# hrMWELex – a MWE lexicon of Croatian extracted from a parsed gigacorpus

**Nikola Ljubešić[1], Kaja Dobrovoljc[2], Simon Krek[3], Marina Peršurić Antonić[4], Darja Fišer[4]**

[1] Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
I. Lučića 3, HR-10000 Zagreb
nljubesi@ffzg.hr

[2] Trojina, Institute for Applied Slovene Studies
Dunajska 116, SI-1000 Ljubljana
kaja.dobrovoljc@trojina.si

[3] Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana
simon.krek@ijs.si

[4] Faculty of Arts
Aškerčeva 2, SI-1000 Ljubljana
mpersuric@gmail.com, darja.fiser@ff.uni-lj.si

## Abstract

The paper presents the process of building the hrMWELex lexicon of multiword expressions extracted from the 1.9 billion-token parsed corpus of Croatian. The lexicon is built with the newly developed DepMWEx tool which uses dependency syntactic patterns to identify MWE candidates in parse trees. The extracted MWE candidates are subsequently scored by co-occurrence and organized by headwords producing a resource of more than 30 thousand headwords and 12 million MWE candidates. The evaluation of the lexicon showed an overall precision of just over 50% and quite varying precision over specific syntactic patterns. Finally, opportunities for the refinement and enrichment of this recall-high resource by distributional identification of non-transparent MWEs and cross-language linking are presented.

### hrMWELex – Leksikon hrvaških večbesednih zvez, izluščenih iz skladenjsko označenega milijardnega korpusa

V prispevku predstavimo postopek izdelave leksikona hrMWELex, ki smo ga izluščili iz korpusa hrvaških besedil, ki je skladenjsko označen in vsebuje 1,9 milijarde besed. Leksikon smo zgradili s pomočjo orodja DepMWEx, ki za prepoznavanje kandidatov večbesednih zvez v odvisnostnih drevesih uporablja odvisnostne skladenjske vzorce, jih rangira in organizira glede na jedrno besedo. Izluščen leksikon vsebuje 30.000 jedrnih besed in 12 milijonov večbesednih zvez. Evalvacija leksikona pokaže natančnost luščenja, ki presega 50%, pri čemer natančnost pri različnih skladenjskih vzorcih zelo niha. Na koncu prispevka predstavimo možnosti za izboljšave in razširitev bogatega leksikona s pomočjo prepoznavanja netransparentnih večbesednih zvez s pomočjo načel distribucijske semantike ter možnosti povezovanja večbesednih zvez z ustreznicami v drugih jezikih.

## 1. Introduction

Multiword expressions (MWEs) are an important part of the lexicon of a language. There are various estimates on the number and therefore importance of MWEs in languages, but most claims point to the direction that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words (Baldwin and Kim, 2010).

There are two basic approaches to identifying MWEs in corpora: the symbolic approach, which relies on describing MWEs through patterns on various grammatical levels, and the statistical approach, which relies on co-occurrence statistics (Sag et al., 2001). Most approaches take the middle road by defining filters through the symbolic approach and rank the candidates passing the symbolic filters by the statistical approach.

The two most frequently used grammatical levels used for describing MWEs are the one of morphosyntax and syntax (Baldwin and Kim, 2010). While morphosyntactic patterns (Church et al., 1991; Clear, 1993) are much more used since they have already yielded satisfactory results, there is a number of approaches that use the syntactic grammatical level as well (Seretan et al., 2003; Martens and Vandeghinste, 2010; Bejček et al., 2013).

In this paper we describe an approach that relies on syntactic patterns to identify MWE candidates. Our main argument for using the syntactic grammatical level is that on languages with partially free word order, such as Slavic languages, morphosyntactic patterns often have to rely on hacks, like allowing up to $n$ non-content words between fixed words or classes, thereby keeping the precision under control while at the same time trying not to loose too much recall. Still, a significant amount of recall is lost since often only the most frequent order of constituents of an MWE is taken into account.

On the other hand, an argument against using syntax for describing MWEs is the precision of the syntactic analysis which is around 80% for well-resourced Slavic languages while morphosyntactic description of well resourced Slavic languages regularly passes the 90% bar.

