



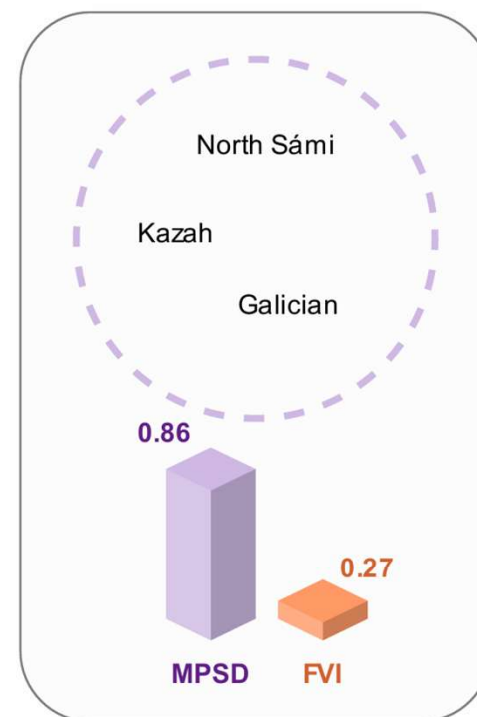
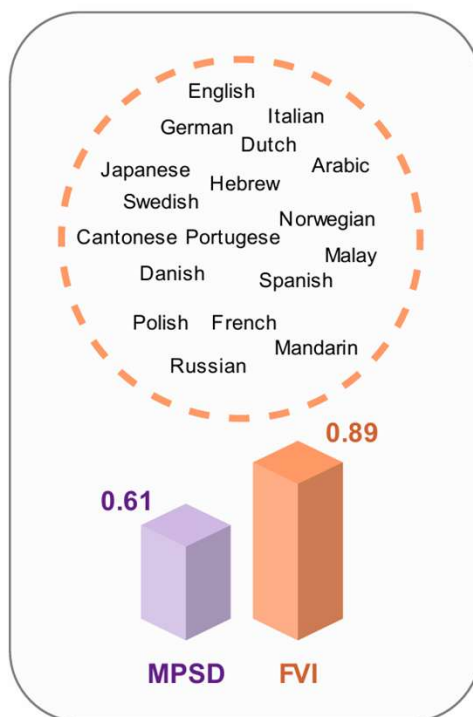
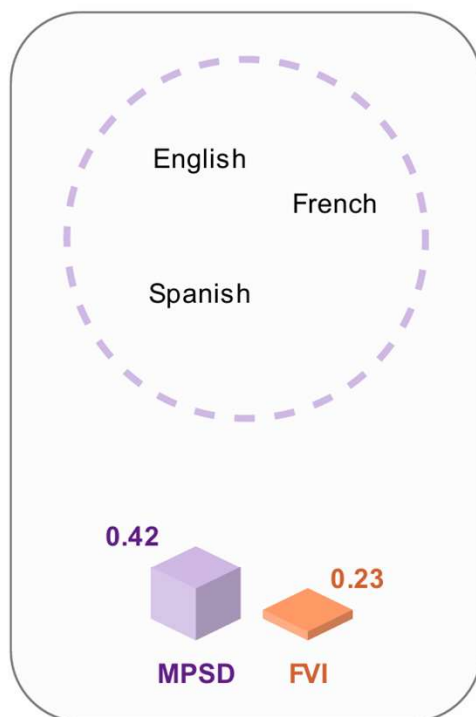
Poster session A

chair prof. dr Ranka Stanković

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

Are these language samples "typologically diverse"?

A survey of what "typological diversity" means in NLP research. Previously presented at EMNLP 2024.

