Towards Automatic Definition Extraction for Serbian

Ranka Stanković, Cvetana Krstev, Rada Stijović, Mirjana Gočanin, Mihailo Škorić

Towards Automatic Definition Extraction for Serbian

Stanković Ranka[1], Krstev Cvetana[1], Stijović Rada[2], Gočanin Mirjana[2], Škorić Mihailo[1]
[1] University of Belgrade, Serbia
[2] Institute for the Serbian Language of SASA, Serbia

Abstract

The paper presents preliminary results of the automatic extraction of candidates for dictionary definitions from unstructured texts in the Serbian language with the aim of accelerating dictionary development. Definitions in the Serbian Academy of Sciences and Arts (SASA) dictionary were used to model different definition types (descriptive, grammatical, reference-based and synonym-based) having different syntactic and lexical features. The research corpus consists of 61,213 definitions of nouns, which were analysed using Serbian morphological e-dictionaries and local grammars implemented as finite state transducers in an open-source corpus processing suite Unitex. The 21 models developed up to the present moment cover 57% of dictionary definitions, 83% of which were fully recognized. The analysis has shown that many definitions have a structure that can be modelled, as evidenced by the statistics of definitions grouped by type. These models were used to retrieve noun definitions from a 1.4-million-word corpus containing 25 primary and secondary school textbooks covering various domains. The obtained results were thoroughly analysed, and guidelines were offered for their improvement.

Keywords: definition modelling, definition extraction, Serbian, automatization of dictionary-making, local grammar

1 Introduction

In the age of electronic lexicography, in which the goal is fast production of dictionaries and the products derived from them, special attention is paid to automatic or semi-automatic performance of some tasks. The use of electronic dictionaries can speed up the work on dictionary production by automatically associating some grammatical information to lemmas, namely, word class, word forms and different types of markers (Krstev 2008). Since the research related to extraction of dictionary examples has shown that information extraction from a corpus can be used to speed up the work on the Serbian Academy of Sciences and Arts Dictionary of Serbo-Croatian Literary and Folk Language (SASA Dictionary) and other dictionaries (Stanković et al. 2019; Kitanović et al. 2021), we focused our present research on the extraction of the sentences contained in the definition. The extraction also implies recognition of paradigmatic lexical relations, e.g. synonyms, antonyms, hypernyms, hyponyms. The problem of automatic extraction of definitions from the text has not been thoroughly researched for the Serbian language so far. The assumption is that its solution could significantly contribute to the acceleration of the development of the SASA dictionary, as well as other dictionaries.

Definition extraction is a relevant task in different areas. In addition to dictionary writing, it is also important in the domain of question answering, namely responding to 'What-is' questions, as well as for ontology development. The context in which we will apply definition extraction is semi-automatic creation of dictionaries. In this paper, we present an approach for definition modelling and extraction, relying on the existing Serbian dictionaries (morphological and descriptive), as well as the results of the preliminary experiments in automatic extraction of definitions from unstructured Serbian text.

Definition extraction task can be formalized either as a sentence classification task (i.e., containing term-definition pairs or not) or a sequential labelling task (i.e. identifying the boundaries of terms and definitions). The previous work on definition extraction can be classified as follows: 1) the rule-based approach with linguistic rules and templates that capture patterns to express term-definition relations; 2) the feature engineering approach relying on the statistical machine learning models with syntax and semantic features; 3) the deep learning approach, using word embeddings via multiple layers of neural networks (Veyseh et al., 2020).

Barnbrook (2002) claimed that definition is a basic activity of language, of particular importance to linguists because of its use of language to describe itself. He described the subset of general language used in definition sentences and the development of a taxonomy of definition types, a grammar of definition sentences and parsing software which can extract their functional components. We were inspired in our work by his definition type taxonomy (structural pattern) and definition language grammar.

The second section of this paper discusses the problem of dictionary definition modelling and methods for automatic extraction of candidates for dictionary definitions. The third section provides an analysis of lexical and syntactic characteristics of dictionary definitions in Serbian based on the corpus of definitions of nouns from the five digitized volumes of the SASA dictionary and presents the models developed for noun definitions taking the form of local grammars that can be used for recognition and extraction. These models were applied to a corpus consisting of textbooks and the results achieved in definition extraction are presented in the fourth section, with examples of integration of definitions into different dictionaries. The achieved results are analysed and plans for further research presented in the concluding section.

2 Methodologies for Definition Extraction

A According to the standard "ISO 1951:2007, Presentation/representation of entries in dictionaries — Requirements, recommendations and information" a definition is "A statement that describes a concept and permits its differentiation from other concepts within a system of concepts.". According to the standard "ISO 1087:2019 (en) Terminology work and terminology science — Vocabulary", a definition is "representation of a concept by an expression that describes it and differentiates it from related concepts". This standard distinguishes intentional definition, that conveys the intention of a concept by stating the immediate generic concept and the delimiting characteristic(s) and extensional definition, that enumerates all the subordinate concepts of a superordinate concept under one criterion of subdivision. Intentional definitions are preferable to other types of definitions because they clearly reveal the characteristics of a concept within a concept system: they should be used whenever possible.

