Poster session B

chair prof. dr Ranka Stanković

UniDive 3rd general meeting, Tuesday 29 January 2025, Budapest

WG1

UniDive

Funded by the European Union

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

A Corpus of Persian Sentences Annotated with Verbal Multiword Expressions: Development and Guidelines

Primary Data		Main Changes		Final Data
#Sentences: 5617 **#VMWE:** 5383 **#LVC.full:** 5365 **#VID:** 17 **#IRV:** 1	>	1. The Category of the NV in LVCs 2. VMWEs as Light Verbs 3. Agreement on the NV 4. Prefix Verbs	>	**#VMWE:** 5504 **#LVC.full:** 4603 **#VID:** 567 **#VPC.full:** 238 **#VPC.semi:** 86 **#IRV:** 10

The classification for Persian VMWEs

1. Light Verb Constructions (LVC)
2. Verbal Idioms (VID)
3. Verb-particle constructions (VPC)
4. Reflexive verbs (IRV)
5. Multi-Verb Constructions (MVC)

Vahide Tajalli, Mehrnoush Shamsfard, Yalda Yarandi, Mahtab Sarlak, Arezoo Haghbin
NLP lab, Shahid Beheshti University, Tehran, Iran

Morphosyntactic evaluation for text summarization in morphologically rich languages

Batuhan Baykara, Tunga Güngör
Boğaziçi University, Computer Engineering, Istanbul, Turkey

- Text summarization
  - Evaluation metrics (ROUGE, METEOR, etc.)
  - Does not take morphosyntactic structure of words into account
  - Problem when the generated summaries contain words in different forms

- Contributions:
  - Several variants of the commonly used evaluation metrics
    - that take into account the morphosyntactic properties of the language
  - Correlation analysis
    - to see how well the score obtained with each metric correlates with the human score

Cross-Dialectal Perspectives on Pomak

Challenges of Pomak Language:

- Highly under-resourced and endangered, spoken primarily in Bulgaria, Greece, and Turkey.
- Exhibits phonological, lexical, and syntactic diversity.

Research Phases:

- **POS Tagger Development:** Focused on Pomak spoken in Turkey; built on the linguistic framework by Karakaş (2022).
- Graph-based neural parser with BiLSTM embeddings (Dozat et al. 2017).
- **Corpus Creation:** Developed a 650-sentence corpus; addressed data scarcity via cross-lingual transfer learning from Pomak UD Treebank

Key Results:

