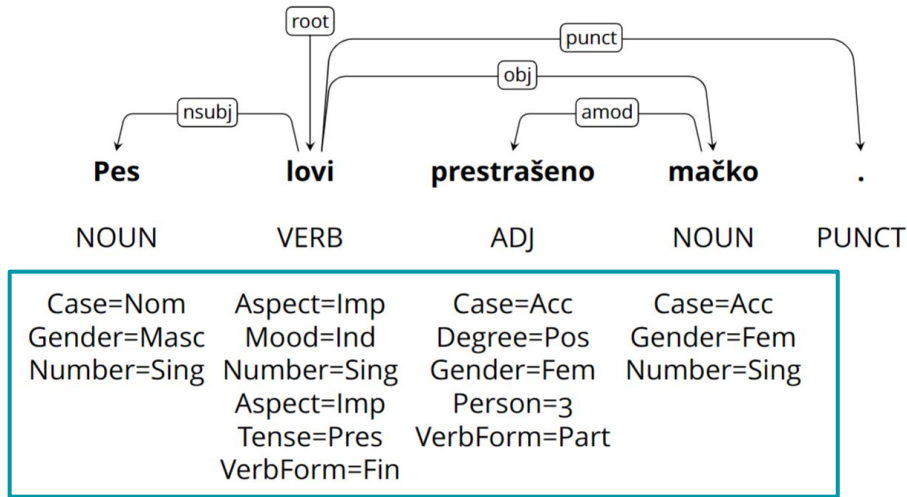
- Part of speech categories (17 tags)
- Morphological features (24 features)
- Syntactic dependencies (37 relations)



(English: *A dog chases a scared cat.*)

Annotation scheme

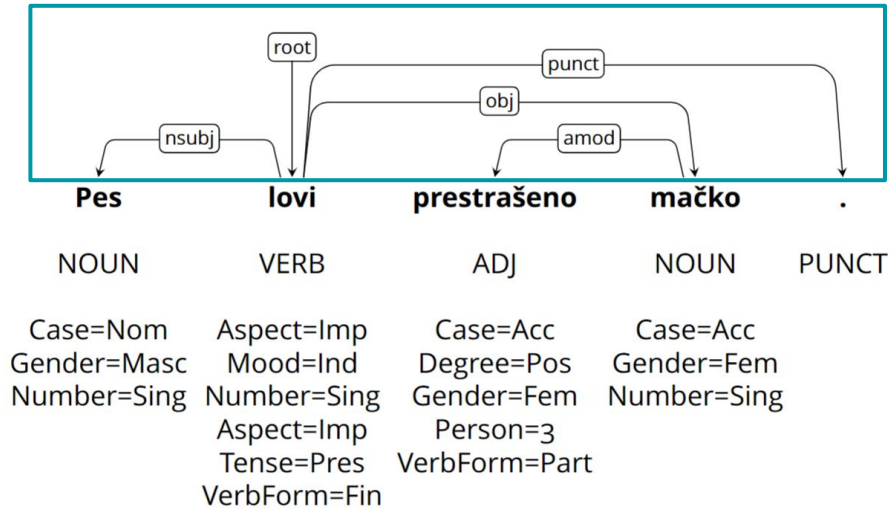
- Part of speech categories (17 tags)
- Morphological features (24 features)
- Syntactic dependencies (37 relations)



(English: *A dog chases a scared cat.*)

Annotation scheme

- Part of speech categories (17 tags)
- Morphological features (24 features)
- Syntactic dependencies (37 relations)



(English: *A dog chases a scared cat.*)

