Challenges in corpus annotation of copulative perception verbs

Alon Fishman
The Open University of Israel

UniDive

Goal: Find (cross-linguistic) patterns of form-function mappings in a class of verbs

Forms:

(i) looks weird.

(ii) wygląda dziwnie. (Polish)
looks weirdly

(iii) nir'et muzar / -a. (Hebrew)
looks.F weirdly / weird.F

Functions:

(a) X has a weird look.

(b) X's look suggests weirdness.

Hebrew 'look'	Adverb	Adjective
'wonderful'	512	23
'excellent'	658	53
'big'	5	392
'simple'	4	481
Russian 'look'	**Adverb**	**Adjective**
'wonderful'	1200	14
'excellent'	1260	19
'elastic'	5	118
'obvious'	9	241

Annotation challenges:

- Multiple polysemies
- Intra- & cross-linguistic variability
- Annotation variability

Is your language missing?

звучать, näyttää, brzmieć,

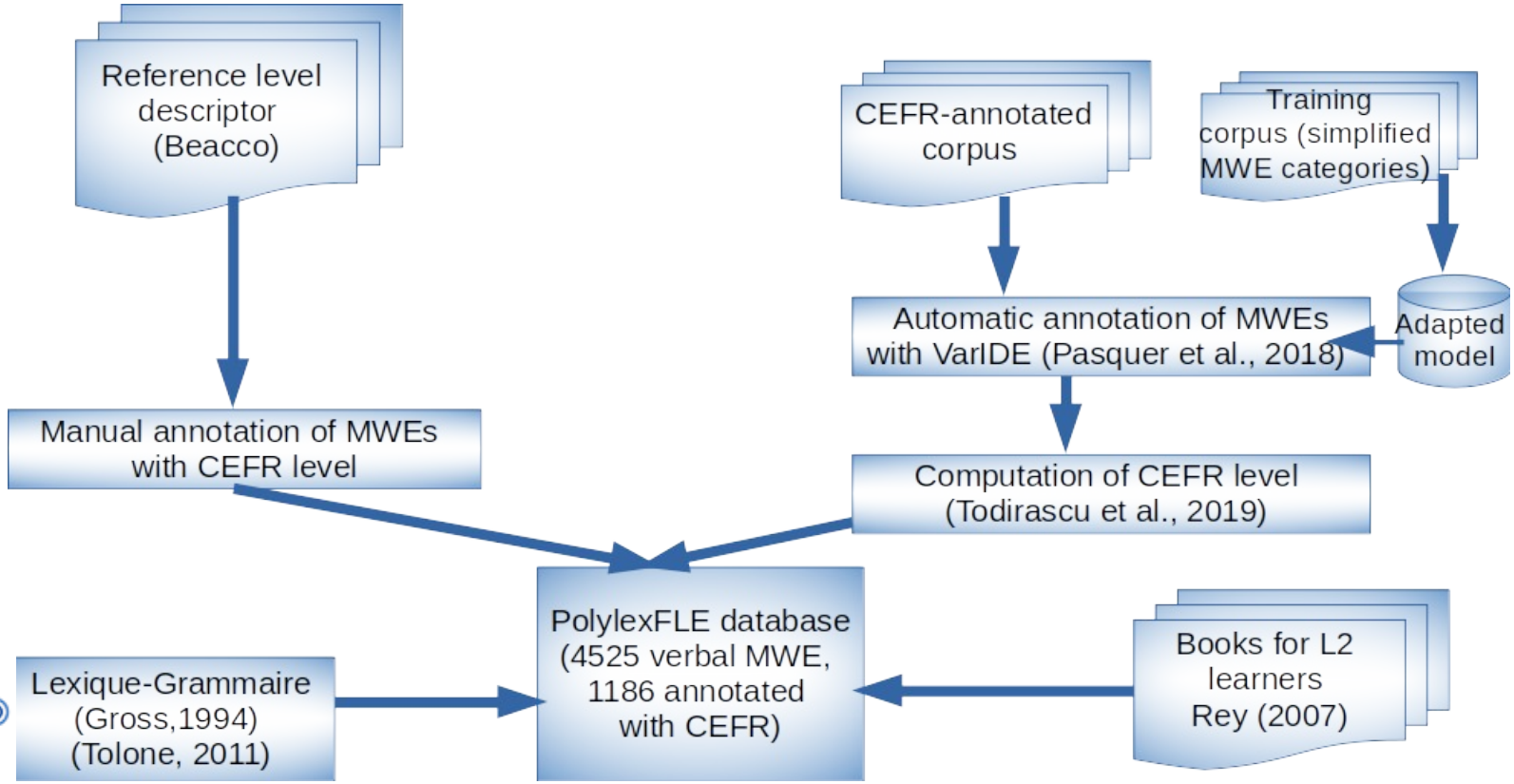
sonar, klingen, avoir l'air

Annotating French MWEs for French L2 learning

Amalia Todirascu, University of Strasbourg
(todiras@unistra.fr) (WG1)

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
UniDive
Funded by the European Union
lilpa linguistique, langues, parole

- **Aims : create a CEFR-level corpus and build a MWEs lexical database for French L2 learners, annotated with CEFR level (related work with ANR project STAR-FLE)**
- **MWEs are difficult for L2 language learners (Bahn and Eldaw, 1993), (Siyanova, 2017)**
- **Simplified typology of verbal MWE (Todirascu et al, 2019): idioms, collocations, fixed expressions**



anr

Features of Annotation of Verb Complements in Different Stages and Varieties of Armenian

(Research project idea)

Anna Danielyan, Marat Yavrumyan

Aims of the Project

- To develop a combined annotation system and a tagset of linguistic universals for verb complements for different historical stages and varieties of Armenian, utilizing the NLP-applicable universality of terminologies and methodologies (typologically oriented grammatical theories and dependency grammar).
- To promote the improvement of digital sets of tags and relations developed for Armenian.