Most approaches that use the syntactic grammar layer for extracting MWEs, like (Pecina and Schlesinger, 2006) and the recently added feature in the well-known SketchEngine (Kilgarriff et al., 2004), take into account only MWEs consisting of two nodes, therefore missing the big opportunity syntax offers in defining much more complex patterns that could not be defined on the morphosyntactic level at all.

Until now, there were no efforts in producing large-scale MWE resources for Croatian. First experiments include (Tadić and Šojat, 2003) who use PoS filtering, lemmatization and mutual information to identify candidate terms as a preprocessing step for terminological work, (Delač et al., 2009) who experiment on a Croatian legislative corpus while developing the TermeX tool for collocation extraction and (Pinnis et al., 2012) who use the CollTerm tool, part of the ACCURAT toolkit, for extraction of terms as the first step in producing multilingual terminological resources. All the mentioned approaches use morphosyntactic patterns for identifying candidates and do not produce any resources. The only resource for Croatian that does rely on syntactic relations is the distributional memory DM.HR (Šnajder et al., 2013), whose primary goal is distributional modeling of meaning.

In this paper we describe our tool that enables writing complex dependency syntactic patterns for identifying MWE candidates and the resulting recall-oriented MWE resource obtained by applying the tool to a 1.9 billion-token parsed corpus of Croatian. As no such lexicon currently exists for Croatian and because it is unrealistic to expect heavy investment in similar resources in the near future, our goal is to build a universal resource that will be useful in a wide range of HLT (human language technologies) applications as well as to professional language service providers and the general public. We therefore aim to strike a balance between recall and precision, giving a slight preference to recall in the hope that, on the one hand, human users can deal with the errors efficiently, and applications on the other can resort to post-processing steps in order to mitigate negative effects of noise in the resource.

The paper is structured as follows: in the next section we describe the DepMWEx tool used in building the resource, in Section 3 we describe the resource in numbers and give its initial evaluation, in Section 4 we discuss further possibilities like calculating semantic transparency and taking a multilingual approach, and conclude the paper in Section 5.

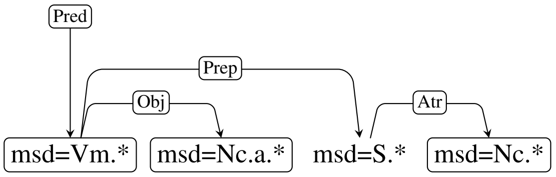

Figure 1: An example of the pattern tree corresponding to the MWE *tražiti rupu u zakonu*, *raditi račun bez konobara* (literally *to write the check without the waiter*), *raditi od buhe slona* (literally *make an elephant out of a fly*, *overexaggerate*) etc.

## 2. The DepMWEx tool

Our DepMWEx (Dependency Multiword Extractor) tool[1] consists of a Python module (defining the Tree and Node classes) and Python scripts that, given a grammar and a dependency parsed corpus, produce a list of strongest collocates for each headword.

### 2.1. The grammar

The grammar consists of a set of grammatical relations, each of which can be described with one or more so-called pattern trees.

Patterns trees are hierarchical structures in which each node contains a boolean function that defines the criterion a node in the parse tree of a sentence must satisfy to fill up that node. An example of a pattern tree, corresponding to the MWE *tražiti rupu u zakonu* (literally *search for a hole in the law*), which will be our working example in this section, is given in Figure 1. This pattern tree describes parse subtrees that have a predicate as a main verb which has direct object and prepositional phrase attached to it. The framed nodes represent headwords, i.e. for the example *tražiti rupu u zakonu*, this MWE candidate will be added to the headwords *tražiti#Vm*, *rupa#Nc* and *zakon#Nc*.

### 2.2. Grammatical relation naming

The name of the grammatical relation of our MWE example is “gbz sbz4 u sbz6”, which is a notation taken over from the Slovene Sketch grammar (Kosem et al., 2013). That grammar is defined over morphosyntactic patterns, and, for reasons of compatibility, this Croatian grammar is based on that notation. The acronym denotes the part of speech (“gbz” being verb, “sbz” noun, “pbz” adjective and “rbz” adverb) while the number denotes the case, and “sbz4” stands for a noun in the accusative case. Finally, one can observe that in the grammatical relation the preposition is lexicalized, which is taken over from the Sketch grammar formalism.

