SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian

Ranka Stanković[1], Branislava Šandrih[1], Rada Stijović[2], Cvetana Krstev[1], Duško Vitas[1], Aleksandra Marković[2]

[1] University of Belgrade, Studentski trg 1, Belgrade, Serbia
[2] Institute for Serbian Language, SASA, Knez Mihailova 36, Belgrade, Serbia
E-mail: ranka@rgf.rs, branislava.sandrih@fil.bg.ac.rs, rada.stijovic@isj.sanu.ac.rs, cvetana@matf.bg.ac.rs, vitas@matf.bg.ac.rs, aleksandra.markovic@isj.sanu.ac.r

Abstract

In this paper we present a model for selection of good dictionary examples for Serbian and the development of initial model components. The method used is based on a thorough analysis of various lexical and syntactic features in a corpus compiled of examples from the five digitized volumes of the Serbian Academy of Sciences and Arts (SASA) dictionary. The initial set of features was inspired by a similar approach for other languages. The feature distribution of examples from this corpus is compared with the feature distribution of sentence samples extracted from corpora comprising various texts. The analysis showed that there is a group of features which are strong indicators that a sentence should not be used as an example. The remaining features, including detection of non-standard and other marked lexis from the SASA dictionary, are used for ranking. The selected candidate examples, represented as feature-vectors, are used with the GDEX ranking tool for Serbian candidate examples and a supervised machine learning model for classification on standard and non-standard Serbian sentences, for further integration into a solution for present and future dictionary production projects.

Keywords: Serbian; good dictionary examples; automatization of dictionary-making; feature extraction; machine learning

1. Introduction

1.1 The aim of the paper

This paper outlines an approach to providing support for building different kinds of monolingual descriptive dictionaries of the Serbian language. The approach was motivated by the need for modernization of the dictionary-making process for the dictionary of the Serbian Academy of Sciences and Arts (SASA), a large monolingual thesaurus of Serbian, as well as for the production of new dictionaries of Serbian. The SASA dictionary is still developed traditionally, and its modernization could serve various different goals: speeding up the dictionary-making process, but also the development of a lexical database as the source for building new dictionaries of Serbian.

In the e-lexicography era, with the imperatives of faster dictionary-making and "smart lexicography", special attention is devoted to semi-automatic selection of dictionary examples from corpora, and the presented approach supports the selection of dictionary examples making the process of dictionary development faster and more productive.

1.2 The role of dictionary examples

Dictionary examples play an important role in dictionary entries and they constitute, according to some authors, a "key microstructural element" of a dictionary (Kosem, 2017: 183). A good example is valuable from the aspects of both language reception and production. Examples have different roles, some of which are mentioned by S. Atkins and M. Rundell: they can complement the definition and help the user understand the meaning of the headword/lexical unit (their informative value); they should show the typical and natural way of behaviour of a word: syntactic patterns, collocations, as well as its colligational preferences – preferred form(s) of the paradigm, or the position(s) in the sentence; and since examples should help the understanding of the definition, they must be easy to understand – which means that their syntactic structure should be simple and their lexis not too difficult and uncommon. Informativeness, typicality with naturalness, and intelligibility are basic criteria for good dictionary examples (see more on these criteria in Atkins & Rundell 2008: 458–461).

However, many metalexicographers point out that it is not easy to find good dictionary examples in corpora. Kilgarriff et al. (2008: 429) note that reading concordances is "an advanced linguistic skill", and "the point of reading concordances – to pick up the common patterns that a word occurs in – is itself an abstract and high-level task". This task is difficult even for trained lexicographers. In addition, finding good examples is time-consuming. The corpora are very big nowadays, the number of concordances one gets for a keyword is often too large, and it is impossible to read all of them. All this was the motivation for the development of GDEX, a tool designed for extraction of good dictionary examples (Kilgarriff et al., 2008), now used not only by lexicographers, but also in language teaching and learning.

1.3 SASA-Dataset

The SASA dictionary is conceived as a thesaurus, meant primarily for native speakers. Its primary goal is to help understanding words from different kinds of texts (receptive use of dictionary). It covers a large portion of the vocabulary of the Serbian language, standard and vernacular, for the last 200 years. In Zgusta's terms, it is a combination of the standard- and overall-descriptive dictionary (Zgusta, 1971: 212), which means that all marked lexis (dialectal, archaic or dated, jargon, etc.), as well as non-standard phonetic, morphological and syntactic forms and types of complements are labelled.

Each dictionary entry contains (or may contain) several subentries (one subentry for each lexical unit), and their descriptive definitions (sometimes definitions by synonyms). Every definition is followed by several (2 to 6) illustrative examples (examples are listed chronologically), with precise bibliographic references.

The first volume of this dictionary was published in 1959 (the project itself has been underway since the last decades of the 19th century), and the last, 20th volume was published in 2017 (the total number of volumes planned is 35). This is a long-term, time-consuming project.

Although the process of dictionary-making continues in the traditional way, there have been several initiatives for its modernization and acceleration. Digitization of the published volumes began in 2016, and the first exploitation of two digitized volumes was reported in Stijović and Stanković (2017). Dictionary entries from five volumes were automatically parsed and stored as a structured text in a lexical database, which offers the opportunity to use this data for extraction of different kinds of knowledge, as well as knowledge about examples.

This data-driven approach, combined with lexicographic expert knowledge, is the basis for the improvement of dictionary example selection which will be useful both for the production of different dictionaries of Serbian and the forthcoming volumes of the SASA dictionary.

Section 2 describes some steps towards modernization of the dictionary-making process and the development of the digital version of SASA dictionary, starting with retro-digitization process, followed by several ideas about modernization of dictionary-making and the description of the current, traditional practice of dictionary example selection. Section 3 presents a part of the feature distribution analysis of examples from five SASA dictionary volumes, while a comparison with feature distribution in sentence samples extracted from corpora is given in Section 4. The research focused on the development of the initial components of a model for example selection is presented in Section 5, followed by ideas for future work and some concluding remarks at the end of the paper.

2. SASA Dictionary

2.1 SASA Dictionary retro-digitization

The first ideas how to modernize the work on the SASA dictionary came many years ago (Sabo & Vitas, 1989). These ideas were later revitalized and various possibilities for updating the work on this dictionary were considered (Vitas & Krstev, 2015; Ivanović et al., 2016). The modernization of work finally began only in 2016 with digitization of printed volumes (Stijović & Stanković, 2017). Out of 20 volumes already

published, three were available as MS Word files, two as pdf files and others only in paper form. At the same time, a formal description of dictionary entry was produced, and a lexical database model was developed (Stanković et al., 2018).

The conversion of the SASA dictionary from unstructured text into a lexical database consisted of a thorough analysis of formatting conventions that were used for typesetting dictionary entries, as well as identification of triggers (such as special words, abbreviations or punctuation marks) used to introduce specific information. This analysis enabled the recognition of the entry structure: headword group, grammatical data, etymology, lexical units (senses), multiword expressions and proverbs (if any). Each lexical unit may contain linguistic labels (domain, style, time etc.), syntax patterns, definitions, related words, examples of usage, followed by bibliographic references.