A lexicographic definition is the identification of the semantic content of a certain lexeme, with relevant elements of realization. The semantic content consists of an archiseme, which carries information about the lexeme belonging to a wider lexical-semantic group, and a lower-ranking seme, which carries information about individual characteristics of a lexeme, based on which one lexeme differs from another in the same lexical-semantic group. The basic types of lexicographic definition are descriptive, synonymous, and combined - descriptive and synonymous. A descriptive definition without synonyms is used to define the basic meanings of simple, non-derived words (which have no synonyms), and a synonymous one, which consists of words of close or the same semantic structure, when there is a near synonym of that lexeme or when that token is marked with wider uses (Gortan-Premk 2014: 131–132). Combined definitions are used in large descriptive dictionaries - the descriptive part identifies the term directly and lists the elements of meaning, and the synonym refers to the semantic content indirectly (Gortan-Premk 1980: 111–112; Gortan-Premk 1983).

The problem of automatic definition extraction is a relevant task in different areas of natural language processing including Question Answering systems that synthesize answers to questions of the type "What is…" based on one or more sources, and dictionary writing and ontology development (Navigli & Velardi 2010). The approaches to solving this problem are often based on the development and application of lexical-syntactic patterns. For example, in English, the following patterns are often found in definitions (Jin et al. 2013):

<term> defined (as|by) <definition>
define(s)? <term> as <definition>
definition of <term> <definition>
<term> a measure of <definition>
<term> is DT <definition> (that|which|where)
<term> comprise(s)? <definition>
<term> consist(s)? of <definition>
<term> denote(s)? <definition>
<term> designate(s)? <definition>
<definition> (is|are|also) called <term>
<definition> (is|are|also) known as <term>
"part of/in/on/…", "type of", "is <hyperonym> …", …

However, Navigli and Velardi point out that the application of such patterns can give a weaker response (they do not recognize enough definitions) and less precision (they recognize sentences that are not definitions), due to very variable syntactic structures that define the terms. For text definition modelling, they suggest using WordClass Lattices (WCLs), a generalization of word grids and an alternative to lexical-syntactic patterns for which they believe both precision and recall are improved. The lattice is a directed acyclic graph, a subclass of finite automata, aiming to preserve the main differences between different sequences, while at the same time removing redundant information. Pattern comparison is based on the use of an asterisk (wildcard *), which facilitates the clustering of sentences. Each sentence class is then generated by a grid of word classes (each class is either a high-frequency word or a word's part-of-speech). A key feature of the approach is the ability to identify definitions and isolate hypernyms.

The task of automatic definition extraction can be formalized in different ways. One of the possibilities is to consider it as a problem of classifying sentences into those that are potential candidates for defining a term and those that are not, namely, as a problem of determining whether a sentence contains a term-definition pair or not. Automatic extraction of definitions can also be regarded as an annotation task where it is required to identify the boundaries of concepts and their definitions. For example, in the sentence "**Virus** je *program ili kôd koji se sam replikuje u drugim datotekama s kojima dolazi u kontakt.…*" (A virus is a program or code that replicates itself in other files it comes in contact with) the headword and its definition would be marked with XML tags as follows: *<headword>Virus</headword> je <def> je program ili kôd koji se sam replikuje u drugim datotekama s kojima dolazi u kontakt.</def>*

Morphosyntactic patterns applied to the computational linguistics corpus are used to extract candidates for definitions for Slovenian and English (Pollak et al. 2012) by automatic recognition of terminology and semantic annotation of WordNet meanings. The tool uses patterns of the form "NP is NP", "NP refers to NP", "NP denotes NP", where NP refers to the noun phrase. The authors also use the Slovenian WordNet to detect definitions: sentences that potentially contain definitions begin with a word from the WordNet and at least one more word belonging to the same chain of hypernym/hyponym relations. The system passes information about the preferred position of the term for which a definition is sought: whether it must be at the beginning of the sentence, after predefined initial forms, which is the default choice, or whether it can be anywhere in the sentence.

Spala and colleagues (2019) associate a detailed annotation scheme with the corpus in order to explore diverse structures

of term definitions in free and semi-structured texts. In addition to the basic concept (Term) and its main definitions (Definition), sentence segments containing pseudonyms or additional names (Alias Term) are also annotated and associated with the basic term. Likewise, noun phrases (Referential Term) that refer to the previously marked term, secondary definitions (Secondary Definition) with additional information, qualifiers (Qualifier) that specify any conditions under which the definition applies and alike are also annotated. The tagged segments are connected by appropriate relations.

Ristić et al. (2018) analyse vertical chaining of concepts present in their lexicographic definitions using the thematic field 'house, building' as an example, pointing to a series from the highest levels of conceptual categorization, complex and simple primitives, to lower levels of hypernyms (PLACE> AREA) [WHERE], AREA > SPACE [SOMETHING THAT EXTENDS (BOUNDLESSLY) IN ALL DIRECTIONS]). They also state that this type of research could contribute to introducing advanced search of digitized editions of descriptive dictionaries of Serbian.

The context in which the definitions for automatic extraction are analyzed and formalized in our paper is the support of dictionary drafting (Kilgarriff & Rychlý 2010), which implies the development based on corpora. The results were achieved by using electronic dictionaries of the Serbian language and local grammars developed based on the results of some previous similar research, particularly (Barnbrook 2002), analysis of definitions in existing dictionaries and previous research into Serbian (Krstev et al. 2015).

