

A Step toward Machine Recognition of Complex Sentences

Marko Orešković¹, Juraj Beničić², Mario Essert²

¹National and University Library in Zagreb, Hrvatske bratske zajednice 4, Zagreb, Croatia
²Faculty of Mechanical Engineering and Naval Architecture, Ivana Lučića 5, Zagreb, Croatia

Abstract – This paper presents theoretical and technological background of a model for machine recognition of complex sentences. It is based on the Syntactic and Semantic Framework (SSF) which implements fundamental linguistic fields network resources and encyclopedias. It can be used to extract subject, predicate and object, as well as other sentence's parts (e.g. NP/VP/PP), and in some cases even semantic roles. In compound sentences the machine can easily recognize independent sentences, whereas in complex sentences the machine recognizes the main clause and the related subordinate clauses as well as sentence types (subject, object, predicate, etc.). Using stored patterns various theories can be tested.

Keywords – complex and compound sentences, machine recognition, syntactic patterns, frequency analysis, computer model.

1 Introduction

This paper deals with several new methods for extraction of complex sentences. The term 'complex sentence' denotes a sentence consisting of two or more 'simple sentences'. For traditional Croatian grammar [1], and also not so obsolete functional grammar [2], i.e. lexical functional syntax [3], the

basic structure of the sentence is formed from functional elements: subject, object, predicate, and adverbials, which, have the role [4] of transferring the information (spoken or written) between the sender and the recipient. These functions can be extended with various complements.

Another approach is formalistic which was introduced by the Generative grammar [5], i.e. 'phrase structure grammar', and is based on the observation of the pronounced parts as statements based on parts of speech. The statement or phrase is a word or set of words which acts as a whole. Phrases are dependent units, but by mutual interconnecting (according grammatical rules) can become independent. Independent phrases which render complete meaning are called clauses. A necessary, but not sufficient condition for clause is to contain verbal phrase. One or more clauses giving an independent meaningful entity is called a sentence.

The main clauses are such clauses which, as sentences, can stand alone, whereas subordinated clauses cannot. Two or more main clauses connected with conjunctions or coordinators, form a complex sentence which, due to the independence of the individual parts, is called a compound sentence. In this series, coordinators of independent clauses may be conjunctive, disjunctive or negative, and may be represented by a comma only. It is important that any independent clause is also standalone (i.e. that can be understood independently of the others).

Subordinate clauses are dependent and cannot stand alone, and together with a main clause form an entity which is called a complex sentence. Subordinated clauses act as embedded entities, whether they substitute or complement a part of the sentence.

A more systematic categorization of clauses is performed through syntactic functions in sense of predicate-argument strings, where the functional categorization S-P-O (subject, predicate, object) is described by clauses and their types, for example, WH (who, whom, why, where) and SV (subject-verb constructions), thus, clauses act as arguments, adjuncts or predicative expressions. A special type of clauses is called relative clauses which are divided

DOI: 10.18421/TEM74-20


<https://dx.doi.org/10.18421/TEM74-20>

Corresponding author: Marko Orešković,
National and University Library in Zagreb, Hrvatske
bratske zajednice 4, Zagreb, Croatia
Email: moreskovic@nsk.hr

Received: 04 September 2018.

Accepted: 06 November 2018.

Published: 26 November 2018.

 © 2018 Marko Orešković, Juraj Beničić, Mario Essert; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.

The article is published with Open Access at www.temjournal.com

from the main clauses by relative pronouns and use its morphological form to realize their syntactic function. In other words, it means that clauses in compound sentences act as verb/noun modifiers.

2 Computer analysis of complex sentences

The usage of computers in analysis of sentences, and especially in complex sentences, is a rather demanding task from both aspects - algorithmic design and programmatic implementation. It is necessary to provide the machine with all needed information to make the sentence machine-readable, that is, it needs to have the same characteristics it has for a human when he reads or writes it. Therefore, it is necessary to extract all words from the sentence, and then tag each word with proper part of speech tags, grammatical categories, etc. Only then, at the level of binary information that represents the individual word in the sentence, parsing of phrases and clauses is performed, which is a foundation for their interrelationships study. Assigning of tags to some word is usually performed by statistical methods with partial knowledge of the morphosyntactic characteristics of the word in the observed language. It is obvious that this approach cannot deal with word ambiguity which every language has.

