Towards a Knowledge-Aware AI
A. Dimou et al. (Eds.)

doi:10.3233/SSW220022

Distinguishing the Types of Coordinated Verbs with a Shared Argument by Means of New ZeugBERT Language Model and ZeugmaDataset

Helena MEDKOVÁ [a,1] and Aleš HORÁK [b]
[a] *Masaryk University, Faculty of Arts, Brno, Czech Republic*
[b] *Masaryk University, Faculty of Informatics, Brno, Czech Republic*

Abstract. Sentences where two verbs share a single argument represent a complex and highly ambiguous syntactic phenomenon. The argument sharing relations must be considered during the detection process from both a syntactic and semantic perspective. Such expressions can represent ungrammatical constructions, denoted as zeugma, or idiomatic elliptical phrase combinations. Rule-based classification methods prove ineffective because of the necessity to reflect meaning relations of the analyzed sentence constituents.

This paper presents the development and evaluation of ZeugBERT, a language model tuned for the sentence classification task using a pre-trained Czech transformer model for language representation. The model was trained with a newly prepared dataset, which is also published with this paper, of 7,849 Czech sentences to classify Czech syntactic structures containing coordinated verbs that share a valency argument (or an optional adjunct) in the context of coordination. ZeugBERT here reaches 88 % of test set accuracy. The text describes the process of the new dataset creation and annotation, and it offers a detailed error analysis of the developed classification model.

Keywords. natural language understanding, coordinated verbs with shared argument, zeugma, BERT language model, dataset

1. Introduction

Coordinated structures are a widely occurring phenomenon in the language, yet they pose problems to syntactic parsers [22,13] to correctly analyze the dependents for all conjuncts because of high ambiguity, and even to their exact realisations in grammar formalisms [25,3] and treebank annotation [5]. One such problematic structure is called zeugma, usually regarded as a figure of speech. Zeugma refers to a coordination of two expressions joined together by a single ambiguous expression where each conjunct is simultaneously related to a different meaning of the joining word. An example of such intended use of zeugma is depicted by the sentence

[1] Corresponding Author: Helena Medková, Masaryk University, Faculty of Arts, Brno, Czech Republic; E-mail: gerzova@phil.muni.cz.

(1) She drew a gun and a picture of a gun.

Here, the verb "drew" bears two different senses: a) "to pull out a weapon," and b) "to paint a picture." The whole sentence is thus an ambiguous expression yoking together two expressions with different meanings. The resulting construction appears strange from the semantic point of view [24] and, as such, it attracts attention. Besides neglecting the collocability of expressions, zeugma may also indicate an erroneous violation of syntactic rules leading to ungrammatical sentences. Our motivation is to detect such structures comprehensively, which can be valuable for refining syntactic parsers or improving grammar checking proofreader modules.

This paper presents a new linguistic model, ZeugBERT, developed specifically for the sentence analysis of coordinated verbal phrases with a (possible) shared argument. Previous approaches solved this task by rule-based techniques [17,1]. ZeugBERT is designed to solve the classification task of distinguishing sentences with coordinated verbal phrases into three classes: 1) coordination of verbal phrases that do not share a constituent (*coordSent*), 2) coordination of verbal phrases with a shared argument (*coordComArg*), and 3) coordination of verbal phrases with a shared argument that crosses the argument structure for both verbs, the rhetorical concept called zeugma (*coordZeug* class).

For the purpose of ZeugBERT development, we have created a manually tagged dataset of Czech sentences, denoted as ZeugmaDataset, used for training and testing the model. The dataset is publicly available at the website[2] of Natural Language Processing Centre, Masaryk University.

2. Coordinated structures

Coordination is a syntactic relation between two or more conjuncts that have equal syntactic status, primarily in terms of functional likeness [8]. With the ZeugBERT model, we concentrate on three kinds of verb coordinations with a possible shared argument. The specification of these types is explicated in the following subsections.

2.1. Coordination of verbs with the shared argument

The observed structure within this phenomenon is a shared noun phrase (NP) or prepositional phrase (PP) in the right (or left) periphery of coordination of the two verbs conjoined by *and*, *or* conjunctions as in the typical example (2) [7]:

(2) Vosy **vykusují a vysávají** přezrálé **ovoce**. (Wasps bite and suck out overripe fruit.)