3 Analysis and Recognition of Definitions in the SASA Dictionary

The first step in our research was a thorough analysis of various lexical and syntactic features of definitions in the Serbian Academy of Sciences and Arts (SASA) dictionary; this part of a dictionary entry is presented in italics. The definition structure in the SASA dictionary is informally outlined in the Guidelines for Dictionary Processing. The guidelines suggest that the descriptive definition comes first, followed by the definition by substitution with related words; ex: **безвлашће**… *стање без државне власти, анархија, безакоње* (powerlessness, ... state without power of a state, anarchy, lawlessness). The closest senses can be separated just by semicolons, thus representing a single definition; eg: **годишњак**…*периодична публикација која излази једанпут годишње; штампани годишњи извештај неке установе.* [yearbook … periodical published once a year; printed annual report of an institution].

We will illustrate a few models given in the guidelines: if a noun is to be defined by a relative clause, the definition should begin with the expression "онај који ..." ("one who ..."), as opposed to adjectives where definitions begin with "који..." ("which ..."). For abstract nouns ending with *-ост*, *-ство*, *-ота*, *-ођа*, *-ина*, etc. first a general definition is given in the form: osobina, stanje ili svojstvo (an attribute, a state or a feature) *онога који је* (of one who is) */ онога што је* (which is), followed by the adjective from which the abstract noun was derived. This general definition may occasionally be supplemented by two or at most three synonyms.

For the construction of the system for automatic extraction of definitions, it was necessary to formalize models of definition types, for which the Guidelines for Dictionary Processing alone were not enough. It was necessary to create a corpus that would provide the means for conducting an analysis of examples of definition structures. Inspired by the endeavours of large lexicographic houses and research centres that have recognized the possibilities provided by lexicographic databases, especially in supporting modern dictionary use (Rundell 2014), we formalized the microstructure of a SASA dictionary entry (Stijović and Stanković 2017). This made possible the design of a lexical database that can store a structured record of a dictionary article in a relational structure, and the development of a software solution that transforms the unstructured text of a Word document into a relational database (Stanković et al. 2018).

These preliminary steps led to the construction of a research corpus consisting of definitions from the five digitized volumes of the SASA dictionary (SASA Dictionary 2019; Stanković et al, 2018). The corpus contains 61,213 noun definitions, 19,708 verb, 12,312 adjective, 2,564 adverb definitions and 2,967 definitions of other word classes. The main aim of developing models of clauses used in SASA dictionary definitions is to enable formalization of definitions for some future dictionaries - to establish which models are the most frequently used, which are preferable, and which definitions unnecessarily deviate from the common patterns.

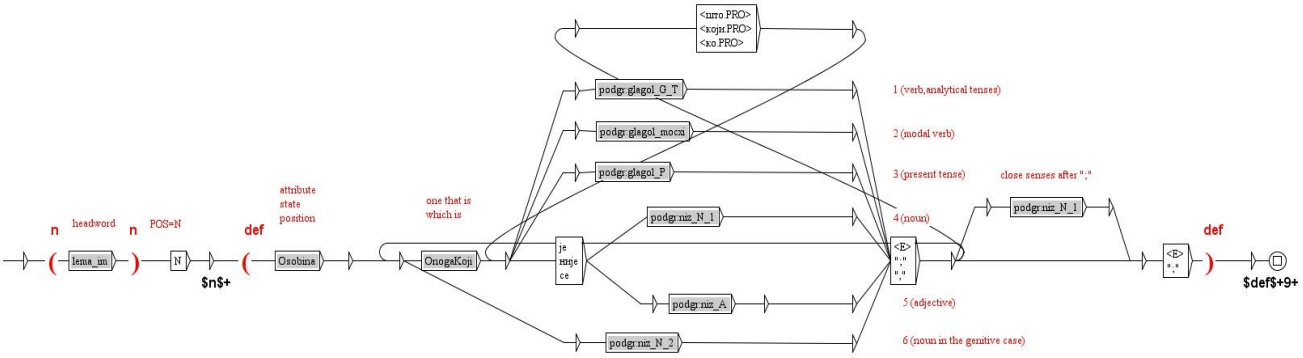
The definitions of nouns from the SASA dictionary were analysed using Serbian morphological e-dictionaries and local grammars in the form of Finite-State Transducers (Krstev 2008) and implemented in the Unitex corpus processing suite[1]. Electronic morphological dictionaries of Serbian intended for automatic processing have been undergoing development for many years now, and their current size and content ensure successful use in different real-world applications in the field of Serbian language processing. These dictionaries contain automatically generated word forms of more than 200,000 lemmas[2], and their content covers both general vocabulary and proper names - personal, geopolitical, names of organizations and the like. Moreover, the dictionary contains multi-word units, which are recorded in traditional dictionaries as syntagms or phrases. The basic unit of these dictionaries is a word form associated with its lemma (usually the headword of a traditional dictionary entry), Part-Of-Speech, markers that describe its syntactic, semantic, usage, dialectal and domain properties, followed by the codes of its possible morphosyntactic realization.

Finite state transducers (FSTs) are abstract mathematical constructions that allow modelling of local grammars to describe some linguistic constructions, for example, noun phrases. A finite state transducer "passes" through the text it analyses to compare a text chunk with the model it represents. In the case of successful recognition, a final state transducer produces some result, which can be a modification of the source text by adding tags for types of recognized

[1] Unitex/GramLab - Lexicon-Based Corpus Processing Suite (https://unitexgramlab.org/)

[2] A part of this lexicon is publicly available for use within the Unitex system

words or a recognized syntactic structure (Vitas & Krstev 2012). Finite state transducers are visualized by graphs for easier development and use. A local grammar and its corresponding graph that models (and recognizes) definitions of nouns that represent attributes and/or state is given in Figure 1. Below the graph, six definitions are listed that are recognized by the corresponding paths in the graph.