UD categories and guidelines

- Well-documented online: <https://universaldependencies.org/guidelines.html>

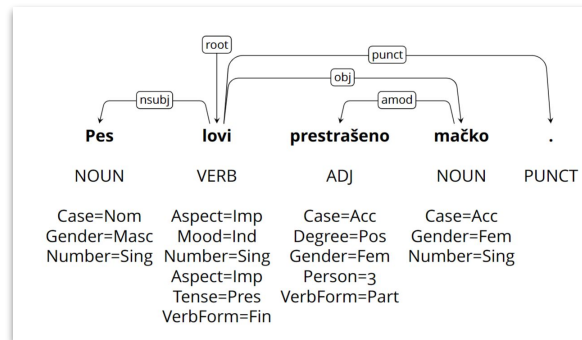
Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Lexical features*	Inflectional features*	
	<i>Nominal*</i>	<i>Verbal*</i>
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Deixis	Polarity
	DeixisRef	Person
	Degree	Polite
		Clusivity

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	Headless	Loose	Special	Other
conj cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

CONLL-U Format

- One word per line, 10 columns separated by tabs.
- Sentences separated by blank lines.
- Comment lines begin with a hashtag (e.g. # sent_id =)



1: Token ID

2: Surface word form

3: Uninflected word form

4: Universal part of speech

5: Language-specific tag (e.g. MULTEXT-East)

6: Universal morphological features

7: ID of head word

8: Universal relation to head

9: enhanced dependency

10: any other annotation

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Pes	pes	NOUN	Ncmsn	Case=Nom Gender=Masc Number=Sing	2	nsubj	_	_
2	lovi	loviti	VERB	Vmpr3s	Aspect=Imp Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0	root	_	_
3	prestrašen o	prestrašen	ADJ	Appfsa	Case=Acc Degree=Pos Gender=Fem Number=Sing VerbForm=Part	4	amod	_	_
4	mačko	mačka	NOUN	Ncfsa	Case=Acc Gender=Fem Number=Sing	2	obj	_	_
5	.	.	PUNCT	Z	_	2	punct	_	_

Core UD fields

























Optional fields

Data repository

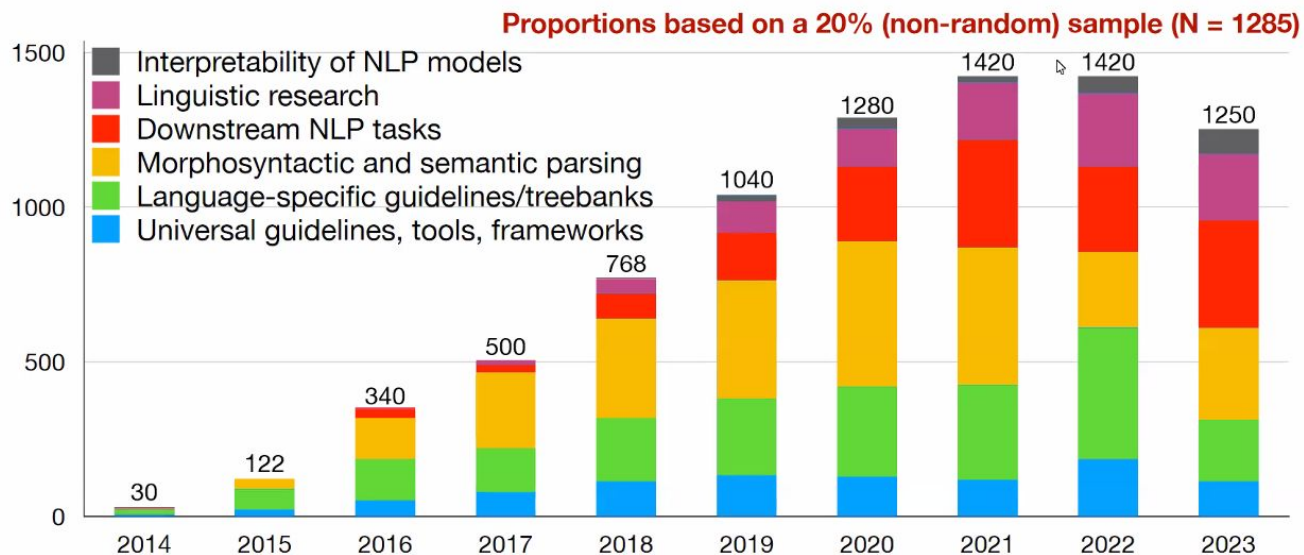
- Official UD dataset release every 6 months
 - LINDAT/CLARIN repository
 - Development: <https://github.com/UniversalDependencies>
- UD v2.13 in numbers:
 - **259 treebanks**
 - **148 languages**
 - 30 language families
 - 1.8 million sentences
 - 30.8 million words
 - 577 contributors

