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

Helena Medková

Faculty of Arts, Masaryk University Brno,
Czech Republic
gerzova@phil.muni.cz

Abstract. Coordinate structures represent a specific linguistic problem relating to questions of sentence boundaries and multiple sentence element [1]. A particular difficulty lies in processing at the level of automatic syntactic analysis of the sentence. To deal with the outlined issue, we decided to use the machine learning classification method, for which it is necessary to prepare a sufficiently large amount of data. This paper presents the methods and procedures we used to build a dataset focused on the phenomenon of verb coordinations that may share an argument in context.

Keywords: Coordination · Zeugma · Syntactic analysis · UDPipe 2 · Brat annotation tool · VerbaLex

1 Introduction

The phenomenon of coordinate structures is a challenging task in natural language processing as it can be a complex problem also for a human annotator.

Difficulties can arise because of the parts of sentence ellipsis, which makes such constructions semantically ambiguous and complete reconstruction of the meaning or the author's intention is not always entirely possible. We show the example of multiple interpretation possibilities on the sentence (1) from corpus czTenTen17 [2]:

(1) ***Obřad*** *má* ***zachránit*** *a* ***přinést duším*** *posvátný* ***klid****.* (*The ceremony have to save and bring sacred peace to the soul.*)

In the Czech sentence, we cannot reliably determine if the *ceremony* is the subject that grammatically agrees to the verb (*mít / have to*) or if it is the object of the verb *save*. The coordination could also be ungrammatical if we read the indirect object in the dative (*duším; souls*) as an argument of both coordinated verbs. In practice, such structures tend to be excluded from automatic processing because of their difficulty to handle. [3]

In this paper, we present the dataset building process and a description of the methods we used. The dataset focuses on two predicate coordinations that

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 125–133, 2021.

share at least one argument in the context of the sentence. As an ungrammatical equivalent of such structures, we consider a zeugma.

An annotated dataset will allow us to use supervised machine learning methods and train a classifier to recognize verb coordinations with a shared argument. Furthermore, it will be possible to compare the benefits of a different approach (than rule-based) to the problem.

2 The coordinate structures

The typical example is the coordination of two verbs that bind the same object (2) [2]. The sentence elements that would be repeated in two sentences are thus brought into the same syntactic position by their deletion from the surface structure in one of the coordinated sentences [4]. These structures allow the writer to avoid the redundancy of words in the sentence when syntactic rules are fulfilled.

(2) *Tím* ***zmírňuje*** *a* ***odstraňuje*** ***pískání*** *a* ***hučení*** *v uších.* (*It reduces and eliminates whistling and tinnitus.*)

We can also find formally equivalent structures in sentences in which the two predicates do not share anything (3) [2].

(3) *Jde o léky*[...], *které alergické* ***příznaky zmírňují*** *a* ***brání zhoršení*** *nemoci.* (*The medicines*[...] *that relieve allergic symptoms and prevent the disease from worsening.*)

The non-grammatical alternative to the structures above is binding two expressions by a single dependent element, where the syntactic rules are not met. The expressed syntactic dependency of the constituent contradicts the required syntactic dependency demanded by one of the conjuncts [5]. See sentence example(4) [2].

(4) *Balzám má* ***zmírňovat*** *a* ***předcházet otokům*** *v oblasti očí* [...]. (*The balm is supposed to relieve and prevent swelling in the eye area* [...].)

3 Data collection

We worked with Sketch Engine tool to collect the data, choosing the corpus czTenTen17 [2] as the source of linguistic material for the dataset. We searched the corpus with CQL queries focusing on structures containing verb coordination with specific context restrictions.

```
1. [tag="k1.*"][tag="k5.*"][word="nebo|a"][tag="k5.*"]
[tag="k1.*"]
2. [tag="k1.*"][tag="k5.*"][word="nebo|a"][tag="k5.*"]
[tag="k7.*"][tag="k1.*"]
```

The first two CQL queries seek after structures where the immediate context, i.e. the 1st position by KWIC [6], contains a noun on the left and either a noun or a preposition and a noun on the right. Furthermore, we removed passive forms from the search by the negative filter because, in such structures, the object moves to the subject position and the representation of the noun-verb relation changes from a government to an agreement.

```
3. [tag="k1.*c1.*"][word=".*" & tag!="kI.*"]{0,3}[tag="k5.*"&
tag!="k5.*mN.*" & lemma!="být"][word="nebo|a"][tag="k5.*"&
tag!="k5.*mN.*" & lemma!="být"][word=".*" & tag!="k[157I].*"]
[word=".*" & tag!="k[157I].*"]{0,1}[word=".*"&
tag!="k[157I].*"]{0,1}[tag="k1.*"] within <s/>
```

```
4. [tag="k1.*c1.*"][word=".*" & tag!="kI.*"]{0,3}[tag="k5.*"&
tag!="k5.*mN"][word="nebo|a"][tag="k5.*" & tag!="k5.*mN"]
[word=".*" & tag!="k[157I].*"][word=".*" & tag!="k[157I].*"]
{0,1}[word=".*" & tag!="k[157I].*"]{0,1}[tag="k7.*"]
[tag="k1.*"] within < s/>
```