Another approach is to use the lexicon, which will have all words tagged with all possible tags (properties), and by using already prepared syntactic and semantic patterns, determine which word from the lexicon needs to be used in a particular case. Presently, our Syntactic and Semantic Framework (SSF) lexicon holds over 800,000 words and over 130,000 multiword expressions which are tagged with hierarchical structure of grammatical (WOS) and semantical (SOW) tags. Building of quality syntactic patterns and their recognition in sentences is based on that, ontology like, tree structure.



Figure 1. The word cube (in Croatian/hrv. 'kocka') and the related tags in the SSF's lexicon

The SSF contains enriched information from various network resources, e.g. Croatian Language Portal (<http://hjp.znanje.hr/>), The Miroslav Krleža

Institute of Lexicography online encyclopedia (<http://www.lzmk.hr>), CroWN – Croatian version of the WordNet, and also, due to its presence in the Global Linguistic Linked Open Data Cloud (<http://lod-cloud.net>), the world largest encyclopedia BabelNet (<http://live.babelnet.org/>) – on 284 languages, 1.307.706.673 lexical and semantic relations, 72.542.300 definitions, etc.

As shown in Figure 1., the SSF's lexicon along with the word itself, also contains additional information (e.g. words lemma and internal morphological structure, various tags like definitions in which every word is additionally linked to the lexical representation of it). WOS and SOW tags are organized as hierarchical T-structure [6] with ontological role. In addition to the classical lexicon, the SSF also contains subatomic lexicon of syllables, morphemes and syllable morphemes, which is useful in analysis of texts from the phonological-morphological aspect, and molecular lexicon of multiword expressions (collocations, phrases, etc.) which is mostly used in syntactic and semantic analysis.

3 Independent and dependent clauses

Two or more simple sentences which are formed from one or more clauses can be sequenced or merged into a larger sentence - an independently composite sentence. In the case of sequencing, sentences are separated by comma (so called asyndetic independent sentences), whereas in the process of merging, conjunctions are used for bonding and these sentences are called syndetic independently composite sentences. The process of extraction of independent clauses can be performed in these steps:

1. From the selected corpus take a sentence one by one. For the machine, the first problem is to extract sentences since the punctuation is not the universal delimiter (e.g. abbreviations may contain punctuation, and the machine must know that in such cases they do not represent the sentence ending).
2. Once the sentence is extracted, the next problem is how to differ comma in asyndetic sentences and the comma which delimits words from some phrase. The similar applies to the conjunction *and* since its role is twofold (e.g. in the sentence 'John and Mary are picking flowers' where it joins subjects or in the sentence 'John cries and Mary is surprised', where it delimits clauses).
3. Therefore, it is necessary to extract sentence parts first (at least subject and predicate), and then do the analyses of punctuations which delimit sentences, and determine sentence type.

The program works correctly for any number of clauses which is shown in Figure 2.

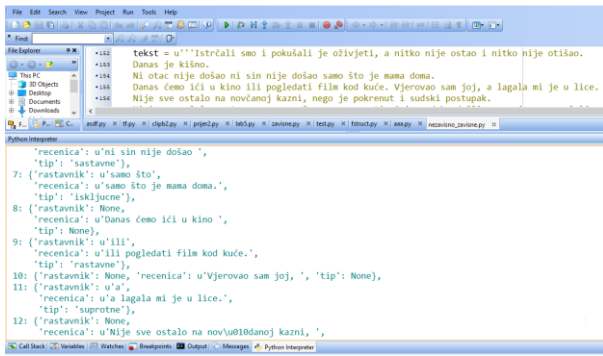


Figure 2. Extraction of independent complex sentences

Extraction of dependent sentences is much more complicated since the traditional Croatian grammar still mixes syntactic and semantic categories. On the other hand, the position of clauses in a dependent sentence (whether it is in front of, or behind, or in the middle of it), has focused the computational analysis towards building of syntactic patterns which can contain both WOS and SOW tags.

4 Regular expressions and syntactic patterns

Regular expressions (REGEX) are meta tags which are used to define and extract structure from unstructured environment.