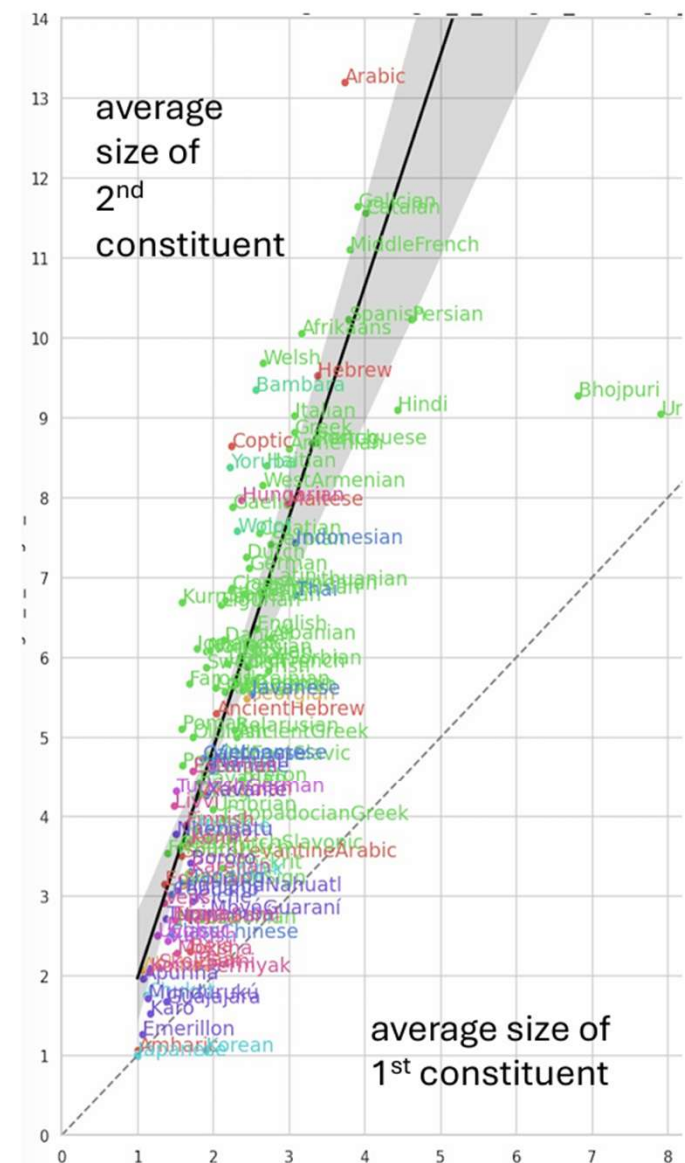


The 2.5x factor: A linguistic “short- before-long” riddle across the 160 UD languages

Kim Gerdes

Sylvain Kahane

This verb has two dependents on its right side





Conversion of the School Dictionary of Croatian into a Sense Inventory for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Case Study

Siniša Runjaić¹, Jaka Čibej²

¹Institute for the Croatian Language, Croatia

²Centre for Language Resources and Technologies, University of Ljubljana, Slovenia

Relevant to WG2

You think (and feel) that your language should be represented in the ELEXIS-WSD Parallel Sense-Annotated Corpus too, but you have some **questions**:

- What is the ELEXIS-WSD corpus? What is its purpose?
- Ok, now I know what the corpus is... What are the mandatory prerequisites for joining?
- Ok, the translation of the original English WikiMatrix subcorpus, i.e. getting 2,024 validated sentences in any language seems like a relatively easy task, but how about sense inventory?
- Ok, most of the original 10 tailored data for their sense inventories from national Wordnet, but what if we don't have access to it? Can it still be done some other way?



WG1
Work in progress

Annotation of Old Egyptian Multiword Expressions in PARSEME

Roberto A. Díaz Hernández
University of Jaén

UniDive

cost

Aim of this research work

The aim of this project is to annotate Old Egyptian multiword expressions in PARSEME. It is a work in progress that will allow the creation of a new lexical resource for the study of Egyptian. The PARSEME methodology will be applied to the analysis of MWEs in Old Egyptian in order to study the diachronic use of its MWEs.

Methodology

Old Egyptian MWEs will be annotated via the editor of the FoLia Linguistic Annotation Tool (FLAT) using a CoNLL-U file containing the **UD Egyptian-UJaen treebank**. An Old Egyptian MWE candidate will be first classified according to the morphology of the head word as a verbal multiword expression (VMWE), a nominal multiword expression (NMWE) or a functional multiword expression (FMWE).

Conclusion

This is an innovative and multidisciplinary project. It will not only contribute to computational linguistics, confirming the tenet that MWEs are a universal phenomenon, but it will also shed light on a topic to which little attention has been paid by Egyptian philology.



Universidad de Jaén

Funded by
the European Union

EGIPTOLOGÍA
Universidad de Jaén



You are an expert linguist helping to identify multi-word expressions (MWEs) in a large corpus. A multi-word expression is a sequence of words that form a single unit of meaning and cannot be easily deduced by the meanings of individual words. Here is the information about a potential MWE:

Candidate Phrase: {n_gram}
PMI Score: {pmi_score}
Frequency in Corpus: {frequency}

Example Sentences:

1. "{sentence_1}"
2. "{sentence_2}"
3. "{sentence_3}"

****Questions:****

1. Does the candidate phrase overlap in meaning or structure with any known MWEs? If so, which one(s)?
2. Could this phrase be considered a new variation or extension of an existing MWE? Why?
3. If it is novel, does it demonstrate properties of an MWE such as idiomaticity or collocational fixedness?
4. How likely is it that this expression is becoming a trend in social media language? Rate this likelihood on a scale from 1-5.
5. Based on the given information, would you classify this candidate as:
 - **Novel MWE**
 - **A variation of an existing MWE**
 - **Not an MWE**

Explain your decision with examples and reasoning.

Annotator Model: Gemma 9b

Annotator and Judge Model Prompts:

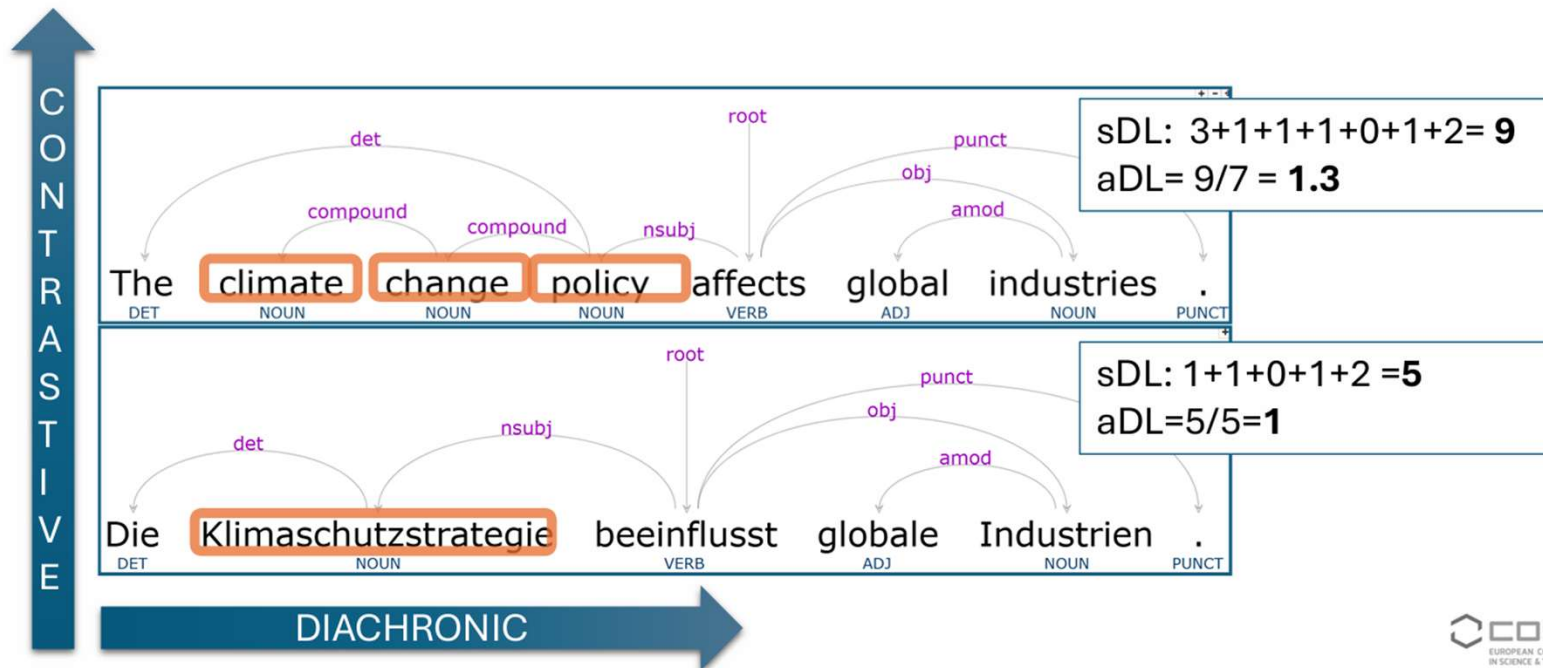
You are a judge evaluating the response of a linguist who has classified a multi-word expression (MWE) in a large corpus. Based on that, can you provide a label for the candidate phrase from the following options:
["**Novel MWE**", "**Variation of an existing MWE**", "**not an MWE**"]?
Please provide only the label without any additional information.

**Judge Model:
Qwen 2.5 72B**

Morphological differences affect dependency length: the case of compounds in English and German