The third and fourth CQL queries seek after verb coordinations where the immediate context, i.e., positions 1–3 from KWIC [6], contains a noun in nominative on the left, and a noun or preposition besides a noun on the right side. Within the immediate context on the left, we removed punctuation by the negative filter, and on the right side, we removed prepositions, verbs and punctuation on positions 1–3.

4 Linguistic data preprocessing for a manual annotation

To build a gold-standard annotated dataset, we used the web-based text annotation tool Brat [7] that supports, for instance, two basic types of annotations. It allows adding a label to a specific word (text span annotations) and adding relations among words in a sentence (relation annotations).

4.1 Data preprocessing

Since developing our text markup methodology for annotations in Brat would be inefficient, we took advantage of the UDPipe 2 [8] that works with CoNLL-U formatted files. It parses the input text file into sentence segments, giving each word a set of features (lemma, part-of-speech tag, morphological tag, dependency relation).

For the conversion of the UDpipe 2 (CoNLL-U) format to the standoff format for Brat, we use the ConllXtostandoff.py program [9] that creates .txt files containing the original sentences and .ann files with annotations from the CoNLL-U format, which Brat graphically displays.

Brat enables text annotation editing if particular labels are defined in the configuration files. We designed a script makeconffiles.py that extracts a required set of files (annotation.conf, tools.conf, visual.conf) from the output of UDPipe 2.

UDPipe 2 uses the positional morphological tag system [10], universal dependency tags [11] and universal dependency relations [12], which are developed for consistent grammar annotations across many languages.

The annotation.conf file defines universal positional tags (NOUN, ADJ, ADV and other) at the text span annotation level and universal dependency relations (*nsubj*, *obj*, *conj* and other) at the relation level. For our purposes, the essential dependency relation is coordination (*conj*). In dependency relations, it is a relation between two elements connected by conjunctions *and*, *or*. The head of this relation is the first conjunct, while the other elements depend on it [13].

4.2 Replacing the relation *conj* between coordinated verbs

We rename a syntactic relation *conj* in the coordinations, where both conjuncts have a common argument to *coordComArg*. If the argument does not grammatically correspond to the syntactic pattern of one of the conjuncts, we mark this defective structure as zeugma with label *coordZeug*. If conjuncts do not share any part of the sentence (except subject), we label the relation as *coordSent*. The original *conj* tag represents other types of coordinations.

4.3 The standard dataset – statistics

The manually tagged dataset consists of 2610 segments sorted by the number of ten to 261 files. One segment is a part of a sentence as parsed by the UD Pipe 2 tool. We randomly pick sentences from language material that we gained from corpus czTenTen17 [2]. The resulting statistics shows table 1.

Table 1: Statistics of the manually annotate dataset

Data set statistics	Count
Segments	**2610**
CoordComArg	**682**
CoordSent	**1506**
CoordZeug	**22**

5 Annotation automatization

Manual annotation of raw text is a time-consuming process, and the usage for machine learning requires thousands of annotated cases of the desired

structures. We decided to design a script for relabeling relations in the UD Pipe 2 output to speed up annotations. We defined rules for the detection of zeugma and verb coordinations with or without a common argument.

5.1 Rules drafts

Based on the manual annotations experiences, the first step was to formulate theoretical rules for the automatized retrieval of the *coordComArg* and *coordZeug* structures. In addition, we define the *coordSent* relation as any other verbal coordination that the rules for distinguishing *coordComArg* and *coordZeug* do not cover. We describe these rules in the following two subsections.

CoordComArg This rule defines the verb coordination with a possible common argument. The prerequisite for the *coordComArg* structure is an identical valency of the verbs (see coordination below (5) [2]).

(5) *Lada v současnosti* ***vyvíjí a vyrábí své vlastní automobily***. (*Lada is currently developing and producing its cars.*)

To that purpose, we need to create a list of possible valency complements from a valency lexical database and, for each complement, a list of verbs that can bind with it. We assume that if two coordinated verbs are in the same list and simultaneously have suitable complements in the neighbouring context but not in their own, we consider this complement as shared.