2.2 Towards modernization of SASA dictionary-making

Transformation of the digitized text of the SASA dictionary into various standard structured formats and a lexical database was implemented using a custom software solution, with the primary goal to speed up the linear production process of the dictionary. This enabled the use of the lexical database for research purposes. After successful import of two volumes: the 1st and 19th into the database (Stanković et al., 2018), the process continued with another three volumes: 2nd, 18th and 20th.

Dictionary entries are represented by lexical entry elements in the database, with one or more lexical senses (units) that are further illustrated by examples. Each example is followed by information about the bibliographic source, the author, and optionally about the location, and indirectly related to information about the headword of dictionary entry, its part of speech and linguistic labels assigned to the headword and lexical unit. A classification of labels is also incorporated in the database to provide clustering of dictionary (sub)entries using several criteria: by domain (for terminology and specialized vocabulary), by region (dialect), register, style etc. Interlinking of related words is envisaged as more explicit, on the level of lexical units (senses), which will enable the reuse of dictionary content that already exists in the database.

The fine-grained structure of the database enabled the creation of a dataset of examples supported by a set of related information: headword/lexical unit the example is related to, part of speech, and linguistic labels. The dataset of examples derived from the SASA dictionary is a dataset of good dictionary examples that can serve various purposes: it can be used to procure examples for the SASA dictionary as well as for new dictionaries, but it can also be used for the development of a machine system for example selection.

From the analysis of samples of dictionary examples, metrics and example feature distribution can be derived, which can reduce the search space for relevant examples, for example, by setting the upper and lower limits for sentence length, based on the

most common length of example (in words, tokens and characters). Also, having in mind that this dictionary includes citations from a 200-year period, a time boundary can be set when extracting examples for some future dictionary of modern language.

About 12% of all examples in the digitized volumes of SASA dictionary contain lexis marked as obsolete (label заст.), 7% as dialect (дијал.), 4% as irregular (некњ.), 2% as vernacular (нар.), 2% as ephemeral (необ.) and the remaining 2% marked with labels for other types of non-standard lexis, in total 29%. These figures are approximative, since some examples contain lexis marked with several labels, and for this analysis only the first of them was taken into account.

2.3 The current practice of dictionary example selection

2.3.1 Criteria for example selection

Finding appropriate examples in a citation bank as big as the one for the SASA dictionary[1] (about five million paper slips, hand- or typewritten, only recently scanned and partially annotated with headwords) is a difficult and time-consuming job – a lexicographer has to read hundreds, sometimes even thousands of citations (for example, there are 2,830 citations for the preposition *po 'on', 'over', 'by'*). When choosing illustrative examples for lexical units (LU) in the SASA dictionary, lexicographers are not guided by linguistic criteria alone. We will describe here briefly some of other criteria, primarily extralinguistic ones. The corpus of examples in paper form, used for this monolingual thesaurus, was made up of excerpts from resources written in Serbo-Croatian (SC), from the beginning of the 19th century to the present day, as well as about 300-word collections (for details see Stanković et al., 2018). Written texts, as well as word collections, come from what used to be the SC language territory. According to the Style Guide[2], lexicographers have to choose two to six examples for each LU, taking into account the following facts: a) each example should clearly show the meaning of the LU; b) they have to be from different parts of SC language territory; c) they should be from different periods, and listed chronologically, the oldest being the first, while the examples from word collections are given at the end, after all the examples from published sources; d) they should be written by renowned writers. What is not written in the Style Guide (and lexicographers learn it by word of mouth) is that

[1] It is important to emphasize that the citation bank for the SASA dictionary constantly gets up-dated and thus continues to grow – in the course of dictionary-building lexicographers continually consult reference literature (encyclopedias, different kinds of dictionaries, manuals etc.), and some of the recently published books, text-books etc. are also excerpted. The SASA dictionary contains only a small portion of these citations because of the described selection criteria.

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

it is not advisable to use more than one example of the same author. An exception to this rule can be made if there are not enough examples by other authors.

Only a few linguistic criteria are mentioned in the Style Guide. They can be paraphrased as follows: 1) each chosen example should show different relations of the headword with other words (the rection, for example); 2) it is recommended that every example represents a finished syntactic whole – with a subject and a predicate. It is even possible to add a missing sentence constituent, but it has to be in square brackets, as a mark of this kind of editorial intervention. (Though the excerpts are in the form of full sentences, the context they provide is sometimes insufficient, and it is necessary to provide a wider context.) As the first criterion is very important, it needs a more detailed explanation. Namely, the role of the examples is to convey the information about valency and rection of the headword in an implicit way (explicit syntactic information, if required by the Style Guide, is placed before the definition). Since the Style Guide for the SASA dictionary was written during the 1950s, there is no mention of collocations or of using examples to show the most frequent ones.

2.3.2 Editorial interventions and a control corpus

Sometimes it happens that additional examples are needed for a sense or lemma. There are two scenarios in such a case: 1) Experienced lexicographers may rely on their knowledge and invent an illustrative example. If such an example is typical for the standard language, the source is marked by the abbreviation Ed. 'Editor'. The example for the noun *pivnica*, 'pub' is of this kind: *Najbolje je točeno pivo u češkim pivnicama (Ред.). 'The best is draft beer in Czech pubs'* (Ed.). 2) An editor may also provide an example from the non-standard language, which usually means that he/she comes from a specific region; in such a case, the source is marked by the abbreviation of the editor's name.

Editor's intuition may and should be supported by the corpus data. It is common for lexicographers to look for examples in the corpus of contemporary Serbian (SrpKor, developed by D. Vitas and a group of collaborators from University of Belgrade, http://www.korpus.matf.bg.ac.rs/korpus/), which is being used as a control corpus, but they rarely refer to it, although all concordances are associated with data about the source (Vitas & Krstev, 2012; Utvić, 2014).

2.3.3 Allowed and recommended interventions on examples from the corpus

Examples from the corpus may be modified by lexicographers. It is advisable to shorten sentences that are too long, and this kind of intervention should be marked by an ellipsis ("…"). It is allowed to omit all irrelevant sentence constituents (different kinds of modifiers, words in enumerations etc.) or even a whole subordinate clause, if it is not important for illustrating the LU. Here is an example from which the beginning of

the sentence, as well as the relative clause were omitted, being irrelevant for the verb headword: *[omitted: U VII., VI. i V. razredu veliki broj slabih učenika došao je otuda, što su] mnogi učenici [omitted: , koji su iz matematike cele godine imali dobre ocene,] na ispitu [inserted: su] podobivali slabe ocene* '[omitted: In the seventh, sixth and fifth grade the number of bad students increased since] many of the students [omitted: , who had good grades in Mathematics during the school year,] got poor grades on the exam'. The same example shows an inserted part in square brackets, namely "su", the simple present tense form of the verb *to be*, 3rd person plural, which was removed with the first omission. This insertion enabled the editor to form a correct sentence shorter than the original one.