WG1



Diego Alves Luigi Talamo Pauline Krielke

UniDive

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

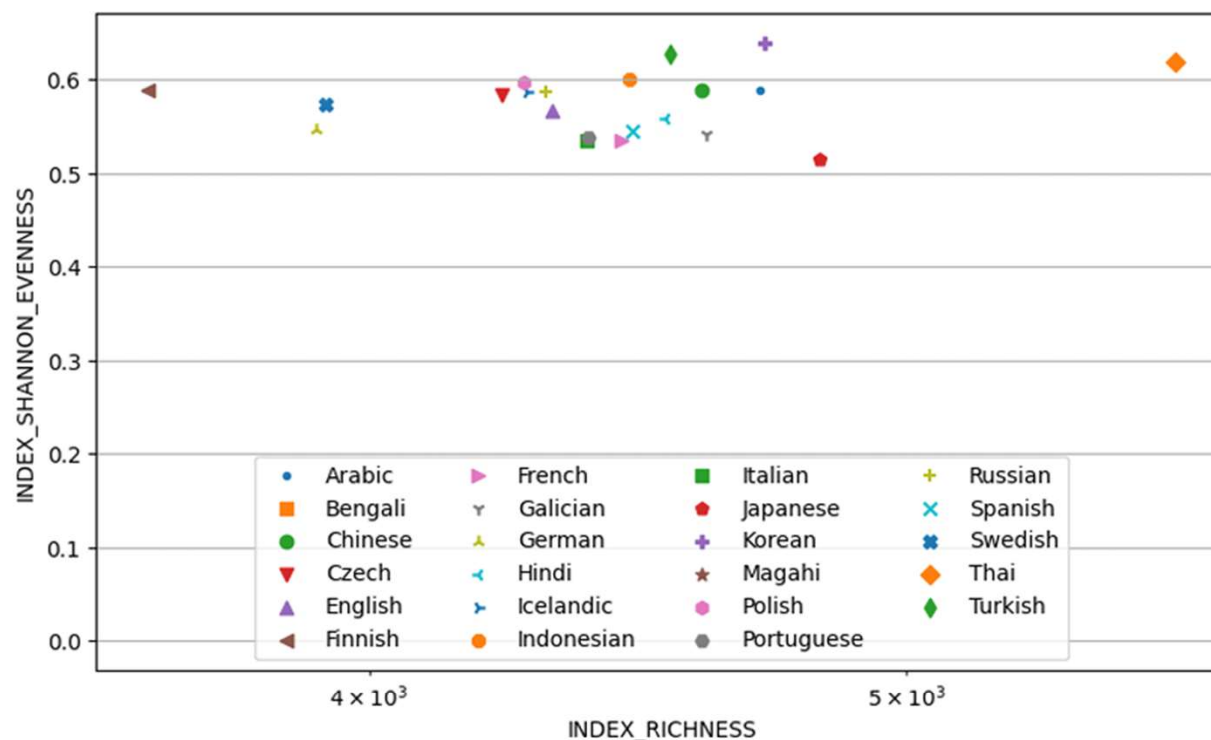
iDeal
SFB 1102

Funded by
the European Union

UNIVERSITÄT
DES
SAARLANDES

HUN REN HUNGARIAN RESEARCH
CENTRE FOR LINGUISTICS

DELTA: A new pipeline for measuring diversity across various linguistic levels



Periphrastic Verb Forms

	UPOS	VerbForm	Mood	Aspect	Tense	Voice	Number	Person	Phrase
Několik	DET								
jsem	AUX	Fin	Ind	Imp	Pres	Act	Sing	1	*
jich	PRON								
našel	VERB	Part		Perf	Past	Act	Sing		*
jsem našel	VERB	Fin	Ind	Perf	Past	Act	Sing	1	[2, 4]
Znalazť-	VERB	Fin	Ind	Perf	Past	Act	Sing		*
-em	AUX			Imp			Sing	1	*
ich	PRON								
kilka	DET								
Znalazťem	VERB	Fin	Ind	Perf	Past	Act	Sing	1	[1, 2]

_{cs} Několik jsem jich našel. / _{pl} Znalazťem ich kilka. 'I found several of them.'

Superframes: a proposal for a schema for universal semantic role annotation

Superframes: Proposal for a Universal Schema for Semantic Role Annotation

hhu, UniDive Kilian Evang, Heinrich Heine University Düsseldorf, Germany COST Funder by the European Union

Idea

- annotate argument and modifier relations with the same inventory of labels
- annotate without the need for a frame lexicon
- annotate atop UD
- frame everything in terms of coarse binary static frames ("superframes")

Inventory of "Superframes"

Frame	Arg1	Arg2
SITUATION	theme	situater
ACCOMPANIMENT	accompanier	accompanied
COMPARISON	compared	reference
CONCESSION	asserted	conceded
SAME	same	same-as
LOCATION	has location	landmark
CONTAINMENT	contained	container
CONTACT	on-surface	surface
WRAPPING/WEARING	wrapper	wearer
PROPERTY	has property	property
CLASS	has class	class
DYN	has aspect	aspect
CHG	changed	chg aspect
CONT	continues	cont aspect
DEBIT	debitated	debit aspect
HABIT	habitual	habit aspect
INT	initiated	int aspect
PREV	presented	prev aspect
EXISTENCE	exists	existence
FOCUS	has focus	focus
IDENTIFICATION	identified	identifier
MOD	has mode	mode
REC	recurs	rec mode
NEG	negated	neg mode
POSS	possible	poss mode
QUANTITY	has quantity	quantity
RANK	has rank	rank
STATE	has state	state
MESSAGE	has message	message
SEQUENCE	follows	followed
CAUSATION	result	causer
REACTION	reaction	trigger
SENDING	sent	sender
CONDITION	has condition	condition
EXCEPTION	has exception	exception
EXPERIENCE	experienced	experienced
EXPLANATION	explained	explanation
MEANS	performed	means
REPRODUCTION	reproduced	original
PART-WHOLE	part	whole
STUFF	stuff	made-of stuff
SPECIFIC-GENERIC	specific	generic
SOCIAL-ROLE	has social role	social situation
POSSESSION	possessed	possessor
TIME	has time	time

Static Predicates

Kim *owns* a house. The house *belongs* to Kim.

Dynamic Predicates

Sandy *gives* Kim a house. Kim *breaks* the vase. Kim *breaks* the vase.

Adpositional Modifiers and Adverbial Clauses

La maison de Kim. a man with a beard. Kim left after paying.

Adverbial/Adjectival Modifiers and Relative Clauses

It happened *because* a redneckery car. the house that Kim bought *because* it.

Nonlocal Dependencies

Kim *thought* to *send* Kim and Sandy *arrived* at the house.

Idioms and Light Verb Constructions

Kim *kicked* the bucket. Kim *gave* the bucket a kick with their foot.

