Syntactic analysis of the Slovak sentence

Michaela Vočková and Stanislav Krajči

Institute of Computer Science, Pavol Jozef Šafárik University in Košice, Slovakia,
michaela.vockova@student.upjs.sk

Abstract: The natural language processing is recently a very discussed topic in computer science. The main idea is an understanding of human languages by computers. In this work-in-progress paper, we propose the algorithm for creation of a tree structure of the Slovak sentence. The tree structure of a sentence represents the relationships and dependencies between words in a sentence. The root of the tree is a predicate. Understanding a structure of sentence is important for other natural language processing tasks, such as semantic analysis. There are many different types of sentences in the Slovak language, which we took into account for creating the algorithm. For example, a multiple sentence member, compound sentence, compound predicate and others. Our algorithm correctly analysed 85 sentences from 100 different sentences.

1 Introduction

Natural language processing is part of artificial intelligence and linguistics, focusing on understanding human language by computers. There are different tasks in natural language processing:

- Automatic summarization provides summaries or detailed information of text of a known type.
- Co-reference resolution refers to a sentence or more extensive set of text determining which word refers to the same object.
- Discourse analysis refers to the task of identifying the discourse structure of a text.
- Machine translation refers to automatic translation of text from one human language to another.
- Morphological segmentation refers to separate words into individual morphemes and identifies the class of the morphemes.
- Named entity recognition describes a stream of text and determines which text items relate to proper names.
- Optical character recognition gives an image representing printed text, which helps determine the corresponding or related text.
- Part of speech tagging describes a sentence, determines the part of speech for each word.

Some of the tasks can be used as a subtask for more complex assignments [1].
Semantic and syntactic parsing is also part of natural language processing, aiming to provide internal relations between words. There are two approaches for finding the structure of sentence: constituent parsing and dependency parsing. Constituent parsing provides a constituent tree where nodes are phrases. The goal is to find these phrases and their relations. The approaches of constituent parsing include the chart-based and the transition-based models. Both have statistical and neural models. Dependency parsing is using bilexicalized dependency grammar, which contains all semantic and syntactic dependencies. Dependency parsing models are divided into two groups: graph-based models and transition-based models, both of which have their own statistical or neural network approaches [2].
This work-in-progress paper proposes the improvement of algorithm for creation of a tree structure of the Slovak sentence [19]. This algorithm is not based on statistical data from the corpus, but takes raw data from Tvaroslovník. It is a database of all forms of Slovak words. The tree structure of a sentence can represent the relationships and dependencies between words in a sentence. The root of the tree is a predicate. The tree structure for Slovak sentence: `Hodina dnes začala malým kvízom.`[1] is shown in Figure 1.

2 State of Art

Institute of Formal and Applied Linguistic at Charles University in Prague has created the Prague Dependency Corpus, which is an excellent contribution to natural language processing. Several tools have been developed to find out a sentence structure or work on other natural language processing tasks based on this corpus or Universal Dependency Treebank. For example [3]:

- Netgraph – this is a graphically oriented client-server application for searching in an annotated corpus.
- TrEd – an editor used to search for a syntactically annotated sentence structure.
- Morfo – a system for morphological analysis of the Czech language.

[1] The class starts with a small quiz today.

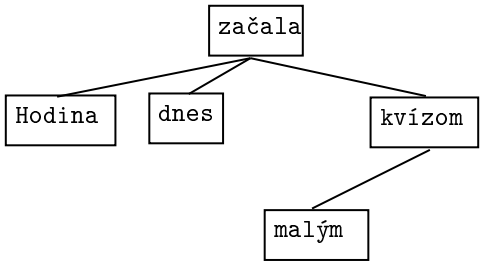


Figure 1: Example of sentence tree structure for `Hodina dnes začala malým kvízom.`[1]

- MorfoDita – a free tool for morphological analysis of natural language texts.
- Moses – a statistical machine translation system that automatically allows training translation models for any language pair.
- UDPipe – a trainable channel for tokenization, labeling, lemmatization, and relationship analysis. Institute developed two version of UDPipe [4], [5].