2.3.4 Summary of interventions

The dataset from five dictionary volumes comprises ~60,000 dictionary entries with ~105,000 lexical units (senses). Around 11,500 dictionary entries have headwords with several (numbered) lexical units. In the observed dataset, 70% of data entries have examples. According to the analysed dataset, approximately 71% of the examples were not shortened, 22% were shortened once, 6% twice, and 1% more than twice. Words were inserted (to clarify the meaning or to complement what is missing) in 7% of observed examples, while 93% were without any insertion. In total: 66% of the examples were not modified, 20% had one shortening and no insertions, 6% more than one shortening and no insertions, while 5% had an insertion but were not shortened and 2% had both insertions and shortenings. The number of editorial examples was relatively small, and we have not used these in our test set.

2.3.5 What should a good example contain?

As Atkins and Rundell (2008) point out, there is plenty of evidence when a lexicographer works with corpus data, trying to record how a word behaves, but not all of it is relevant for the description of a word's behaviour. The concept of lexicographic relevance is based on Fillmore's theory of frame semantics. The idea behind the concept is that a proper way to describe a word means that all the constructions it participates in should be identified as well as "all those through which its full semantic potential is to be expressed" (Atkins & Rundell, 2008: 252) should be recorded in the lexicographic database. The concept of lexicographic relevance was illustrated by the analysis of verbs, nouns and adjectives, since any word of this kind "cannot be used correctly if the constructions in which it participates are not known" (*ibid.*). Frame semantics links the meaning of a word with the syntactic contexts in which it occurs. To determine what is relevant for the semantic analysis implies identifying lexicographically relevant sentence constituents for verbs, nouns and adjectives.

An important conclusion by Atkins and Rundell (2008: 272) is that grammatical

contexts for discovering relevant information about keywords may differ depending on their part of speech. For example, if the keyword is a noun, lexicographically relevant co-constituents are its modifiers (the prototypical modifier of a noun in Serbian is an adjective phrase) and complements. If the keyword is an adjective, it is important, too, to consider its modifiers (for example, an adverb) and complements (noun phrases or prepositional phrases). For a verb keyword, it is important to note all its complements (objects, subject and object complements etc.).

The notion of lexicographic relevance may also be applied to the selection of good dictionary examples. The constituents important for proper analysis of an LU are also important for its illustrative examples. All relevant modifiers and complements, which affect the meaning of the LU, should be contained in the illustrative example. If a noun has a complement that affects its meaning, the complement should be represented in the example: *Tada se javila u njega velika ljubav i velika podobnost **za slikarstvo*** (paraphrase: 'In that moment he felt a great affection and a great talent **for painting**')[3]. If a keyword is a verb that in one of its senses takes a subject or object complement, then, of course, this complement has to be represented in the example: *On me smatraše **izgubljenom ovcom*** 'He considered me **a lost sheep**'.

It is important to emphasize that "lexicographic relevance relates to what is relevant for an LU, and not to a lemma" (Atkins & Rundell, 2008: 150). We find similar considerations in Popović (2003). The author also believes that modernization of the description of both syntax and lexicography of Serbian standard language is needed. He points out that it is necessary to establish a relation between syntactic and lexicographic description. As for dictionaries, they should take into account the syntactic distribution of lexemes. Words from major word classes should be treated as central for certain types of syntactic units and syntactic information should be given systematically.

2.3.6 Is a context given in one sentence example enough for all word classes?

Some additional remarks are necessary. Conjunctions in Serbian are often at the initial position of the sentence, demarcating its beginning and delimiting it from the context that precedes it (Popović, 2004: 276–277). In such a case, semantic identification of the conjunction requires the context of the sentence that precedes the one beginning with the conjunction. For example, the conjunction *i* 'and', in one of its senses in the SASA dictionary, signals that an utterance comes as a conclusion, explanation, etc. of the sentence it follows. In this case it is necessary to adduce both the sentence beginning with a conjunction and the one before it: *Obeća, da će ovih dana otići. I održa reč* ('He

[3] A similar, bad example, missing this kind of noun complement, is mentioned in Atkins & Rundel (2008: 460): *One woman in every two hundred is **a sufferer*** (of what?).

promised he would leave one of these days. And he kept his promise').

3. The features of dictionary examples

3.1 The role of example features

In order to facilitate example selection an extraction tool for representative sentences was developed – Good Dictionary EXamples, GDEX (Kilgarriff et al., 2008), used today not only by lexicographers, but also in language teaching and learning. In this paper we present research aimed at the development of a GDEX method for Serbian that ranks corpus sentences and suggests the most appropriate ones.

As the gold standard for the development of our method, dictionary examples from five out of twenty volumes of the SASA dictionary (Stijović & Stanković, 2017), presented in Section 2, were used. The main reason for choosing examples from this dictionary as the gold standard was the fact that they were manually selected by experienced lexicographers[4]. In the first phase we automatically analysed various lexical and syntactic features of the gold standard examples, classified them and compared the results with the control corpus (both gold and control corpus are presented in Section 4). The initial set of features was inspired by Kilgarriff et al. (2008) and Kosem (2017), and guided by recapitulation of features given in Kosem et al. (2019).

3.2 Feature extraction

Feature extraction is enabled by the development of a web service inspired by the work described in Kilgarriff et al. (2008), Kosem (2017), and Kosem et al. (2019), which can presently extract 41 features. The developed service receives a text snippet as a string (in our case a sentence), which can have additional metadata attached (e.g. source, keyword/headword, labels), and returns a dictionary[5] structure comprised of feature names and their values. The list of requested features can also be customized. The system is envisaged to process both the sentences from corpora and dictionary examples extracted from the lexical database. In the text that follows, the term sentence will refer to both dictionary examples and sentences from the control corpus (Section 4).

The implemented set of features is described by metadata, i.e. several attributes are assigned to each feature: code, description, processing level (char, word, and sentence), headword dependency (yes/no), weight (for weighted sum and use in our future model

[4] They were chosen according to the principles described in previous sections (2.3.1 to 2.3.6) of this paper. Since these examples have been subject to multiple check-ups (the dictionary-making process goes through several phases), they can be considered a gold standard.

[5] In Python terminology.

for ranking), type (categorical or quantitative), types of graphical representation and visualization parameters (range, bins). The feature list is not conclusive, and in the future, as a result of the present analysis, other features could be added, and additional metadata assigned to features, such as an eliminatory data range, preferred data range and the like.

For this research a subset of 14 features is taken into consideration:

- Character-based:
  - sentence_length: Number of all characters
  - no_digits: Number of digits
  - no_weird_chars: Number of characters ("#$%&\'()*+-/:;<=>?@[\\]^_`{|}~‘„”…)
  - no_commas: Number of commas
  - no_punctuation: Number of all punctuation marks

- Token-based:

no_all_tokens: Number of all tokens (contiguous sequences of characters e.g. words, numbers, punctuation marks; produced using NLTK's recommended tokenizer [6] *nltk.word_tokenize* (Bird et al., 2009))

  - avg_token_len: Average token length
  - max_token_len: Max token length
  - no_all_words: Number of all words (contiguous sequence of letters)
  - avg_word_len: Average word length
  - no_capitalised_words: Number of words that begin with uppercase, which are not at the beginning of the sentence
  - no_rare_tokens: Number of tokens with frequency threshold in the referent corpus
  - avg_freq_in_corpus: Average word frequency in the referent corpus

- Syntactic features:
  - no_pronouns: Number of tokens tagged as pronouns

The set of features that were computed, but not taken into account in this paper, includes: count of blacklisted words, does the headword occur more than once, the number of lemmas that appear multiple times, does the sentence contain between 15 and 40 tokens, number of tokens that contain both alphabetic and numeric characters, number of tokens tagged as proper names, POS-tag of the first word in the sentence, the position of the headword in the sentence, does the sentence begin with a word from a stoplist, etc. The only features that are specific for this exact research are counts of

[6] http://www.nltk.org/

ellipsis (deletions from the original sentences), inserted segments and lexicographic labels, but they are used for example classification, not for ranking (see Section 4).