1. **адаптираност** N *стање онога што је прилагођено, прилагођеност, подешеност*. (the state of what is adjusted, the adjustment, the setting)
2. **оповргљивост** N *особина онога што се може оповргнути*; (a property of what can be refuted)
3. **вољкост** N *особина онога што ствара добро расположење*, (an attribute of that which creates a good mood)
4. **паланчанство** N *стање, особина онога који је паланчанин, паланачко порекло, паланачки дух и сл.* (condition, characteristic of the one who is from a small town, a small-town origin, a small-town spirit, etc.)
5. **опречност** N *особина, стање онога што је опречно, супротност, противречност*; (a property, a state of what is opposite, the opposite, the contradiction)
6. **богомољство** N *особина, својство богомоља, претерано ревносна побожност*. (a trait, a property of a worshiper, overzealous piety)

Figure 1: A local grammar that models (and recognizes) definitions of nouns that represent attributes and/or state.

From the point of view of automatic recognition of definitions in the SASA dictionary, definitions can be divided into three main groups. The first group consists of highly schematized definitions that refer to other dictionary entries, e.g. pointing to a common or literary form ("see…"), or to a derived word that retained the meaning of the basic word ("diminutive of …"). These definitions are easy to model. The second group consists of "definitions" of some types of proper names, e.g. names of holidays, saints, monasteries, etc. These definitions are similar to the explanations given in encyclopaedias, they are expressed freely, and are difficult to model with local (shallow) grammars. One example is **Брашанчево** N *католички црквени празник који пада у други четвртак после Духова* (a Catholic church holiday that falls on the second Thursday after Pentecost). The third group consists of proper definitions that are schematized to a certain extent. For instance, feminine gender nouns derived by gender motion have a schematized definition, e.g., **бунтовница** N *жена бунтовник* (woman rebel). However, these definitions are often expanded with additional information, e.g., **верница** N *жена верник, припадница неке религије* (a woman believer, a member of a religion).

We have developed 21 definition models covering definitions from all three groups to some extent: 7 for the definitions from the first group, one for the definitions from the second group and 13 for the definitions from the third group. When developing these models by means of local grammars with FSTs we tend to capture all types of definitions, namely descriptive, reference-based and synonym-based. The coverage of noun definitions in the SASA dictionary by the developed local grammars is represented in Table 1. We have mentioned before that our corpus of definitions consists of 61,213 definitions of nouns. During the development of local grammars, we reduced this set by excluding definitions containing unknown words (not recorded in the e-dictionaries used by local grammars) since local grammars even if appropriate for them could not be successful. We thus obtained a set of 51,729 definitions.

The analysis of definitions has shown that the SASA dictionary contains a large number of schematized definitions (34% of all definitions) and they have mostly been fully recognized (from 91 to 100% depending on type). As expected, definitions of encyclopaedic entries are difficult to capture (only 26% of recognized definitions were fully recognized) and it is possible that quite a number of them were not recognized at all. Definitions schematized to a certain extent were covered with a differing success rate depending on their type, ranging from 38% for nationalities to 87% for "type of…" definitions. Some of the definition types look very specific, for instance, model 19 describing devices, tools etc. This type is specific because it uses a closed set of words for describing artefacts which are further distinguished by prepositional phrases that describe their purpose. Once developed, this type of definition will be used in the models that still have to come to light.

Local grammars that model definitions of nouns (and other types of words) will contribute to the creation of dictionaries and other lexical resources in various ways. For instance, when creating a dictionary, they will be used to check the compliance of definitions with the adopted forms. In addition, they enable automatic linking of dictionary entries in the database, which was illustrated by the examples in Table 1. Moreover, definition models will enable not only automatic linking of the entries in the SASA database, but also enrichment and linking of other lexical resources for Serbian (morphological e-dictionaries and WordNet). Finally, as will be shown in the next section, local definition grammars can also be used to extract definitions from domain corpora.