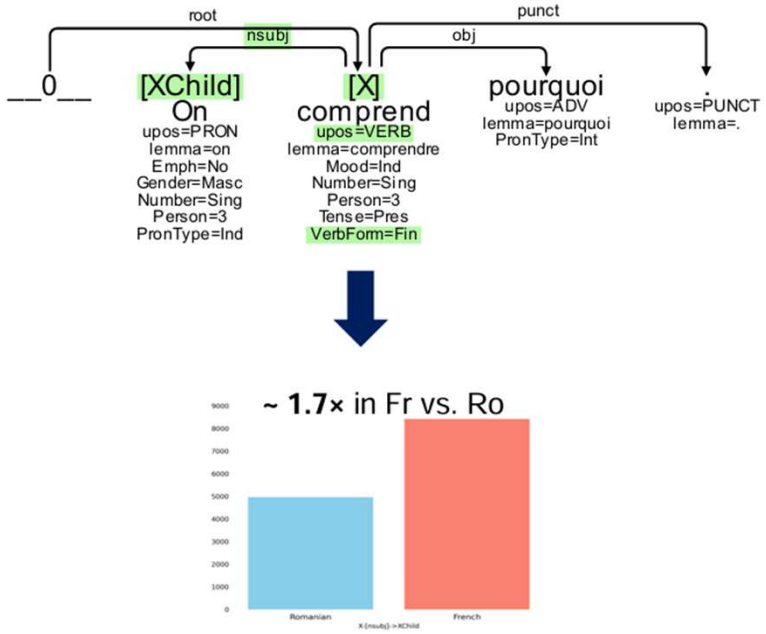
Performance Improvements

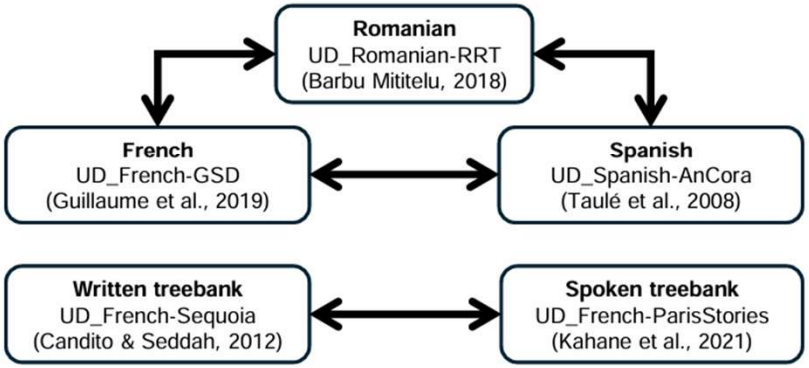
- **Word-based unlabeled attachment score (UAS)** Improved to **68%**
- **Labeled attachment score (LAS)** Improved to **62%**

Building Quantitative Contrastive Grammars from Syntactic Treebanks

Santiago Herrera et al.

Mining **CONTRASTIVE PATTERNS** across comparable treebanks using simple ML techniques





Perceptions on MWE lexicons use in NLP by the User Community: features, challenges and recommendations

Raquel Amaro, Voula Giouli, Gražina Korvel, Irina Lobzhanidze,
Verginica Barbu Mititelu, Giedrė Valūnaitė Oleškevičienė

Working Group: WG2 (Lexicon-corpus interface)

Objective: Finding gaps in current MWE lexicons and provide recommendations for improvement

Methodology: Structured questionnaire shared via mailing lists, networks, and social media

Findings:

- Demographics: 134 responses; 86% academic, 10% corporate, 8% government
- Usage: MWE lexicons widely applied in parsing, MT, NER
- Challenges:
  - Inconsistent definitions and annotations
  - Limited machine-readability and language coverage
  - Scarcity of real-world MWE examples

Recommendations:

- Coverage: expand to include dialects, specialized domains, and dynamic updates
- Information richness: enhance annotations and examples linked to corpora
- Interoperability: develop universal MWE typologies for cross-linguistic applications

Lexicons enhance MWE identification

■ Context:

Two words with unexpected behavior

ملحه علي ركبته [his salt on his knees] ≠ Salt + Knees

Mieć muchy w nosie [To have flies in your nose] ≠ Flies+noise

Get angry

■ Objectif

Unseen VMWEs: Identifying MWEs that have not been seen in training datasets.
Idiomatic Ambiguity: Differentiating literal from figurative meanings.

■ Results

Lexicon integration for MWE identification for Arabic and Polish using:

- **Ar**: LexAR + AraBERTLite
- **Pl**: Verbel + Mtlb-Struct
- **A classifier (PIEC) to distinguish idiomatic vs. literal MWEs**

UniDive 3rd General Meeting

Developing Digital Tools for Aromanian Language Use and Distribution

Marija Pendevska[1], Branislav Gerazov[2], Branko Prlja[3]
[1]Komercijalna Banka AD Skopje, [2]UKIM FEEIT, [3]Arno.mk

Planned Activities:

- Capture knowledge for synchronous and asynchronous media.
- Use English and regional Balkan languages as linguistic bridges.
- Develop speech technology for real-time language access.

Work in Progress:

- Development digital tools aimed at enhancing community participation and involvement.
- Inclusion of additional bridge languages to expand reach.

Annotating Constructions with UD

the experience of the Italian Constructicon

Ludovica Pannitto[1], Beatrice Bernasconi[2], Lucia Busso[3], Flavio Pisciotta[4], Giulia Rambelli[1], Francesca Masini[1]

[1]Alma Mater Studiorum - University of Bologna, [2]University of Turin, [3]Aston University, [4]University of Salerno