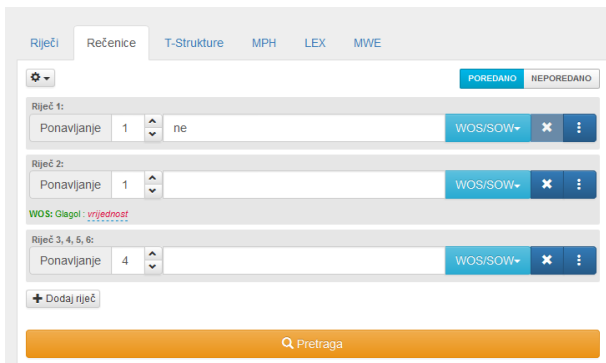


Figure 3. The pattern generator in the SSF

The SSF has a special module which enables user to create such patterns without any technical knowledge of regular expressions. The module is made for linguistic experts who thus have the ability to create patterns, e.g. a pattern that will find the subject in the sentence: a noun or noun phrase in the nominative, an infinitive verb if the noun is absent or adjective if the noun is missing. Of course, if none of these three or more variants doesn't return the subject word, it is a non-subjective sentence. Creation of a pattern in the module is conceived and realized in a very intuitive way. The user fills the WOS/SOW tags for each word, along with the part of the word that

needs to be matched. If the user wants one or more words to be ignored in the matching process he can enter a regular expression '+' in the text box (or leave it empty). By defining patterns that will have additional refinement for any of the categories (e.g. person, gender or number), those syntactic entities that meet congruence can be extracted, which is extremely valuable in (future) verifications of grammatical correctness of texts. To achieve all that, it is necessary to systematically and professionally develop a set of patterns for Croatian language.

For this paper, the set of syntactic and semantic patterns were constructed, which roughly correspond to the main features of a larger number of dependent clauses and their categorization by domestic grammars. The algorithm is developed only for dependent (subordinate) sentences that have one main and one subordinate clause, regardless of their composition. Figure 4. shows such patterns in the SSF module which are used to extract dependent sentences in the Croatian language. Similar patterns can be made for any natural language.

Table showing syntactic patterns for subordinate clauses in the Croatian language. The table has columns for pattern ID, WOS/SOW tags, and the corresponding regular expression. The patterns are numbered from 208 to 232.

Figure 4. Syntactic patterns for subordinate (dependent) clauses in the Croatian language

There are certain differences in grammars: Težak and Babić [1] have 'comparative' dependent sentences and don't have 'appositional', whereas Bičanić et al. [7] have the opposite. Therefore, as the test sample, 10 sentences from each of these types are taken as well as the types that are common to them. In this way we deal with 14 types, and the corpus of 140 sentences with additional 10 sentences which do not belong to any of these types (simple and independent sentences).

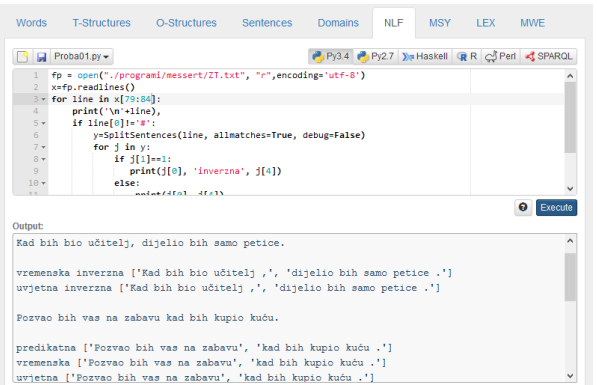


Figure 5. Extraction of subordinate/dependent sentences

For 150 sentences of that corpus the following results were obtained (Figure 5.):

1. Out of 140 dependent sentences (subordinate clauses), 135 (96%) were correctly recognized and extracted, as some patterns are not well written (for predicate or appositional clauses)
2. Although recognition was successful, there is still a need to work on pattern improvement, because the same sentence appears in several categories. The first reason is that patterns include only WOS patterns (and not SOW as well). The second reason lies in the lexicon in which some words are not properly tagged or aren't tagged at all.
3. For 10 independent sentences the program did not categorize any of them which is correct.

5 Extraction of dependent sentences over the SSF module

According to the traditional Croatian grammar and multi-criteria classification of dependent sentences, it is common to distinguish these types: subjective, object, predicate, attributive, appositional and adverbial with its subtypes: time, place, causal, comparative, consequential, permissive and conditional.

The algorithm for extraction of these types is as follows:

1. Construct syntactic patterns for particular types of sentences.
2. For the loaded sentence from the corpus, determine whether there are two (or more) verbs.
3. For each sentence with (at least two) clauses to check whether there are any of the conjunctions or the conjugal groups between them.
4. Check for a possible inversion (subordinate, and main).
5. Determine the type of sentence based on the pattern and its variants.