Plot Annotation (WIP)

Text	Type	Language	Annotator 1			Annotator 2			Annotator 3				
			Sentences	Predicates	Hours	Predicates	Hours	Sentences	Predicates	Hours	Sentences	Predicates	Hours
The Little Prince	Fiction	English	95	838	102	82	64	13.1	838	64	95		
PUD	News	German	8	100	29.75	3.4	21	205	12	17.1			
PUD	News	English	8	113	29.75	3.8	21	212	12	17.7			
IMB	Fiction	German	197	2001	780	10	255	3044	200	15	198		
											2305	136	15

<https://github.com/textbooster/superframes> - Supported by HHI Strategische Forschungsbereiche

- annotate arguments and modifiers with the same inventory of relations
- annotate without the need for a frame lexicon
- annotate atop UD

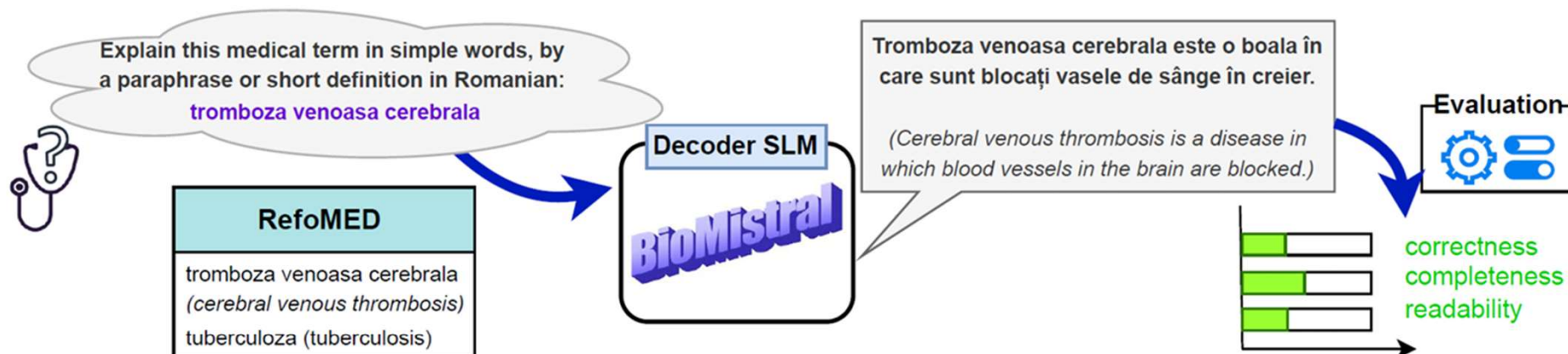
Explain this Medical Term in my Language: A Case Study of Small Language Models for Medical Paraphrase Generation



WG2, WG3

Ioana Buhnla

ATILF, CNRS - University of Lorraine, Nancy, France



- **Invented or foreign words:** binevălenie ("stări de binevălenie și de depresie"); feelings of well being and depression
- **Incorrect grammatical structures:** Gender mistakes

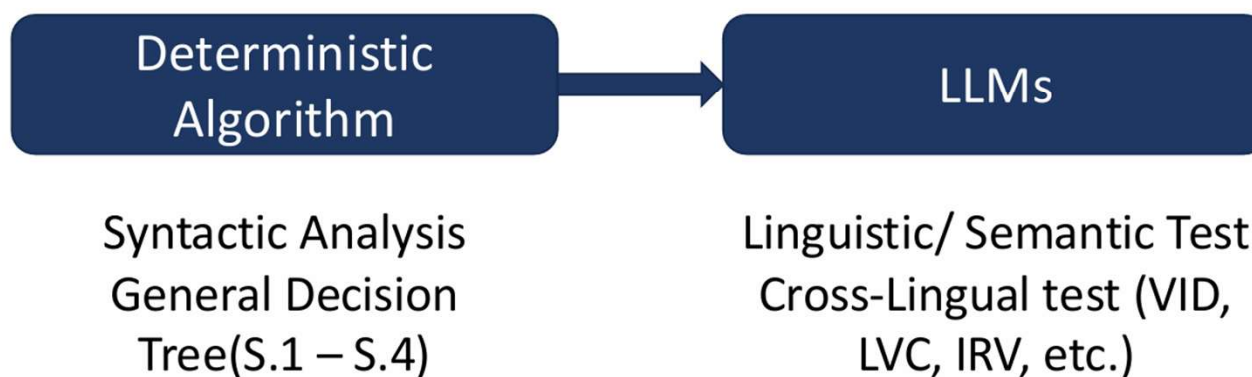
- **The SLM acts as a proofreader:** Palpitații (palpitații în română); Palpitations (palpitations in Romanian)
- **English VS Romanian prompt:** The SLM gives better answers when the medical terms **do not have** diacritical marks

Towards Implementing the PARSEME annotation guidelines with parsing and LLMs

Hanbyul Song, Agata Savary, Mathieu Constant

- ❖ **Research Goal:** Develop an interpretable approach for VMWE identification while enhancing the identification of previously unseen VMWEs.