UniDive

EXPERIMENTAL LAB
lilec.lab@unibo.it

Funded by the European Union

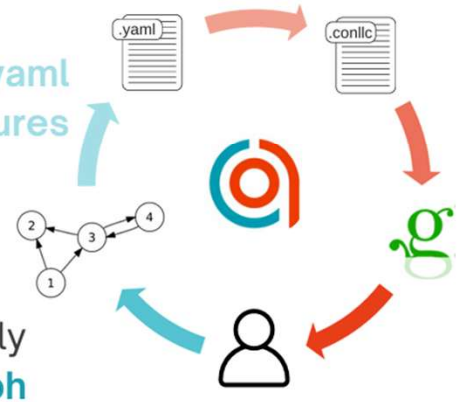
cost

WG2

The Italian Constructicon (ItCon) workflow

1. new *cxns* are described in a yaml file through their main features

2. and formalized in the UD-compatible CoNLL-C format

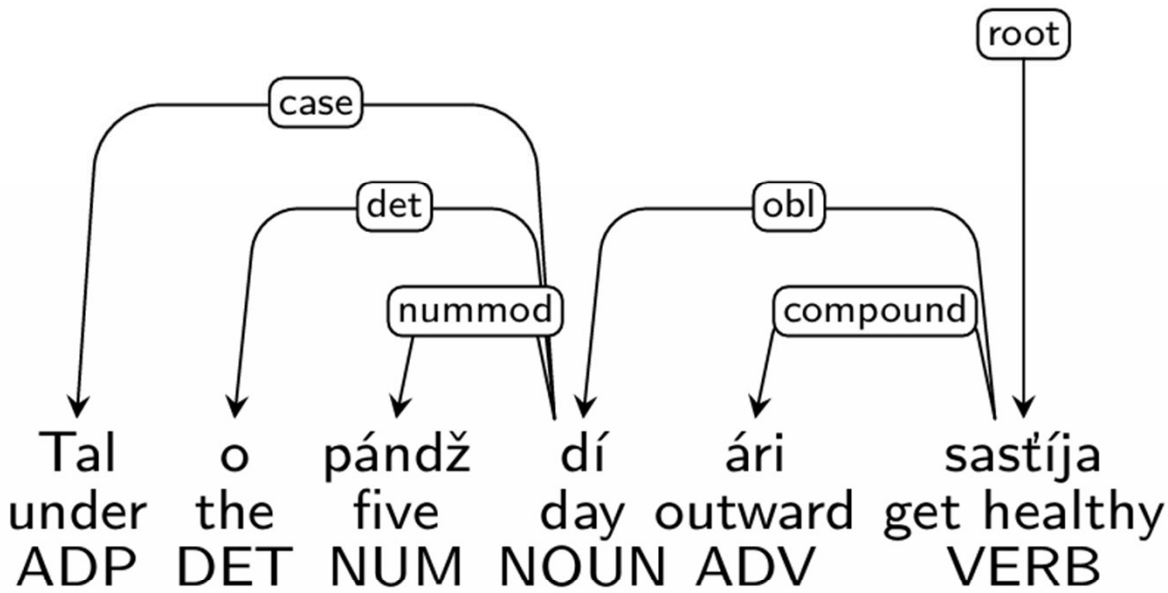


3. the CoNLL-C is translated into grew queries to retrieve instances of each *cxn*

4. examples are manually checked by a linguist

5. *cxns* are semi-automatically organized into a directed graph

Universal Dependencies for Selice Romani



He/She recovered in five days.

Kaja Dobrovoljc, Jaka Čibej

UNIVERZA V LJUBLJANI

JSI

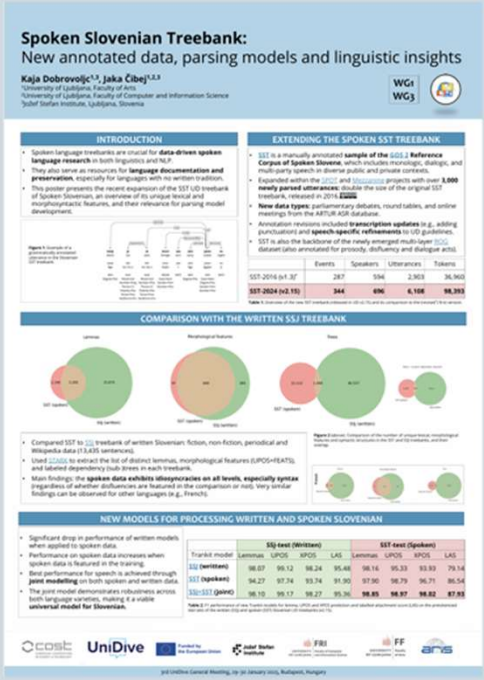
Spoken Slovenian Treebank:

New annotated data, parsing models and linguistic insights

Key Insights:

- **Expanded Spoken Data Treebank:** Over 3,000 new utterances, including parliamentary debates and online meetings.
- **Mode-Agnostic Parsing Model:** Joint modelling on spoken and written data achieves SOTA results on both modalities.
- **Bottom-up Idiosyncracy Identification:** Spoken data reveals distinct lexical, morphological, and syntactic features in comparison to writing.