The Czech dependency grammar theory considers such structures as the result of unifying transformation (fusion), which can be viewed as elliptical structures [11,10]. It means that the argument, that would be repeated in the sentence is elided from the surface structure of the first (or second conjunct) to avoid redundancy. From the generative grammar perspective, the pattern of the sentence (2) represents the Right node raising (RNR), where the object of the first conjunct is moved to the end of the coordination [8]. Such structure is a typical representative of the studied phenomenon and also the most

[2] https://nlp.fi.muni.cz/projects/zeugma

frequent configuration in the dataset. Two coordinated verbs *vykusují a vysávají* (bite and suck out) share a single noun phrase *overripe fruit* as their dependent. According to Gerdes and Kahane [6], a shared dependent is governed by both heads of verbal phrases, nevertheless it creates a prosodic unit with the nearest conjunct. The relation here is defined as a pure dependency, while the relation between the other conjunct and the shared dependent as an inherited dependency.

2.2. Zeugma

The second examined structure is the coordination of verbs that share a complement, however, their argument structure does not correspond to each other as can be seen in Example (3). The Czech verb *zmírňovat* (to relieve) binds with an obligatory accusative object, unlike the verb *předcházet* (to prevent) which binds a dative object. This phenomenon is denoted as zeugma [19] and it is usually considered as an ill-formed structure. The ungrammaticality typically rises on the side of the coordination where the dependency between a conjunct and a shared dependent is inherited, i.e. where the conjunct and the dependent do not create one syntactic (prosodic) unit [6]. According to Karlík [10], the dependent element adopts the form required by the adjacent expression.

(3) ***Zmírňují a předchází bolestem** šíje. (*They relieve and prevent neck pain)[3]

(4) *Cestující **nastupovali a vystupovali z vlaku**. (*Passengers were getting on and exiting the train.)

The shared dependent can be (besides a complement as presented in Example (3)) also an adjunct, but the dependency strength between the verbs and the adjunct is much weaker than in the case of verbs and complements [7]. It is also the source of frequent issues when deciding whether or not is the dependent shared or not. Sentence (4) is an example of a zeugma, where the shared dependent is the adjunct in form of the prepositional phrase [PP *z vlaku*] (from the train).

2.3. Coordination of verbs without the shared dependent

The last case is represented by coordinated verbal phrases without a shared dependent (i.e. not just the coordination of the heads) [9]. In the ZeugmaDataset, such sentences contain a coordination of two main clauses where the conjuncts are usually two clauses with verbs on their boundaries. The case, where the verbs stand on the borders of the clauses is represented by Example (5).

(5) **Ptáci** kolem **umlkli** a **zvedl se** lehký **vánek**. (The birds around became silent and a light breeze arose.)

(6) **Pánové doktoři** ze sálu **vtipkovali** a **ujistili mě o své šílenosti**. (The doctors in the room joked and assured me of their insanity.)

Another option for coordinated verbs is a compound sentence, where the verbs share the left side of their valence structure, i.e. the subject part of their dependency structure

[3] The * (star) here explicitly marks a sentence which is considered ill-formed.

Table 1. Statistics of the training, testing and validation subsets (numbers of sentences)

	train_set	test_set	valid_set	whole dataset
coordSent	3,062	875	437	4,374
coordComArg	1,744	499	248	2,491
coordZeug	689	197	98	984
all classes	5,495	1,571	783	7,849

(as can be seen in Example (6). In the generative grammar, this type of clause-level coordination is known as a Forward Conjunction Reduction (FCR), where the right-hand conjunct [VP *ujistili...*] (they assured) inherits the subject [NP *Pánové doktoři...*] (the doctors) [12].

3. ZeugmaDataset – manually annotated dataset of coordinated structures

In this section, we want to introduce the ZeugmaDataset that was specially created with the aim to distinguish the three sentence classes: *coordSent*, *coordComArg*, and *coordZeug*. ZeugmaDataset consists of corpus-based samples of Czech coordinated structures with two coordinated verbs from the largest Czech web corpus csTenTen17 [21], which was crawled in 2017, therefore it reflects the language reality until that year.

