Automatic Detection of Zeugma

Helena Medková

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

Abstract. The term zeugma denotes the linguistic phenomenon that means yoking together expressions with different argument structure, or with semantically incompatible meanings. For that reason, its detection is in many cases relatively difficult subject, mainly because three perspectives of language have to be considered: syntax, semantics and pragmatics. The presented paper describes a possible approach to the detection of this kind of structure. The output of the research will be part of the new online proofreader for Czech.

Keywords: zeugma, grammar checker, VerbaLex.

1 Introduction

Zeugma is the linguistic phenomenon occurring across many languages. Generally, linguists consider it as a stylistic figure in which two or more expressions with different meanings are forced together as in the example (1) [10]:

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

For that reason, the resulting structure sounds strange or funny. Several linguistic approaches also used so-called zeugma test for the ambiguity recognition. However, this method has proved to be not entirely reliable because an ambiguous expressions do not always lead to zeugmaticity and vice versa [10].

In the Czech language zeugma means in many cases a coordination of elements sharing a common argument. Nevertheless, the dependent expression complies only with one's element argument structure, so the resulting construction is non-grammatical [6].

This paper focuses mainly on non-grammatical sentences where the expressions have different argument structure. An illustrative example is in the sentence (2) where the coordinated elements comprise of predicates. The verb *pocházet* (*come*) should be accompanied with the preposition *z* (*from)* since the only obligatory complement of a verb *pocházet* (*come*) is the genitive prepositional phrase instead of local.

A proper formulation of such structure is a pronominal coreference (2a), where a personal pronoun substitutes the object of the second verb. This strategy also intensifies if the verbs express temporal succession or their meaning is not analogous [4].

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2020, pp. 79–86, 2020. © Tribun EU 2020

(2) *Pocházím a bydlím v Praze.* (*I come and live in Prague*)

(2a) *Pocházím* (*od / z*) *Prahy a žiju v ní.* (*I come from Prague and live in it.*)

When processing this task, it is necessary to distinguish semantics that sets verbal argumentative structure. Let us take the verb *navazovat*[1] as the example. In meaning *to follow up* something, it is essential to use the prepositional phrase. Otherwise, if it is used in a sense *to establish something*, it requires the object in the accusative (cf. sentences (3), (4)).

(3) *Navazuje a rozšiřuje publikaci…* (*Follows up and extends publications…*)

(4) *Mohou navazovat a prohlubovat své kontakty mezi sebou …* (*They can establish and deepen their contacts with each other…*)

Zeugma recognition is challenging for verbs with an implicit (zero) object because of utterance ambiguity. The knowledge of the author's intentions in the utterance context is fundamental for its correct interpretation. In these cases, it is complicated to determine, in particular, whether the coordinated verbs have a common argument.

While in the example (5) it is possible to find out quite clearly from the context of the verb *číst* (*read*) that the verbs share the same object, in the example (6) it is not accessible to judge. Without further knowledge of the plotted situation, it is not possible to determine precisely whether the pupils read the same text with which they later worked in the lesson.

(5) *… byť Chrome umí číst a pracovat s lokálními médii…* (*… although Chrome can read and work with local media…*)

(6) *Žáci v předchozí hodině četli a pracovali s textem na téma životní prostředí – třídění odpadů.* (*Pupils in the last lesson read and worked with the text about the environment – waste sorting.*)

2 Rule-based Detection of Zeugma

This paper presents a rule-based approach to zeugma detection. It involves creating a grammar with specific rules that check if there is a proper object in a proper form for a particular verb in the context of the sentence.

If the rule does not find a grammatical addition, parser marks structure as zeugma and send it to output.

The rules are focused mainly on the verb that has an incorrect phrase in a given sentence. In most cases, rules detect the unsuitable dependent located at the postposition of the coordination (as in the examples (2), (3), (5)). It implies that it is mainly the first verb in the conjunction, which binding needs to be corrected.

[1] Two meanings: *to follow up*, *to establish something*

The first rules were manually created within the diploma thesis [1] for 83 verbs. The disadvantage of this approach was that the rules covered only the non-grammatical structures for the verbs specified in the variables.

However, in the Czech lexicon, there are thousands of verbs. E.g. the lexical database VerbaLex [2] contains 10469 verbal lemmas and 19247 verbal patterns. It would be tedious to manually create rules for all the verbs contained in Czech lexicon.

Therefore, new grammar was generated based on manual grammar according to the verbal patterns in VerbaLex [2].

2.1 VerbaLex Processing

The intermediate step to obtain automatically generated grammar in this approach is to create a better processable data structure from the database. At first, it is necessary to get all single verbs from VerbaLex synonymous series

dok[2]: ověřit/ověřit si[3]:1 ned: ověřovat/ověřovat si:1

and its relevant parts of the frames (with an argument structure after VERB in right periphery)

+AG(kdo1[4];<person:1>;obl)+++VERB+++INFO(co4[5];<fact:1>;obl)

from the database into a dictionary):

```
{'ověřit': {'transitivity': '=canbepassive: yes',
            'first_objects_frames':
            [['INFO(co4;<fact:1>;obl)\n'],
            ['INFO(že;<info:1>;obl)'], [...]]}}
```

Grammar is generated after this processing phase.