Relevant for: WG1, WG3, WG4



UniDive

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

MWE-Annotator: Automatic Identification of MWEs in Dutch – Jan Odijk / Gosse Bouma (WG1+2)

- MWE-Finder: [Odijk et al 2024]
  - identifies 1 MWE in a large corpus
  - Web application for linguists/lexicographers
  - Provides >11k Dutch MWEs in canonical form
- MWE-Annotator
  - Identifies each of the 11k MWEs in a large corpus
  - Creates annotations
  - Command-line tool

Georgian Treebank in Universal Dependencies Framework: Annotation and Parsing with UDPipe

Irina Lobzhanidze[1], Erekle Magradze[2], Svetlana Berikashvili[3], Anzor Gozalishvili[4], Tamar Jalaghonia
Ilia State University

Funded by the European Union

Motivation and Significance

- Georgian, a complex morphosyntactic Kartvelian language, was underrepresented in the Universal Dependencies (UD) framework.
- Addressed the challenges of split-ergativity, free word order, and complex inflectional morphology to enrich global linguistic resources.

Key Contributions

First Georgian Syntactic Treebank within UD:

- Annotated 3,164 sentences (56,239 tokens) from diverse genres and domains.
- Included data from the Georgian Language Corpus (GLC) and Wikipedia, ensuring variety and linguistic depth.

Adaptation of UD Standards:

- Developed annotation guidelines aligning Georgian morphosyntactic features with UD principles.
- Created language-specific documentation for public use.

Annotation Process

Addressed syntactic constructions:

- Simple Clauses: Predicate with primary arguments.
- Coordinated Clauses: Main or subordinate clauses in a coordinate structure.
- Subordinate Clauses: Core and non-core dependents in clausal structures.

Ensured data quality:

- Compilation of Guidelines for syntactic functions.
- Manual and automated validation.

Model Training and Results

Frequent misinterpretations regarding the model output:

- Gold data included more complex structures, while the parser often oversimplified;
- Challenges caused by split-ergativity in distinguishing subjects and objects marked with Case=Nom or Case=Dat;
- Inconsistencies in modifier assignments based on positional emphasis or sentence context.

The results highlight the strength of the model in basic parsing but also reveal challenges with Georgian's free word order and case-marking system.

Challenges related to the syntactic treebank:

- Mapping Limitations: Georgian morphosyntactic features like diathesis-related tags (e.g., autoactive, inactive) were not fully compatible with UD standards;
- Annotation Accuracy: syntactic dependencies like flat:foreign and flat:name required manual corrections;
- Complex Structures: difficulties in annotating valency-changing operations, and arguments marked by various cases (e.g., nominative, ergative, dative).

WG 1

UniDive

3rd General Meeting

Hungarian Research Centre for Linguistics

Bucarest, Hungary, 29-30 January 2025

https://unidive.lisn.upsaclay.fr/

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Treebank for Characterisation: Syntax of Speakers in Roman Tragedy