What do we have?
2 corpora for Eastern Armenian and 1 for Western Armenian in UD (also, a corpus of Classical Armenian developed lately by another team)
More than 40 verb complements in traditional grammar VS 10 relations in UD corpora for Armenian

Challenges

- Inconsistencies and gaps in traditional approaches concerning core-noncore distinction and voice determination, that need to be clarified in order to obtain high-quality corpus annotation for NLP purposes.

Issues and solutions
Accusative-like dependents for middle and passive verbs, determination of dependency relations of ditransitive causative verbs. The issue of causative voice.
Some solutions found due to reconsidering some aspects in tradition and the inventory of syntactic relations in UD while working on the corpus of Western Armenian developed later than the corpus of Eastern Armenian.

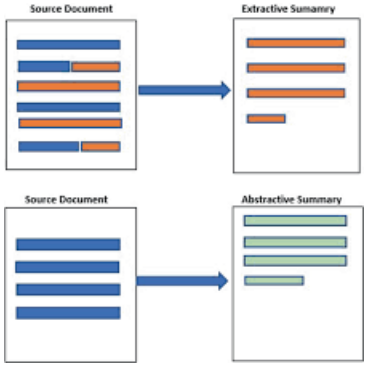
Perspectives

- To provide a suitable basis for bringing out inter-linguistic parallelism across varieties of Armenian compiling a general set of tags and relations for verb complements. This will be a step towards further bigger projects for creating diachronic and/or parallel corpora for Armenian.

Abstractive Text Summarization Datasets, Models, and Tokenization Approaches for Turkish and Hungarian

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

- Text summarization
  - Extractive text summarization
  - Abstractive text summarization

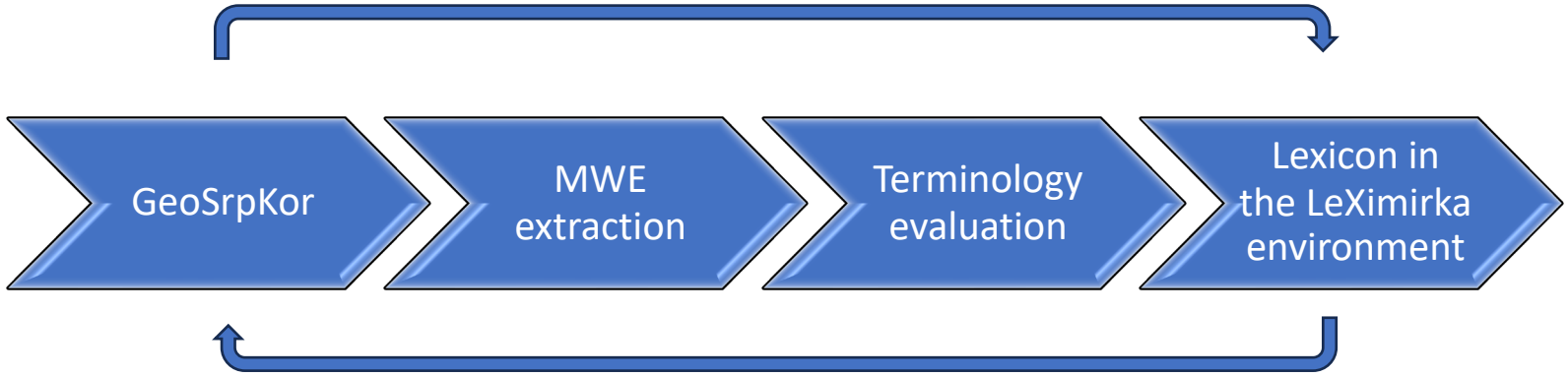


- This work is related to abstractive text summarization

- Contributions:
  - Two large-scale publicly available summarization datasets for Turkish and Hungarian
  - Strong baselines for both datasets
  - Comparing pointer-generator model (commonly-used baseline model for summarization) with BERT-based models
  - Two morphological tokenization methods
    - SeparateSuffix
    - CombinedSuffix

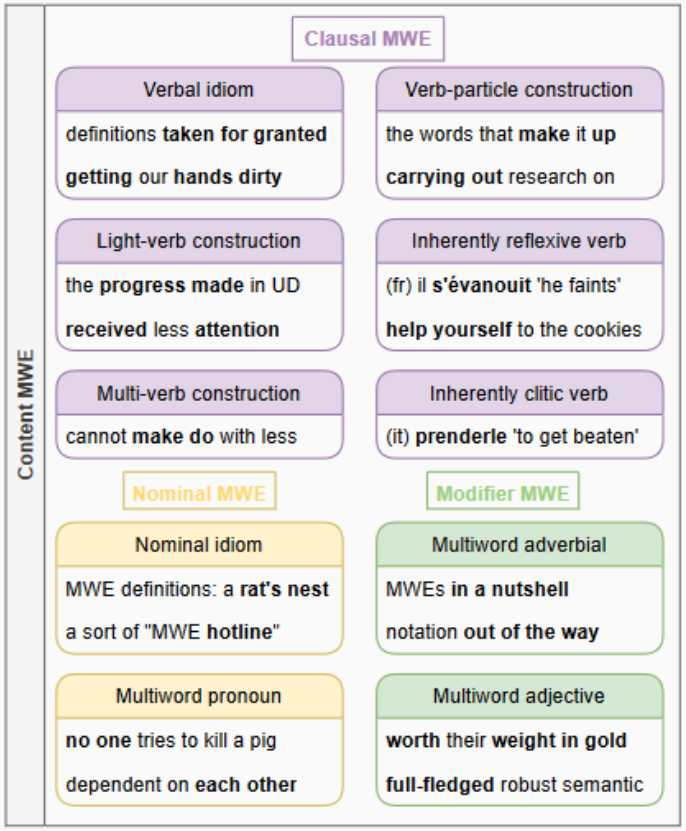
Bridging the Geological Lexicon and Corpus with Focus on MWEs Extraction