A direct search using the Corpus Query Language (CQL) for two verbs connected with the conjunction *and* or *or*[4] revealed 15,703,841 verb coordinations in the corpus. In the successive step, all passive constructions were removed with a negative filter leading to 15,206,270 sentences out of which 7,849 were randomly selected for further processing. The resulting dataset was then manually annotated by a single annotator with the three class labels.

The preceding version of the dataset was prepared as a benchmark dataset for the rule-based detection of zeugma [15]. The dataset contained 2,762 sentences of which 1,081 were positive cases of zeugma and 1,681 sentences covered non-zeugmatic coordinations of verbs. In the current ZeugmaDataset, the class of non-zeugmatic coordinations was split into two other classes (verbs with a shared argument, `coordComArg`, and compound sentences, `coordSent`). The whole dataset was extended, cleaned and refined with the methods that are described in [16]. The UDPipe2 web service [20] was used for tokenization, lemmatization, and morphological and syntactic analysis. The dataset for download thus follows the UDPipe2 CoNLL-U format[5] with the sentence class annotations marked as specific dependency relation of the verb in the DEPREL field.

3.1. The statistics of the dataset

The whole ZeugmaDataset is divided into three balanced subsets for training, testing and validation in a ratio of 70:20:10, see Table 1. The split data were selected manually, especially for the *coordZeug* class it was important to choose the sentences uniformly, to avoid an occurrence of the same coordinated verbs in the subset instead of variability.

[4] The exact query was `[tag="k5.*"] [word="a|nebo"] [tag="k5.*"]`, i.e. a $verb_1$ *a*|*nebo* $verb_2$, where "*a*" and "*nebo*" stand for "and" and "or".

[5] `https://universaldependencies.org/format.html`

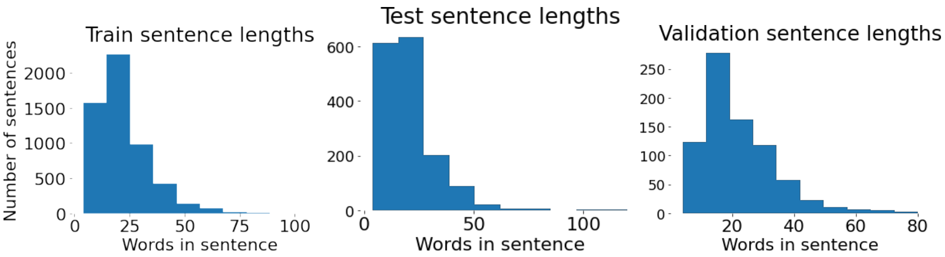

Figure 1. Sentences lengths of the training, testing and validation subsets

Table 2. The count of unseen verb lemmata versus the count of unique coordinated verb lemmata in the train, test, and valid subsets.

Class	Test dataset		Valid dataset	
	unseen in train	total	unseen in train	total
coordSent	836	865	412	433
coordComArg	406	469	212	241
coordZeug	90	138	35	81

Since various details may impact the whole learning process, we made a quantitative comparison of the most frequent sentence lengths of the language data to see if any subsets contained disproportionately short or long sentences. As shown in Figure 1, sentences with lengths of 20–30 words are the most frequent in all subsets. The average length of simple sentences in Czech fiction is 5–8 words [23] so for the presented data, the character of the training sentences should be taken into account, and the fact that they are uncorrected compound sentences from the web corpus.

The dataset split was also driven by the requirement of introducing unknown verb lemmata [6] in the validation and testing subsets to avoid possible lexical bias in training. The learning task was particularly difficult for the zeugma class since this phenomenon appears more frequently in the corpus with some specific verb coordinations. Therefore, we ensured that the coordinations of verbs in the test and validation sets were as unique as possible (see Table 2 for detailed statistics).

3.2. Rules for the annotation of the shared argument

Since a high degree of ambiguity is a characteristic feature of coordinated structures, distinguishing between the classes can be challenging. Therefore, in this section, we outline the procedures we followed to annotate the dataset.

An essential guideline in this matter is using the *Vallex 3.0* valency dictionary [14], according to which we check the types of arguments for specific verbs. We identify a coordination with shared argument when there is a semantically compatible verb complement (required by the verb's argument structure), a typical adjunct (often occurring in combination with the verb), or another optional element of the verbal structure according to the valency dictionary in the left-hand (or right-hand) context of the coordination.