Distribution of linguistic patterns

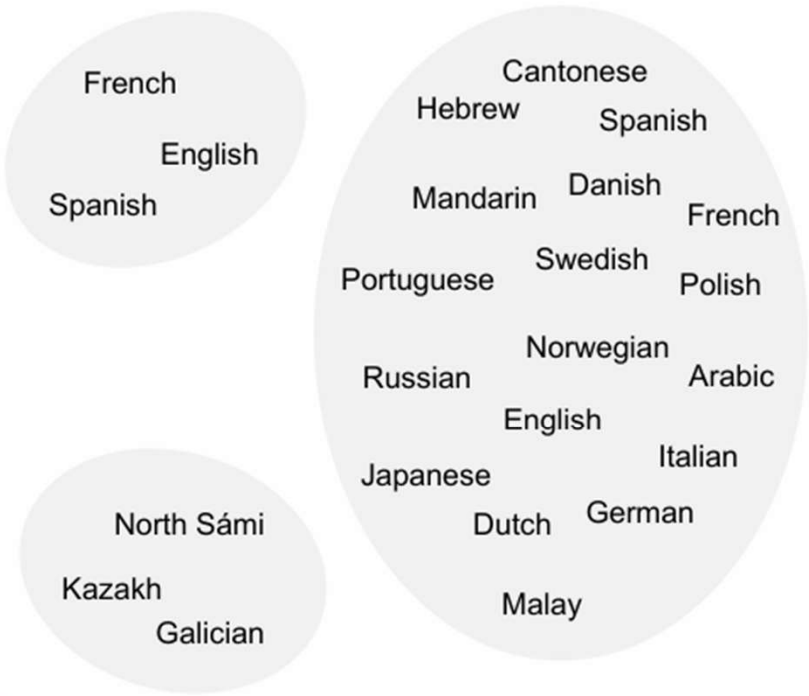
Modellisation of metadata in CoNLL-U – annotation of speakers –

federica.iurescia@unicatt.it
giovanni.moretti@unicatt.it

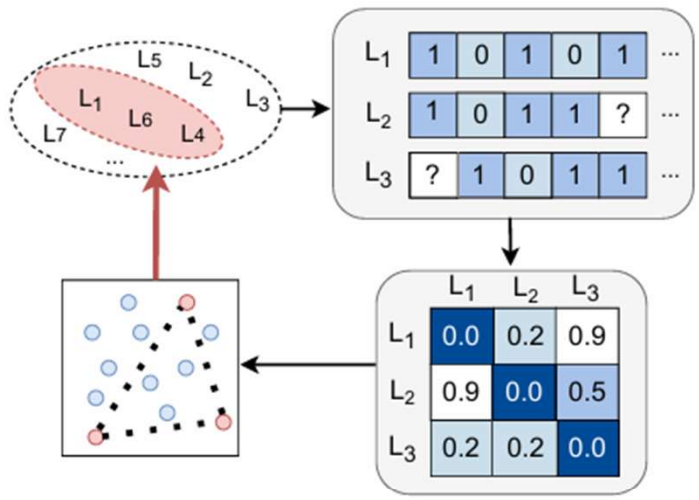
UNIVERSITÀ CATTOLICA del Sacro Cuore
CIRCSE
Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione

How Can I Select Diverse Evaluation Languages?

Equally Diverse?



A Solution?



Authors: Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux & Johannes Bjerva

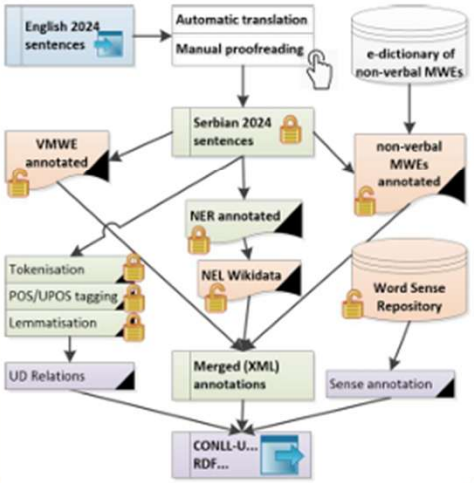
Relevant working groups: WG3, WG4

Progress in SR-ELEXIS Semantic Annotation: Focusing on Multiword Expressions, Named Entities, and Sense Repository

Cvetana Krstev, Ranka Stanković, Aleksandra Marković, Milica Ikonić Nešić

ELEXIS-WSD

SR sense-annotated corpus - in progress (UniDive WG2.T2)



COMPLETED TASKS

- tokenization
- POS-tagging
- lemmatization
- NE annotation

IN PROGRESS

- more MWEs (710: 653)
- filling gaps in SrpWN for WSD-SR
- postediting done for 1,526 new synsets (70,97%)