Biljana Rujević, Cvetana Krstev, Mihailo Škorić



A taxonomy proposal for multiword expressions

Carlos Ramisch – Aix Marseille Univ, France



- Nominal idioms vs. compounds
- Named entities and terms
- Adjectival vs. adverbial MWEs
- Selected prepositions in adverbials (***in adition*** *to*) vs. adpositions (***in spite*** *of*) vs. determiners ***a lot*** *of apples*
- . . .

Feedback wanted

Come share challenging examples in your languages!

UD Syntax for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Pilot Study

Carole Tiberius[1], Jaka Čibej[2], Jelena Kallas[3], Kertu Saul[3], Kadri Muischnek[4], Simon Krek[5]

[1]Instituut voor de Nederlandse Taal, The Netherlands,
[2]Faculty of Arts, University of Ljubljana, Slovenia,
[3]Institute of the Estonian Language, Estonia,
[4] University of Tartu, Estonia,
[5]Jožef Stefan Institute, Slovenia

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

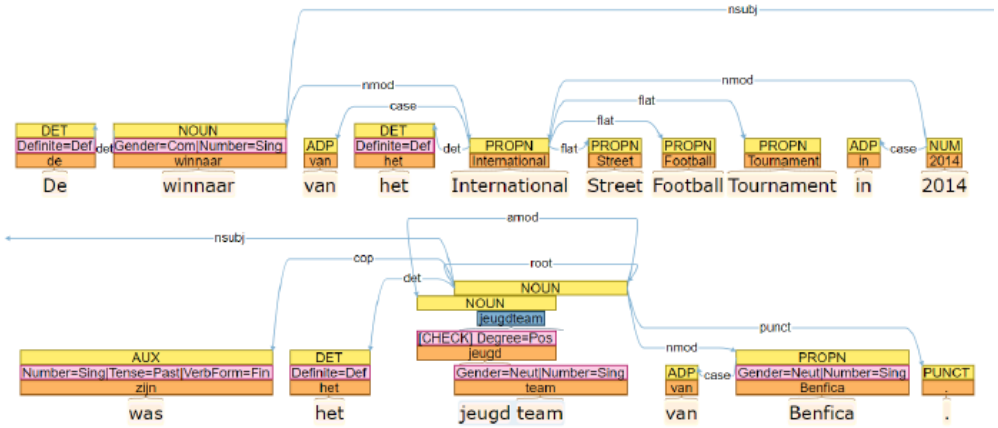
Manually-curated lexical-semantic resource combining corpora and sense inventories

Relevance: **WG1** and **WG2**

Extension of dataset within UniDive with:

- new languages and
- new annotation layers:
  - annotation of multiword expressions following the PARSEME annotation guidelines
  - annotation of named entities
  - syntactic parse structure following Universal Dependencies

Manual validation of UD parse: Dutch and Estonian



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

UniDive

Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora

WG2

- Interlinking MWE lexicon entries with their occurrences in corpora
- publishing of aligned and annotated corpus as LD employing NIF

Otolex-lemon
Decomp
Vratrans
FRaC

ELEXIS-WSD Parallel Sense-Annotated Corpus

```
<http://url> a nif:ContextCollection ;
    nif:hasContext <http://url/enwsd> .

<http://url/enwsd> a nif:Context,
        nif:OffsetBasedString ;
    nif:beginIndex "0" ;
    nif:endIndex "49" ;
    nif:isString "He is named after
    the astronomer Galileo Galilei." .

<http://url/enwsd#offset_0_49_0> a
nif:OffsetBasedString, nif:Phrase ;
nif:anchorOf "Galileo Galilei." ;
nif:beginIndex "33";
nif:endIndex "49" ;
nif:referenceContext <http://url/enwsd> ;
nif:taMsClassRef/itsrdf:taIdentRef wd:Q307;
itsrdf:taClassRef dbo:Person,wd:Q5,
    <http://nerd.eurecom.fr/ontology#Person>.
```

```
:le_blood_pressure a ontolex:LexicalEntry,
                     ontolex:MultiwordExpression;
  ontolex:canonicalForm [ontolex:writtenRep "blood pressure"@en];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:sense [ontolex:reference
       <https://dbpedia.org/page/Blood_pressure>];
  decomp:constituent :cm_blood;
  decomp:constituent :cm_pressure;
  rdf:_1 :le_blood; # lexical
  rdf:_2 :le_pressure.  # entries
```

Nexus Linguarum

ranka@rgf.rs
christian.chiarcos@uni-a.de
milica.ikonic.nesic@fil.bg.ac.rs

BACKGROUND 1 — High cost + time requirement

OBJECTIVE 2 — Generate idiomatic instances with LLMs

METHOD 3 — Prompt GPT4 for context, then generate samples

WG1, WG3

IDIOM CORPORA CONSTRUCTION VIA LARGE LANGUAGE MODELS

ITU NLP — NATURAL LANGUAGE PROCESSING RESEARCH GROUP

EVALUATION 4 — Sequence labelling task on idiom detection

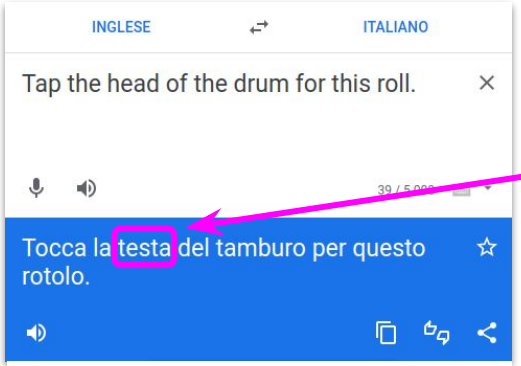
RESULTS 5 — Not in human-level yet promising