CoordZeug We assume that verbs yoked by another sentence element in such structures require a different valency complement (6) [2].

(6) *Analyzujte, jak organizace* ***rozhoduje a komunikuje změny.*** (*Analyze how the organization makes decisions and communicates changes.*)

We will use the list of valency complements again to follow the assumption that zeugma will most likely arise in coordinations with different verbal valency patterns if, in the context of the first, or the second verb, in the sentence, the appropriate complement is not found.

5.2 Implementation of the rule drafts

We generated a dictionary from the lexical database of Czech verb valency frames VerbaLex, [14] where the keys of the dictionary consist of any first obligatory complements of verbs in the database. The values of these keys contain lists of verbs that can have such complements according to the database. We saved the data structure in a .json file.

Further, we wrote a preprocess_relations.py script that takes as input the UDPipe 2 output in ConLL-U format. The program first goes through the

input file and searches for verb coordinations, storing them in a list of tuples containing the ids of the sentences and verbs and lemmas.

The program also stores important sentence features for each word (word id, lemma, word type, tag, binding position, dependency relation) in a dictionary where the key is the sentence id (sent_id) and the value is a list of tuples.

The script does not handle reflexive verbs, as it is impossible to determine without any other rule whether the clitic "se" is part of the verbal or noun phrase from the context of the sentence.

Program stores a context for each coordination in the list of two tuples that represent the left and right context. The context span is regularly five positions from each verb (KWIC <-5, 5>). The tuples store the numbers of the cases of such sentence positions where the nouns PRON (pronoun), NOUN, DET (determiner) and PROPN (proper noun) occur.

Coordinations are further processed using a dictionary generated from VerbaLex. Each verb obtains the list of arguments based on the dictionary. If the verbs can have the same argument structure, accordingly to the dictionary, and do not have a suitable complement in their context, they are stored with the ids to the list of common argument verb coordinations.

Similarly, we handle zeugmatic coordinations. If the verbs do not occur in the same list in the VerbaLex dictionary, and at the same time one of the verbs does not have a suitable binding in the context, the sentence id and verb id are saved into the list.

The output of the whole program is a newly processed CoNLL-U format file, renaming original *conj* relation to *coordComArg* and *coorZeug* according to the created lists. The *coordSent* relation matches the coordinations that do not cover the lists for *coordComArg* and *coorZeug*.

6 Comparing automatic and manual annotations

We tested the annotation preprocessing program on the dataset that we manually annotated in Brat, which covers mainly grammatically correct structures and on the dataset created for evaluating zeugma detection [15], where the zeugma occurs in significantly higher numbers. Table 1 and table 3 illustrate the evaluation of the program.

Table 2: Evaluation of automatical annotation on dataset focused on correct verb coordination. CoordCA – coordComArg, coordSe – coordSent, coordZe – coordZeug.

		Actual					
		coordCA	**coordSe**	**coordZe**	**Precision**	**Recall**	**F-score**
Predicted	**coordCA**	396	279	6	58,15 %	55,70 %	56,90 %
	coordSe	298	1106	7	78,38 %	73,05 %	75,62 %
	coordZe	17	129	11	7,01 %	45,83 %	12,15 %

We gained 58,15 % precision in detecting the common argument of two verbs and 55,70 % coverage based on the data. According to these results, we can assume that the program could significantly speed up the annotation. We could refine the program with more sophisticated going through the VerbaLex database and more accurate processing of context coordination (e.g., it does not consider whether punctuation is present in the context so that the program may consider a noun in another sentence as the verb object). Furthermore, the absence of several verbs (for example, *ignore*, *overthrow*) in VerbaLex causes false negatives and the failure to process coordination.

The rules for the detection of the zeugma proved to be ineffective. Most of the false-positive cases is caused by naive searching of the coordinations context and also by ellipses. In sentence seven [2], we see a typical example of a mislabeled zeugma. According to VerbaLex, the verbs *depart*, *leave* have obligatory complements that do not match. However, the verb *depart* has no complements in its context, so the coordination is evaluated as a zeugma.

(7) *Vojáci* ***odcházejí a nechávají Achilla...*** (*The soldiers are departing and leaving Achilles.*)

Based on the results of the automatic annotations, we found that some coordination went unnoticed in the dataset with manual annotations. With the annotation preprocessing, we managed to get better results for coordinations with a common argument compared to the original data, as shown in Table 3.