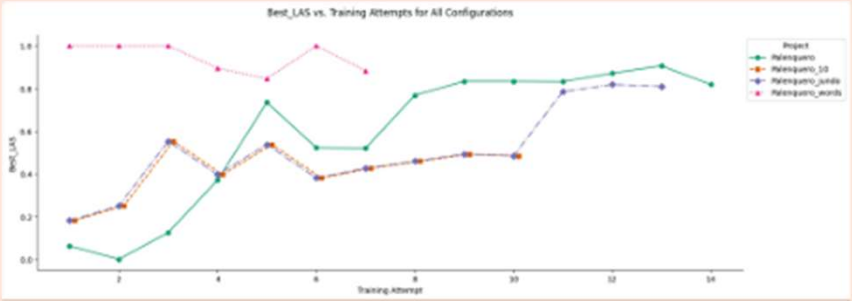
TO BE DONE

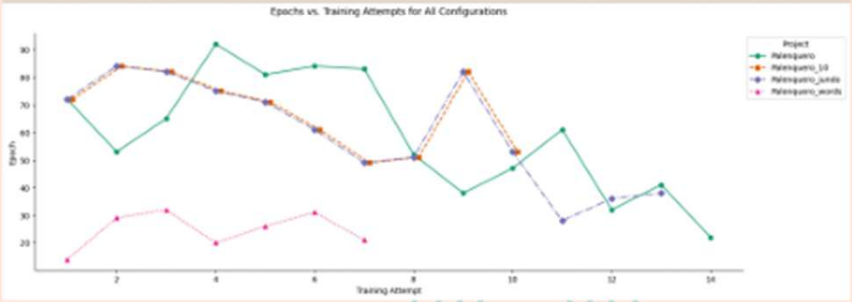
- evaluation & correction of MWE & VMWE annotation;
- completion of NE linking;
- SR finalization, semi-automatic meaning assignement, syntactic annotation.

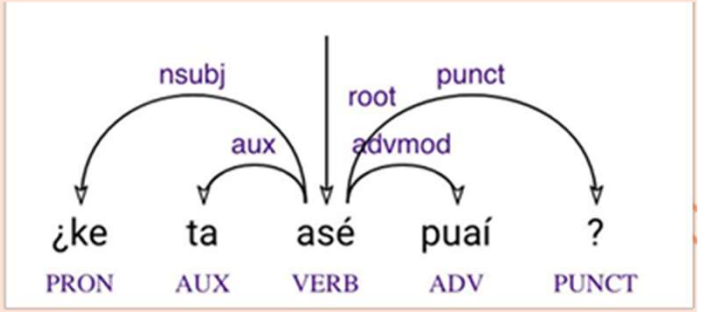
MULTIPLE CONFIGURATIONS FOR AUTOMATED POS TAGGING AND PARSING OF PALENQUERO CREOLE (COLOMBIA)

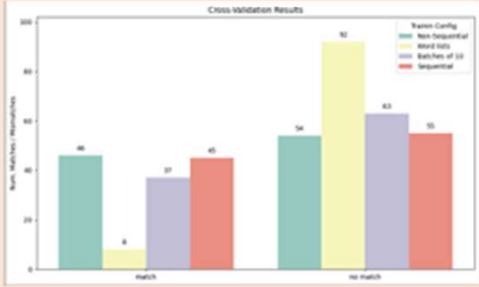
Daniel J. Casas — daniel.jimenezcasas@upf.edu — WG1

- This study examined four training configurations to test which one performed best at POS tagging and parsing of Palenquero.
- Mixed methods yielded better results when combining ruled- & non-ruled based approaches.