Combination of a deterministic algorithm and LLMs

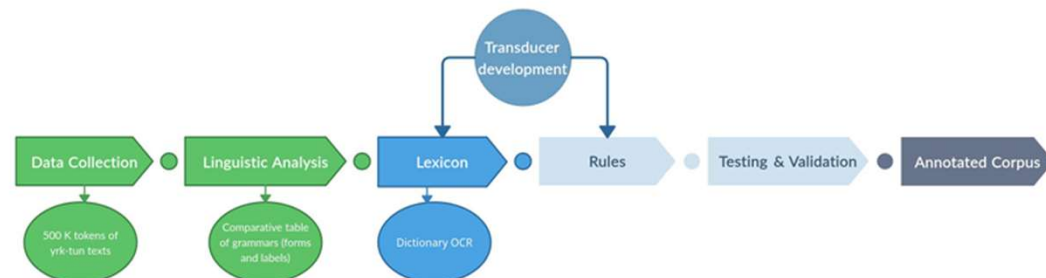


Developing a Helsinki Finite-State Transducer for Tundra Nenets

Furkan Akkurt
Boğaziçi University

Nikolett Mus
Hungarian Research Centre for Linguistics

- **Tundra Nenets:** An endangered Samoyedic language of Siberia, known for its rich morphology.
- **Goal:** Syntactically annotated corpora for linguistic and computational use.
- **Approach:** FST construction for automatic morphological annotation.
- **Why It Matters:**
 - **Facilitation of Linguistic Research:** Reveals unique linguistic features of Tundra Nenets, enriching linguistic theories.
 - **Preservation and Documentation:** Develops resources to preserve the language and support education.
 - **Digital Inclusion:** Enables tools like spell checkers and language platforms for Tundra Nenets.



CROSS-LINGUAL TRANSFER IN MULTILINGUAL LANGUAGE MODELS: IMPACT OF LINGUISTIC SIMILARITY AND PARAMETER-EFFICIENT FINE-TUNING

Fred PHILIPPY^{1,2}

¹ Zortify S.A., Luxembourg

² University of Luxembourg, Luxembourg

fred@zortify.com



WG3

UniDive

cost
COOPERATION
IN SCIENCE
AND TECHNOLOGY

Funded by
the European Union

Cross-lingual transfer is unbalanced across language pairs

We identify **5 types of factors** that contribute to cross-lingual transfer

We aim to decouple linguistic similarity and transfer performance

Source Language	bg	de	el	en	es	fr	it	nl	pl	pt	ru	sv	th	tr	ur	vi	zh	AVG
bg	71.20	69.52	69.74	67.49	75.91	72.44	71.72	61.48	69.54	50.16	52.04	62.73	59.16	70.28	69.92	66.22		
de	66.69	76.51	71.64	67.96	76.99	73.33	72.83	62.50	71.86	49.76	53.89	62.26	69.48	71.50	70.24	67.08		
el	67.23	71.22	76.83	69.06	78.94	75.39	74.31	64.27	71.02	49.34	57.19	63.95	62.50	71.46	71.94	68.29		
en	66.33	69.90	70.36	74.97	75.87	73.77	71.68	61.86	68.84	51.84	56.85	62.50	60.20	70.56	70.04	67.09		
es	65.35	69.48	71.50	66.51	62.79	75.01	73.83	60.92	69.54	50.18	54.73	61.62	58.64	70.96	69.44	66.70		
fr	66.13	71.30	72.16	69.00	79.24	78.04	74.93	62.75	71.36	50.26	54.91	63.01	60.00	72.32	71.40	67.79		
it	66.19	70.74	72.32	68.90	79.48	75.57	77.39	62.06	70.32	51.34	54.55	63.07	60.32	70.86	70.60	67.58		
nl	64.27	68.34	69.40	66.87	72.26	71.26	70.52	67.09	68.28	49.22	55.03	62.79	63.31	69.44	70.04	65.88		
pl	67.15	72.10	71.84	68.58	78.28	74.25	73.75	63.11	74.57	49.88	56.09	64.09	60.50	71.20	72.20	67.83		
pt	62.14	62.89	67.41	64.47	74.29	69.14	68.68	58.61	64.67	66.23	51.04	58.40	56.05	66.33	66.03	63.62		
ru	61.14	65.27	64.53	63.27	68.66	66.93	66.85	56.15	64.21	49.96	65.69	56.23	54.71	66.19	65.75	62.37		
sv	65.29	67.78	69.76	66.15	73.39	71.56	70.16	62.30	67.64	51.02	56.31	71.16	59.66	68.92	68.96	66.00		
th	59.50	63.83	64.33	62.24	66.10	65.11	64.41	61.58	64.99	45.01	49.26	57.84	62.65	63.79	65.67	61.22		
tr	65.49	69.46	70.40	67.49	76.61	73.53	72.42	61.96	70.06	49.76	57.41	61.74	60.22	75.13	71.92	66.90		
ur	65.45	69.30	70.38	67.21	76.79	73.03	72.65	63.29	70.74	48.54	56.29	63.07	60.74	71.28	76.15	66.96		
vi	65.23	69.18	70.15	67.35	75.83	72.56	71.74	61.86	69.24	50.83	55.40	62.30	59.88	70.01	70.02			
zh																		
AVG																		

1. Linguistic Similarity

2. Lexical Overlap

3. Model Architecture


4. Pre-Training Settings

5. Pre-Training Data

Cross-Lingual
Transfer
Performance


Language
Distance

Impact on
Representation
Space



Do you know
what the
«black hole»
of linguistics is?