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

▶		Abaza	1	<1K	🗨️	Northwest Caucasian
▶		Afrikaans	1	49K	🗨️🗨️	IE, Germanic
▶		Akkadian	2	25K	🗨️🗨️	Afro-Asiatic, Semitic
▶		Akuntsu	1	1K	🗨️🗨️	Tupian, Tupari
▶		Albanian	1	<1K	🗨️	IE, Albanian
▶		Amharic	1	10K	🗨️🗨️🗨️	Afro-Asiatic, Semitic
▶		Ancient Greek	3	456K	🗨️🗨️🗨️	IE, Greek
▶		Ancient Hebrew	1	39K	🗨️	Afro-Asiatic, Semitic
▶		Apurina	1	<1K	🗨️🗨️	Arawakan
▶		Arabic	3	1,042K	🗨️🗨️🗨️	Afro-Asiatic, Semitic
▶		Armenian	2	94K	🗨️🗨️🗨️🗨️	IE, Armenian
▶		Assyrian	1	<1K	🗨️🗨️	Afro-Asiatic, Semitic
▶		Bambara	1	13K	🗨️🗨️	Mande
▶		Basque	1	121K	🗨️	Basque
▶		Beja	1	1K	🗨️	Afro-Asiatic, Cushitic
▶		Belarusian	1	305K	🗨️🗨️🗨️🗨️	IE, Slavic
▶		Bengali	1	<1K	🗨️	IE, Indic
▶		Bhojpuri	1	6K	🗨️	IE, Indic
▶		Bororo	1	1K	🗨️	Bororoan
▶		Breton	1	10K	🗨️🗨️🗨️🗨️	IE, Celtic
▶		Bulgarian	1	156K	🗨️🗨️🗨️	IE, Slavic
▶		Buryat	1	10K	🗨️🗨️	Mongolic
▶		Cantonese	1	13K	🗨️	Sino-Tibetan
▶		Catalan	1	553K	🗨️	IE, Romance
▶		Cebuano	1	1K	🗨️	Austronesian, Central Philippine
▶		Chinese	7	309K	🗨️🗨️🗨️🗨️	Sino-Tibetan
▶		Chukchi	1	6K	🗨️	Chukotko-Kamchatkan

Research based on UD treebanks



Using Treebanks in NLP

Use case 1: Parser development and evaluation

- Multilingual **tools for grammatical annotation** from raw text to UD
 - Single tool for many annotation layers
 - Segmentation, tokenization, lemmatization, POS tagging, feature prediction, parsing
 - Single tool for many languages
 - Major advances through CoNLL Shared Tasks 2017-2018
 - Invaluable for low-resource languages
- Two main approaches
 - Transition-based vs. Graph-based parsing



Use case 1: Parser development and evaluation

- Typical evaluation setup for UD annotation:
 - **train-dev-test** data splits already featured in the official release of UD treebanks
 - **F1** scores for the full pipeline
 - **UAS** = unlabelled attachment score (% of words with correct **head** prediction)
 - **LAS** = labelled attachment score (% of words with correct **head and label** prediction)

Metric \ System	Tokens	Words	Sents.	Lemmas	UPOS	XPOS	UFeats	UAS	LAS
Trankit (English EWT)	98.67	98.67	90.49	96.65	96.47	96.75	97.25	91.29	89.4

Use case 2: Model understanding

- Syntactic probing: technique used to investigate **what linguistic information is encoded** in the hidden layers of neural network models
- Language models like BERT embed entire syntactic trees implicitly in their word representation spaces, demonstrating deep understanding of language structure

A Structural Probe for Finding Syntax in Word Representations

What does BERT learn about the structure of language?