FUTURE WORK 6 — Additional languages and prompting techniques

Advances in Natural Language Processing: Bridging Text and Knowledge via Grounding and Innovative Applications

Edoardo Barba · *UniDive - 2nd General Meeting | WG-{1,3,4}*

1



Lexical Ambiguity

Machine Translation

2

I like singing and dancing

Noisification

I like *cantare* and *bailar*

1) Reconstruction

2) Translation

I like singing and dancing

J'aime chanter et danser

Transfer learning via noisification and reconstruction of the original input sentence / translation in a target language

3

I like singing and dancing

Word Sense Disambiguation

BabelNet

ballare danzare bailar danser

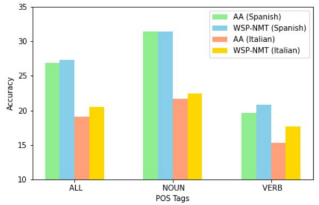
cantare cantar chanter

Dataset Creation:

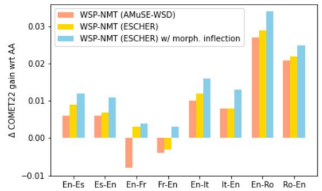
Word Sense Disambiguation + BabelNet

4

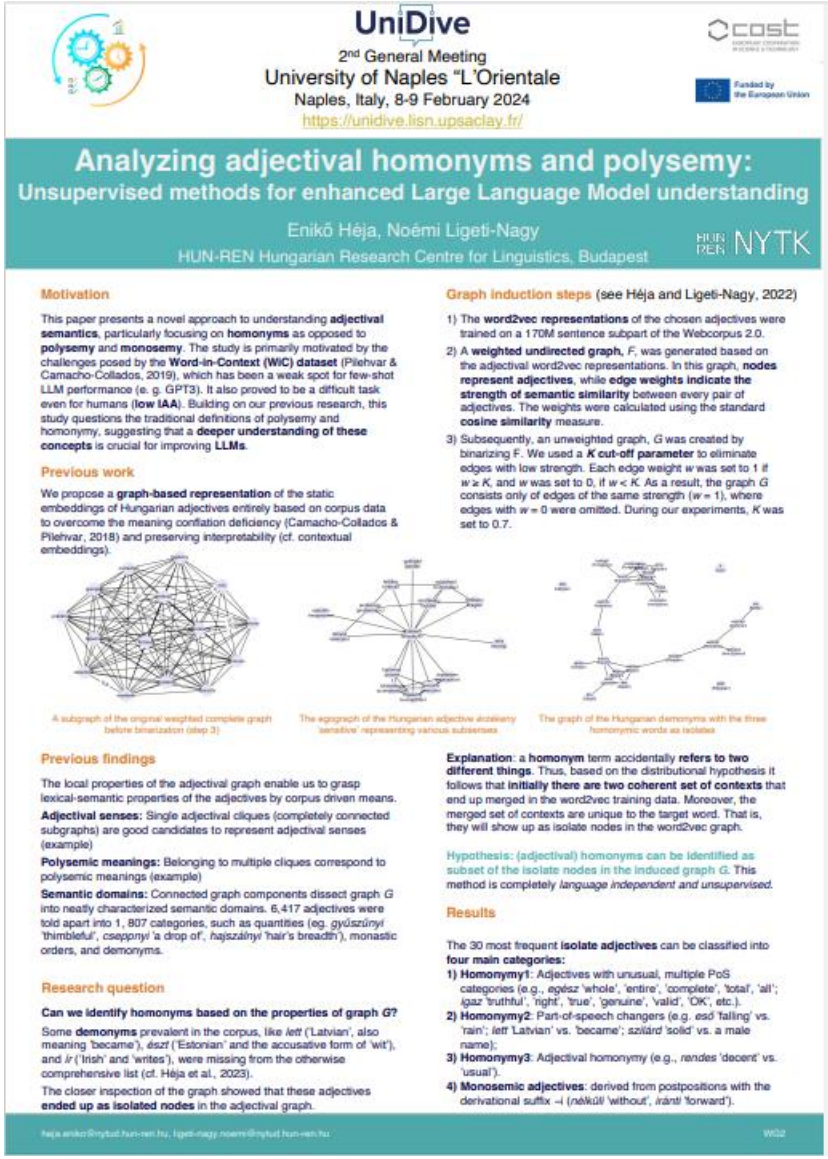
After standard MT training



Improvements on the disambiguation capabilities



Improvements in Machine Translation



UniDive
2nd General Meeting
University of Naples "L'Orientale
Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

Analyzing adjectival homonyms and polysemy:
Unsupervised methods for enhanced Large Language Model understanding

Enikő Héja, Noémi Ligeti-Nagy
HUN-REN Hungarian Research Centre for Linguistics, Budapest

Motivation

This paper presents a novel approach to understanding **adjectival semantics**, particularly focusing on **homonyms** as opposed to **polysemy** and **monosemy**. The study is primarily motivated by the challenges posed by the **Word-in-Context (WiC) dataset** (Pilehvar & Camacho-Collados, 2019), which has been a weak spot for few-shot LLM performance (e. g. GPT3). It also proved to be a difficult task even for humans (**low IAA**). Building on our previous research, this study questions the traditional definitions of polysemy and homonymy, suggesting that a **deeper understanding of these concepts** is crucial for improving **LLMs**.