The Natural Language Processing Centre at Masaryk University in Brno is mainly engaged in research into the processing of the Czech, English, and Slovak languages. They deal with morphological, syntactic, and semantic analysis and the creation of corpora and dictionaries. The institute has created several tools that work with morphological, syntactic, and semantic analysis. Examples include [6]:

- Majka – morphological analyzer for Slovak, Czech, Polish, Swedish, German language.
- The Sketch engine – a tool used to search for information from text corpora.
- CZ accent – a tool for adding accents to text.
- Synt and SET – parsers used to determine the structure.
- Visual Browser - Java software that visualizes data into RDT format.

Institute of Theoretical and Computational Linguistics at Charles University develops computational tools for automatic language processing, for example, syntactic annotation of Czech corpora or grammar-based treebank of Czech language. [7].
Similar to the Czech language, there are several tools, dictionaries, and conferences in natural language processing research in Slovak languages. Language Institute of L'udovít Štúr offers a wide selection of dictionaries. These include a [8], [9], [10] and much more [11]. It also provides the Slovak National Corpus. It is an electronic database, mainly containing Slovak texts from 1955 from different styles, genres, thematic areas, region and other. Language Institute of L'udovít Štúr developed tools for searching words in Slovak National Corpus and working with them. For example, DEVELOPER visualizes an occurrence of one or two words in the corpus. DIAKRITIK corrects the diacritics, and KOLOKAT visualizes distances between two terms in the corpus [12]. Every two years, the institute organizes a conference SLOVKO on natural language processing [13]. In 2017, D. Zeman presented an article Slovak Dependency Treebanks in Universal Dependencies about converting the syntactically annotated part of the Slovak National Corpus into the annotation scheme known as Universal Dependencies. Universal Dependencies is an international standard and also the largest database of freely available dependency treebank[14]. Database of Slovak words and their forms Tvaroslovník was created at Pavol Jozef Šafárik University at Košice [15], [16]. Master thesis [17] deals with the creation of an algorithm for finding the structure of the sentence.

3 Dictionaries

It is necessary to have more information about words to create a sentence structure. Therefore we are using the dictionary Tvaroslovník and Valency dictionary for our algorithm of syntactic analysis.

3.1 Tvaroslovník

Tvaroslovník is a database of all forms of all Slovak words from [8] and [9]. Every row contains information about form of the word, its part-of-speech and grammatical categories of the word. Data in Tvaroslovník was collected from the dictionary of Slovak language. Database contains approximately 220,000 words and 24,000,000 records of words and all their forms. All data and information are saved in one table. There is a list of columns:

- *idWord* – unique identification number for word,
- *idForm* – unique identification number of word's form,
- *form* – a form of a word,
- *part-of-speech*,
- *categories* – grammatical categories, there are different for every part-of-speech.

Table 1 shows an example of records for the word `hodina` [2].

3.2 Valency dictionary

Valency dictionary contains two types of the most common covalence between words. First is covalence between verb and preposition or verb and the most common case of the following term. Covalence between noun and preposition is the second type of valency dictionary. To built the valency dictionary, we took noun and verbs from Tvaroslovník and covalencies with prepositions and cases were automatically created from examples in Krátky slovník slovenského jazyka [18]. Dictionary cointans columns:

- *idWord* — unique identification number for word from Tvaroslovník,
- *preposition* — preposition which follow after noun or verb,
- *case* — case of word after noun or verb.

Table 2 illustrates examples from dictionary of covalence.

4 Tree structure of sentence

We presented the main idea of the algorithm for finding the tree structure in the article [19]. For the algorithm, we expanded the table of relations and added cases of Slovak sentences, which we describe in the subsection Special cases of sentences. Table 3 illustrates the new relationship table, and algorithm 1 describes the pseudocode for the main idea of the tree finding algorithm.

4.1 Special cases of sentences

Slovak is a flexible language and has many peculiarities that we took into account when creating the method.

- **Multiple sentence member:** The first is multiple sentence members. We find out whether there is a conjunction or a comma in the sentence during searching for initial possible relations. If so, we look at the word before and after the conjunction if it is the same part of speech and has the same grammatical categories. After fulfilling the condition, we add a relation between conjunction and the words to the possible relations. The conjunction then takes over the grammatical categories of the words it connects. For example, in sentence `Noviny a časopisy píšu o celebritách.` [3] words `noviny` and `časopisy` are same sentence member, therefore there are relations `noviny` and `a` with priority 12 and `časopisy` and `a` with priority 12 in the list of possible relations. Word `a` participates as noun in nominative case.