Do Neural Language Models Show Preferences for Syntactic Formalisms?

Artur Kulmizev
Uppsala University
artur.kulmizev@lingfil.uu.se

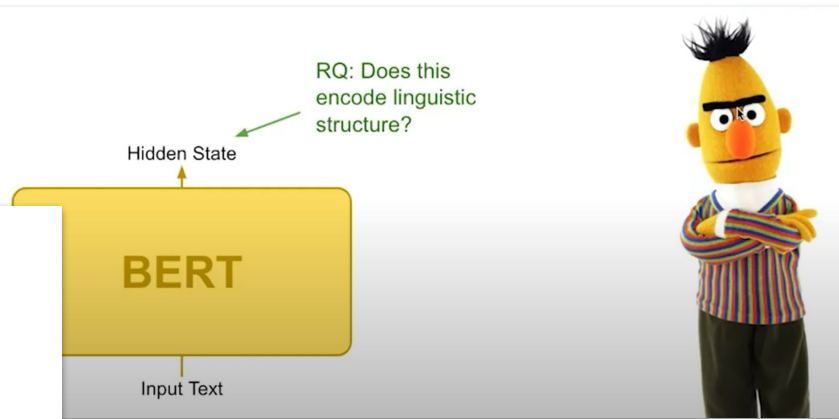
Vinit Ravishankar
University of Oslo
vinitr@ifi.uio.no

Mostafa Abdou
University of Copenhagen
abdou@di.ku.dk

Joakim Nivre
Uppsala University
joakim.nivre@lingfil.uu.se

Abstract

of language models, often establishing strong parallels between the two (Benedict et al., 2019; Abner



Use case 3: Downstream NLP tasks

- E.g. machine translation, sentiment analysis, relation extraction, question answering
- With the rise of LLM, there is less need for supervised syntactic parsing in downstream applications.
- However, explicit syntactic structure can still be useful for
 - High-precision tasks
 - Resource efficiency
 - Hybrid systems

Is Supervised Syntactic Parsing Beneficial for Language Understanding Tasks? An Empirical Investigation