Graph (group)	Type	No of recognised definitions	To the end	Example
1 (1)	see	10849	9830 (91%)	**акушерка** N *в.бабица* (accoucheuse N v. midwife)
2 (1)	surname	3672	3670 (100%)	**Андрић** N *презиме* (Andrić N surname)
3 (1)	verbal noun	913	912 (100%)	**плакање** N *гл. им. од плакати* (crying N verb. n. from to cry)
4 (3)	type of	1435	1250 (87%)	**бисер** N *врста пенушавог вина* (pearl N a kind of sparkling wine)
5 (1)	diminutive	1629	1575 (97%)	**ветрић** N *дем. и хип. од ветар;* (small wind N dim. and hyp. of wind)
6 (1)	first name or a nickname	914	857 (94%)	**Пачавра** N *лични надимак* (Pačavra N personal nickname)
7 (3)	women	596	483 (81%)	**бедевуша** N *танка, висока незграпна жена.* (bedevuša N a thin, tall clumsy woman)
8 (1)	augmentative	464	454 (98%)	**бубетина** N *аугм. и пеј. од буба.* (big bug N aug. and pej. from bug)
9 (3)	attribute or condition	507	379 (75%)	**плиткоумност** N *особина, својство онога који је плиткоуман,* (shallowness N a trait, a property of one who is shallow-minded)
10 (3)	geographic name	436	280 (64%)	**Абисинија** N *ранији назив за Етиопију.* (Abyssinia N former name of Ethiopia.)
11 (3)	inhabitant	214	178 (83%)	**Папуанка** N *припадница домородачког становништва Папуанаца* (Papuan woman N a member of the indigenous population of the Papuans)
12 (3)	people	21	8 (38%)	**Бугари** N *јужнословенски народ* (Bulgarians N South Slavic people)
13 (2)	proper names (facilities, deities, astral bodies, holidays)	94	24 (26%)	**Вечерњача** N *народни назив за планету Венеру* (Večernjača N folk name of the planet Venus)
14 (1)	collective nouns	103	101 (98%)	**браћа** N *зб. им. од брат* (brothers N col. noun derived from brother)
15 (3)	animals	2115	1564 (74%)	**арнаут** N *назив за помамна, љута коња;* (arnaut N name for a mad, angry horse;)
16 (3)	plants	1371	1005 (73%)	**брусница** N *шумска ниска жбунаста биљка…* (cranberry N low shrubby forest plant…)
17 (3)	plant and animal organs	84	56 (67%)	**папоњак** N *мали, закржљали клип кукуруза.* (paponjak N small, stunted corn cob)
18 (3)	the one who is… that what is…	2221	978 (44%)	**балавац** N *онај који је недорастао, незрео,* (snot-nosed kid N one who is not grown-up, immature **пикантерија** N *оно што је пикантно* (piquancy N what is spicy)
19 (3)	tool, device, machine…	340	168 (49%)	**алтиметар** N *справа за мерење надморске висине* (altimeter N device for measuring altitude)
20 (3)	set, part…	2874	1984 (69%)	**бифтек** N *комад говеђег меса од леђне печенице* (beefsteak N a piece of roast beef from the back
21 (3)	prepositional clause	715	343 (48%)	**баврљуга** N *у дечјој бројаници* (bavrljuga N in a children's tongue twister)
Total		**31567 (61%)**	**26099 (83%)**	2125 definitions recognised by more than 1 graph
Different		**29442 (57%)**	**24347 (83%)**	

Table 1: The analysis of recognized definitions from the SASA dictionary; the defined word is given in small caps, connected entries are given in bold, while trigger words enabling the connection are given in italic.

4 Corpus Analysis

4.1 Creating a Textbook Corpus

For testing the automatic extraction of definitions from unstructured text by means of local grammars presented in the previous section, we created a corpus of 25 primary and secondary school textbook covering the following domains: biology, history, logic, music, computer science, physics and design and technology (technological education). The biology domain is covered by seven textbooks: one for primary school, four for high school, and two for agricultural professional / vocational school; history domain is represented by four primary school textbooks; computer science is represented by one primary school textbook and tree high school ones, while physics is represented by one primary school textbook. General technique is represented by three textbooks for professional/vocational technical schools relating to the following subjects: power engineering, mechanical engineering for agriculture and tools and mechanization. Logic and philosophy, on the other hand were covered by two high school textbooks focusing on humanities subjects . The domain of music is represented by two music high school textbooks: History of Music and Century of Jazz. The corpus is being developed and 35 textbooks have been used for the purposes of this experiment. However, the current version contains 31 textbooks and it is expected to grow in the near future. The textbooks were scanned, optical character recognition was performed, and recognition errors were manually corrected (though a certain number of OCR errors remained). The corpus consists of 85,628 sentences, 3,4M tokens (165K different) and 1,4M words (165K different).

The corpus was processed by using electronic dictionaries of the Serbian language that recognized approximately 238,000 word forms, 5,000 multi-word units, while 5,700 word forms remained unrecognized. Among the unrecognized words are the remaining optical character recognition errors, foreign names and abbreviations (Facebook, WWW, HTML, SMS, RNA, ATR, HIV), as well as a number of words belonging to domain-specific terminology, such as: biology - *eukariote* (eukaryotes), *citoplazma* (cytoplasm), chemistry - *acetilhlorin* (acetylchlorine), *organele* (organelles), *hloroplas*t (chloroplast), music - *diksilend* (dixieland), etc. The majority of unknown words are the result of OCR errors that remained in the text, but we are working on correcting them.[3]

4.2 Recognition of Candidates in the Textbook Corpus

To determine whether it is possible to recognize definitions of domain-specific terms in the domain corpus text, a subset of local grammars presented in Section 3 was applied to the corpus consisting of 25 textbooks; namely we excluded models for schematized definitions that link entries in the dictionary, since they are not relevant to the unstructured texts. Table 2 shows examples of successful, or partly successful, definition recognition. It should be noted that in some cases only the initial parts of the definition were recognized, for example, full definition for *drama* is **Драма** је *врста књижевног дела које настаје да би се изводило на позорници* (Drama is a type of literary work that is created to be performed on stage) (the underlined sequence was not recognized, see example 4. in Table 2). On the other hand, some long and useful definitions were completely retrieved, e.g. **Научни систем** је *јединствен скуп знања (чињеница, закона, теорија) која су међусобно повезана и сређена на основу извесних принципа* (A scientific system is a unique set of knowledge (facts, laws, theories) that are interconnected and arranged on the basis of certain principles.).Automatic recognition of complete definitions is a complex task, because it requires grammars that perform a complete parse. However, even partial recognitions can be useful for pointing to the sentences that potentially contain definitions.

