*MWELex – MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora

Nikola Ljubešić
University of Zagreb, Faculty of Humanities and Social Sciences, Ivana Lučića 3
E-mail: nikola.ljubesic@ffzg.hr, http://nlp.ffzg.hr/

Kaja Dobrovoljc
Trojina, Institute for Applied Slovene Studies, Dunajska 116, SI-1000 Ljubljana
E-mail: kaja.dobrovoljc@trojina.si

Darja Fišer
Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana
E-mail: darja.fiser@ff.uni-lj.si

Keywords: Slovenian, English, Croatian, multilingual lexical repository

Received: May 1, 2015

*The paper presents *MWELex, a multilingual lexical of Croatian, Slovene and Serbian multi-word expressions that were extracted from parsed corpora. The lexica were built with the custom-built 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 23 to 48 thousand headwords and 3.2 to 12 million MWE candidates per language. Similarly, precision over specific syntactic patterns varies greatly, 0.167-0.859 for Croatian, 0.158-1.00 for Slovene. The possible extension of the tool is demonstrated on a simplistic distributional-based extraction of non-transparent MWEs and cross-lingual linking of the extracted lexicons.*

*Povzetek: V prispevku predstavimo večjezični leksikon *MWELex, ki vsebuje hrvaške, slovenske in srbske večbesedne zveze, ki smo jih izluščili iz skladenjsko označenih korpusov. Leksikon smo zgradili s pomočjo lastnega orodja DepMWEx, ki za prepoznavanje kandidatov večbesednih zvez v odvisnostnih drevesih uporablja odvisnostne skladenjske vzorce, jih rangira in organizira glede na jedrno besedo. Leksikon vsebuje med 23 in 48 jedrnih besed in med 3,2 in 12 milijonov večbesednih zvez. Možnosti razširitve orodja pokažemo s pomočjo preprostega, na načelih distribucijske semantike temelječega luščenja večjezičnih netransparentnih večbesednih zvez iz izluščenega večjezičnega leksikona.*

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 have been no efforts in producing large-scale MWE resources for Croatian, Serbian or Slovene. The first experiments in Croatian 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 term extraction as the first step in producing multilingual terminological resources. All these 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.

A detailed account of the lexicographic treatment of corpus-based phraseology is given by Gantar [Gantar and Peterlin, 2006]. A comprehensive linguistic analysis of the potential and limitations of pattern-based extraction of MWE from a reference corpus was performed by Arhar [Arhar Holdt, 2011]. Semi-automatic procedures to extract MWEs for the Slovene Lexical Database have been proposed by Kosem et al. [Kosem et al., 2013a] while Krek and Dobrovoljc [Krek and Dobrovoljc, 2014] have conducted a pilot study in which they compare the performance of word-sketch-based vs. parser-based collocation extraction.

In this paper we describe a custom-based 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 parsed corpora of Croatian, Slovene and Serbian. As no such lexicon currently exists for the three languages included in the experiment presented in this paper, 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.

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 pattern trees.

Patterns trees are hierarchical structures in which each node contains a boolean function. This function defines the criterion that a node in the parse tree of a sentence must satisfy in order 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 the main verb which has a direct object and a prepositional phrase attached to it. The framed nodes represent headwords, e.g. *tražiti rupu u zakonu*, to which the MWEs will be added, namely *tražiti#Vm*, *rupa#Nc* and *zakon#Nc*.

The expressiveness of the formalism is substantial, allowing for boolean functions in specific nodes to include restrictions not only on the value of a specific node, but the remaining nodes in the pattern tree as well. One example of using this level of expressiveness is the restriction of the agreement in gender, number and case between nouns and their modifiers, which is a common linguistic phenomenon.

Another example where this level of expressiveness is exploited is the phenomenon in all three languages used in this experiment where nouns with numeral modifiers take the genitive case and not the semantically intended accusative case (semantically encoding the patient, beneficiary etc.) such as in the Croatian example *Podučavam studente* (accusative case, "I teach students") and *Podučavam pet studenata* (genitive case, "I teach five students").

2.2 Grammatical relation naming

The name of the grammatical relation of our MWE example is "gbz sbz4 u sbz6", which is a notation adopted from the Slovene Sketch grammar [Kosem et al., 2013b]. That grammar is defined over morphosyntactic patterns, and, for reasons of compatibility, all three grammars used in this experiment are based on that notation. The acronym denotes

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

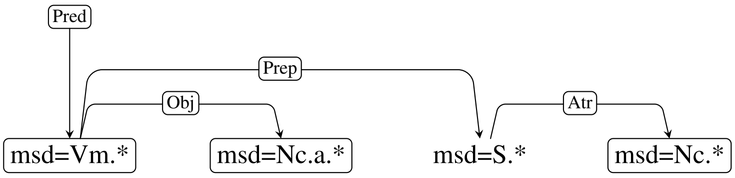

Figure 1: An example of the pattern tree corresponding to the Croatian MWE *tražiti rupu u zakonu*, *raditi račun bez konobara* (literally "to write the check without the waiter"), *raditi od buhe slona* (literally "make an elephant out of a fly", "overexaggerate") etc.