Goran Glavaš

University of Mannheim
Data and Web Science Group

goran@informatik.uni-mannheim.de

Ivan Vulić

University of Cambridge
Language Technology Lab

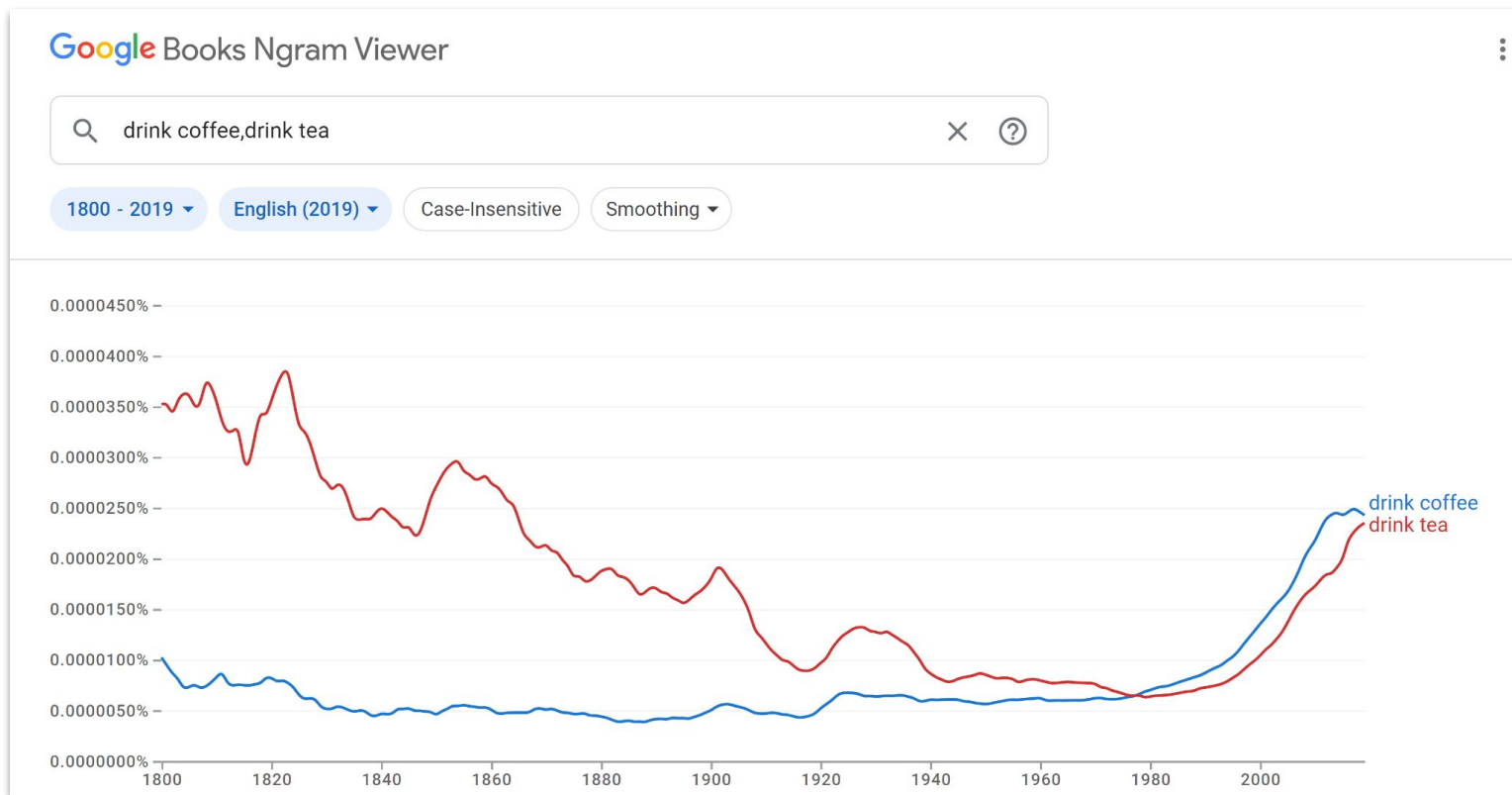
iv250@cam.ac.uk

Abstract

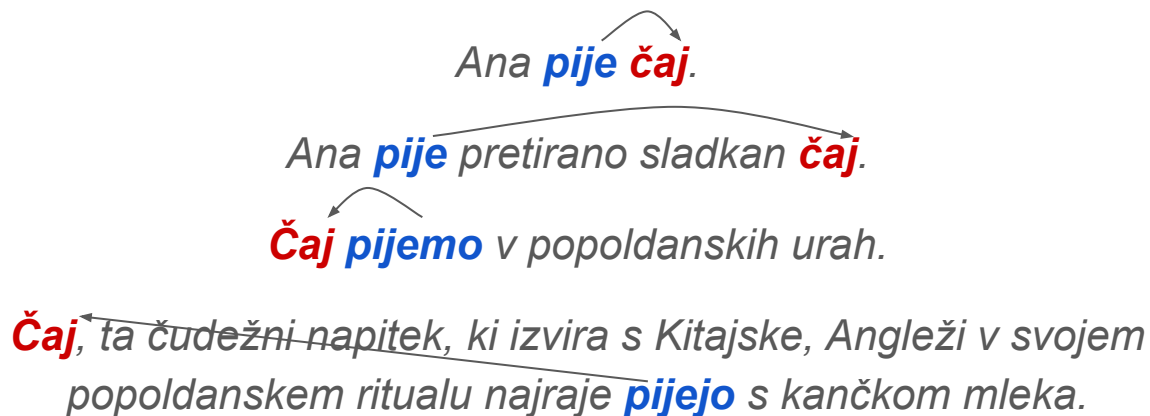
Traditional NLP has long held (supervised) syntactic parsing necessary for successful higher level semantic language understanding

rather strong common belief that high-level semantic language understanding (LU) crucially depends on explicit syntax. The unprecedented success of neural language learning models based on trans-

Use case 4: Information Extraction



Use case 3: Information Extraction



All examples are instances of the same simple tree -- **drink tea** -- which can easily be retrieved from parsed data.