2.2 Generated Grammar Organization

The grammar contains rules for four possible sentence arrangements, where zeugma can be recognized. Each of the four arrangements aggregates verbs with the same obligatory complements at the position of the first subject. There are 649 such aggregates in grammar, containing together 5300 verbs. There is also a potential for further expansion when we resolve the component issues associated with the processing.

[2] Aspect: pf. / impf.

[3] to verify

[4] nom_anim

[5] acc_inanim

An example of the generated rule. Rule illustration represents one aggregate with four potential structures that can be recognized.

In the variable *$verb(lemma)* are merged all verbs with the first obligatory accusative object in valency pattern according to VerbaLex.

The transitive verbs (in lexical database have label *=canbepassive: yes*, require another condition (*$verb(tag not): k5.*mN.**) to pass out all passive forms. Tags *bound* and *rbound* signal segment boundaries.

```
TMPL: bound $context* $verb (word a) (tag k5.*) $prep $noun
$context* rbound AGREE 2 4 mgn MARK 2 4 5 6 <zeugma>

TMPL: bound $context* $verb (word a) (tag k5.*) $refl $prep
$noun $context* rbound AGREE 2 4 mgn MARK 2 4 5 6 7 <zeugma>

TMPL: bound $context* $verb (word a) (tag k5.*) $refl $noun
$context* rbound AGREE 2 4 mgn MARK 2 4 5 6 <zeugma>

TMPL: bound $context* $verb (word a) (tag k5.*) $noun
$context* rbound AGREE 2 4 mgn MARK 2 4 5 <zeugma>

$verb(lemma): rdousit pohněvat setřepat ověřit ověřovat ...
$verb(tag not): k5.*mN.*
$prep(tag): k7.*
$context*(tag not): k3.*yF.* .*c4.*
$noun(tag not): k3.*yF.* .*c4.*
$noun(tag): k[123].*
```

The method is currently unreliable when the null object (6) occurs, and also in cases of the elided object as seen on (7). The antecedent of the verb *odmítnout* (*refuse*) is expressed in the first clause of the compound sentence. Therefore the rule recognizes it as zeugma.

> (7) *Po bulharské okupaci Makedonie dostal pozvání do probulharské vlády, ale odmítl a zapojil se do komunistického odboje.* (*After the Bulgarian occupation he got an invitation to probulharic government, but he refused and joined to communistic resistance.*)

Nevertheless, there is a possibility to recognize whether two verbs in a sentence have a common object. One of the solutions involves statistics using collocations. As can be seen in the example (7), the verb *odmítnout* (*refuse*) is not semantically compatible with the complement *do komunistického odboje* (*into the communist resistance*) of the verb *zapojit* (*join*).

With the knowledge of the verb and object collocability, it will be possible to determine whether the object belongs to both predicates in coordination or only to one. It will help to get more accurate results.

An alternative solution to this is the creation of a text preprocessing tool that can learn how to classify coordinations, where the dominant expression shares a common addition.

The tools of Nature language processing centre (CZPJ FI MU) are applied for detection purposes. Namely the morphological tagger Majka [9], disambiguator Desamb [8] and SET parser [7] are used for syntactical analyses.

3 Standard Data Set for Zeugma Detection

The data set consists of the sentences manually selected from the cztenten17 [5] corpus. As the zeugma is relatively challenging to find, an error was intentionally made in 17 sentences. Although there was an effort to get all of the sentences authentic, two clauses were utterly made up. Labels *korpus, upraveno,* * (*corpus, edited,* *) indicate the origin of the sentences.

The basis of data set forms 750 sentences with zeugma (83 various verbs) collected within the diploma thesis *Automatic detection of non-grammatical constructions in Czech* [1].

The original data set was annotated and expanded to 1013 positive cases of zeugma. To each verb in incorrectly coordinated construction were also added ca. 20 negative ones (giving a total number of 1681). 1137 of those sentences containing coordination or zeugma also have concurrent ones, which provide additional context to them. The whole data set now comprises of 5313 sentences.

Table 1. Data set – statistical data

Data set statistics	Count
Sentences in data set	**5313**
Sentences with correct coordination	**1681**
Sentences with Zeugma	**1013**
Words	**79379**
Verbs	**84**

In the data set, there are only zeugmas formed by verbs connected by the conjunction 'a'. Therefore there is a necessity to enlarge and balance it in the future. Nevertheless, the data set makes possible monitoring the quality of rules even now.

3.1 Evaluation of Grammars

The python script automatically evaluates grammar quality. As the input files, it takes .csv data set and the SET output file. The SET parser output consists of the defective coordinations labelled as <*zeugma*> and the whole sentence with label <*sentence*>:

<zeugma> (k5eAaImIp3nS): zkoumá a bádá nad artefakty
<sentence> (kIx.): Archeolog zkoumá a bádá nad artefakty .[6]

The program converts the SET output into the matrix with actual values (1 – positive and 0 – negative), as it also similarly evaluates the data set with predicted values. The script then creates the confusion matrix and counts the required measures.