The definitions recognized in the textbook corpus are in some cases similar to the definitions in the SASA dictionary, e.g. definitions for the orchid family (example 1 in Table 3), while in other cases they differ, stressing a different attribute, e.g. definitions of *alcoholic* (example 2 in Table 3). In some cases, there are several definitions for different senses of a lemma, while textbooks offer yet another one, e.g. definitions of *winch* (example 3 in Table 3). On the other hand, three related but different definitions were retrieved from the corpus for the bird family (example 4 in Table 3; note that this lemma has not yet been covered by the SASA dictionary). In some cases, definitions were retrieved for domain specific terms that are not listed in the SASA dictionary, e.g. the definition of *bentonite* (example 5 in Table 3).

[3] The automatically lemmatized and POS tagged corpus of textbooks (Stanković et al. 2020) is available on the corpus search platform of the Jerteh Society for Language Resources and Technologies https://noske.jerteh.rs/#dashboard?corpname=SkolKor. The search system for monolingual and multilingual corpora is based on the NoSketch Engine platform (Rychly 2007), which is open-source software - a simplified version of the commercial Sketch Engine tool. Each textbook is supplied with metadata: title, autor, publisher, year of publishing, subject, school level (primary, secondary) and school class. As a guest, a user can presently search several corpora under NoSkatchEngine more corpora will be available in the near future.

domain	scope	recognized	correct	examples
biology	full	44	40	**Биљне ваши** сy *ситни инсекти који се хране биљним соковима.* (Plant lice are tiny insects that feed on plant juices.) (15) **Протеини** сy *група органских молекула са разноврсним функцијама*... (Proteins are a group of organic molecules with various functions …) (20)
	partial	97	94	**Бреза** је *брзорастућа врста чије се дрво* (Birch is a fast-growing species whose tree) (4)
history/ logic/ music	full	10	10	**Авари** сy *номадско племе турског порекла*. (The Avars are a nomadic tribe of Turkish origin.) (12)
	partial	55	34	**Драма** је *врста књижевног дела које настаје...* (Drama is a type of literary work that is created …) (4)
computer science/ design and technology (technological education).	full	13	12	**Рачунарска мрежа** је *скуп међусобно повезаних рачунара који комуницирају и размењују информације*. (A computer network is a set of interconnected computers that communicate and exchange information.) (20)
	partial	57	45	**Програмски језик** је *скуп правила којима се рачунару*… (A programming language is a set of rules that ... to a computer…) (20)

Table 2. Examples of definitions recognized in the textbook corpus.

comparison	source	examples
1. similar	SASA	**орхидеја** N *у мн.: фамилија вишегодишњих зељастих украсних биљака* (orchid N in pl.: family of perennial herbaceous ornamental plants)
	textbooks	**Орхидеје** сy *вишегодишње зељасте биљке,...* (Orchids are perennial herbaceous plants, ...)
2. different	SASA	**алкохоличар** N *човек који се одао пићу*; (alcoholic N a man who indulges in drinking;)
	textbooks	**Алкохоличар** је *особа која показује зависност од алкохола, који штетно делује на организам,...* (An alcoholic is a person who exhibits alcohol addiction, which has a harmful effect on the body, ...)
3. multiple in SASA	SASA	**витло** N *назив за разне направе које се окрећу*; (winch N name for various rotating devices;) **витло** N *обртна направа у виду елисе (у машини).* (winch N rotating device in the form of a propeller (in a machine).)
	textbooks	**Витло** је *уређај који углавном служи за подизање или вучу терета.* A winch is a device that is mainly used for lifting or towing loads.
4. multiple definitions	textbooks	**птице** сy *далеко најбројнија група копнених кичмењака,...* (birds are by far the most numerous group of terrestrial vertebrates, ...) **Птице** сy *једина основна група данашњих кичмењака која има*… (Birds are the only basic/core group of today's vertebrates that has…) **Птице** сy *најмлађа основна група копнених кичмењака.* (Birds are the youngest basic/core group of terrestrial vertebrates.)
5. not in SASA	textbooks	**Бентонит** је *врста глине (хидратисани алуминијум…* (Bentonite is a type of clay (hydrated aluminum ...)

Table 3. A comparison of definitions in the SASA dictionary and those extracted from the textbook corpus.

The productivity of the developed models is represented in Figure 2. It is noticeable that model 20 (defining sets, groups, parts, etc.) is by far the most successful in retrieving definitions from the textbook corpus. Presently it covers 9% of all noun definitions in the SASA dictionary surpassing all other models, except those applied to schematized definitions.