The least problematic are verb complements with an obligatory argument on the right-hand side of the coordination, as illustrated in the sentence (7a). We also observe

[6] The dictionary verb form in the infinitive.

the co-occurrence of coordinate verbs on the left side of the coordination, see sentence (7b), although they are less frequent in the dataset.

(7a) [. . .] umožňují zlepšovat ~~kvalitu pitné vody~~ a měnit kvalitu pitné vody. ([...] they allow to improve ~~drinking water quality~~ and change drinking water quality.)

(7b) [. . .] **kvalitu pitné vody** umožňují **zlepšovat a měnit** ~~kvalitu pitné vody~~. ([...] the quality of drinking water they allow to improve [the quality of drinking water] and change ~~the quality of drinking water~~.)

(8) [. . .] jeho vlákna **izolují a regulují teplotu**. ([...] its fibers isolate and regulate temperature.)

In some cases, the common accusative argument identification is a matter of sentence interpretation. We have decided to classify verb coordinations in the way which avoids creating figurative senses.

A borderline example is illustrated by the sentence (8), where both verbs, *to isolate* and *to regulate*, have obligatory complements but are not fully collocable from a semantic point of view. According to corpus evidence, a typical strong collocation in accusative to the verb *isolate* is "*teplo*" (heat) (with the logDice score [18] of 7.7), and not "*teplota*" (temperature). However, it is questionable to what extent the conjunction is semantically incompatible since the expected complement for the verb *isolate* is semantically concrete and *temperature* is an abstract concept. The shared argument in this sentence is thus not a nonsense but rather a conceptual confusion. *To regulate temperature* here forms a syntagm, the subject of the verb *isolate* is considered to be implicit in this case.

(9) *Cestující nastupovali a vystupovali [PP z vlaku]. (*The passengers were entering and exiting [PP from the train]).

(10) Ugrofinské jazyky se také mohly vyvinout a rozšířit až [PP po skončení doby ledové]. (Ugrofin languages might also develop and spread [PP after the end of the Ice Age].)

Expressions may bind to verbs as complements in their syntactic structure, see corpus sentence (7a), or adjoin the verbs (9).

In Example (9), we can observe a direction crossover (from where, to where) between verbs. According to *Vallex 3.0*, the prepositional phrase is obligatory; therefore, we consider the adjunct as shared.

In some cases, however, it is not clear whether the prepositional phrase is a shared adjunct or whether it only adjoins the conjunct on the left or right side of the coordination, see Example (10). In such cases, it depends primarily on the interpretation of the sentence, so we consider adjuncts to be shared adjuncts even potentially.

Since the strength of the adjunction in these additions tends to be more a matter of scale [7], it is not always easy to label an adjunct as shared or unshared. In these cases, therefore, the intuition of the annotator plays a significant role.

Table 3. Comparison of the previous rule-based technique and the current ZeugBert model to zeugma detection

	precision	**recall**	**F1-score**	**support**
Rule-based zeugma detection with preceding dataset	0.98	0.38	0.55	1013
Rule-based zeugma detection with current test subset	1.00	0.30	0.46	197
ZeugBert with current test subset	0.82	0.83	0.83	197

Table 4. The results of the ZeugBERT model with the test set data

	precision	**recall**	**F1-score**	**support**
coordSent	0.92	0.92	0.92	875
coordComArg	0.85	0.85	0.85	499
coordZeug	0.82	0.83	0.83	197
accuracy			**0.88**	1,571
macro avg	0.86	0.87	0.86	1,571
weighted avg	0.88	0.88	**0.88**	1,571

4. The ZeugBERT language model

The main goal was to design a tool that recognizes ill-formed structures among coordinated verb phrases. We initially developed a rule-based method with 83 manually created rules for the zeugma detection task. The disadvantage of this approach was the specificity to particular verbs, a low recall, complicated and time-consuming extensibility of the rules, and zero reflection of the semantics. Another problematic aspect was the strong prerequisite of building the rules on top of the (correct) output of a morphological analyzer [15].