Are you interested in exploring
a potential way to make the
interior of a black hole more
accessible to external
observers?



If you are
interested,
you can

[come to our poster]

The “KIPARLA Forest” treebank of spoken Italian

an overview of initial design choices

Ludovica Pannitto, Caterina Mauri

Alma Mater Studiorum - University of Bologna



UniDive



EXPERIMENTAL LAB

lilec.lab@unibo.it



Funded by
the European Union



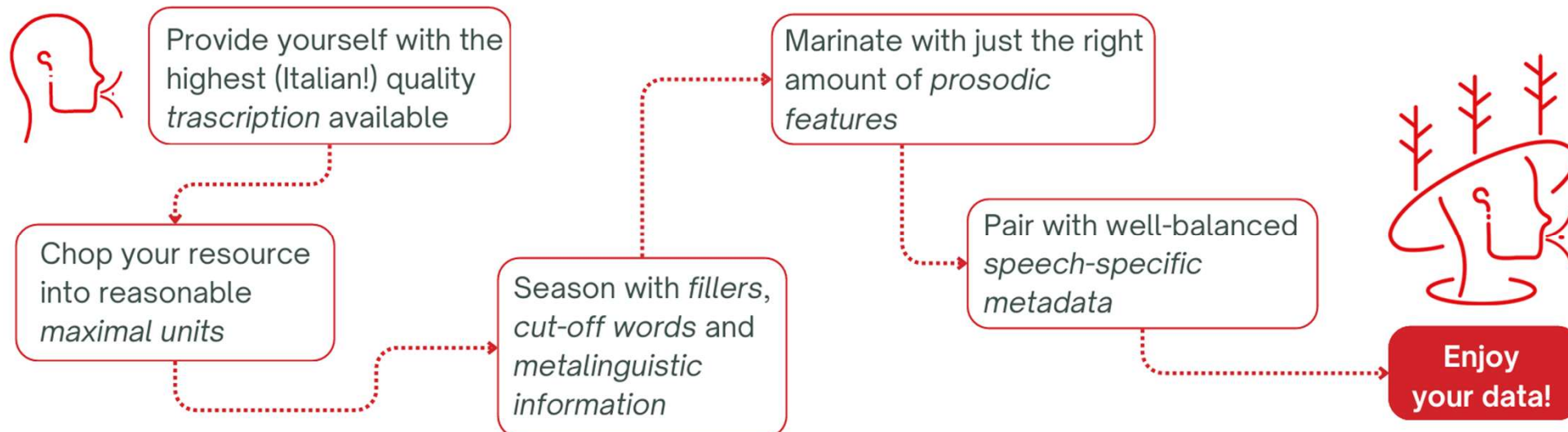
cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

WG1

WG4



How to **prepare** a treebank to observe linguistic cooperation?



Constructing a PARSEME-NL corpus: part I

Carole Tiberius, Lut Colman & Marit Janssen

UniDive WG1 (WG2)

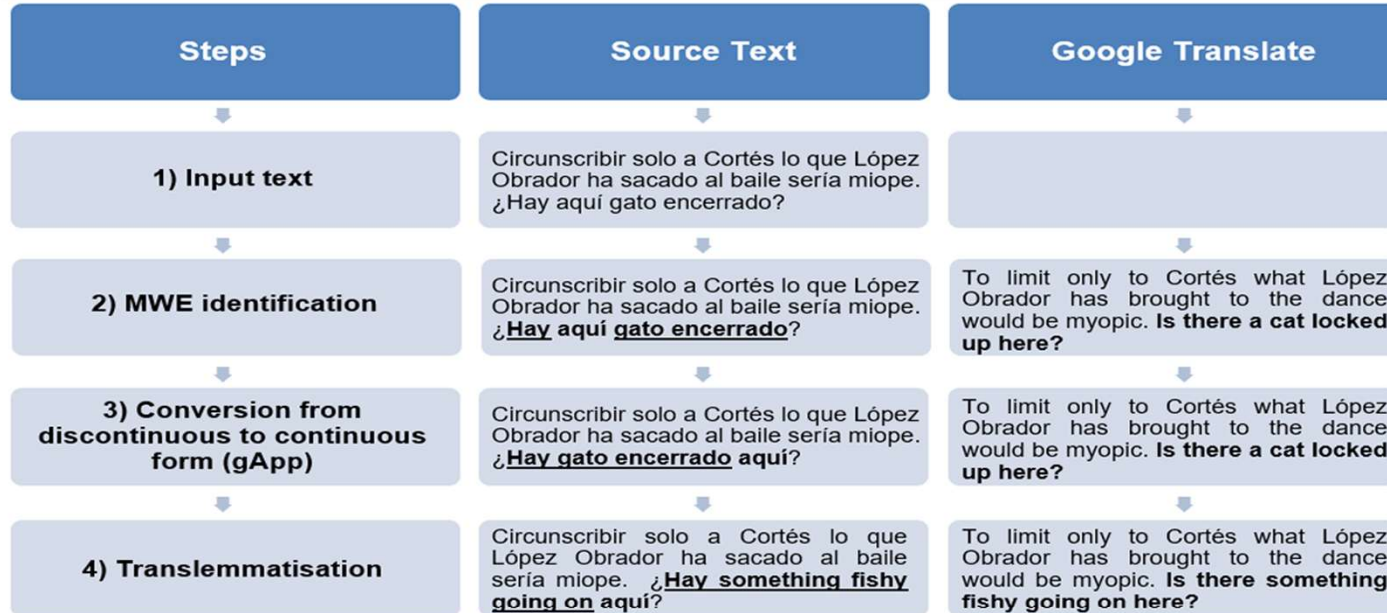
- UD_Dutch-Alpino
- VMWE categories:
VID, IRV, VPC.FULL, VPC.SEMI, LVC.FULL, LVC.CAUSE, MVC, IAV
- Challenges posed by the VPCs
 - Semantic idiosyncrasy of particle verbs in Dutch
 - Particle verbs embedded in other VMWEs
e.g. in IRV (*zich afspelen* 'to be set')
 - Particle verbs with a verbal constituent that does not occur as independent verb e.g. *bootsen* in *nabootsen* 'to imitate'