The analysis showed that a group of features can be used as filter features, namely, as strong indicators that a sentence should not be used as an example. Sentences that have at least one non-zero value for any feature belonging to this group are categorised as negative samples (e.g. blacklist_count, contains_web_or_email). Features that were not taken into consideration in this analysis were mostly dependent on the headword, e.g. its position in the example. They were classified as headword dependent and will be part of future analysis.

3.3 API for feature extraction

The extraction of features is implemented as a web service[7]. This web service is also used for other tasks, such as text classification and corpus cleaning.

An example of the activation of this web service using *curl* in Unix is the following:

```
curl -d '{"data": "We are demonstrating the usage of our feature extractor!", "lang":"en", "kwic": "usage", "feature_names": ["sentence_length","avg_word_len", "no_all_tokens"]}' -H "Content-Type: application/json" -X POST http://147.91.183.8:12347/features
```

and the fields are:

- data (string) – mandatory, contains text for which features are being extracted
- lang (string) – optional (the default value is “sr” for Serbian, but most of the features can be extracted for English, as well)
- kwic (string) – optional (only for headword-dependent features)
- feature_names (list of strings) – optional (if omitted, returns list of all feature values)

For the given example, the output would be:

```
{"sentence_length": 56, "avg_word_len": 5.222, "no_all_tokens": 10}
```

[7] Extraction of GDEX features, http://gdex.jerteh.rs/.

4. Feature analysis

4.1 The gold and the control dataset

Each example extracted from the SASA dictionary for the gold dataset is supplied with a list of supporting information: volume, dictionary headword, headword's part of speech, linguistic labels (some of which are mentioned in Section 2.2), type of editorial intervention (if any) on the example (shortening or insertion) and a code for the bibliographical source. The size of the gold corpus is 133,904 examples, comprising 1,711,231 words or 10,577,723 characters. Within the gold dataset three types of partitioning were used: 1) by published volume (labelled D01, D02, D18, D19 and D20), 2) by type of lexis/language (labelled with DSS for standard Serbian and DNS for non-standard Serbian) and 3) by part of speech (POS) of the headword/keyword (N – nouns, V – verbs, A – adjectives, ADV – adverbs and X – other).

DSS partition contains sentences in contemporary language with examples that were not modified by editors. We presume that they would be good examples for some future dictionary of contemporary Serbian. DNS contains examples in languages other than standard Serbian (Church Slavonic, Čakavian, Kajkavian), and lexis marked with labels some of which are mentioned in subsection 2.2 (obsolete, dialect, non-standard, vernacular, ephemeral, loanwords, slang). A small number of examples with uncertain boundaries of dictionary entry elements, usually in phrases and proverbs, were excluded from the research, as well as examples from poetry that have the " | " delimiter between verses.

In addition to the corpus made of examples, we prepared a control dataset derived from various texts, which was used as a sample corpus for dictionary example extraction. The control dataset of example candidates was obtained from the digital library Biblisha[8] (Stanković et al., 2017), SrpKor – the corpus of contemporary Serbian (Vitas & Krstev, 2012; Utvić, 2014) and Serbian ELTeC Collection[9]. It consists of several text collections of different types, which reflect text variability. For the first collection with contemporary novels (labelled CN), the sentences were extracted from seven novels written by contemporary Serbian writers and from sevens novels written in German and translated to Serbian. In order to represent domain knowledge, two scientific journals (labelled SJ) were used: *The Journal for Digital Humanities Infotheca*[10] and *Underground Mining Engineering*[11]. The sample labelled DP, with 17 issues of the daily

[8] http://jerteh.rs/biblisha/

[9] *Distant Reading for European Literary History* (COST Action CA16204) https://distantreading.github.io/ELTeC/srp/index.html

[10] http:/infoteka.bg.ac.rs/index.php/en

[11] http://ume.rgf.bg.ac.rs/index.php/ume

newspaper *Politika* published in 2001–2010, was retrieved from SrpKor. A part of the Serbian ELTeC was used, which contains 10 novels and excerpts from 15 novels that were all published 100 or more years ago (labelled ON for old novels). The system for Serbian text processing, based on comprehensive e-dictionaries and local grammar in the form of finite-state automata (Krstev, 2008) was used for sentence segmentation.

Concordances were extracted using appropriate regular expressions, to serve as candidate examples for corresponding headwords in volumes to come. They were bound by sentence delimiters and left/right context of up to 500 characters. The size of the control corpus was 30,104 sentences, comprising 908,980 words or 5,841,700 characters. A sample of 2,752 candidate examples (taken from all parts of the control corpus) was manually evaluated by two lexicographers: they evaluated 1,434 examples as inadequate (useless), 723 as inadequate but improvable with major changes, 441 as good examples in which only minor changes are required, and 154 as very good examples for which no changes are required.

4.2 Feature distribution in the gold dataset of good examples

The comparison by volumes did not show any significant deviations. All feature distributions were similar, as expected, given the same guidelines and methodology used for all published volumes in the last 70 years. Figure 1 presents frequency distribution by number of words in the examples. On the left side each volume of SASA dictionary is represented by a histogram with parts of speech in different colours. The right-hand side shows histograms of partitions of the control dataset.