The evaluation of the rule-based approach with the preceding version of the dataset (see Section 3) revealed an F1-score of 0.55 with high precision of 0.98 but at the cost of a significantly low recall of 0.38. To offer a fair comparison, we have now reevaluated the rule-based approach with the current ZeugBert test subset, and it reached almost the same results: F1-score of 0.46, i.e. precision of 1.00 and recall of 0.30 (see Table 3).

To achieve satisfactory results, we have employed machine learning methods currently widely prefered in NLP tasks solutions. In the development of the ZeugBERT model, we have adopted the concept of the Bidirectional Encoder Representations Transformers (BERT) language model [4], specifically the SlavicBERT pre-trained model[7] [2] for Czech (and other Slavic languages). The ZeugBERT mode has been fine-tuned to the sentence classification task with the ZeugmaDataset described above and the *huggingface transformers* library [26].

The input to the training process was formed by the word sequences of the sentences without any special preprocessing. The model was trained for 3 epochs. the batch-size 16, 500 warm-up steps and the weight decay of 0.01.

Compared to the rule-based approach, the ZeugBERT method proved to be very efficient in distinguishing between zeugma and other types of coordinated verbs. Table 4 summarizes the ZeugBERT performance with the testing set where it achieved an overall accuracy of 88 %.

[7] `https://huggingface.co/DeepPavlov/bert-base-bg-cs-pl-ru-cased`

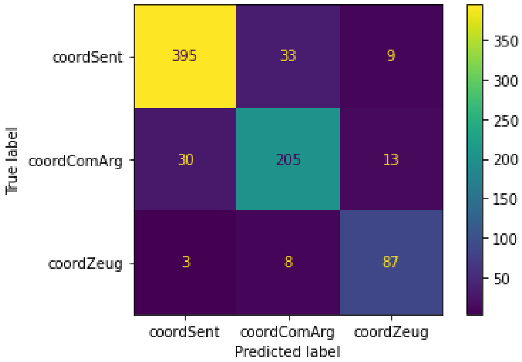

Figure 2. Confusion matrix of ZeugBERT results with the validation subset (numbers of classified sentences)

The model detects zeugma with 83 % accuracy, and we noticed a significant improvement in the recall of detected cases, which also reaches 83 % with the test set data. Based on these results, the ZeugBERT method proved to be an effective solution to the problems of low recall and difficult extensibility of the previous rule-based approach to zeugma detection. Since the rule-based method distinguished just zeugma vs non-zeugma classes, we cannot compare its results for the full three class setup of the ZeugmaDataset and compute a comparable overall accuracy, macro average and weighted average measures. The ZeugBERT model detects the *coordSent* class (the most frequented class in the dataset) with the highest f-score of 92 %. For the *coordComArg* class, the model reaches comparable recognition results as with zeugma.

5. Error analysis

We have examined in detail the cases where the model made a misclassification with the validation set sentences, see Figure 2 for the confusion matrix. An exhaustive error analysis outlining the possible causes of the errors made by the model is presented and discussed in the following subsections, with summary lists in Tables 5–7.

By observation of misclassified sentences, we have identified several regular structure configurations where the model made a mistake, even though in some cases the reasons for the misclassification could not be reliably recognized. Such sentences are labelled as *Uncertain causation* in the tables.

5.1. The compound sentences, coordSent

The model made one of the major traceable errors in distinguishing between structures where sentences potentially share (or do not share) an adjunct (19 %). A typical example of such sentence is presented in (11). "*Long-term diarrhoea*" could cause mortality in wildlife, but it is not necessarily the (only) cause of weight loss. Therefore, errors of this nature are not entirely misclassifications because sentences of this type are semantically ambiguous, and we could potentially consider the adjunct as shared.