Figure 6. shows search results for adverbial dependent sentences.

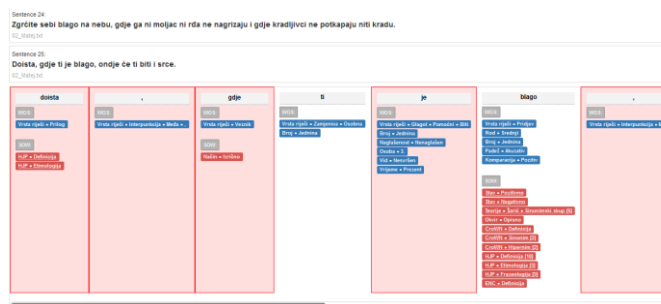


Figure 6. The results of extraction of adverbial subordinate/dependent clauses

Compared to the application of the Python program described in Section 3, this approach gave about 20% better results on the same sample test sentence, but its real value is in the overall approach, easier pattern creation and interaction. Recursive use of multiple corpus patterns ensures greater precision and better filtration results.

6 Syntactic patterns as a dictionary

Although used to identify types of subordinate clauses, these patterns have a much greater importance and a much wider application. Their strength is that at the same time in the analysis they may take care of a word order and their ontological structure to the level of morph/syllable, and also overcome problems with multiword expressions (collocations, idioms, terminological expressions, etc.). It is therefore logical to think about the permanent storage of such patterns and creation of their lexicon. The future lexicon of syntactic patterns in the SSF will have a twofold role:

- a) it will be used in the process of parsing when the document is loaded in the SSF,
- b) it will be used in syntactic and semantic researches.

The first role will rely on basic patterns, for example the congruity of the adjective pattern and the noun or the congruity of an adverb and verb. For example, in the Croatian language, an adverb cannot be followed by a verb, even if the machine in the text hrv. 'iz hrama' might erroneously conclude that it is a adverb + third person of the verb 'hramati'. If this 'impossible' pattern is included in the lexicon, together with information that it never appears in the Croatian language, then the machine will find more words in the lexicon which may apply (e.g. a feminine gender of a noun *hram* in instrumental case which is also *hrama*).

The second role depends on the type of research, for example, typological research in the left-right asymmetry of a natural language with respect to placements in a series of specific types of words left or right of another type (nouns or verbs). These researches were actual in the 60's of the last century [8] but appear in various versions even today [9].

Table 1. shows that by Greenbergs universalities in sentences (of most indoeuropean languages) the most common combination is:

Dem – Num – A – N

or inverse:

N – A – Num – Dem

where **Dem** denotes demonstrative pronoun, **Num** denotes a number, **A** denotes adjective, and **N** denotes a noun.

Table 1. Results of extraction of time adverbial clauses [9]

✓	a	Dem	Num	A	N	MANY
✓	b	Dem	Num	N	A	many
✓	c	Dem	N	Num	A	FEW
✓	d	N	Dem	Num	A	few
∅	e	Num	Dem	A	N	-
∅	f	Num	Dem	N	A	-
∅	g	Num	N	Dem	A	-
∅	h	N	Num	Dem	A	-
∅	i	A	Dem	Num	N	-
∅	j	A	Dem	N	Num	-
✓	k	A	N	Dem	Num	FEW
✓	l	N	A	Dem	Num	few
∅	m	Dem	A	Num	N	-
✓	n	Dem	A	N	Num	FEW
✓	o	Dem	N	A	Num	many
✓	p	N	Dem	A	Num	FEW
∅	q	Num	A	Dem	N	-
✓	r	Num	A	N	Dem	FEW
✓	s	Num	N	A	Dem	few
✓	t	N	Num	A	Dem	few
∅	u	A	Num	Dem	N	
∅	v	A	Num	N	Dem	
✓	w	A	N	Num	Dem	FEW
✓	x	N	A	Num	Dem	FEW

With the help of the SSF it is possible for the selected corpus to conduct the same research, as well as many other researches that are currently being carried in the world, for example [9]:

The order of attributive adjectives:

A_{size} > A_{color} > A_{nationality} > N

The order of adverbs:

Adv_{no more} > Adv_{always} > Adv_{completely} > V

The order of circumstances in adverbials:

Time > Place > Aspect > V

The order proposal orientation and places:

P_{direction} P_{location} NP

The order of TAM (tense-aspect-mode) morphemes:

Sentiment Time Aspect V,

also the order of restructuring auxiliary verbs, the order of dative and accusative encyclicals, etc. Similar researches were conducted by Grosu [10].