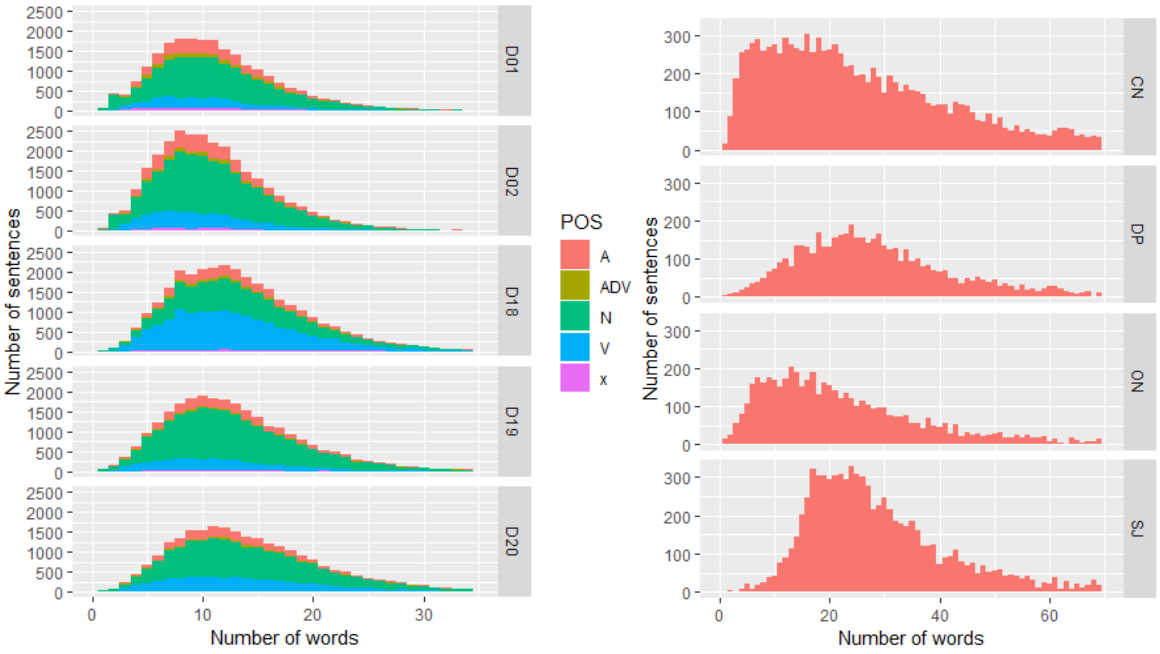

Figure 1: The histogram of the number of words in examples.

The histograms on the left show that sentences in the last three volumes tend to be slightly longer than in the first two, and that nouns (green) are the most numerous words. In volume D18 the number of verbs (blue) is considerably greater than in the other volumes, which can be explained by numerous verbs in this volume beginning with *o,* derived by the productive prefixes *od-* (allomorph *ot-*) and *o-*.

Comparison of lengths of examples for different parts of speech in the SASA dictionary shows that examples for adjectives and nouns tend to be longer than those for adverbs and verbs. Figure 2 presents corresponding boxplots, where the box represents the interquartile interval (IQR) with lower (Q1) and upper quartile (Q3), the middle bold line being the median (Q2), and the rhombus in the middle of the box presenting the average value, with POS on the x-axis and sentence/token length in characters on the y-axis. Dots present outlier examples longer than Q3+1.5*IQR.

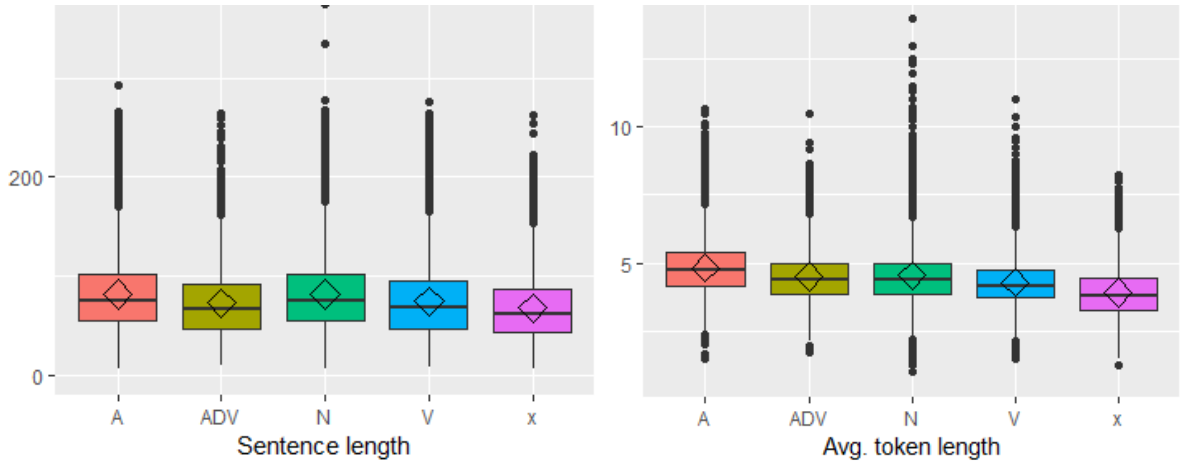

Figure 2: Boxplots showing sentence/token length per POS in the SASA dictionary.

4.3 Feature distribution on both corpora

Figure 3 presents a boxplot diagram of sentence length statistical values per partition (volume and text collection). It can be observed that the sentences in the control dataset partitions are longer than in any volume of the dictionary, that the dispersion for contemporary novels (CN) is the highest, that the average length of sentences in journals and daily papers is similar, and that old novels (ON) have shorter sentences than contemporary ones (CN).

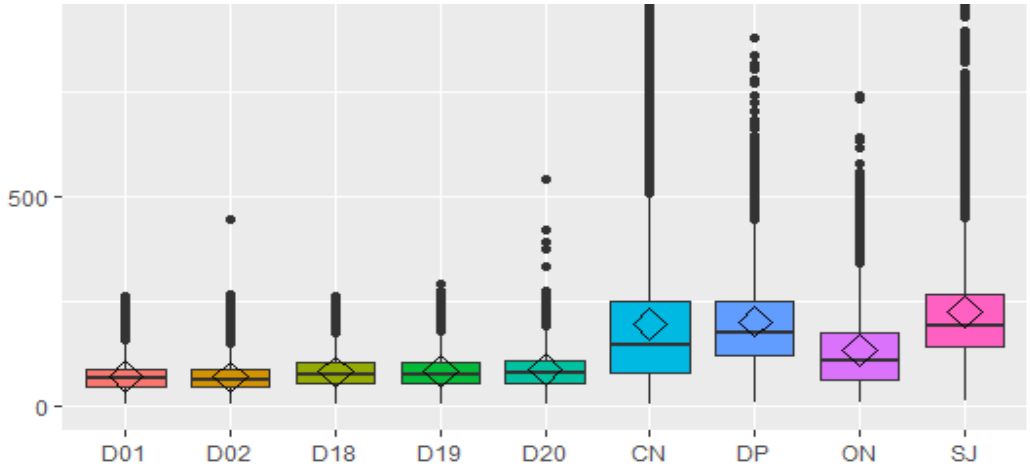

Figure 3: Boxplot of sentence (example) length (in number of characters) per partition.

The distribution of punctuation marks (normalized on sentence size) is presented in Figure 4 on the left: dictionary examples have less punctuation marks than the control corpus. The average word length is similar for all dictionary volumes, slightly shorter for novels and much longer for daily papers and even more for journals (Figure 4, right), probably due to the use of specific terminology, as expected.