UNIVERSITÄT DES SAARLANDES

UniDive

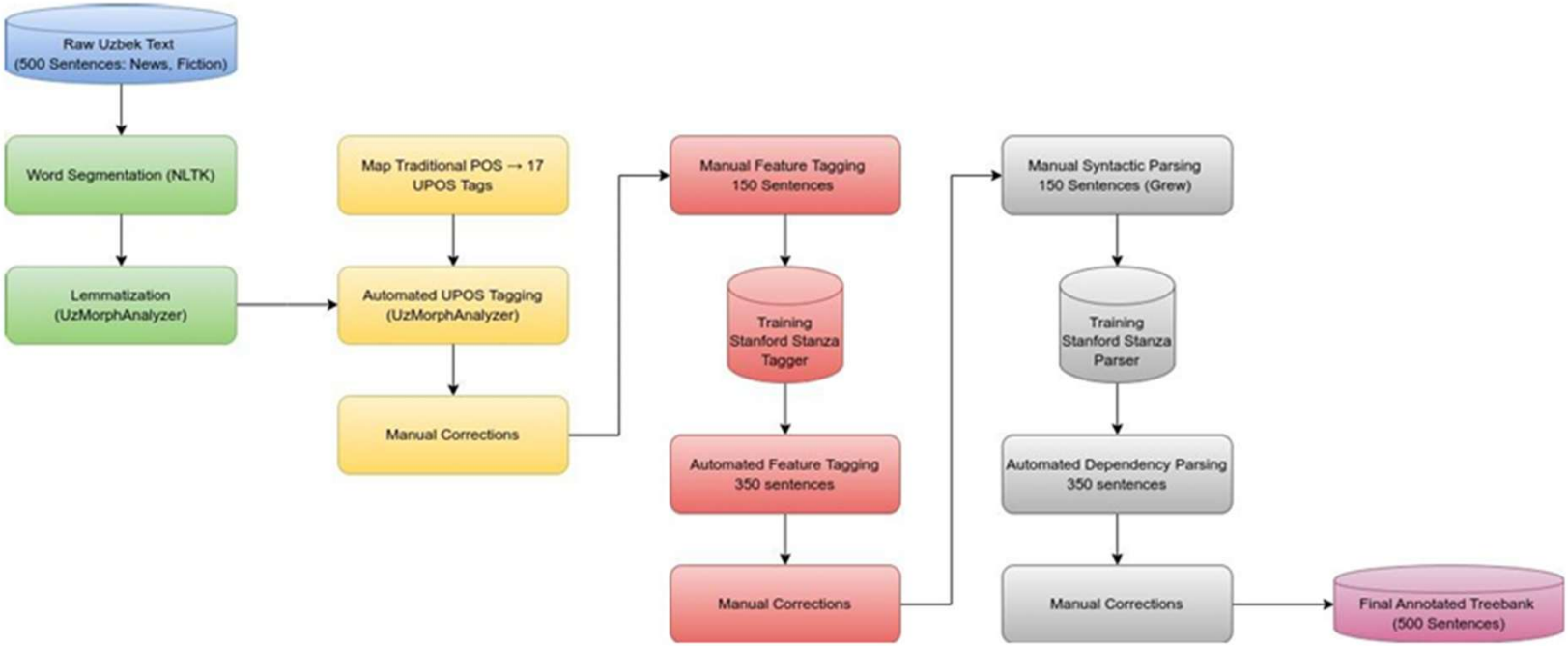
3rd General Meeting
Hungarian Research Centre for Linguistics
Budapest, Hungary, 29-30 January 2025

Funded by the European Union

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Universal Dependencies Treebank for Uzbek

Arofat Akhundjanova
Saarland University



Working Group 1

UniDive · cost · Funded by the European Union · Universitat Rovira i Virgili · Research Group on Mathematical Linguistics

3rd General Meeting - Budapest

Complete work

A Computational and Quantitative Reassessment of Greenebrgian Language Types

Antoni Brosa Rodríguez – Universitat Rovira i Virgili – antoni.brosa@urv.cat

2.11

U1. In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.

```
%1-SVO
pattern {V [upos=VERB]; V -[subj]-> S;  S [upos=PROPN|NOUN]; V -[comp:obj]-> O; O [upos=PROPN|NOUN]; V >> S; S << O; V<<O}
without {V -[punct]-> P; P [lemma="?"|"!"]}
```

Language	BWO	Adposition
Afrikaans	SVO	Prepositional
Belarussian	SVO	Prepositional
Dutch	SVO	Prepositional
German	SVO	Prepositional
Greek	SVO	Prepositional
Tamil	SOV	Postpositional
Urdu	SOV	Postpositional

U1 Results

U1	True	Wrong
ori	90.80%	9.20%
pn	72.84%	27.16%
pnN	79.31%	20.69%
any	83.91%	16.09%
q_ori	100%	0%
q_pn	95.06%	4.94%
q_pnN	98.85%	1.15%
q_any	98.85%	1.15%