Not as complex as it seems

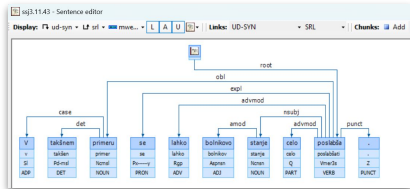
- Annotation guidelines are [well-documented online](#) and easily comprehended
- Common NLP applications mostly only involve a subset of UD tags:
 - e.g. **VERB**, **NOUN**, **ADJ**, **ADV** part-of-speech categories
 - e.g. relations for nominal predicate arguments (**nsubj**, **obj**, **iobj**; **obl**, **advmod**)
- Many tools for processing CONLL-U files are available
 - [conllu](#) python library
 - [pyconll](#) python library
 - [nltk.corpus.reader.conll](#) module
- Theoretically not ideal, but ‘good enough’ for many applications.

UD Landscape for Slovenian

Data and tools

- Two manually annotated UD treebanks for Slovenian
 - **SSJ** (news, non-fiction, wikipedia): **13k sentences** (267k tokens)
 - Part of the SUK training corpus: additional 750k tokens with lemmas, UPOS and XPOS
 - **SST** (spontaneous speech): **6k sentences** (76k tokens)
- Several tools facilitating their analysis or the creation of new treebanks

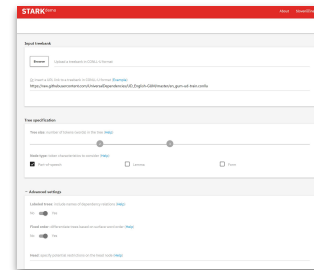
Q-CAT tool for manual treebank annotation



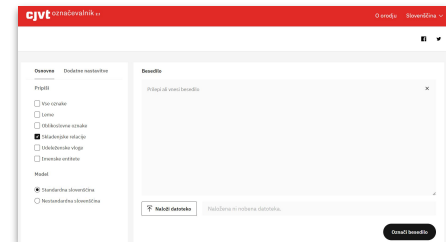
Drevesnik service for treebank querying



STARK tool for bottom-up tree extraction



Označevalnik service for automatic text annotation

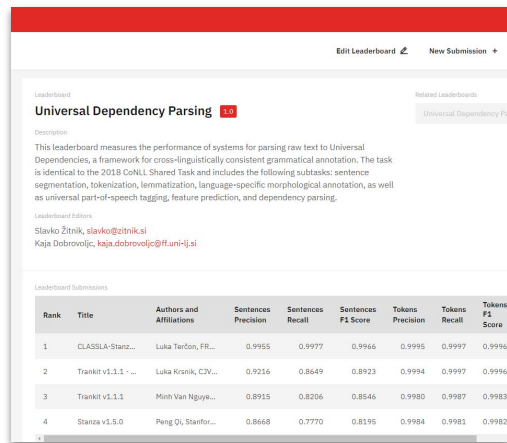


Data and tools

- Locally developed **state-of-the-art annotation models** for Slovenian
 - [CLASSLA-Stanza](#): models trained on latest SUK/SSJ + tagging control + Sloleks lookup
 - [Trankit](#): [models](#) trained on latest SSJ and SST

	Lemma F1	UPOS F1	XPOS F1	LAS F1
Trankit	98.17	98.98	97.97	94.22
○ CLASSLA-Stanza	99.15	99.13	98.29	91.13

- [SloBench](#) leaderboard for UD annotation
 - Hidden Slovenian UD test set
 - Evaluation based on the official CoNLL 2018 ST script
 - **New competing systems are welcome!**



Universal Dependency Parsing 1.0

This leaderboard measures the performance of systems for parsing raw text to Universal Dependencies, a framework for cross-linguistically consistent grammatical annotation. The task is identical to the 2018 CoNLL Shared Task and includes the following subtasks: sentence segmentation, tokenization, lemmatization, language-specific morphological annotation, as well as universal part-of-speech tagging, feature prediction, and dependency parsing.