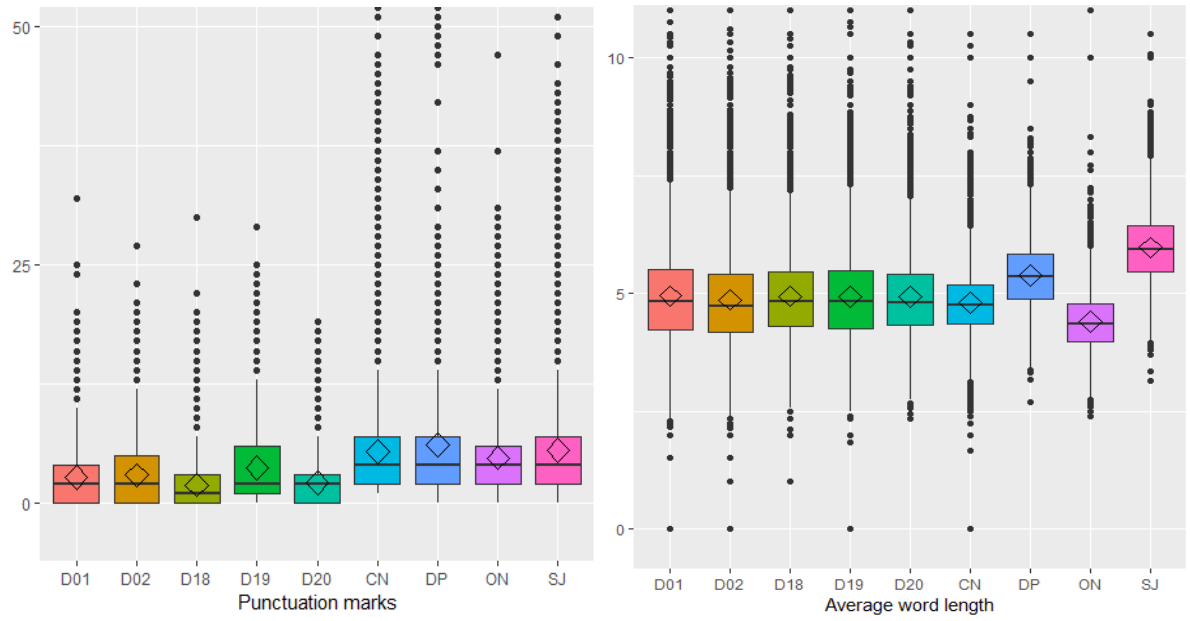

Figure 4: Boxplots for number of punctuation marks and average word length per partition.

According to the corpora, sentences in novels have more pronouns than examples in the SASA dictionary (Figure 5, left). The first two volumes have a very low median, which corresponds to the lexicographers' practice of choosing examples with nouns because they are easier to understand. Sentences extracted from daily papers and scientific journals also have very few pronouns, which can be explained by a greater need for precision in scientific and journalistic language.

In order to approximate and predict the ability of a user (with a specific profile) to understand a specific example, a "frequency indicator" was calculated for each example/sentence (Figure 5, right), as the average frequency of each word in it. The underlying assumption is that the more frequent the words in the example, the greater the possibility that the user will understand it. Word frequencies were obtained from SrpKorp2013 (Utvić, 2014). Examples from novels have higher frequency indicators, while these indicators are lower for examples from journals. The first two volumes of the SASA dictionary have a wider span of frequency indicators than other volumes (as expected, due to the type of the lexis contained in each volume; for example, the majority of the lexis beginning with a, contained in the first volume, is of foreign origin, while the second volume contains lexis mostly labelled as regional, obsolete, ephemeral, etc.).

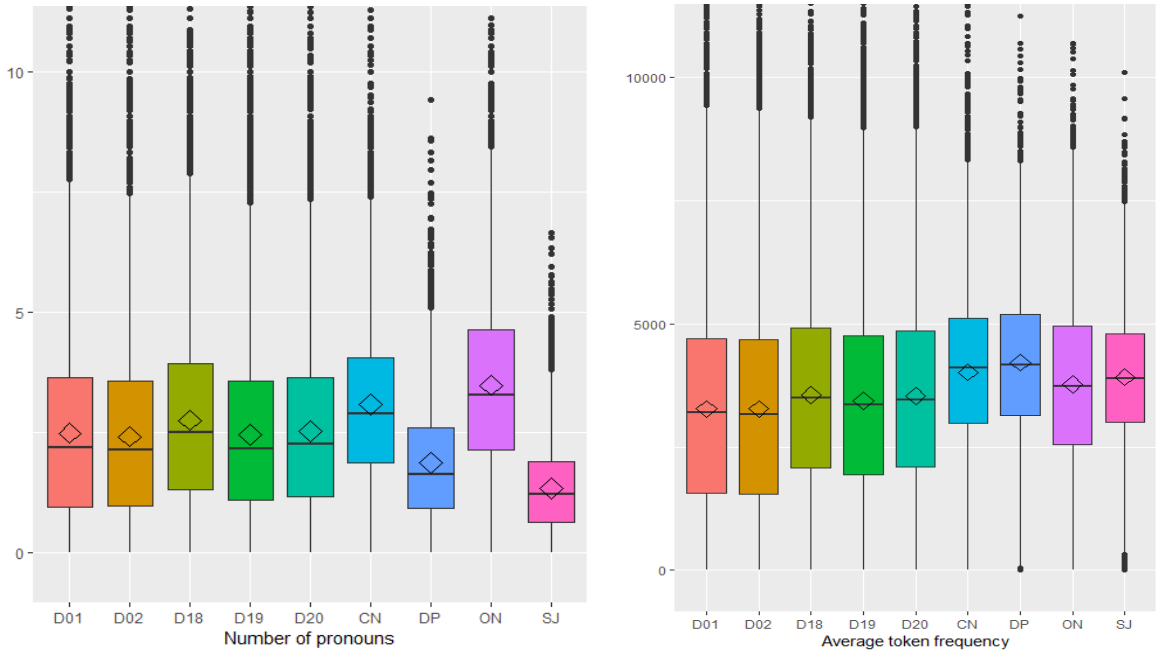

Figure 5: Boxplot of number of pronouns and token frequency per partition.

Figure 6 (left) shows that standard Serbian (DSS) and non-standard (DNS) in the dictionary have a similar distribution of the number of words in the examples, which means that there is no difference in this respect between good examples illustrating standard or non-standard lexis. On the other hand, the evaluated dataset has a wider range for inadequate examples (DNS (NO)), while a similar distribution with those in the dictionary. The results for other features also show that there are no significant differences between examples in DSS and DNS.

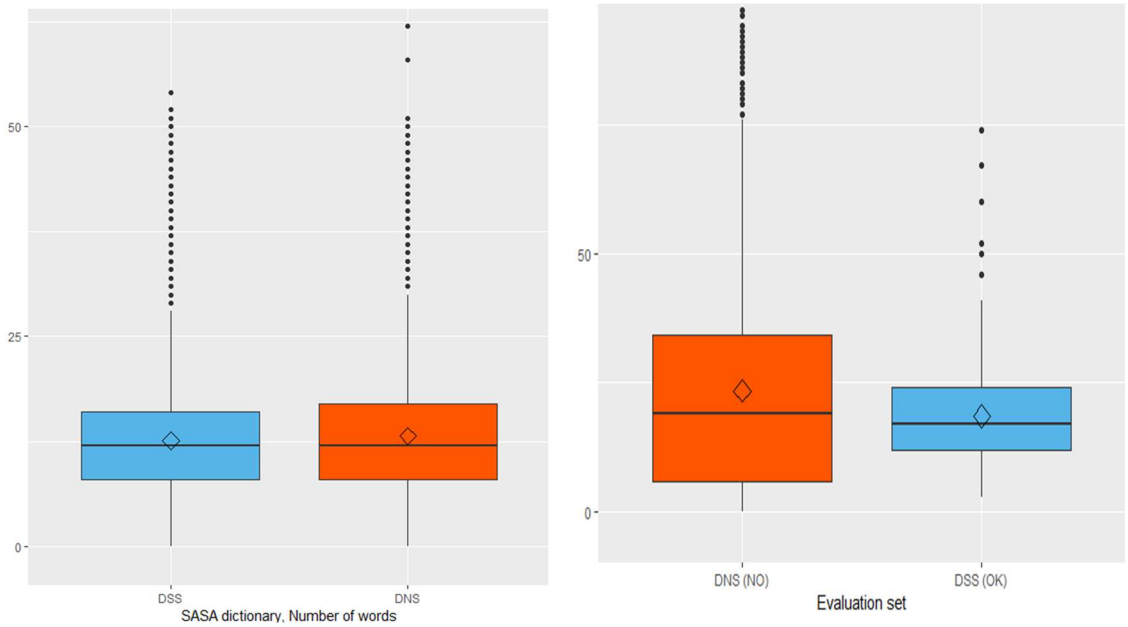

Figure 6: Boxplot of number of words per language type partitions.