Previous work

We propose a **graph-based representation** of the static embeddings of Hungarian adjectives entirely based on corpus data to overcome the meaning conflation deficiency (Camacho-Collados & Pilehvar, 2018) and preserving interpretability (cf. contextual embeddings).

Graph induction steps (see Héja and Ligeti-Nagy, 2022)

1) The **word2vec representations** of the chosen adjectives were trained on a 170M sentence subpart of the Webcorpus 2.0.
2) A **weighted undirected graph,** *F*, was generated based on the adjectival word2vec representations. In this graph, **nodes represent adjectives**, while **edge weights indicate the strength of semantic similarity** between every pair of adjectives. The weights were calculated using the standard **cosine similarity** measure.
3) Subsequently, an unweighted graph, *G* was created by binarizing F. We used a ***K* cut-off parameter** to eliminate edges with low strength. Each edge weight *w* was set to 1 if $w \geq K$, and *w* was set to 0, if $w < K$. As a result, the graph *G* consists only of edges of the same strength ($w = 1$), where edges with $w = 0$ were omitted. During our experiments, *K* was set to 0.7.

A subgraph of the original weighted complete graph before binarization (step 3)

The egograph of the Hungarian adjective *érzékeny* 'sensitive' representing various subsenses

The graph of the Hungarian demonyms with the three homonymic words as isolates

Previous findings

The local properties of the adjectival graph enable us to grasp lexical-semantic properties of the adjectives by corpus driven means.

Adjectival senses: Single adjectival cliques (completely connected subgraphs) are good candidates to represent adjectival senses (example)

Polysemic meanings: Belonging to multiple cliques correspond to polysemic meanings (example)

Semantic domains: Connected graph components dissect graph *G* into neatly characterized semantic domains. 6,417 adjectives were told apart into 1, 807 categories, such as quantities (eg. *gyűszűnyi* 'thimbleful', *cseppnyi* 'a drop of', *hajszálnyi* 'hair's breadth'), monastic orders, and demonyms.

Research question

Can we identify homonyms based on the properties of graph *G*?

Some **demonyms** prevalent in the corpus, like *lett* ('Latvian', also meaning 'became'), *észt* ('Estonian' and the accusative form of 'wit'), and *ír* ('Irish' and 'writes'), were missing from the otherwise comprehensive list (cf. Héja et al., 2023).

The closer inspection of the graph showed that these adjectives **ended up as isolated nodes** in the adjectival graph.

Explanation: a **homonym** term accidentally **refers to two different things**. Thus, based on the distributional hypothesis it follows that **initially there are two coherent set of contexts** that end up merged in the word2vec training data. Moreover, the merged set of contexts are unique to the target word. That is, they will show up as isolate nodes in the word2vec graph.

Hypothesis: (adjectival) homonyms can be identified as subset of the isolate nodes in the induced graph *G*. **This method is completely *language independent* and *unsupervised*.**

Results

The 30 most frequent **isolate adjectives** can be classified into **four main categories:**

1) **Homonymy1**: Adjectives with unusual, multiple PoS categories (e.g., *egész* 'whole', 'entire', 'complete', 'total', 'all'; *igaz* 'truthful', 'right', 'true', 'genuine', 'valid', 'OK', etc.).
2) **Homonymy2**: Part-of-speech changers (e.g. *esd* 'falling' vs. 'rain'; *lett* 'Latvian' vs. 'became'; *szilárd* 'solid' vs. a male name);
3) **Homonymy3**: Adjectival homonymy (e.g., *rendes* 'decent' vs. 'usual').
4) **Monosemic adjectives**: derived from postpositions with the derivational suffix *-i* (*nélküli* 'without', *iránti* 'forward').

heja.eniko@nytud.hun-ren.hu, ligeti-nagy.noemi@nytud.hun-ren.hu

Analyzing adjectival homonyms and polysemy

Enikő Héja – Noémi Ligeti-Nagy

HUN REN HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

- **Unsupervised graph-based** method
- **Adjectival** semantics: homonymy (previous study: polysemy)
- **Homonyms** appear as isolate nodes in the graph; 4 distinct categories
- **Language-independent**