Table 3: Statistics actualization of the dataset focused on verb coordinations with a shared argument

Dataset with preprocessed annotations	Count	Manualy annotated dataset	Count
Segments	**2610**	Segments	**2610**
CoordComArg	**1551**	CoordComArg	**1506**
CoordSent	**712**	CoordSent	**682**
CoordZeug	**22**	CoordZeug	**22**

As we see in Table 4 the precision of zeugma recognition improved many times on the dataset focused mainly on the zeugma phenomenon. However, this is a result of the unbalance of the dataset. Therefore, it might be beneficial to merge the two datasets; the rule evaluation results could then be more reliable. We could increase the recall of the rules by including reflexive verbs in the preprocessing and by designing a special rule to recognize such coordinations that may have the same binding in specific contexts.

The evaluation of the rules on the zeugma-focused dataset showed decrease in precision and recall scores for the *coordComArg* and *coordSent* relations rules. In the dataset where ungrammatical constructions are much more frequent, the

Table 4: Evaluation of automatical annotation on dataset focused on zeugma phenomenon. CoordCA – coordComArg, coordSe – coordSent, coordZe – coordZeug.

		Actual					
		coordCA	**coordSe**	**coordZe**	**Precision**	**Recall**	**F-score**
Predicted	**coordCA**	508	161	422	46,56 %	52,59 %	49,39 %
	coordSe	436	563	305	43,17 %	67,91 %	52,79 %
	coordZe	22	105	282	68,95 %	27,95 %	39,77 %

simple processing of valency frames from VerbaLex and naive passing through the context of verb coordination might have been more evident.

7 Summary and future work

This paper presented approaches we applied for building a dataset focused on coordinate structures of two verbs. The aim is to create a gold-standard dataset that can be used for training and testing a classifier for zeugma and verb coordinations with a shared argument recognition using machine learning methods.

We described the possibilities of speeding up the manual annotation process with automatic preprocessing, which could help create an extensive dataset with thousands of positive cases.

The outlined preprocessing showed promising results on tested data. However, annotation accuracy can be increased by improved coordination context managing, additional inclusion of reflexive verbs in the processing, and refined work with the valency frame database.

Therefore, we will continue editing and expanding the dataset in terms of content using the presented methods.

Acknowledgements. This work was supported by the project of specific research *Využití strojového učení při detekci společného argumentu v koordinovaných strukturách* (The application of machine learning methods to shared argument detection in verbal coordination structures); project no. MUNI/A/1184/2020).

References

1. Panevová, J., Gruet Škrabalová, H.: Elipsa. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny. URL:https://www.czechency.org/slovnik/ELIPSA (2017)
2. 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)
3. Lopatková, M., Mírovský, J., Kuboň, V.: Gramatické závislosti vs. koordinace z pohledu redukční analýzy (in Czech, Grammatical dependencies vs. coordination

from the perspective of reduction analysis). In: V. Kůrková et al. (Eds.): ITAT 2014 with selected papers from Znalosti 2014, CEUR Workshop Proceedings Vol. 1214, Praha, Univerzita Karlova. (2014) 65
4. Karlík, P., Gruet Škrabalová, H.: Koordinace (in Czech, Coordination). In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny. https://www.czechency.org/slovnik/KOORDINACE (2017)
5. Karlík, P.: Zeugma. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny. https://www.czechency.org/slovnik/ZEUGMA (2017)
6. Cvrček, V.: Kvantitativní analýza kontextu (in Czech, Quantitative context analysis). Praha, Nakladatelství Lidové noviny. Studie z korpusové lingvistiky (2013) 25
7. 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)
8. 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
9. Auffarth, B., Pyysalo, S, Ninjin: ConllXtostandoff: Script to convert a CoNLL X tabbed dependency tree format. https://github.com/nlplab/brat/blob/master/tools/conllXtostandoff.py (2006)
10. Hajič, J.: Popis morfologických značek – poziční systém (in Czech, Description of morphological tags - positional system). http://ucnk.ff.cuni.cz/doc/popis_znacek.pdf (2000)
11. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Proceedings of LREC. https://aclanthology.org/L12-1115/ (2012)
12. de Marneffe, C.-M., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C. D.: Universal Stanford Dependencies: A cross-linguistic typology. In: Proceedings of LREC. (2014)
13. Universal Dependencies contributors: Introduction. https://universaldependencies.org/u/dep/conj.html (2021)
14. 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)
15. Medková, H.: Automatic Detection of Zeugma. In: Horák, A., Rychlý, P., Rambousek, A.: Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, Brno: Tribun EU (2020) 79-86.