Histograms and boxplots were supported by a data summary of calculated features, which offered the guidelines for data cleaning and control corpus preparation. We performed the preprocessing of both datasets we are using (SASA examples and control) and produced data summaries. These were analysed by lexicographers, on the basis of which parameters for potential example cleaning were deduced and threshold values for them were defined. Table 1 presents the data summary from SASA dictionary for five representative features.

Percentile	**Sentence length**	**N° of digits**	**N° of words**	**Avg. word length**	**N° of stop words**	**N° UCase inside sent.**
5th	28	0	5	3.6	0	0
40th	64	0	10	4	3	0
Median	73	0	12	4.8	4	0
65th	87	0	14	5.2	5	0
95th	150	0	25	6.6	10	2

Table 1: Data summary from SASA dictionary for selected features.

5. Preliminary model for identifying good dictionary examples

The future system for semi-automatic identification of good dictionary examples relies on the results of the outlined analysis and includes already developed modules for detection of good examples, as well as for detecting those that are not appropriate examples for standard language use. Filtering and ranking of examples can be

performed using rules obtained from analysed data (feature vectors) combined into a single score. The development of the GDEX function is inspired by the state of the art implementation[12] for which the following functions were developed: *blacklist(), greylist() and optimal_interval()*. For each feature the function *optimal_interval* uses four key percentiles from the gold SASA dataset (as shown in Table 1)[13], where feature values lower than the first and higher than the last are assigned a score of 0.01, in the middle interval scores are 1, and between them a linear interpolation function is used. The four percentiles were computed for different key values, but final results will be deduced after a broader evaluation campaign, with parallel evaluation and adequate interrater agreement. For the *greylist* function only two key values are used (5^{th} and 95^{th} percentiles): values lower of the 5^{th} are assigned a score of 1, higher than 95^{th} a score of 0, and between them linear interpolation is used. Besides the solution with multiple assessments of features, we have also used the analytic hierarchy process (AHP), where each feature value is converted to a numerical value from 0 to 100 and a numerical weight (priority) is assigned to it (the sum of all weights being 1), which gave us better results. The precision calculated on the evaluation set for the first 100 ranked examples was 0.77, for the first 200 it was 0.70, for 400 it was 0.65, for 1,000 it was 0.6, etc. We believe that the results can be improved with additional rules, since the evaluators have noticed that some patterns and some types of sentences can indicate their inadequacy. For example, if the adverb of time or place is not the headword to be illustrated by the example, sentences beginning with these adverbs are not good examples, because they often need the preceding context (*Onda sam otputovao. 'Then I left'*).

Sentences are ranked by a GDEX weighted sum of feature score values, which is then mapped to a user-friendly final score from 1 (poor, lowest 20%) to 5 (good, 20% highest), representing their suitability to serve as examples.

Sentences from the prepared dataset, represented as feature-vectors, were used as the dataset for a supervised Machine Learning (ML) model, which was then used in a GDEX classifier for contemporary Serbian sentences. Since the dataset of examples was unbalanced, with twice as many DSS examples as DNS examples, we have randomly extracted 44,808 (out of 89,096) examples with standard lexis from the DSS dataset and labelled them as 'OK' (positive class) and the same number of examples (44,808) from the DNS set with non-standard lexis (labelled as 'NO' – negative class). Since the manually evaluated sample was small it was replicated five times, yielding 7,165 'NO' and 6,585 'OK' examples.

We used the AdaBoost (Rätsch et al., 2001) algorithm's implementation in Weka (Eibe et al., 2016), a suite of machine learning software. The trained model was evaluated in a 10-CV (cross-validation) setting, with the default Weka parameters for this algorithm.

---

[12] https://www.sketchengine.eu/syntax-of-gdex-configuration-files/

[13] The 40th and 65th percentiles of the SASA dictionary for number of words are the same as the values in the example given to the Sketch Engine.

In the first decision step, the most distinctive feature, as expected, was *abbrev* (the indicator of the existence of a linguistic label). Namely, the corresponding rule is: "if the *abbrev* linguistic label is missing, there is a 92% chance that the sample is positive". The confusion (error) matrix represents the features in predicted and actual classes: true positive 7,475 (0.68); false positive 3,581 (0.32); true negative 9,729 (0.87); false negative 1,451 (0.13). This result can be considered satisfactory; however, there is a serious issue – the existence of a linguistic label *abbrev* cannot be expected for corpora in general. Therefore, we wanted to build another classifier that uses other features.

The first step is feature analysis and feature selection. We first determined and visualised a Pearson correlation matrix that contains the correlation of features to manually assigned labels, where green represents a strong positive correlation, red a strong negative correlation, and yellow no correlation. After removing irrelevant features (those that have a very low correlation with *label*, like *avg_word_len*, or those that are highly correlated with each other, such as *max_word_len* and *max_token_len*), we represented each sample with the shorter feature vector (Figure 7).