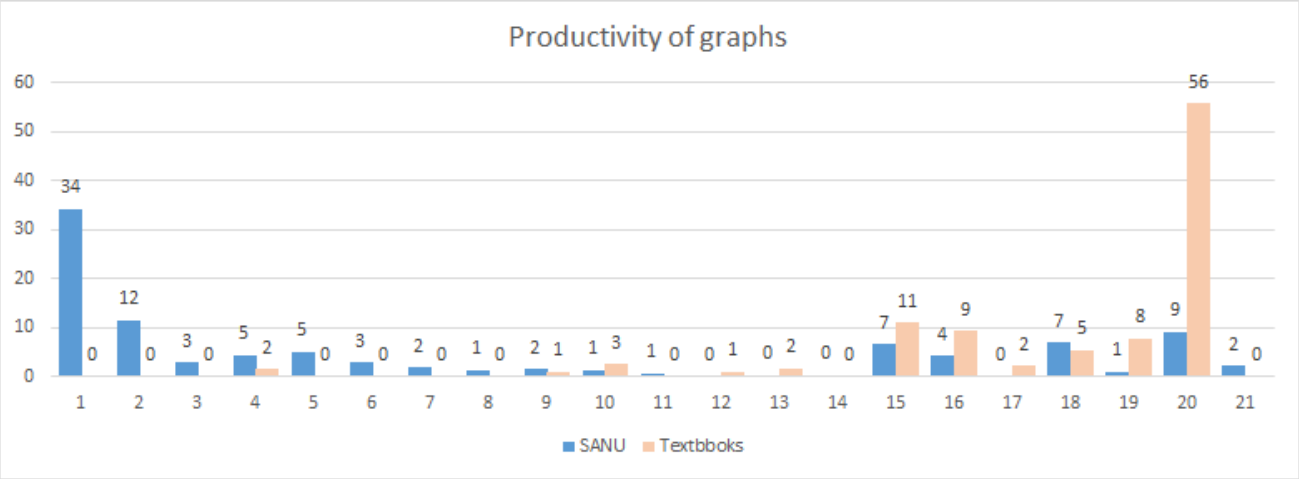


Figure 2. Productivity of models on the SASA dictionary and the corpus of textbooks.

The results obtained by this experiment can be considered preliminary and promising, but far from satisfactory. It is necessary not only to refine local grammars, but also to develop additional models. Besides, more possibilities to connect a word to its definition should be explored; namely, on this occasion we used only one pattern: **word** is/are *definition*. It can be seen in Table 2 that especially partially recognized definitions are not always correct, which is most prominently present in the history, logic, and music domains. This means that more strict conditions (like gender and/or case agreement) should be used for extraction from unstructured texts than are necessary when modelling dictionary definitions.

5 Conclusion

The paper presents preliminary results of the automatic extraction of candidates for dictionary definitions from unstructured texts in the Serbian language, with the aim of accelerating the development of dictionaries. In order to model different types of dictionary definitions, we used a corpus consisting of definitions taken from the five digitized volumes of the SASA dictionary. The analysis has shown that a large number of noun definitions have a structure that can be modelled (57%). Our long-term objective is definition extraction from unstructured texts. To that end, we prepared a corpus consisting of primary and secondary school textbooks to which we applied the models developed for the SASA dictionary.

While the rule-based approach is intuitive and has high precision, it suffers from the low recall issue. In order to overcome it, we plan to investigate other methods to complement this approach. Hill et al. (2016) report the effectiveness of both neural embedding architectures and definition-based training for developing models that understand phrases and sentences, with two applications of these architectures: reverse dictionaries that return the name of a concept given a definition or description and general-knowledge crossword question answers.

Inspired by the work of Noraset et. al (2017), Bosc & Pascal (2018), our model will be improved by learning to compute word embeddings by processing dictionary definitions and trying to reconstruct them. We will use Dict2vec, based on natural language dictionaries that builds new word pairs from dictionary entries, so that semantically related words are moved closer, and negative sampling filters out pairs whose words are unrelated in dictionaries. (Tissier et al., 2017) We will investigate whether definition models can improve standard word embeddings.

6 References

A Style Guide for Dictionary-Making, Belgrade: SASA Institute for Serbo(-Croatian) (manuscript), 1959 and (supplemented) 2017 [Упутство за обраду Речника, Београд: Институт за српск(охрватск)и језик САНУ (рукопис), 1959. и (допуњено) 2017]

Bosc, T. & Pascal V. (2018). Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1522-1532.

Barnbrook, G. (2002). *Defining Language, A local grammar of definition sentences, Studies in Corpus Linguistics*, (Vol. 11). John Benjamins Publishing.

Gortan Premk, D. (1980). O gramatičkoj informaciji i semantičkoj identifikaciji u velikom opisnom rečniku. [On grammatical information and semantic identification in a large descriptive dictionary.], *Naš jezik*, XXIV/3, pp. 107–114.

Gortan Premk, D. (1983). Synonymous array in lexicographical definition. [Sinonimski niz u leksikografskoj definiciji.] In Naučni sastanak slavista u Vukove dane, 12/1, pp. 45–50.

Gortan Premk, D. (2014). Definisanje u srpskoj leksikografiji. [Definition in Serbian lexicography]. In *Savremena srpska leksikografija u teoriji i praksi*, R. Dragićević (ed.), Beograd: Filološki fakultet, pp. 131–139.

Hill, F., Cho, K., Korhonrn, A. & Bengio, Y. (2016). *Learning to understand phrases by embedding the dictionary*. Transactions of the Association for Computational Linguistics, 4, 17-30.

Jin, Y., Kan, M. Y., Ng, J. P., & He, X. (2013). Mining scientific terms and their definitions: A study of the ACL anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 780-790.

