MUNI
ARTS

# Building a Dataset for Detection of Verb Coordinations with a Shared Argument

Helena Medková
Faculty of Arts, Masaryk University,
Arna Nováka 1, 602 00 Brno

# Introduction

- supported by the specific research project
  - The application of machine learning methods to shared argument detection in verb coordination structures); project no. MUNI/A/1184/2020
- part of phd thesis, focused on the detection of non-grammatical structures
- rule-based approach

# Common argument of two coordinated verbs I

- a syntactic phenomenon bordering a sentence and a multiple sentence element [5],
- ***Obřad*** *má* ***zachránit*** *a* ***přinést duším*** *posvátný* ***klid*** [1], (The ceremony have to save and bring sacred peace to the soul)

# Common argument of two coordinated verbs II

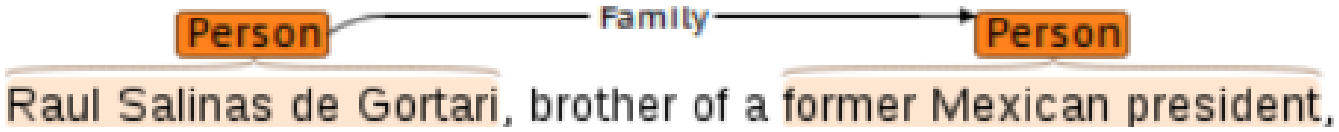
- the coordination of verbs with a shared argument:
  - *Tím **zmírňuje** a **odstraňuje** **pískání** a **hučení** v uších*. [1], (It reduces and eliminates whistling and tinnitus.)
- ungrammatical analogy – zeugma:
  - ***Balzám** má **zmírňovat** a **předcházet otokům.*** [1], (The balm is supposed to relieve and prevent swelling in the eye area)
- coordination of two senteces
  - Jde o léky[...], které alergické **příznaky zmírňují** a **brání zhoršení** nemoci. [1] ,(The medicines[...] that relieve allergic symptoms and prevent the disease from worsening.)

# Data collection

- Sketch Engine,
- corpus czTenTen17 [1],
- CQL queries targeting structures containing verb coordination:
- Ex.:
  - [tag="k1.*"][tag="k5.*"][word="nebo|a"][tag="k5.*"][tag="k1.*"]

# Preprocessing of the linguistic data for a manual annotation I

- Brat web-based text annotation tool
  - annotation at the level of words and the relations between them [2],
  - https://brat.nlplab.org/index.html






MUNI ARTS

# Preprocessing of the linguistic data for a manual annotation II

- preprocessing with UDPipe 2 [3]
  - CoNLL-U format files,
  - positional system of morphological tags,
  - universal dependency tags,
  - universal dependency relations,
  - http://lindat.mff.cuni.cz/services/udpipe/run.php

# Preprocessing of the linguistic data for a manual annotation III

- conversion from CoNLL-U format to Brat format
- configuration files generation:
  - annotation.conf, tools.conf, visual.conf
- renaming the *conj* relation:
  - coordComArg
  - coordZeug
  - coordSent



MUNI ARTS

# Manually tagged dataset statistics

| | count |
|---|---|
| segments | 2610 |
| coordSent | 1506 |
| coordComArg | 682 |
| coordZeug | 22 |



MUNI ARTS

# Annotation automatization

- speeding up the annotation process,
- exploitation of the The Valency Lexicon VerbaLex [4],
- script with defined rules for determining zeugma and coordinations of verbs with (or without) a shared argument,
- a new file in CoNLL-U format,
- original *conj* relation is relabeled to *coordComArg*, *coordSent* and *coordZeug* according to defined rules

# Evaluation of the rules – manually tagged dataset

| | | Actual | | | | | |
|---|---|---|---|---|---|---|---|
| | | CoordComArg | CoordSent | CoordZeug | Precision | Recall | F1-score |
| Predicted | CoordComArg | 396 | 279 | 6 | 58,15 % | 55,70 % | 56,90 % |
| | CoordSent | 298 | 1106 | 7 | 78,38 % | 73,05 % | 75,62 % |
| | CoordZeug | 17 | 129 | 11 | 7,01 % | 45,83 % | 12,15 |



MUNI ARTS

# Evaluation of the rules – dataset to zeugma detection

| | | Actual | | | | | |
|---|---|---|---|---|---|---|---|
| | | CoordComArg | CoordSent | CoordZeug | Přesnost | Pokrytí | F1-score |
| Predicted | CoordComArg | 508 | 161 | 422 | 46,56 % | 52,59 % | 49,39 % |
| | CoordSent | 436 | 563 | 305 | 43,17 % | 67,91 % | 52,79 % |
| | CoordZeug | 22 | 105 | 282 | 68,95 % | 27,95 % | 37,77 % |



MUNI ARTS

# Future work

- speeding up the program to preprocess relations (for more extensive data)
- consideration of possible refinements of the rules
- creation of a manually annotated dataset (comprising at least 10,000 or more coordinated structures)
- training a classifier

# Bibliography

[1] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen corpus family. In: 7th International Corpus Linguistics Conference CL. pp. 125-127. (2013)

[2] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J.: brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations Session at EACL 2012. (2012)

[3] Straka, M.: UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Stroudsburg, PA, USA,(2018) 197-207

[4] Hlaváčková, D., Horák, A.: VerbaLex - New Comprehensive Lexicon of Verb Valencies for Czech. In: Computer Treatment of Slavic and East European Languages. p.~107-115, 6 pp. Bratislava, Slovakia: Slovenský národný korpus (2006)

[5] Panevová, J., Gruet Škrabalová, H.: Elipsa. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.\url{URL: https://www.czechency.org/slovnik/ELIPSA} (2017)