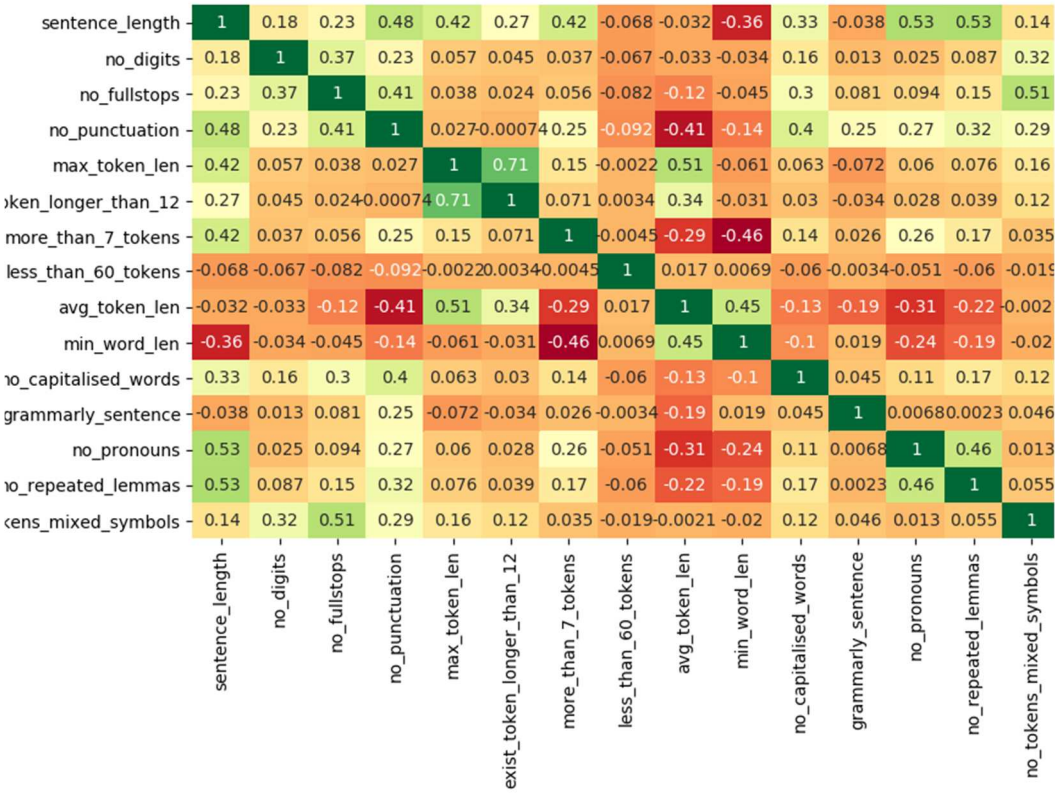

Figure 7: Pearson correlation matrix.

The gold dataset was split into a training and a validation set (20% of the dataset). The results of the Logistic Regression (Hosmer et al., 2013) classifier are given in Table 2, where NO stands for non-standard and OK for standard language.

	Precision	Recall	F1-score	Number of samples
NO (NS)	0.84	0.68	0.75	11,056
OK (SS)	0.73	0.87	0.79	11,180
ALL	0.78	0.77	0.77	22,236

Table 2: Results of the logistic regression binary classifier.

All metrics show better results for the negative class. Out of 11,056 negative samples in the validation set, 7,520 were classified as negative (68%, true negative), and the remaining ones as positive (23%, false positive). From 11,180 positive samples, 9,727 were classified as positive (87%, true positive), and the remaining ones as negative (13%, false negative).

The feature extractor is freely available, while the GDEX ranking and trained ML model are available for authorized users. The future system for semi-automatic identification of good dictionary examples implies the development of more modules, e.g. a user interface for feature extraction and for GDEX parameter fine tuning, but the evaluation of the first results of the developed core components is encouraging.

6. Future work and concluding remarks

The first results are encouraging, and they motivate further detailed analysis of other computed features and the introduction of new ones. Improvement of the weighted measure of features will follow, with a combination of expert knowledge and data training results.

Implementation of other features and criteria will be integrated into the web application and selections of parameters and features to be calculated will be enabled. Full system integration will combine the use of a lexical database with corpora exploitation via the developed web service and software. Since the work on digitization of other volumes of the SASA dictionary is continuing, more data is expected to bring more refined conclusions.

There is obviously a lot of room for improvement of the trained model, e.g. with the introduction of new features, by adding more samples, or using other state-of-the-art neural network architectures. Another future step is the model's evaluation on a control dataset – extraction and ranking performance is going to be tested by more lexicographers, with parallel evaluation and interrater agreement checking. Finally, we also plan to introduce flexible mapping of computing scores – from 1 (worst) to 5 (best) – and score our examples using them. This can be performed either by looking at the rules and constructing an equation, or by a trained classifier.

7. Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #178009, #III 47003 and #178003.

8. References

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bird, S., Loper, E. & Klein E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Eibe, F., Hall, M. A. & Witten, I. (2016). *The Weka Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, Fourth Edition.

Hosmer Jr., D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Vol. 398. Hoboken, NJ: John Wiley & Sons.

Ivanović, N., Jakić, M. & Ristić, S. (2016). Građa Rečnika SANU – potrebe i mogućnosti digitalizacije u svetlu savremenih pristupa. In S. Ristić et al. (eds.) *Leksikologija i leksikografija u svetlu savremenih pristupa*, Beograd: Institut za srpski jezik SANU, pp. 133–154. [The material of the Dictionary of the SANU - the needs and possibilities of digitization in the light of contemporary approaches (in Cyrillic)].

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress, EURALEX 2008. Barcelona: Universitat Pompeu Fabra*, pp. 425–432.

Kosem, I. (2017). Dictionary examples. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana: University of Ljubljana, Faculty of Arts.

Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2019). Identification and Automatic Extraction of Good Dictionary Examples: the Case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119–137.

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

Popović, Lj. (2003). Integral sentence models and their importance for lexicographic description and corpus analysis [Integralni rečenični modeli i njihov značaj za lingvistički opis i analizu korpusa]. *Naučni sastanak slavista u Vukove dane*, 31(1), pp. 201–220. (In Serbian, cyrillic.)

Popović, Lj. (2004). *Red reči u rečenici* [Word order in sentences]. Beograd: Društvo za srpski jezik i književnost Srbije. (In Serbian, Cyrillic.)

Rätsch, G., Onoda, T., & Müller, K. R. (2001). Soft margins for AdaBoost. *Machine learning*, 42(3), pp. 287–320.

Sabo, O. & Vitas, D. (1998). Mogućnost osavremenjivanja izrade rečnika na primeru

Rečnika srpskohrvatskog književnog i narodnog jezika SANU i Instituta za srpskohrvatski jezik. In *IV međunarodni naučni skup „Računarska obrada jezičkih podataka",* Portorož: Institut Jožef Stefan, pp. 375–384 [Possibility for modernizing the development of the dictionary on the example of the Dictionary of the Serbo-Croatian literary and vernacular language SASA and the Institute for Serbo-Croatian].

SASA Dictionary: Речник српскохрватског књижевног и народног језика САНУ, I–XX (The Dictionary of the Serbo-Croatian Standard and Vernacular Language) (1959–2017). Београд: Институт за српски језик САНУ и САНУ.

Stanković, R., Krstev, C., Vitas, D., Vulović, N. & Kitanović, O. (2017). Keyword-Based Search on Bilingual Digital Libraries. In A. Calì, D. Gorgan & M. Ugarte (eds.) *Semantic Keyword-Based Search on Structured Data Sources. COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9, 2016, Revised Selected Papers*, pp. 112–123. DOI:10.1007/978-3-319-53640-8_10.

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 J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana, pp. 941–949. Available at: https://euralex.org/category/publications/euralex-2018/.

Stijović, R. & Stanković, R. (2017). Digital edition of the SASA Dictionary: a formal description of the microstructure of the SASA Dictionary [Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU]. *Naučni sastanak slavista u Vukove dane,* 47(1), pp. 427–440. (In Serbian, Cyrillic.)

Utvić, M. (2014). The construction of reference corpus of contemporary Serbian [Izgradnja referentnog korpusa savremenog srpskog jezika] (Doctoral dissertation, University of Belgrade).

Vitas D. & Krstev C. (2015). Blueprint for the computerized dictionary of the Serbian language [Nacrt za informatizovani rečnik srpskog jezika]. *Naučni sastanak slavista u Vukove dane*, 44(3), pp. 105–116. (In Serbian, Cyrillic.)

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

Zgusta, L. (1971). *Manual of Lexicography.* Praha: Academia.