Which part of the grammatical relation is the actual headword the MWE candidate occurs under is labeled by uppercasing that grammatical relation element, so under the verb *tražiti#Vm*, the MWE candidate *tražiti rupu u zakonu* will appear under the grammatical relation “GBZ sbz4 u sbz6”.

[1] `https://github.com/nljubesi/depmwex`

### 2.3. Candidate extraction

The candidate extraction procedure is the following: over each parsed sentence from the corpus, each pattern tree makes an exhaustive search for sentence subtrees that satisfy its constraints. All subtrees corresponding to a pattern tree of a specific grammatical relation are written to standard output as (subtree, grammatical relation) pairs.

### 2.4. Candidate scoring

Once all (subtree, grammatical relation) pairs are extracted from the corpus, co-occurrence weighting is performed and MWE candidates are organized by their headwords and their grammatical relations. For now only the log-Dice measure (Rychlỳ, 2008), the association measure used in the Sketch Engine, is implemented in the tool. A selection of the resulting output for the headword *tražiti#Vm* is given in Table 1.

## 3. Resource description

### 3.1. The corpus

The lexicon was extracted from the second version of the Croatian Web corpus hrWaC (Ljubešić and Klubička, 2014), containing 1.9 billion tokens. The corpus was annotated with morphosyntactic, lemmatization and dependency parsing models built on the SETimes.HR manually annotated corpus (Agić and Ljubešić, 2014).

### 3.2. The grammar

The grammar for Croatian used in the DepMWEx tool was modified from the grammar for Slovene, which is based on the Slovene sketch grammar used in the SSJ project.[2] At this point the grammar consists of 63 grammatical relations defined through the same number of patterns trees. The constituents of the pattern trees are nouns in 53 relations, verbs in 33 relations, adjectives in 15 relations and adverbs in 11 relations.

### 3.3. The resulting lexicon

The resulting lexicon was filtered by the available lexical resources for Croatian, the Croatian morphological lexicon[3] and the Apertium morphological lexicon for Croatian.[4] Two frequency thresholds were enforced during the extraction process: the MWE candidate had to be of frequency 5 or higher, and the lexeme had to form at least 5 MWE candidates satisfying the first threshold. Entries for 46,293 lexemes (19,041 nouns, 11,183 adjectives, 7,028 verbs and 2,058 adverbs) were produced containing all together 12,750,029 MWE candidates. The relationship between the number of grammatical relations, the number of MWE candidates and the respective part of speech of the head is depicted in Figure 2. It shows that nouns are the most productive part of speech, being followed by verbs, adjectives and adverbs.

| tražiti#Vm | logDice | freq |
|---|---|---|
| GBZ sbz4 | | |
| pomoć#Nc | 8.358 | 9410 |
| odšteta#Nc | 7.958 | 1949 |
| odgovor#Nc | 7.851 | 4339 |
| povrat#Nc | 7.775 | 1952 |
| ostavka#Nc | 7.763 | 1900 |
| zvijezda#Nc | 7.503 | 2490 |
| smjena#Nc | 7.354 | 1385 |
| rješenje#Nc | 7.116 | 3127 |
| posao#Nc | 7.071 | 6353 |
| naknada#Nc | 7.031 | 1713 |
| sbz1 GBZ sbz4 | | |
| prodavač#Nc način#Nc | 8.457 | 330 |
| tužiteljstvo#Nc kazna#Nc | 7.295 | 147 |
| čovjek#Nc mudrost#Nc | 6.932 | 114 |
| čovjek#Nc pomoć#Nc | 6.840 | 108 |
| sindikat#Nc povećanje#Nc | 6.801 | 104 |
| tužitelj#Nc kazna#Nc | 6.575 | 89 |
| prosvjednik#Nc ostavka#Nc | 6.057 | 62 |
| čovjek#Nc odgovor#Nc | 6.001 | 60 |
| žena#Nc muškarac#Nc | 5.893 | 58 |
| radnica#Nc pomoć#Nc | 5.832 | 53 |
| rbz GBZ | | |
| uporno#Rg | 7.589 | 715 |
| stalno#Rg | 7.579 | 1434 |
| GBZ sbz4 za sbz4 | | |
| ponuda#Nc podizanje#Nc | 10.831 | 587 |
| rješenje#Nc problem#Nc | 7.465 | 60 |
| sredstvo#Nc ideja#Nc | 6.995 | 39 |
| stan#Nc najam#Nc | 6.871 | 36 |
| naknada#Nc šteta#Nc | 6.869 | 36 |
| obračun#Nc život#Nc | 6.756 | 33 |
| GBZ po sbz5 | | |
| vrlet#Nc | 6.118 | 7 |
| internet#Nc | 5.612 | 227 |
| džep#Nc | 5.487 | 36 |
| kontejner#Nc | 5.334 | 29 |
| oglasnik#Nc | 4.718 | 10 |
| kvart#Nc | 4.714 | 21 |
| inercija#Nc | 4.623 | 5 |
| forum#Nc | 4.263 | 115 |
| knjižara#Nc | 4.181 | 8 |