WG3

Probing Coreference Models of Romance Languages using Multiword Expressions

Evelin Amorim
evelin.f.amorim@inesctec.pt

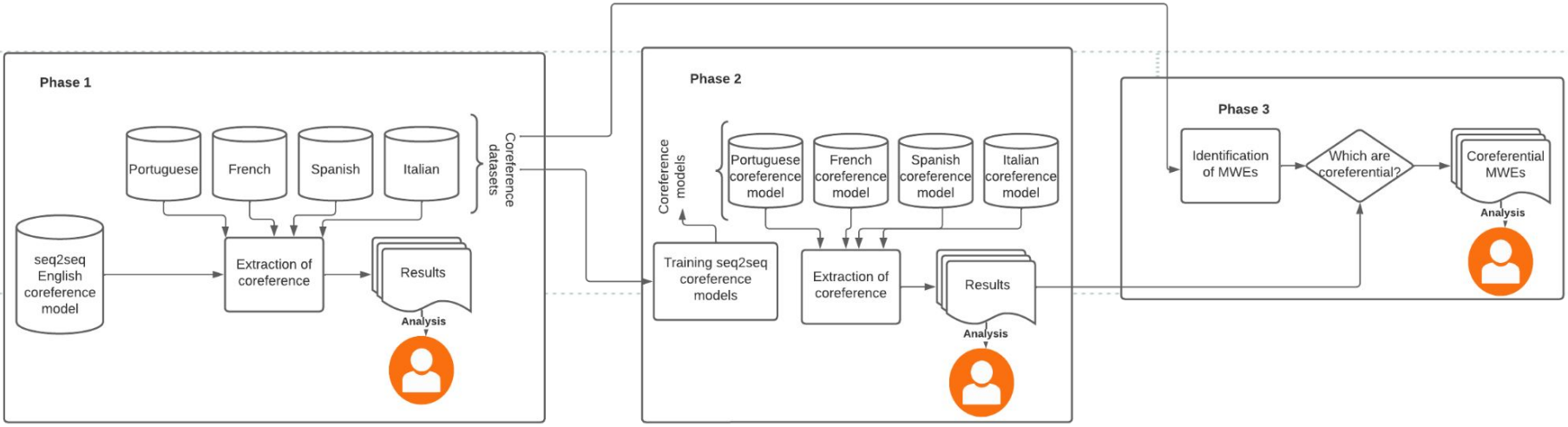
- Coreference: *Who* is the entity mentioned by some expression?

  O usuários antigos eram os que mais reclamavam.

  Old users were the ones who complained the most.

Glossary: Os usuários mais antigos/Old Users, eram/were, os/the, que/who, mais/the most, reclamavam/complained

- How does MWE influence the current SOTA English coreference model and how does this seq2seq model behave in Romance Languages?



- Outcomes:Understanding of the influence of MWEs in coreference models and improvement of coreference models for Romance languages

Contacts:
evelin.f.amorim@inesctec.pt

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

2nd General Meeting
University of Naples "L'Orientale"
Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

Multilingual semi-automated identification and annotation of multiword expressions

Ilan Kernerman
Lexicala by K Dictionaries

WG2 (WG3)

AIM

Develop tools and datasets for the automatic identification of MWEs and their annotation and integration in multilingual language settings.

STEPS

(1) Extract MWEs from lexicographic resources and LLMs
(2) Apply cross-lingual embeddings and list the top 80 most frequent MWEs
(3) Auto-extract candidate and manually annotate them
(4) Apply deep learning techniques to discover/detect and tag unseen MWEs
(5) Finetune different LLMs on the training set in few-shot transfer for the languages covered and produce generally useful MWE detectors

KEYWORDS

MWE identification
MWE annotation
machine learning
models
lexicography
multilingual LLMs
zero/few-shot transfer

OUTCOMES

✓A framework for MWE discovery and annotation
✓Trained models for MWE identification
✓UD-based annotated datasets for MWE identification

Thank you to Prof. Šikonja University of Ljubljana - FRIMarko Robnik

VMWEs in the Croatian verb valency database

Ivana Brač, Lobel Filipić, Maja Matijević, Siniša Runjaić
Institute for the Croatian Language

Relevant to WG2, WG1

- reflexive verbs, verbal idioms, and light verb constructions in Croatian general language online dictionaries (*Croatian Language Portal*; *Croatian Web Dictionary – Mrežnik*) and online valency lexicons (*CROVALLEX*; *e-Glava*)
- **aim:** description of VMWEs in the verb valency database *Verbion – Semtactic* (*Semantic-Syntactic Classification of Croatian Verbs* (HRZZ, IP-2022-10-8074))
- **questions:**
  - non-inherently reflexive verb – lemma, sublemma, sense?
  - LVCs – separate verb sense, semantic roles, etc.
  - …

WG1

UniDive

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

2nd General Meeting — University of Naples L'Orientale — 8-9 February 2024

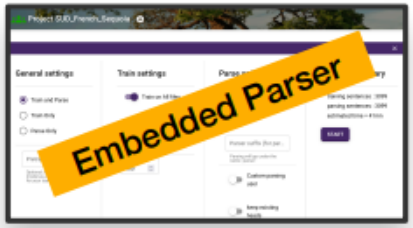
https://unidive.lisn.upsaclay.fr/

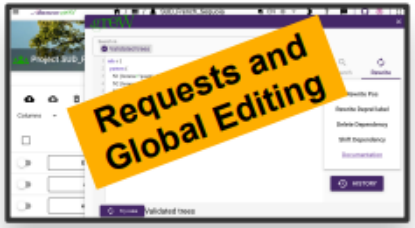
ArboratorGrew: Collaborative Curation for Treebank Manipulation

Khensa Daoudi, Kim Gerdes, Gaël Guibon, Bruno Guillaume, Kirian Guiller

https://arborator.github.io/

ARBORATOR.grew





Autogramm

Loria

Inria

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

modyco

ANR

ANR-21-CE38-0017

Word Segmentation in Universal Dependencies

Free Morphs

- en nice (property)
- en work (action or object)
- en now (property)
- cs pes ‘dog’ (object)
- en ouch (not a root)

Clitics

- en **the** book
- cs Smál **se**. ‘He laughed.’
- es **de la** escuel-a ‘of the school’;
- de **auf der** Brücke ‘on the bridge’

Roots (+Affixes)

- cs plyn ‘gas’ (free root)
- tr ev-ler ‘house-Plur’ (free root with affix)
- it alber-o ‘tree’ (bound root with required affix)
- en re-place-ment (non-required affixes)
- cs Josef-ov-ým ‘Josef-Poss-Ins’ (opt ⇒ req affix)

Compounds (+Affixes)

- en flower-pot
- de Auto-bahn ‘highway’
- cs straš-pytel ‘scaredy-cat’
- el γεω-γραφ-ία / geô-graf-ía ‘geography’ (affix)