Table 2. Data set – comparison of evaluations

Grammar	Precision	Recall	F-score	Accuracy
Manual grammar	**0.982**	**0.381**	**0.549**	**0.765**
Generated grammar	**0.796**	**0.224**	**0.350**	**0.687**

Grammar control requires the most accurate results in the first place. That is the reason why the precision measure is for proofreader [3] purposes more important than the recall. As seen in table 2, manual grammar has on the data set relatively accurate results.

However, testing on the part of czTenTen17 corpora in diploma theses [1] showed precision score 0,633 (without mistakes in morphological tags, it was 0,869). Enlarging data set primarily with negative cases of zeugma will provide significantly more reliable results.

Although automatically generated rules in the grammar are generic, without special conditions for achieving better scores, testing on the data set has shown promising results for further experiments with this approach.

However, there is a number of issues for consideration. The first significant problem is recall. Manually created grammar covers approximately 38 % of defective structures. The aim was to get more precise results for the price of lower coverage. Generically created rules have even ten percent lower recall. This approach allows covering a more considerable amount of Czech verbs. However, the set restrictions in the rules do not enable matching each structure in the data set. It also significantly reduces the recall.

Grammar does not allow finding structures as in example (8). The current rule is limited with the condition, that there must be no suitable addition for the verb *doporučit* (*to recommend*) in the whole sentence context.

(8) *Doporučíme a vybereme s vámi rostliny...* (*We recommend and choose with you the plants...*

(9) *... zde bych doporučil a odkázal na Castanedu* (*... in this place I would recommend and refer to Castaneda*

[6] An archaeologist examines and researches artifacts.

Also in example (9) is shown, that rules do not recognize the zeugma, when the expected object of the first verb is in the accusative, but the predicates share a prepositional phrase with the dominant noun in the accusative (or vice versa (10)).

(10) *Dohlíží a prověřuje faktury....* (*He supervises and checks out the invoices...*)

In the future, it is necessary to focus on enlargement of the rules patterns and also choosing contextual restrictions in a more targeted way.

Testing on the czTenTen17 corpus revealed a large amount of errors in morphological tags. Contextual ellipses or vaguely formulated rules caused another relevant errors.

4 Summary and Future work

This paper has presented the rule-based approach to zeugma detection that proposes a manual grammar, and also grammar automatically generated from the lexicon of verb valencies VerbaLex. The standard annotated data set was created for purposes of this research, that extended earlier data from the master thesis. Further work involves mainly enlargement of the current data set. The inclusion of semantic perception, with the usage of semantic roles in VerbaLex, collocations or machine learning, also has to be considered. The increment of recall and precision of grammar is necessary as well.

Acknowledgements. This work was supported by the project of specific research *Čeština v jednotě synchronie a diachronie* (Czech language in the unity of synchrony and diachrony; project no. MUNI/A/0913/2019) and by the Technology Agency of the Czech Republic under the project TL02000146.

References

1. Geržová, H.: Automatická detekce negramatických větných konstrukcí pro češtinu (in Czech, Automatic detection of non-grammatical constructions in Czech). Master's thesis, Masaryk University, Faculty of Arts, Brno (2019 [cit 2020-10-30]), https://is.muni.cz/th/fuz2y/
2. 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). ISBN 80-224-0895-6.
3. Hlaváčková, D., Hrabalová, B., Machura, J., Masopustová, M., Mrkývka, V., Valíčková, M., Žižková, H.: New Online Proofreader for Czech. In: Horák, Aleš; Rychlý, Pavel; Rambousek, Adam (eds.): Slavonic Natural Language Processing in the 21st Century. p. 79-92, 14 pp. Brno: Tribun EU (2019). ISBN 978-80-263-1545-2.
4. Hrbáček, J.: Společný předmět u dvou slovesných přísudků (in Czech, Two verbal predicates with common object). Naše řeč. **47**(2), p. 118–120. Ústav pro jazyk český AV ČR, (1964). http://nase-rec.ujc.cas.cz/archiv.php?lang=en&art=5022

5. 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, July)
6. Karlík, P.: Zeugma. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy – Nový encyklopedický slovník češtiny, (2017). `https://www.czechency.org/slovnik/ZEUGMA` Last accessed 26 Oct 2020
7. Kovář, V., Horák, A., Jakubíček, M.: Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. In: Human Language Technology. Challenges for Computer Science and Linguistics. p. 161-171, 11, pp. Berlin/Heidelberg: Springer (2011). ISBN 978-3-642-20094-6.
8. Šmerk, P.: K morfologické desambiguaci češtiny (in Czech, Towards morphological disambiguation of Czech). PhD thesis, Masaryk University, Faculty of Informatics, Brno (2007)
9. Šmerk, P.: K počítačové morfologické analýze češtiny (in Czech, Towards Computational Morphological Analysis of Czech). Ph.D. thesis, Faculty of Informatics, Masaryk University (2010)
10. Viebahn, E.: Ambiguity and Zeugma. Pacific Philosophical Quarterly. 99. 10.1111/papq.12229, (2018). `https://onlinelibrary.wiley.com/doi/pdf/10.1111/papq.12229?casa_token=`