Language	OURS
afrikaans	SVO
akkadian	SOV
akuntsu	SOV
amharic	SOV
ancient greek	NDO
ancient hebrew	VSO
arabic	VSO
belarussian	SVO
buryat	SOV
classical chinese	SVO
dutch	SVO
faroese	SVO
galician	SVO
german	SVO
gheg	SVO
gothic	SVO
greek	SVO
hindi_english	SVO
icelandic	SVO
karelian	SVO
kazakh	SOV
kiche	SVO
komi zyrian	SVO
kurmanji	SOV
latin	NDO
maltese	SVO
marathi	SOV
moksha	SVO
north sami	SVO
old east slavic	NDO
old french	SVO
sanskrit	SOV
slovak	SVO
tamil	SOV
turkish_german	SOV
upper sorbian	SVO
western armenian	NDO
western puebla nahuatl	SVO
xibe	SOV
yakut	SOV
yoruba	SVO
yupik	NDO

Proposal

- Different formalisation
- More data
- From labels to quantities
- Different scheme
- Different data management
- Strict interpretation of Greenberg claims
- Flexible interpretation of Greenberg claims

NLPre: A language-centric platform for benchmarking NLPre systems

Martyna Wiącek, Alina Wróblewska

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

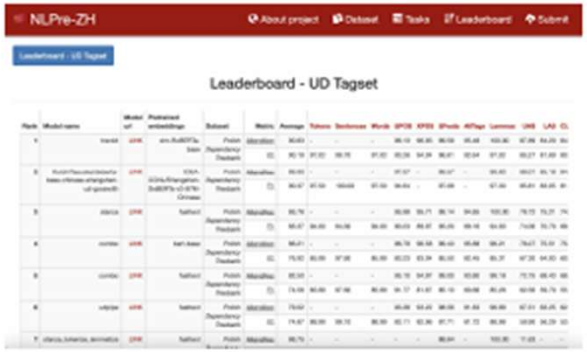
Online benchmarking

Language-centric

Up-to-date leaderboard

NLPre-GA (Irish)

NLPre-PL (Polish)



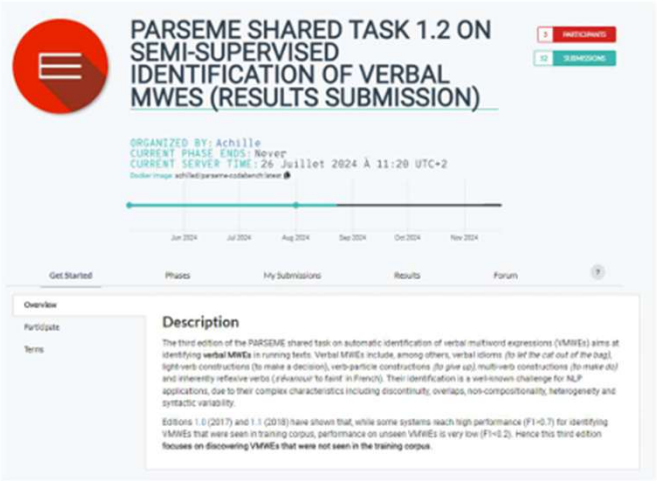
NLPre-ZH (Chinese)

UniDive 3rd GM, Budapest 2025

Porting the PARSEME 1.2 shared task and diversity metrics to the Codabench Platform

Achille Desreumaux, Louis Estève, Agata Savary, Anne-Catherine Letournel
Université-Paris-Saclay, LISN–CNRS, France

- **Codabench**: Open source Machine Learning competition platform administred by the LISN laboratory
- **PARSEME 1.2**: Third iteration of shared tasks on Verbal Multiword Expressions (VMWEs)



- How can Codabench be used to host multilingual shared tasks such as PARSEME 1.2?

- Addition of **3 diversity metrics**
- Exploration of two possibilities:
  - **Result competition** (similar to ad-hoc PARSEME 1.2)
  - **Code competition**

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
cnrs
université PARIS-SACLAY
UniDive
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
Funded by the European Union

Annotating Noun Compound Candidates in Irish Text

WG1, WG4

- Noun compounds display non-compositional behaviour
- Evaluate how NLP applications handle idiomaticity
- No such dataset for **Irish** noun compounds!

ADAPT DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

To Annotate Compositionality…

? Terminology
? Named Entities
? Annotator Expertise
? Confidence of Annotators

What kind of data?
Where to find annotators?

Interesting Constructions

- Productive constructions
  - E.g. *lucht oibre* 'working people'
- Genitive constructions
  - E.g. *tóradh na talún* 'fruits of the earth'
- Mythical creatures
  - E.g. *bean sí* 'banshee'

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
UniDive
Funded by the European Union