(11) Postižená zvěř **hubne a hyne v důsledku** vyčerpávajících dlouhodobých **průjmů**. (Affected animals lose weight and die as a result of exhausting long-term diarrhoea.) [coordComArg → coordSent][8]

Table 5. Error analysis of *coordSent*

	Reason for model confusion	**Count**	**%**
1	Coordination of verbs with a potentially shared adjunct	8	19.0
2	Coordination of verbs with an accusative argument that is semantically incompatible	7	16.7
3	The first verb in coordination without obligatory right-hand argument	8	19.0
4	The potentiality of a zeugma	5	11.9
5	Coordination of verbs with an elided shared argument in surface structure	1	2.4
7	Confusion of the adjunct with an ungrammatically conjoined expression (typically a prepositional phrase)	3	7.1
8	Questionable collocability of potentially shared argument	1	2.4
9	Uncertain causation	9	21.4

(12) Ráda ve volném čase **maluji a čtu knihy**. (I like to paint and read books in my spare time.) [coordComArg → coordSent]

Sentence (12) illustrates the coordination of two verbs with an accusative complement in their argument structure that is not semantically compatible. The model tags this coordination as *coordComArg*. However, the words *paint books* are not semantically compatible. The model could be confused by considering the prepositional phrase *ve volném čase* (*in the spare time*) as a shared adjunct. Therefore, we tested the sentence without the prepositional phrase, but the model still labels it as *coordComArg*.

(13) A Egon Bondy **běhal a kopal do vzduchu** a držel se za koleno: Dostal jsem z toho do lejtka křeč, ale proč do nosu ? (And Egon Bondy was running and kicking in the air and holding his knee: I got a cramp in my calf, but why in my nose?) [coordComArg → coordSent]

(14) [. . .] za současné situace si asi většina lidí **připlatí a koupí K750i**. ([...] in the current situation, most people will probably pay extra and buy a K750i) [coordZeug → coordSent]

(15) Před každým použitím **protřepejte a nanášejte na pokrmy** ze vzdálenosti 20–25 cm, nestříkejte do otevřeného ohně nebo na žhavé předměty. (Before each use, shake and apply to food from a distance of 20–25 cm, do not spray on open flames or hot objects.) [coordComArg → coordSent]

(16) [. . .] až děti **vyrostou a odejdou z domova**. (When children grow up and leave home) [coordZeug → coordSent].

In some cases, the model marked the verb coordination as *coordComArg* in which the first verb is without an obligatory right-hand complement (or adjunct), as illustrated by sentence (13). These cases may also include verbs with a null object.

In some cases, again, there can be an issue with the ambiguity of sentence interpretation (14). The model has identified the structure as a zeugma which potentially could be correct. However, since the first verb can optionally omit the argument from its surface structure, such a structure is not classified as a zeugma in the dataset.

In sentence (15), we exemplify the case where the model has identified the sentence as a verb with a shared argument. Verbs can potentially refer to the same elided object.

[8] Here by [*class_predicted → class_ground_truth*], we indicate the misclassification case where the correct *class_ground_truth* was predicted as *class_predicted* by the ZeugBERT model.

Table 6. Error analysis of *coordComArg*

	Reason for model confusion	**Count**	**%**
1	Valency argument on the left-hand side of the coordination	11	25.6
2	Unrecognized argument on the right-hand side	9	20.9
3	Zeugma confusion with coordination of verbs with prepositional phrases on the right side of the coordination	9	20.9
4	Interpretation issue of the common adjunct	6	14.0
5	Zeugma misclassification based on formal similarity	4	9.3
6	Uncertain causation	4	9.3

However, since the accusative has a complement omitted from the surface structure, the correct classification of such cases is *coordSent*.

Example (16) illustrates the case where the model incorrectly identified a zeugma, typically involving a structure with a prepositional phrase adjacent to the second verb in coordination. Similar to the sentence (8), the model here has identified a common argument for the coordinating verbs, but the collocability of the first verb with the argument is semantically questionable, as we have explained above.

5.2. Verb with a shared argument, coordComArg

In the classification of the *coordComArg* class, the model struggled most with coordinations where the complement of both verbs took place in the left-hand context, in 25.6% of cases, see Example (17). The misclassification here is probably caused by low frequency of such configurations in the dataset.

(17) [...] **příběh dopíši a vydám** na Vánoce, stejně jako minule. (I will finish the story and publish it on Christmas, just like last time) [coordSent → coordComArg]

In contrast, the model did not recognize the accusative argument at the right periphery of the coordination in 20.9 %, see Example (18), even though this is the most frequently represented example of a shared argument in the dataset.

Since many sentences in the dataset are very long, we tried to test only coordination with arguments without further context, and the model then determined the class correctly. Thus, we assume that in these cases, the broad context of the coordination is the cause of the misclassification.