Table 1: Part of the output of the DepMWEx tool for the headword tražiti#Vm

The final resource is encoded in XML and published[5] under the CC-BY-SA 3.0 license.

## 4. Initial resource evaluation

We performed an initial evaluation of the resource by inspecting up to 20 first MWE candidates for each grammatical relation of 12 selected lexemes. The analyzed lexemes were sampled as follows: 3 lexemes were taken for each part of speech, one in the upper, one in the medium and one

[2] `http://eng.slovenscina.eu`

[3] `http://hml.ffzg.hr`

[4] `http://sourceforge.net/p/apertium/svn/HEAD/tree/languages/apertium-hbs/`

[5] `http://nlp.ffzg.hr/resources/lexicons/hrmwelex/`

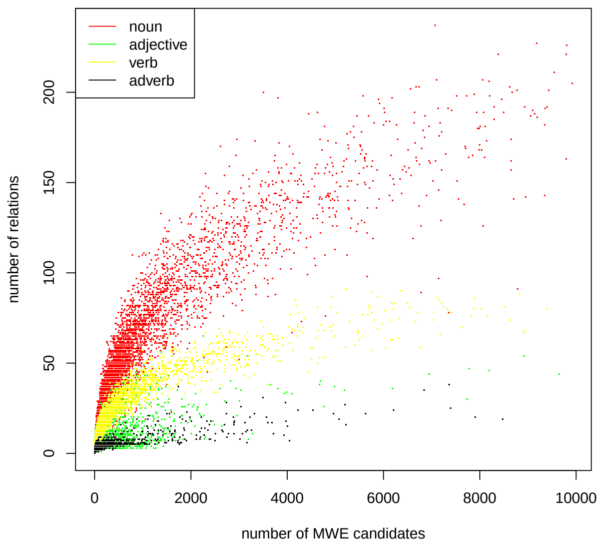

Figure 2: The relationship between the number of relations and MWE candidates by part-of-speech for each lexeme in the resulting lexicon

| lexeme | # evaluated | precision |
|---|---|---|
| burza#Nc | 559 | 0.735 |
| lampa#Nc | 154 | 0.422 |
| lavež#Nc | 34 | 0.324 |
| N | 747 | 0.652 |
| gurati#Vm | 311 | 0.296 |
| razumjeti_se#Vm | 161 | 0.484 |
| tužiti_se#Vm | 77 | 0.26 |
| V | 549 | 0.346 |
| dužan#Ag | 279 | 0.29 |
| legendaran#Ag | 64 | 0.609 |
| svrhovit#Ag | 20 | 0.4 |
| A | 363 | 0.353 |
| naprosto#Rg | 85 | 0.859 |
| trostruko#Rg | 78 | 0.615 |
| jednoglasno#Rg | 62 | 0.806 |
| R | 225 | 0.76 |
| all | 1884 | 0.518 |

Table 2: MWE candidate precision on each of the 12 evaluated lexemes

in the lower frequency range. We had one human annotator at our disposal annotating each MWE candidate as being a MWE or not. The precision obtained on each of the 12 lexemes, along with summaries for each part of speech and all lexemes, is given in Table 2. We can observe that the overall precision of the MWE candidates is just above 50% and that nouns and adverbs are more accurate than verbs and adjectives. Inside each part of speech the MWE candidate accuracies vary significantly and there is no correlation between the frequency range of a lexeme and its precision (the lexemes are ordered by falling frequency).