[2] hour

[3] Newspapers and magazines write about celebrities.

```
input: sentence
output: tree structure of sentence
find all forms for words in sentence from Tvaroslovník;
create list of possible relations;
while list of possible relation is not empty or sentence
 has only one word do
    choose relation with greatest priority;
    add chosen relation to list of final relations;
    remove chosen relation from list
    of possible relations;
    foreach relation in list of possible relation
    do
        if relation has same dependent
        and different superior word as chosen
        relation then
            remove relation from list
            of possible relations;
        end
    end
    remove dependent word of chosen relation
    from sentence;
    if new possible relation is created then
        add new relation to list of possible
        relations;
    end
end
build tree structure from list of final relations;
```

Algorithm 1: Pseudocode for finding tree structure algorithm

- **Multiple verbs in sentence:** Occurrence of several verbs in a sentence is another specification of the sentence. Before we start looking for possible relationships in a sentence, we determine if this is not the case. After determining verbs, we search whether a conjunction or a comma is in the sentence between them. Finding a comma or conjunction classifies a sentence as a sentence. Therefore, we divide the sentence according to the conjunction or comma into subsections with which we work as separate sentences. We connect these sentences with the relationships between the conjunction or comma and the roots of subsentences in the resulting output. Figure 2 shows us example of sentence structure for sentence `Mama číta noviny a otec píše správu.` [4] In a sentence containing more verbs without conjunction or comma between them, we assume that there is a compound verb relation. Therefore, we combine the found verbs with the relation and add them to the list of possible relations. Figure 3 shows us example of such sentence structure for sentence `Ráno začalo pršať.` [5]
- **Same form of word:** Some words have the same form in several cases, so it is sometimes difficult to determine which relationship they can form. We find

[4] Mother is reading newspapers and father is writing an message.

[5] It started to rain in the morning.

idWord	*idForm*	*form*	*part-of-speech*	*categories*
20009	0	hodina	noun	gender: feminine; number: singular; case: nominative
20009	1	hodiny	noun	gender: feminine; number: singular; case: genitive
20009	2	hodine	noun	gender: feminine; number: singular; case: dative
20009	3	hodinu	noun	gender: feminine; number: singular; case: accusative
20009	4	hodina	noun	gender: feminine; number: singular; case: vocative
20009	5	hodine	noun	gender: feminine; number: singular; case: locative
20009	6	hodinou	noun	gender: feminine; number: singular; case: instrumental
20009	7	hodiny	noun	gender: feminine; number: plural; case: nominative
20009	8	hodín	noun	gender: feminine; number: plural; case: genitive
20009	9	hodinám	noun	gender: feminine; number: plural; case: dative
20009	10	hodiny	noun	gender: feminine; number: plural; case: accusative
20009	11	hodiny	noun	gender: feminine; number: plural; case: vocative
20009	12	hodinách	noun	gender: feminine; number: plural; case: locative
20009	13	hodinami	noun	gender: feminine; number: plural; case: instrumental

Table 1: Tvaroslovník

idWord	*preposition*	*case*
6016	null	accusative
6016	proti	dative
31494	v	locative
31494	null	accusative
31494	null	instrumental
62420	null	accusative

Table 2: Examples of covalencies for noun and verbs

all possible relations for the word. In the method where we gradually iterate over the list of possible relations and remove relations with the same dependent word as the currently selected relation, we locate a relation with the same dependent and superior word but with a different priority. We create another list of final and possible relations assigning a relation with a different priority. The method then outputs two trees. Figure 4 illustrates the two possible outputs for sentence Dievča upieklo mame perníkové srdce.[6]

- **Different part-of-speech for same form:** Expect a word having the same form in multiple cases may also have the same form for multiple parts of speech. For example, the word to is a pronoun and particle. We created a list that contains the most commonly used part of speech for these words. If we set the method to find only the most relevant sentence structures, we use only the most often used part of speech for a form.

[6]The girl baked a gingerbread heart for mum.