Rank	Title	Authors and Affiliations	Sentences Precision	Sentences Recall	Sentences F1 Score	Tokens Precision	Tokens Recall	Tokens F1 Score
1	CLASSLA-Stanz...	Luka Teržon, FR...	0.9955	0.9977	0.9966	0.9995	0.9997	0.9996
2	Trankit v1.1.1 -...	Luka Kranik, CIV...	0.9216	0.8649	0.8923	0.9994	0.9997	0.9996
3	Trankit v1.1.1	Minh Van Ngoye...	0.8915	0.8206	0.8546	0.9980	0.9987	0.9983
4	Stanza v1.5.0	Peng Qi, Stanfor...	0.8668	0.7770	0.8195	0.9984	0.9981	0.9982

Data and tools

- Several **automatically parsed reference corpora** for Slovenian ('parsebanks')
 - [Gigafida](#) reference corpus of standard Slovene (1B words)
 - [Trendi](#) monitoring corpus
- Other **resources with UD morphology**
 - [Janes-Tag](#) training CMC corpus (manual)
 - [CLASSLA-web-si](#) (automatic)
 - [Sloleks](#) inflectional lexicon

Current treebank-based research projects

- [SPOT](#): Treebank-Driven Approach to the Study of Spoken Slovene
 - Data-driven **identification of speech-specific syntactic patterns**
- [MEZZANINE](#): Basic Research for the Development of Spoken Language Resources and Speech Technologies for Slovenian
 - Developing an **audio-aware annotation pipeline** based on SST
- PhD research on linguistic characteristics of LLMs (L. Terčon)
 - **Measuring syntactic complexity** in Slovene and English LLM-generated texts
- [PROP](#): Empirical Foundations for Digitally-Supported Development of Writing Skills
 - Analysing (the development of) **syntactic patterns in student writing**



<..> mezzanine



Conclusion

Summary

- **Dependency treebanks** are syntactically parsed corpora, in which the structure of a sentence is described as a set of binary **relations between words**. Such corpora are frequently used in both NLP and linguistics.
- Universal Dependencies has become the standard annotation scheme in NLP, aiming at cross-lingually consistent **morphological and syntactic annotation**.
- Its main advantages include a large manually annotated **multilingual dataset** and a **wide range of tools and services** supporting CONLL-U data analysis and downstream applications.

Conclusion

- The role of treebanks in NLP is shifting, but they are probably here to stay:
 - Fine-tuning for specific linguistic tasks (e.g. parsing)
 - Model interpretability and understanding
 - Transfer-learning and domain-specific applications in low-resource scenarios
 - Educational and research tool -- under-exploited in linguistics
- It is important to ensure an active development of such resources and tools for Slovenian, and everyone can participate.
 - e.g. new, better annotation tools for Slovenian
- Stay tuned for NLP-related [UniDive](#) activities
 - “Universality, Diversity and Idiosyncrasy in Language Technology” (COST Action 2022-2026)
 - Planned in 2025:
 - Shared task on morpho-syntactic parsing
 - Training school on transfer-learning for low-resource languages



Thank you!

Questions?

kaja.dobrovoljc@ff.uni-lj.si

Useful links

- Jurafsky and Martin 2024. Speech and Language Processing. [Chapter 18: Dependency parsing](#)
- <https://universaldependencies.org/>
- UD [tutorial](#) for beginners
- de Marneffe et al. 2021. [Universal Dependencies](#).

- UD for Slovenian:
 - Papers on SSJ and SST treebanks for Slovenian:
 - <https://aclanthology.org/W17-1406/>
 - <https://journals.uni-lj.si/slovenscina2/article/view/12031>
 - <https://aclanthology.org/L16-1248/>
 - Detailed UD guidelines for Slovenian with many examples:
<https://wiki.cjvt.si/books/07-universal-dependencies/page/oznacevalne-smernice>