Tissier, J., Gravier, C., & Habrard, A. (2017). Dict2vec: Learning Word Embeddings using Lexical Dictionaries. In *Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), Sep 2017, Copenhague, Denmark*. pp. 254-263.

Navigli, R. & Velardi, P. (2010). Learning Word-Class Lattices for Definition and Hypernym Extraction. In *Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics*. *Uppsala, Sweden*. pp. 1318–1327.

Kilgarriff, A. & Rychlý, P. (2010). Semi-Automatic Dictionary Drafting, In *A Way with Words: Recent Advances in Lexical Theory and Analysis*. Uganda: Menha Publishers Ltd., pp. 299-312.

Kitanović, O., Stanković, R., Tomašević, A., Škorić, M., Babić, I. & Kolonja, Lj. 2021. "A Data Driven Approach for Raw Material Terminology." *Applied Sciences 11 (7): 2892*, pp. 1-22.

Krstev, C. (2008*). Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade.

Noraset, T., Liang, C., Birnbaum, L., & Downey, D. (2017). Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

Krstev, C., Vitas, D. & Stanković, R. (2015). A Lexical Approach to Acronyms and their Definitions. In *Proceedings of 7th Language & Technology Conference, November 27-29, 2015, Poznań, Poland*, eds. Zygmunt Vetulani & Joseph Mariani, 219–223, Fundacja Uniwersytetu im. A. Mickiewicza, Poznań, pp. 219-223.

Pollak, S., Vavpetic, A., Kranjc, J., Lavrac, B. & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. In: *Proceedings of KONVENS 2012*, *Vienna, September 19, 2012*, pp. 53–60.

Ristić, S., Konjik Lazić, I. & Ivanović, N. (2018) Metajezik leksikografske definicije u deskriptivnom rečniku (na materijalu rečnika srpskog jezika) [Metalanguage of lexicographic definition in descriptive dictionary (on the material of the Serbian language dictionary)]. Južnoslovenski filolog 74/1, 2018, pp. 81–96.

Ristić, S. (2003). Leksikografski metajezik i srpska deskriptivna leksikografija. [Lexicographic metalanguage and Serbian descriptive lexicography.] In: *Naučni sastanak slavista u Vukove dane* 31/1, Beograd: Međunarodni slavistički centar. Filološki fakultet, pp. 119–130.

Rundell, M. (2014). Macmillan English Dictionary: The End of Print? In: *Slovenščina 2.0: empirical, applied and interdisciplinary research 2.2*, 1–14.

Rychly, P. (2007). Manatee/bonito – a modular corpus manager. In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing, 2007*, pp. 65–70.

SASA Dictionary: Речник српскохрватског књижевног и народног језика САНУ, I–XXI [The Dictionary of the Serbo-Croatian Standard and Vernacular Language] (1959–2020). Београд: Институт за српски језик САНУ и САНУ.

Spala, S., Miller, N., Yang, Y., Dernoncourt, F. & Dockhorn, C. (2019). DEFT: A corpus for definition extraction in free-and semi-structured text. In *Proceedings of the 13th Linguistic Annotation Workshop, 2019*, pp. 124–131.

Stanković, R., Stijović, R., Vitas, D., Krstev, C. & Sabo O. (2018). The Dictionary of the Serbian Academy: from the Text to the Lexical Database. In: *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana*. Ljubljana University Press, Faculty of Arts. Čibej, Jaka, Gorjanc, Vojko, Kosem, Iztok & Krek, Simon (eds.), pp. 941–949.

Stanković, R., Šandrih, B., Stijović, R., Krstev, C., Vitas, D. & Marković, A. (2019). SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian. In: *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference, 1–3 October 2019, Sintra, Portugal*, eds. I. Kosem et als., pp. 248–269.

Stanković, R., Šandrih, B., Krstev, C., Utvić M. & Škorić M. (2020). Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In: *Proceedings of the 12th International Conference on Language Resources and Evaluation, LREC* eds. Nicoletta Calzolari et al., pp. 3947–3955.

Stijović, Rada, Ranka Stanković. Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU. [Digital edition of the SASA Dictionary: a formal description of the microstructure of the SASA Dictionary (in Cyrillic)] In: *Naučni sastanak slavista u Vukove dane* 47/1, Beograd: Međunarodni slavistički centar. Filološki fakultet, 2018, 427–440.

Veyseh, A. P B., Dernoncourt, F., Dou, D. & Nguyen, T. (2020). A Joint Model for Definition Extraction with Syntactic Connection and Semantic Consistency. In: *AAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 9098-9105.

Vitas, D. & Krstev, C. (2015) Nacrt za informatizovani rečnik srpskog jezika. [Blueprint for the computerized dictionary of the Serbian language (in Cyrillic).] In: *Naučni sastanak slavista u Vukove dane* – Srpski jezik i njegovi resursi: teorija, opis i primene, 44/3, Beograd: Međunarodni slavistički centar. Filološki fakultet, pp. 105–116.

Vitas, D. & Krstev, C. (2012). Processing of Corpora of Serbian Using Electronic Dictionaries. In: *Prace Filologiczne*, vol. LXIII, Warszawa, 2012, 279–292.

Acknowledgements

This paper is supported by the COST Action CA19102 - LITHME "Language in the Human-Machine Era" https://lithme.eu/. Access to SketchEngine is provided by the ELEXIS project funded by the European Union's Horizon 2020 research and innovation program under grant number 731015.