Dependent	*Superior*	*Priority*	*Required grammatical categories*
verb	auxiliary verb	13	none
noun, adjective, pronoun, numeral	auxiliary verb	13	none
verb	conjunction	12	none
noun	conjunction	12	none
adjective	conjunction	12	none
pronoun	conjunction	12	none
numeral	conjunction	12	none
adverb	conjunction	12	none
adverb	adverb	11	none
adverb	adjective	11	none
pronoun sa, si	verb	10	none
pronoun	adjective	9	none
adjective	noun	8	same gender, case and number
numeral	noun	8	same gender, case and number
pronoun	noun	8	same gender, case and number
noun	noun	7	case of dependent noun is accusative
noun	noun	6	case of dependent noun is genitive
adjective	preposition	5	same case
pronoun	preposition	5	same case
noun	preposition	4	same case
preposition	noun	4	noun and preposition are together in valency dictionary
pronoun	verb	3	case of pronoun is not in valency dictionary and pronoun shouldn't be in the nominative case
noun	verb	3	case of noun is not in valency dictionary and noun shouldn't be in the nominative case
adjective	verb	3	case of adjective is not in valency dictionary and adjective shouldn't be in the nominative case
numeral	verb	3	case of numeral is not in valency dictionary and numeral shouldn't be nominative case
pronoun	verb	2	case of pronoun is in valency dictionary and pronoun shouldn't be in the nominative case
noun	verb	2	case of noun is in valency dictionary and noun shouldn't be in the nominative case
adjective	verb	2	case of adjective is in valency dictionary and adjective shouldn't be in the nominative case
numeral	verb	2	case of numeral is in valency dictionary and numeral shouldn't be nominative case
adverb	verb	2	none
noun	verb	1	noun should be in the first case
adjective	verb	1	adjective should be in the first case
pronoun	verb	1	pronoun should be in the first case
numeral	verb	1	numeral should be in the first case

Table 3: Relations and their priorities

5 Conclusion and future research

To analyze the algorithm for creating a tree structure, we built a dataset with 100 different Slovak sentences. Sentence are taken from fairy-tales and articles on Internet. Dataset contains:

- simple sentences: `Martin zavrtel hlavou.`[7],
- simple sentences with different sentence members: `Chlapec vykročil z tieňa tmavých jedlí na čistinku uprostred lesa.`[8],

[7] Martin waved his head.
[8] The boy walked out of the shadows of dark firs to a clearing in the

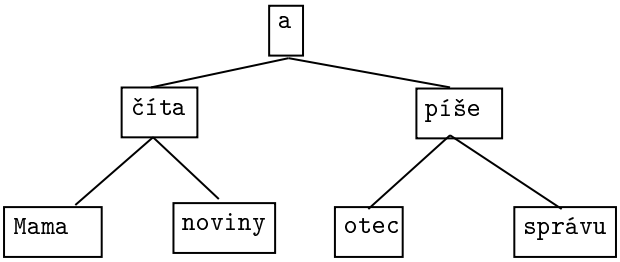


Figure 2: Example of sentence tree structure for `Mama číta noviny a otec píše správu.`[4]

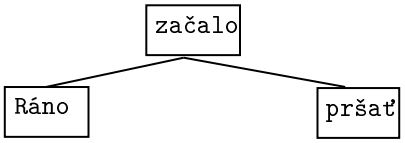
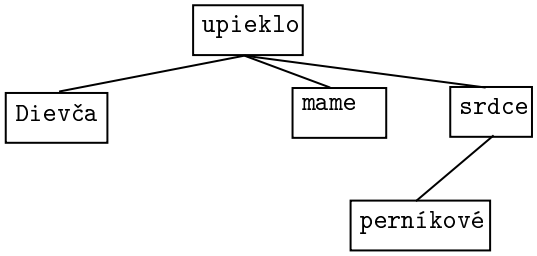
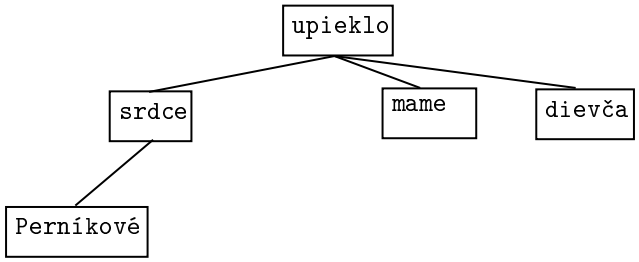