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 Croatian MWE candidate *tražiti rupu u zakonu* will appear under the grammatical relation "GBZ sbz4 u sbz6".

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 in a given language, 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 Croatian headword *tražiti#Vm* is given in Table 1.

3 Resource description

3.1 The corpora

The Croatian and Serbian lexicons were extracted from the web corpora of the corresponding languages, namely the 1.9 billion token Croatian Web corpus hrWaC and the parsed half of the 894 million token Serbian Web corpus srWaC [Ljubešić and Klubička, 2014]. These corpora were annotated with morphosyntactic, lemmatization and dependency parsing models built on the SETimes.HR corpus [Agić and Ljubešić, 2014] of 4.000 sentences.

On the other hand, the 100 million token balanced corpus of Slovene KRES [Erjavec and Logar, 2012] was used for building the Slovene lexicon. Our assumption is that this corpus is better suited for the task of extracting lexical information than the web corpora used for Croatian and Serbian for which there are no other freely available corpora. The KRES corpus was annotated with models trained on the SSJ500k corpus[2] consisting of 11.000 sentences.

3.2 The grammars

The grammars of the three languages used in the DepMWEx tool were based on the Slovene sketch grammar used in the SSJ project.[3] Once the morphosyntax-level grammar was transformed to the corresponding dependency syntax level for Slovene, the grammar was adapted for Croatian and Serbian. At this point the Slovene grammar consists of 75 grammatical relations defined through the same number of pattern trees while the Croatian and Serbian grammars consist of 63 grammatical relations with Slovene-specific relations removed.

3.3 The resulting lexicons

The size of the resulting lexicons is given in Table 2. The size of the Croatian lexicon in the number of headwords is very similar to the size of the Slovene lexicon, although the Croatian corpus from which the lexicon is extracted is almost 20 times the size. The reason for this lies in the fact that in the extraction of the Croatian and Serbian lexicons stricter frequency thresholds were applied due to the expected higher level of noise in web corpora in comparison to the manually built and balanced Slovene corpus. The (subtree, grammatical relation) pair frequency threshold applied on Croatian and Serbian data was 5 while for Slovene the threshold was 2.

There was a second threshold, identical for all three languages, applied on the lexicons, namely that each headword had to contain at least 5 MWE candidates (i.e. above mentioned pairs) satisfying the first frequency threshold to be included in the lexicon.

Finally, the Croatian list of headwords and dependents was filtered through two available morphological lexicons

[2] http://eng.slovenscina.eu/tehnologije/ucni-korpus

[3] http://eng.slovenscina.eu

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: An excerpt of the output of the DepMWEx tool for the Croatian headword *tražiti#Vm*

	lexemes	MWE candidates
hrMWELex	46,293	12,750,029
slMWELex	47,579	6,383,963
srMWELex	23,594	3,279,864

Table 2: The size of the automatically generated lexicons

of Croatian, the Croatian Morphological Lexicon[4] and the Apertium lexicon for Croatian[5]. There was no such lexicon available for Serbian. There was no need for such a filtering process for Slovene since the lemmatization of the corpus is relying on a large morphological lexicon and thereby of very high quality.

The resources, being currently in version 0.5, are encoded in XML and published[6][7][8] under the CC-BY-SA 3.0 license.

4 Resource evaluation

We performed an evaluation of the Croatian and Slovene lexicon by inspecting up to 20 top-ranked MWE candidates for each grammatical relation of 12 selected lexemes for each language. The analyzed Croatian and Slovene lexemes were sampled as follows: 3 lexemes were taken for each part of speech, one in the upper, one in the medium and one in the lower frequency range. One human annotator per language decided whether a MWE candidate was a genuine MWE or not.

Score 1 was assigned to each candidate that represented the appropriate syntactic relationship between the headword and its collocate, regardless of its semantic (un)transparency or syntactic (in)completeness. In other words, if the two-word collocation candidate in question was a syntactically valid lexical realisation of the given grammatical pattern, it was assigned score 1, despite the fact that it was a completely transparent collocation (e.g. *green leaf*) or an idiom (e.g. *green card*). Similarly, the candidate was assigned score 1 also if it formed a semantically complete unit by itself or was only part of a larger multi-word unit (e.g. *zaspati z vestjo*, "to_fall_asleep with conscience", as part of *zaspati z isto/slabo/mirno vestjo*, "to_fall_asleep with clear/guilty conscience"). Although semantically transparent or structurally incomplete two-word units might be of a lesser interest to the community, their recall is more a matter of adjusting the statistical score and/or extending the grammatical patterns to combinations of three or more words rather than a feature of the tool itself.