Paidiom: an MWE-preprocessing algorithm to enhance neural machine translation

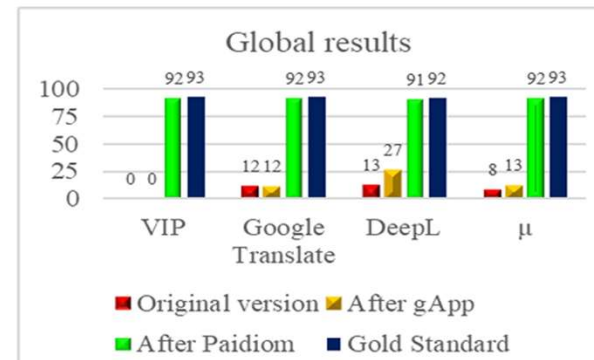
Carlos Manuel Hidalgo-Ternero

University of Malaga, IUITLM (Spain)

How can Paidiom improve neural machine translation?



Experiment
Number of cases: 400
NMT systems: DeepL, Google Translate & VIP
Translation directionality: ES>EN
Idioms: haber gato encerrado ('there is sth fishy going on') ser cuatro gatos ('there is just a bunch of people') dormir la mona ('to sleep off [a hangover]') ganar/costar/pagar cuatro perras ('to earn/cost/pay peanuts')



Error Types in the Latvian Grammatical Error Correction and Fluency Corpus "Norma"

Baiba Saulīte, Roberts Dargis, Kristīne Pokratiece

norma.korpuss.lv

- error-annotated corpus of texts produced by native speakers of Latvian
- documents the most common errors
- valuable for practical use, linguistic research and language technology

Orģināls*

Parasti tie ir ņurnālisti, ar kuru vieglu roku vai muti jaunās reālijas apzīmējums nonāk masu informācijas līdzekļos.

Labotais*

Parasti tie ir ņurnālisti, kas nedomājot jaunās reālijas apzīmējumus lieto plašsaziņas līdzekļos.

Sastatījums

Parasti tie ir ņurnālisti, ar kuru vieglu roku vai muti kas nedomājot jaunās reālijas apzīmējumus nonāk masu informācijā lieto plašsaziņas līdzekļos.

Kļūdas

6.5. Izteicēja izveide x 7.3. Neiederģgs vārds x 10.1. Sekundāra: saistāmība x

Komentārs

☒ Pārbaudģts ☐ Nederģgs

Main Error Category	Sentences	Percentage
7 Use of lexical or structural units	1720	28%
1 Typographical formatting	1140	19%
6 Syntax	1061	17%
5 Punctuation	650	11%
2 Spelling	497	8%
10 Secondary errors	401	7%
4 Word formation	178	3%
8 Text structure	173	3%
3 Derivation	155	3%
9 Other errors	137	2%



Middle Polish Dependency Treebank and its conversion to the UD format (WG1)

Aleksandra Wieczorek

Institute of Polish Language PAS

Alina Wróblewska

Institute of Computer Science PAS

17th-18th c. Polish:

- long, multi-complex sentences
- atypical word order
- discontinuous dependency relations
- frequent Latin insertions
- some grammatical differences

Possible uses of MPDT:

- parsing of the historical Polish corpora
- historical syntax analysis

MPDT:

- 2,000 sentences → 6,000 in 2026
- consistent with Polish Dependency Bank
- conversion to the UD in progress

[Kiedym już poselstwo odprawił,] [którego dosyć *placide et attente* słuchali,] [*ante vota senatorum* odprowadzono mnie do gospody,] [chocia mam sprawę,] [że przedtym przy poślech K. J. M. wotowali] i [nie chciało mi się wynieść], ale [iż to już tak przez kilka sejmików bywało,] [jm. ks. biskup nie dał mi się przykryć] i [ustąpić radził].

Interfacing Wiktionary with Linguse for automatic MWE identification

A case study of language learning

**Automatic
MWE
identification**

Make MWEs discoverable

It can be hard to even know it's there.



**Empirical
validation with
language students**

Provide assistance

Learners can understand the context.

Till Überrück-Fries, Agata Savary, Agnieszka Dryjańska