(18) [...], **sdílet a vyměňovat** si **informace** a **nápady**, [...] (to share and exchange information and ideas) [coordSent → coordComArg]

With the same frequency, the model misclassified coordinated verbs that share a prepositional valency complement on the right-hand side of the coordination as zeugma (19). The causation may be traceable to the ellipsis of the first obligatory accusative complement of the verbs *to request* and *beg*, or the similarity of the syntactic configuration to ungrammatical structures.

(19) "**Žádám a prosím o respekt** soukromí všech, koho se to týká," vzkázala Moore. ("I request and beg for respect of the privacy of all the concerned," Moore said) [coordZeug → coordComArg]

As with the *coordSent* class, we encounter the problem of recognizing and interpreting the common adjunct. In case of Example (20), the prepositional phrase [until lunch] is labelled as a common adjunct. The Czech verbs *přemýšlet* (to think) and *diskutovat* (to

Table 7. Error analysis of *coordZeug*

	Reason for model confusion	**Count**	**%**
1	Unrecognized syntactically or semantically incompatible argument	6	54.5
2	Coordination of a verb with an elided argument of the first verb	2	18.2
3	Binding crossover at the first argument of a ditransitive verb	1	9.1
4	Uncertain causation	2	18.2

discuss) have an elided shared argument *o něčem* (about something). We assume that both these actions in relation to the elided object of discussion finish in the time that the adjunct expresses. However, this is again a subjective interpretation of the sentence.

(20) Proto Patricij s Bbloudem **přemýšleli a diskutovali až do oběda**. (Therefore, Patricius and Bbloud were thinking and discussing until lunchtime.) [coordSent → coordComArg]

5.3. Zeugma constructions, coordZeug

A considerably large group of misclassified ill-formed zeugma structures is formed by verb coordinations which allow for interpreting the sentence in the way where the first verb of the coordination has its valency complement omitted from the surface. An example can be seen in the sentence (21) in which the verb *porozumět* (to understand) forms a syntagm with the expression *problém* (problem) (i.e. *to understand the problem*), but at the same time, it could bind the complement *vám* (*you*) that elided from the surface as expressed in Example (21). Similar cases form 55 % of all errors in this class.

(21a) Abych lépe **porozuměl a prošetřil problém**. (To better understand and investigate the problem.) [coordSent → coordZeug]

(21b) Abych ~~vám~~ lépe **porozuměl a prošetřil problém**. (To better understand ~~you~~ and investigate the problem.)

A less frequent error group consists in a zeugma which the model recognized as a common argument, probably due to the lack of this kind of syntactic configuration in the training data, see Example (22).

(22) "Za toto je odpovědný policejní prezident, proto **žádáme a trváme na** jeho **odchodu**," zdůraznil John. ("The police president is responsible for this, so we demand and insist on his leave," John emphasized.) [coordComArg → coordZeug]

We have also identified a case in which the model classified the coordination as *coordComArg* where the coordinated verbs had a shared grammatically correct second argument in their structure. Nevertheless, the first argument diverged, causing cross-linkage. An example is in sentence (23), where the argument structure of the verb *doporučit* (to recommend) follows an obligatory frame *něco někomu* (something to someone), but the verb *chtít* (to ask) expects the valency of *něco po někom* (something from someone).

(23) Proto vláda přijala usnesení, kde **doporučila a chtěla po Ministerstvu obrany** garanci [. . .] (Therefore, the government passed a resolution recommending and asking the Ministry of Defence to guarantee [...]) [coordComArg → coordZeug]

6. Conclusions

In the presented paper, we have investigated the possibilities of machine learning methods for the task to distinguish grammatical and ungrammatical coordinated structures. A significant contribution is the creation of a benchmark ZeugmaDataset for fine-tuning and evaluating new language models and the new ZeugBERT language model based on it. We have proved that solving this task by deep learning techniques achieves remarkable improvements that ultimately outperform all approaches applied so far on zeugma detection reaching the accuracy of 88 % with the testing set. Additionally, we provide detailed error analysis where we discuss the patterns where the model made an error which will contribute to further improvements of the model's performance.