The SSF has a module that can run programs using one of the popular programming languages (Python, Haskell, Perl, SPARQL), with already developed additional SSF functions (over 40) with which such or similar research can be carried out over a selected corpus. As an example, the function of these two functions can be shown:

1. DetectSPO (sentence)

Extracts subject, predicate and object from the sentence

2. DetectPattern

(sentence, extend=None)

Detects verb's time and type of dependency in the sentence

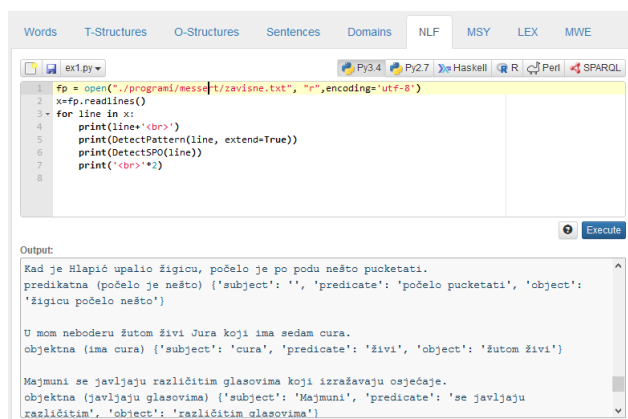


Figure 7. The example of sentence analysis in the SSF

As shown in Figure 7., it is apparent that the machine only partially recognizes the sentences and the word

service in the sentence. The reason is that the tags of some words are incorrect and/or that some patterns are incorrectly written. But what we wanted to show here is that there is a possible model which will enable the machine to use such syntactic analysis.

The future lexicon of syntactic patterns will contain tens of thousands of such patterns, with the possibility to continually upgrade with new ones.

7 Conclusion

This paper presents theoretical and practical implementation of computer module for syntactic analysis which enables the recognition of complex sentences, and future research of statistical methods on them. Recognition of complex sentences is carried out in two ways: the program solution in the Python programming language, and graphical user interface in the syntactic and semantic module. The first one shows an extraction of independent clauses to any level, whereas subordinate/dependent clauses are extracted only to main and subordinate clause (as well as in inversion). The second one shows the creation of syntactic patterns, their generator and recursive execution based on the interactively created regular expressions. Finally, the possibility of using such module in the research of Greenberg's universalities or similar Wiechmann's statistical methods [11] is demonstrated.

References

- [1]. Težak, S. & Babić, S. (2009). *Gramatika hrvatskoga jezika; Priručnik za osnovno jezično obrazovanje*, 17. izdanje, Školska knjiga.
- [2]. Halliday, M. A. K., (1984). *A Short Introduction to Functional Grammar*. London: Arnold.
- [3]. Bresnan, J. & Asudeh, A. & Toivonen, I. & Wechsler, S., (2015). *Lexical Functional Syntax*. 2nd edition. Wiley Blackwell.
- [4]. Berg, T., (2009). *Structure in language: a dynamic perspective*. *Routledge studies in linguistics*; 10, Taylor & Francis
- [5]. Chomsky, N., (1988). Generative grammar. *Studies in English Linguistics and Literature*.
- [6]. Orešković, M., Čubrilo, M., & Essert, M., (2016, September). The Development of a Network Thesaurus with Morpho-Semantic Word Markups. In *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity* (pp. 273-79).
- [7]. Bičanić, A. & Frančić, A. & Hudeček, L. & Mihaljević, M., (2013). *Pregled Povijesti, Gramatike i Pravopisa Hrvatskoga Jezika*, Croatica, Zagreb.
- [8]. Greenberg, J. H., (1963). *Some universals of grammar with particular reference to the order of meaningful elements*. In Joseph H. Greenberg (ed.) *Universals of language*, 73–133. Cambridge, MA: MIT Press.
- [9]. Cinque, G., (2013). *Typological studies: word order and relative clauses*, Taylor & Francis, Routledge, New York.
- [10]. Grosu, A. (2012). Towards a more articulated typology of internally headed relative constructions: The semantics connection. *Language and Linguistics Compass*, 6(7), 447-476.
- [11]. Wiechmann, D. (2015). *Understanding relative clauses: A usage-based view on the processing of complex constructions* (Vol. 268). Walter de Gruyter GmbH & Co KG.