Next, we analyzed the precision of each specific grammatical relation. The precision for each grammatical relation occurring 10 or more times in the 12 lexemes is given in Table 3. The worst performing set of grammatical relations are the "in/ali" (*and/or*) relations which search for the same-POS constituents combined with the *and* or *or* conjunction. Another frequent and poorly performing relation is the one of a noun subject and its main verb predicate when the verb is the head (sbz1 GBZ) while significantly better results (0.64 vs. 0.167) are obtained with the subject as the head of relation (SBZ1 gbz). A similar phenomenon can be observed with the grammatical relation consisting of a main verb and its direct object which is performing very poorly when the verb is considered the head of the relation (GBZ sbz4), but with noun as head (gbz SBZ4), the obtained precision is much higher (0.214 vs. 0.714). This result stresses the fact that some relations are actually not symmetric and that the relations as they are defined now have to be reconsidered in the future.

## 5. Lexicon refinement

At this point we produced a recall-high resource with satisfactory precision, just over 50%, and the next obvious step is additional filtering of the resource with the goal of getting the precision rate up without hurting recall. Besides filtering, classifying the MWE candidates into types of MWEs should be looked into as well.

### 5.1. Semantic transparency

One of the properties of MWEs we are especially interested in is semantic transparency. We have already performed initial experiments in identifying that type of idiosyncrasy by using the distributional approach.

We built context vectors for all MWE candidates that fall under the following grammatical relations: "pbz0 SBZ0", "SBZ0 sbz2" and "VBZ sbz4". Besides building context vectors for MWE candidates, we also built vectors for their heads.

We built context vectors from three content words to the left and right, stopping at sentence boundaries. We took into consideration only MWE candidates occurring 50 times or more, which we consider minimum context information for any prediction. We used TF-IDF for weighting the vector features and Dice similarity for comparing vectors. We obtained the IDF statistic from head context vectors. The full procedure applied in calculating semantic transparency is the following:

1. build the frequency context vector for each MWE and its head
2. subtract the MWE vector frequencies from the headword vector (thereby remove contextual information of that MWE)
3. transform both vectors to TF-IDF vectors
4. calculate the Dice similarity score between each MWE and its head

By inspecting MWE candidates, organized under their heads and ordered by the computed similarity to the head, we observed quite promising results. We give a few examples for the simplest "pbz0 SBZ0" relation:

| relation | frequency | precision |
|---|---|---|
| pbz0 SBZ0 | 94 | 0.809 |
| RBZ gbz | 73 | 0.822 |
| RBZ pbz0 | 65 | 0.923 |
| rbz GBZ | 60 | 0.5 |
| sbz1 GBZ | 60 | 0.167 |
| RBZ RBZ | 52 | 0.558 |
| SBZ1 gbz | 50 | 0.64 |
| GBZ u sbz5 | 49 | 0.204 |
| GBZ0 in/ali GBZ0 | 47 | 0.213 |
| PBZ0 in/ali PBZ0 | 47 | 0.277 |
| GBZ na sbz4 | 46 | 0.283 |
| SBZ0 in/ali SBZ0 | 45 | 0.0 |
| gbz SBZ4 | 42 | 0.714 |
| GBZ sbz4 | 42 | 0.214 |
| rbz PBZ0 | 42 | 0.357 |
| sbz0 SBZ2 | 42 | 0.667 |
| GBZ u sbz4 | 41 | 0.829 |
| SBZ0 sbz2 | 32 | 0.656 |
| RBZ Vez-gbz pbz1 | 27 | 0.704 |
| gbz Inf-GBZ | 25 | 0.64 |
| SBZ0 u sbz5 | 24 | 0.208 |
| gbz na SBZ4 | 23 | 0.652 |
| gbz na SBZ5 | 22 | 0.727 |
| rbz Vez-gbz PBZ1 | 22 | 0.227 |
| SBZ1 gbz sbz4 | 22 | 0.864 |
| sbz0 na SBZ5 | 20 | 0.9 |
| PBZ0 Inf-gbz | 20 | 0.85 |
| gbz s SBZ2 | 20 | 1.0 |
| sbz0 na SBZ4 | 20 | 0.7 |
| sbz0 s SBZ2 | 20 | 0.95 |
| PBZ0 u sbz5 | 20 | 0.05 |
| gbz sbz4 na SBZ5 | 20 | 0.85 |
| pbz0 na SBZ5 | 20 | 1.0 |
| GBZ sbz6 | 19 | 0.421 |
| PBZ0 za sbz4 | 18 | 0.278 |
| SBZ0 na sbz5 | 17 | 0.765 |
| SBZ0 za sbz4 | 17 | 0.529 |
| SBZ0 od sbz2 | 16 | 0.375 |
| PBZ0 sbz6 | 16 | 0.125 |
| PBZ0 prije sbz2 | 15 | 0.6 |
| GBZ sbz4 u sbz4 | 14 | 0.5 |
| PBZ0 na sbz4 | 13 | 0.154 |
| PBZ0 po sbz5 | 13 | 0.308 |
| SBZ0 s sbz6 | 13 | 0.615 |
| GBZ do sbz2 | 12 | 0.417 |
| SBZ0 o sbz5 | 12 | 1.0 |
| PBZ0 na sbz5 | 12 | 0.083 |
| PBZ0 o sbz5 | 11 | 0.182 |
| sbz0 za SBZ4 | 11 | 0.818 |
| GBZ prema sbz5 | 11 | 0.455 |
| sbz1 gbz SBZ4 | 10 | 0.9 |
| SBZ0 u sbz4 | 10 | 0.8 |
| sbz1 GBZ sbz4 | 10 | 0.3 |
| gbz preko SBZ2 | 10 | 1.0 |
| GBZ s sbz6 | 10 | 0.6 |
| PBZ0 od sbz2 | 10 | 0.1 |