Non-Haspelmath Words in UD

- de Liebe-s-brief ‘love letter’
- cs ruk-o-pis ‘manuscript’
- de am ⇒ an dem ‘at the’, im ⇒ in dem ‘in the’
- fr au ⇒ à le ‘to the’

Kilian Evang, Daniel Zeman
evang@hhu.de, zeman@ufal.mff.cuni.cz
Heinrich Heine University Düsseldorf, Charles University

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

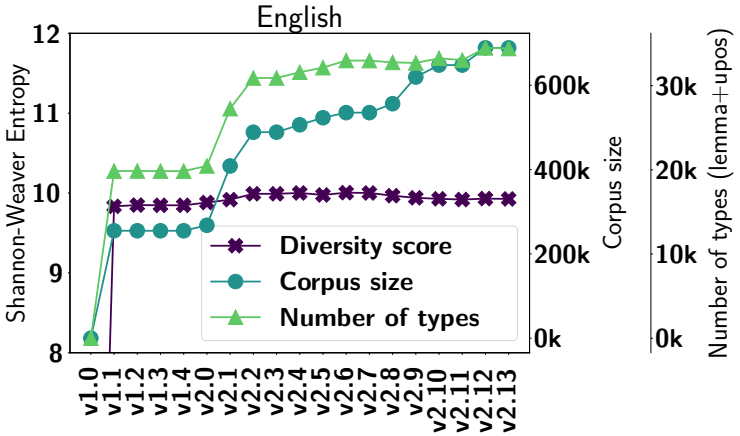
Entropy Behaviour upon Dataset Size Update

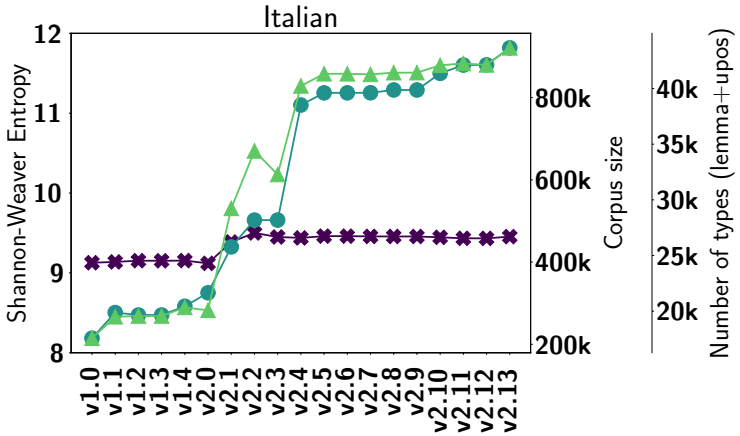
Louis Estève, Agata Savary, Thomas Lavergne

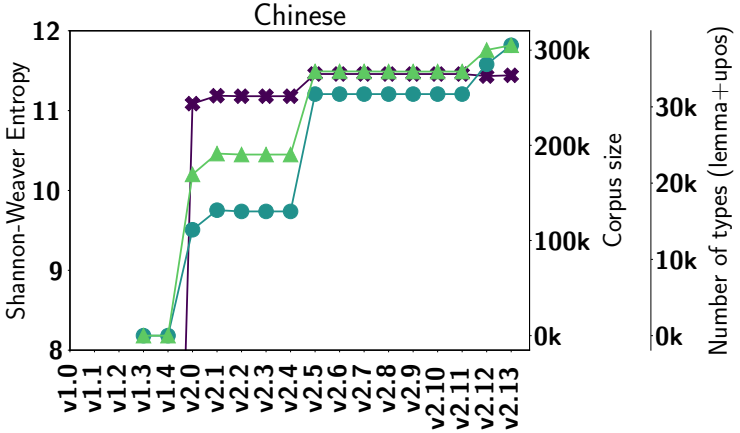
Thursday 8th February 2024

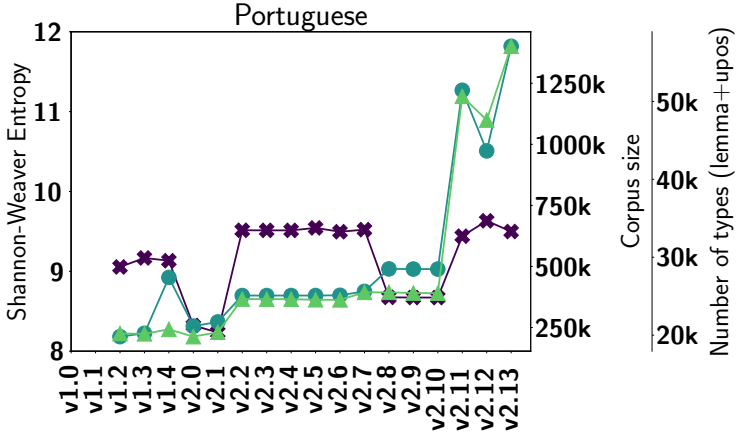
Laboratoire Interdisciplinaire des Sciences du Numérique – LISN, CNRS
firstname.lastname@lisn.upsaclay.fr