The examined phenomena occur across many languages. Even though we focus mainly on detecting non-grammatical structures in Czech, we assume comparable results for the equivalent structures in other languages because of the universality of the classification method. In general, zeugma detection may also be beneficial, for example, for the machine translation output checking.

In the future, we will focus on extending the current dataset to involve the coordination of other phrasal forms than just verb phrases. Since the deep learning method has proven effective for detecting ungrammatical structures, we will continue to develop foundational datasets for detecting other non-grammatical structures such as attraction or verb binding errors.

Acknowledgements

This work was partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101 and by Masaryk University in 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] A. Abeillé. In defense of lexical coordination. *Empirical issues in syntax and semantics*, 6:7–36, 2006.

[2] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[3] B. Crysmann. An asymmetric theory of peripheral sharing in HPSG: Conjunction reduction and coordination of unlikes. In *Proceedings of fgvienna: The 8th conference on formal grammar*, pages 45–64. CSLI Publications Stanford, California, 2003.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] J. Ficler and Y. Goldberg. Coordination Annotation Extension in the Penn Tree Bank. *arXiv preprint arXiv:1606.02529*, 2016.

[6] K. Gerdes and S. Kahane. Non-constituent coordination and other coordinative constructions as Dependency Graphs. In *Depling 2015*, Proceedings of Depling 2015, Uppsala, Sweden, 2015.

[7] J. Hrbáček. K otázce několikanásobného přísudku (On the topic of multiple prepositions). *Naše řeč*, 32(43):4–18, 1960.

[8] R. Huddleston and G. K. Pullum. *Coordination and Subordination*, chapter 9, pages 198–219. John Wiley & Sons, Ltd, 2006.

[9] R. Kaplan and J. Maxwell. Constituent coordination in lexical-functional grammar. 1, 01 2003.

[10] P. Karlík. *Nový encyklopedický slovník češtiny (New Encyclopaedic Dictionary of Czech)*, chapter Zeugma. 2017.

[11] P. Karlík and H. Gruet Škrabalová. *Nový encyklopedický slovník češtiny (New Encyclopaedic Dictionary of Czech)*, chapter Koordinace. Praha, 2017.

[12] G. Kempen. Conjunction reduction and gapping in clause-level coordination: an inheritance-based approach. *Computational Intelligence*, 7(4):357–360, 1991.

[13] S. Kübler, E. Hinrichs, W. Maier, and E. Klett. Parsing coordinations. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 406–414, 2009.

[14] M. Lopatková, V. Kettnerová, E. Bejček, A. Vernerová, and Z. Žabokrtský. *Valenční slovník českých sloves VALLEX (Valency dictionary of Czech verbs VALLEX).* Karolinum, Praha, 2016.

[15] H. Medková. Automatic detection of zeugma. In A. Horák, P. Rychlý, and A. Rambousek, editors, *Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2020*, pages 79–86, Brno, 2020. Tribun EU.

[16] H. Medková. Building a dataset for detection of verb coordinations with a shared argument. In A. Horák, P. Rychlý, and A. Rambousek, editors, *Recent Advances in Slavonic Natural Language Processing (RASLAN 2021)*, pages 125–133, Brno, 2021. Tribun EU.

[17] F. Mouret. A phrase structure approach to argument cluster coordination. In *The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, pages 247–267. Citeseer, 2006.

[18] P. Rychlý. A lexicographer-friendly association score. In *RASLAN 2008*, pages 6–9, Brno, 2008. Masarykova Univerzita.

[19] Y. Shen. Zeugma: Prototypes, categories, and metaphors. *Metaphor and symbol*, 13(1):31–47, 1998.

[20] M. Straka. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, 2018.

[21] V. Suchomel. csTenTen17, a Recent Czech Web Corpus. In P. R. Aleš Horák and A. Rambousek, editors, *Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018*, pages 111–123, Brno, 2018. Tribun EU.

[22] H. Teranishi, H. Shindo, and Y. Matsumoto. Decomposed local models for coordinate structure parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3394–3403, 2019.

[23] L. Uhlířová. O délce věty (About the length of a sentence). *Slovo a slovesnost*, 32(3):232–240, 1971.

[24] E. Viebahn. Ambiguity and zeugma. *Pacific Philosophical Quarterly*, 99(4):749–762, 2018.

[25] M. White. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75, 2006.

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.