Table 3: Precision scores per grammatical relations (sorted by frequency)

- for the head *voda* (water), the most distant MWE candidate is *amaterska voda* (*amaterske vode* refers to a person who moves from professional to amateur), the second one being *Baška voda* (a municipality in Croatia)
- for the head *selo* (village), the two most distant MWE candidates are *Novo Selo Žumberačko* (a municipality) and *špansko selo* (refers to something absolutely unknown to someone, like *it's all Greek to me*)
- for the head *stan* (flat) the least similar MWEs are *vječni stan* (*eternal resting place*, an experimental dark music album and the Catholic metaphor for heaven), *Ninski stanovi* (a municipality) and *tkalački stan* (sewing machine)
- for the head *ured* (office), the most distant MWE is *ovalni ured* (the Oval office)
- for the head *sastanak* (meeting), the most distant MWE is *Brijunski sastanak* (an important meeting during the Croatian independence war)
- for the head *zlato* (gold), among the most distant MWEs are *tekuće zlato* (referring to any liquid which is very valuable) and *crno zlato* (referring to oil)

On the other hand, once we sorted all the results, regardless of their head, the results seem much less usable. Besides non-transparent MWEs, we obtain probable parsing errors, low-frequency entries, entries with very static context etc. Nevertheless, the obtained results can be very useful for a lexicographer inspecting a specific headword and will therefore be added to the new version of the lexicon.

### 5.2. Multilinguality

We have already made first inquiries in the multilingual setting by producing similar lexicons for two other south Slavic languages, namely Slovene [6] and Serbian[7], but using smaller amounts of data. Since the grammatical relations have the same names in grammars of all the languages, we can use *(grammatical relation, dependents)* pairs as features for our context vectors, obtaining therefore a more detailed and selective formalization of the context of a lexeme than in the standard distributional approach as implemented in the previous subsection. We thereby possibly form more potent distributional memories (Baroni and Lenci, 2010) for tasks of inducing multilingual lexicons of closely related languages by using lexical overlap or similarity, as was done in (Ljubešić and Fišer, 2011). It would be interesting to inspect how such a memory compares to the already existing distributional memory of Croatian DM.HR (Šnajder et al., 2013) which takes into account only binary relations.