Entropy Behaviour upon Dataset Size Update



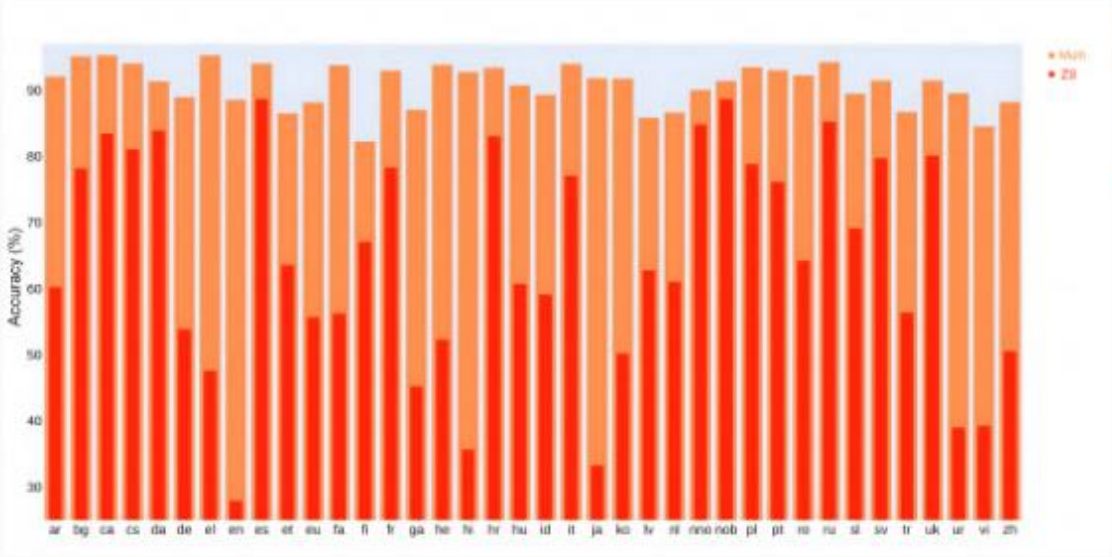






Variability Across Languages in Zero-Shot Multilingual Learning

Manon Scholivet



Lots of **variability** in zero-shot learning. Standard Deviation :

- **3.23** in the *Multi* experiments
- **17.06** in the zero-shot setting

*Is the presence of a **close** language to the target language among the languages in the training set important ?*

→ But what is a "close" language ?

Let's find out !

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

UniDive

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

WG1

Creation Dataset of Token Language Identification for Ukrainian-Russian Code-switching Corpus

Olha Kanishcheva[1,2], Maria Shvedova[1,3]

[1]University of Jena, [2] SET University, [3] National Technical University "KhPI

Our results:

1) A database of about 150,000 tokens containing code-switching between Ukrainian and Russian has been collected.
2) This dataset contains intra-word code-mixing, so-called Surzhik. The dataset is divided at the token level into 5 categories. The next step will be to analyze the obtained dataset and test different classification models on this data.
3) We analyzed the different types of code-switching that occur in our dataset.
4) Some metrics of code-switching have been calculated to show the complexity of the data.

Labels	Description	Tokens
UK	Ukrainian words	93 040
RU	Russian words	30 956
MIX	Ukrainian-Russian hybridized words (Surzhyk)	225
Others	Dialects, other languages, etc.	615
Punct	Punctuation	30 695

On the Inter-Linguistic Disparity of Knowledge Graphs:

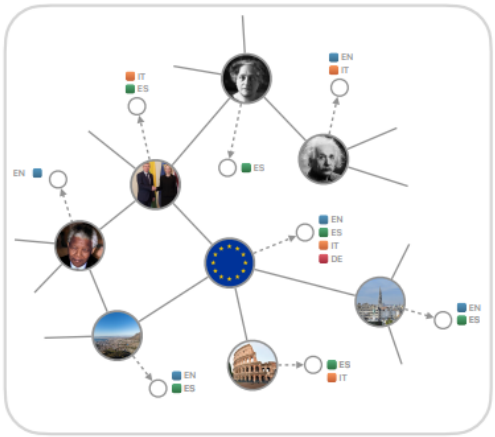
Bridging the Gap between English and Non-English Languages

Simone Conia · *UniDive - 2nd General Meeting | WG-{1,3,4}*

Knowledge graphs (KGs) encode our collective understanding of the world in a structured representation.

Before

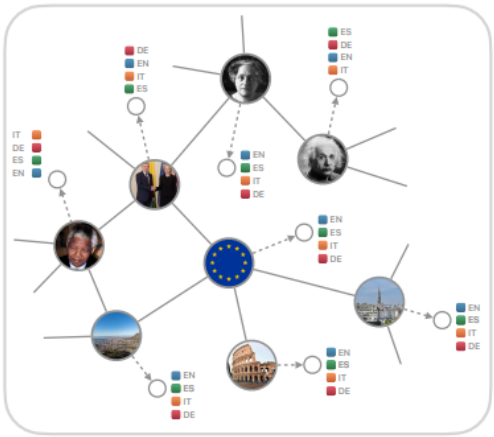
Wikidata



How close can we get to **language parity**?

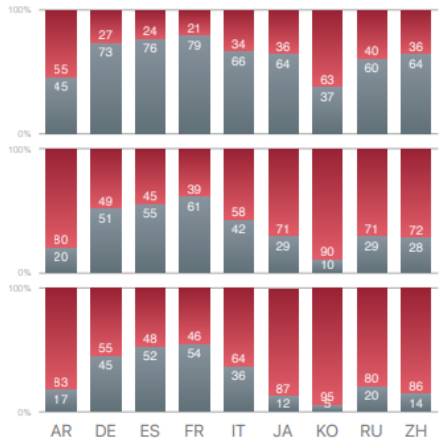
After

Multilingually-enriched Wikidata



Problem: non-English coverage is low.

Wikidata Coverage of non-English entity names vs. English



Our contributions

Data & Benchmarks

WikiKGE-10

35k manually-graded **entity names** across **10 languages**

Systems

M-NTA

Combining **LLMs**, **MT**, and **WS** to generate high-quality names

Applications

KG Completion

Entity Linking

Question Answering