Figure 3: Example of sentence tree structure for `Ráno začalo pršať.`[5]



A



B

Figure 4: Example of two possible outputs for sentence `Dievča upieklo mame perníkové srdce.`[6]

- compound sentences: `Teší sa z jeho krásy a užíva si pokojný relax.`[9],
- sentences with multiple sentence member: `Uprostred hlučného a ubehaného mestečka leží krásny zelený park.`[10],
- sentences with compound predicate: `V mestskej časti si môžu návštevníci užiť kúpalisko.`[11].

We created this dataset manually. To each sentence, we added the required tree structure. As a result, we received 85 identical tree structures. The main difficulties for finding incorrect structure were:

- Digital number in a sentence. For example, `Hrad vznikol pravdepodobne v druhej polovici 13. storočia.`[12]
- Changing the position of words in a nominal predicate. For example, `Vhodná je paralela z čias môjho starého otca.`[13]

In our future work we want to focus on:

middle of the forest.

[9]She enjoys its beauty and enjoys peaceful relaxation.

[10]In the middle of a noisy and deserted town lies a beautiful green park.

[11]Visitors can enjoy the swimming pool in the city.

[12]The castle was probably built in the second half of the 13th century.

[13] A parallel from my grandfather's time is appropriate.

- eliminating the above problems
- testing method on other sentences
- creating a web interface for this algorithm

References

[1] Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: State of the art, current trends and challenges. 2017. arXiv preprint arXiv:1708.05148.

[2] Zhang, M.: A survey of syntactic-semantic parsing based on constituent and dependency structures. Science China Technological Sciences (2020): 1–23.

[3] https://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/html/index.html. (Accessed on 06/10/2021)

[4] Straka, M., Straková, J., Hajic, J.: Prague at EPE 2017: The UDPipe system. 2017. In Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation at the Fourth International Conference on Dependency Linguistics and the 15th International Conference on Parsing Technologies. Pisa, Italy (pp. 65–74).

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

[6] https://nlp.fi.muni.cz/en/NLPCentre. (Accessed on 06/10/2021)

[7] http://utkl.ff.cuni.cz/en/utkl.html. (Accessed on 06/10/2021)

[8] Peciar, Š.: (Ed.) Slovník slovenského jazyka (Vol. 4). Vydavatel'stvo SAV. 1964.

[9] Kraus, J.: Slovník cudzích slov: akademický. Slovenské pedagogické nakladateľstvo. 2005.

[10] M. Považaj. a kol.: Pravidlá slovenského pravopisu. 4. nezmenené vyd. Bratislava. Veda 2013. 592 s. ISBN 978-80-224-1331-2

[11] https://slovnik.juls.savba.sk/. (Accessed on 06/10/2021)

[12] Garabík, R.: Slovenský národný korpus. 2020. Acceseed on https://korpus.sk/.

[13] https://korpus.sk/slovko.html. (Accessed on 06/10/2021)

[14] Zeman, D.: Slovak dependency treebank in universal dependencies. 2017. Journal of Linguistics/Jazykovedný casopis, 68(2), 385–395.

[15] Krajči S., Novotný R.: Tvaroslovník – databáza tvarov slov slovenského jazyka. In zborník príspevkov z pracovného seminára ITAT. 2012.(pp. 57–61).

[16] Krajči S., Novotný R.: Projekt Tvaroslovník – slovník všetkých tvarov všetkých slovenských slov. Znalosti 2012. 2012. 2012. pp. 109–112.Vydavatelství MFF UK.

[17] Hiľovská, J.: Syntaktická analýza slovenskej vety pomocou Tvaroslovníka. UPJŠ.2017.

[18] Kačala, J.: (Ed.) Krátky slovník slovenského jazyka. Veda. 1987

[19] Linková, M., Krajci, S.: Tree structure of Slovak sentences. 2020. In Proceedings of the 20th Conference Information Technologies – Applications and Theory.(pp. 67–74).