We give here one example for the Croatian–Serbian language pair. The Serbian noun *vaspitanje* is not present

[6] http://nlp.ffzg.hr/resources/lexicons/slmwelex/

[7] http://nlp.ffzg.hr/resources/lexicons/srmwelex/

in Croatian, but by observing its strongest MWE candidates, which are for the relation "sbz0 SBZ2" *nastava*, *profesor*, *nastavnik* and for the relation "pbz0 SBZ0" *fizički*, *predškolski*, *građanski*, for a human it becomes obvious that the two Croatian counterparts are *odgoj* and *obrazovanje*, which have very similar entries under the same grammatical relations, such as *uvođenje*, *nastava* and *nastavnik* for the "sbz0 SBZ2" relation and *predškolski*, *zdravstven* and *građanski* for the "pbz0 SBZ0" relation. If a model was constructed by using *(grammatical relation, dependent)* pairs as features and log-Dice as their weights, the models of those two lexemes on the Croatian side would have an overwhelming similarity with the Serbian lexeme in comparison to other lexeme combinations with that Serbian lexeme.

## 6. Conclusion

In this paper we presented the process of building a recall-oriented MWE lexicon of Croatian with the newly developed DepMWELex tool which uses syntactic patterns for MWE candidate extraction. Although MWEs are an important part of a lexicon of a certain language, and often key for proficient knowledge and use of a language, they are still not sufficiently represented in dictionaries, lexicons and other resources. This is especially the case with Croatian and other under-resourced languages. Thus the intention of building this MWE lexicon was to build a MWE resource that has a wide range of use, including HLT applications, professionals and the general public. Such an extensive resource offers a vast array of possibilities of researching the Croatian language and its MWEs. Learners of Croatian, as well as professional translators translating into Croatian as their non-mother tongue lack such a resource.

Since the recall-high approach was taken in producing the resource, the overall precision of the candidates lies slightly above 50%. Nevertheless, there are big differences in accuracies of specific grammatical relation, so a lexicon with precision of $\sim$ 80% can be produced easily by just filtering out the noisy grammatical relations.

The possibility of calculating semantic transparency of MWE candidates with the distributional approach is inspected as well with very promising results on the lexeme level. Using the produced output for modeling the context of a lexeme and using it for cross-language linking is shown off as well.

This work presents just the first step towards a rich MWE resource of not just Croatian, but its neighboring languages as well. Future work on the resource will start with increasing the size of the underlying corpora for the lexicons of Slovene and Serbian and publishing a three-language resource. For that resource to be of maximum value, the possibilities of cross-language linking on both the headword and MWE candidate levels with the distributional approach will be looked into. Finally, focused research on identifying non-transparent MWEs will be undertaken as well.

## 7. References

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Eduard Bejček, Pavel Stranak, and Pavel Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 106–115, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum.

Jeremy Clear, 1993. *Text and Technology: In honour of John Sinclair*, chapter From Firth Principles - Computational Tools for the Study of Collocation. John Benjamins Publishing Company.

Davor Delač, Zoran Krleža, Jan Šnajder, Bojana Dalbelo Bašić, and Frane Šarić. 2009. Termex: A tool for collocation extraction. In Alexander F. Gelbukh, editor, *CICLing*, volume 5449 of *Lecture Notes in Computer Science*, pages 149–157. Springer.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Information Technology*, 105:116.

Iztok Kosem, Simon Krek, and Polona Gantar. 2013. Automatic extraction of data: Slovenian case revisited. In *SKEW-4: 4th International Sketch Engine Workshop*, Talinn, Estonia.

Nikola Ljubešić and Darja Fišer. 2011. Bootstrapping bilingual lexicons from comparable corpora for closely related languages. In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 91–98. Springer.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Scott Martens and Vincent Vandeghinste. 2010. An efficient, generic approach to extracting multi-word expressions from dependency trees. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 85–88, Beijing, China, August. Coling 2010 Organizing Committee.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main Conference*

*Poster Sessions*, COLING-ACL '06, pages 651–658. Association for Computational Linguistics.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, and Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the Terminology and Knowledge Engineering (TKE2012) Conference*, Madrid, Spain.

Pavel Rychlỳ. 2008. A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for nlp. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pages 1–15.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2003. Extraction of multi-word collocations using syntactic bigram composition. In *In Proceedings of the International Conference RANLP'03*, pages 424–431.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Marko Tadić and Krešimir Šojat. 2003. Finding multiword term candidates in croatian. In *Proceedings of Information Extraction for Slavic Languages 2003 Workshop*, pages 102–107.