Score 2, on the other hand, was assigned to each candidate that did not form a valid two-word collocation for the given grammatical pattern due to incorrect pre-processing. This either means that it was assigned an incorrect MSD tag or lemma, which is frequently the case in ambiguous word forms (e.g. noun instead of verb for *stoja* - "stand/stand" or *leglo* -"lie/litter", or adverb instead of neuter adjectives *sanitarno* – "sanitary(ly)", *preventivno* – "preventive(ly)")

[4] http://hml.ffzg.hr

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

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

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

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

Croatian				Slovene		
lexeme	# evaluated	precision	diff	lexeme	# evaluated	precision
burza#Nc	559	0.735		ureditev#Nc	563	0.863
lampa#Nc	154	0.422		krč#Nc	200	0.905
lavež#Nc	34	0.324		varovalo#Nc	49	0.755
N	747	0.652	-0.215	N	812	0.867
gurati#Vm	311	0.296		razmišljati#Vm	293	0.816
razumjeti_se#Vm	161	0.484		zaspati#Vm	197	0.843
tužiti_se#Vm	77	0.26		žagati#Vm	23	0.696
V	549	0.346	-0.475	V	513	0.821
dužan#Ag	279	0.29		odgovoren#Ag	171	0.871
legendaran#Ag	64	0.609		zdravstven#Ag	62	0.645
svrhovit#Ag	20	0.4		medgeneracijski#Ag	21	1.000
A	363	0.353	-0.474	A	254	0.827
naprosto#Rg	85	0.859		nenehno#Rg	101	0.871
trostruko#Rg	78	0.615		dosledno#Rg	69	0.986
jednoglasno#Rg	62	0.806		šepetaje#Rg	23	1.000
R	225	0.76	-0.167	R	193	0.927
all	1884	0.518	-0.336	all	1772	0.854

Table 3: MWE candidate precision and difference between languages on each of the 12 evaluated lexemes

or an incorrect dependency relation or label (e.g. relating an adverbs as an attribute of an adjective instead of as an adverbial of a noun).

The precision obtained on each of the 12 lexemes, along with summaries for each part of speech and all lexemes for both evaluated languages, is given in Table 3. We can observe that the overall precision of the MWE candidates is just above 50% for Croatian but is as high as 85.4% for Slovene. The big difference in precision can be explained in most part by two factors:

1. Slovene has a more mature text pre-processing chain which was trained on more than double the amount of training data

2. the Slovene corpus is manually built (and balanced), while the Croatian corpus (similarly to the Serbian one) is automatically built from the web.

Regardless of the absolute difference in precision, same precision trends can be observed in both languages between different parts-of-speech. Adverbs are the most precise PoS, followed by nouns. Verbs and adjectives have an almost identical and the lowest precision in both languages. As one would expect, the drop in accuracy correlates with the task complexity on a specific part-of-speech (measured through precision, i.e. false positive error), showing a larger precision drop between languages on nouns (21.5%) than on adverbs (16.7%), while on verbs and adjectives the drop is the highest and almost identical (47.4% and 47.5%).

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 4. The worst performing set of grammatical relations in Croatian 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 a relation (SBZ1 gbz). A similar phenomenon can be observed with the grammatical relation consisting of a main verb and its direct object which performs 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. In Slovene, on the other hand, the worst performing grammatical relation is the gbz SBZ2, which matches verb+noun_genitive combinations (e.g. *veseliti se poletja* – "look forward to summer") with as little as 0.158 accuracy. There are several top-performing grammatical relations with all candidates extracted correctly in the Slovene evaluation sample, including the most frequent pbz0 SBZ0 pattern that matches adjective+noun_nominative (e.g. *zdravstveno zavarovanje* – "health insurance").

Croatian			Slovene		
relation	frequency	precision	relation	frequency	precision
pbz0 SBZ0	94	0.809	pbz0 SBZ0	109	1.000
RBZ gbz	73	0.822	rbz GBZ	107	0.953
RBZ pbz0	65	0.923	SBZ1 gbz	86	0.791
rbz GBZ	60	0.5	sbz0 SBZ2	85	0.906
sbz1 GBZ	60	0.167	rbz Inf-GBZ	78	0.974
RBZ RBZ	52	0.558	gbz SBZ4	76	0.750
SBZ1 gbz	50	0.64	rbz PBZ0	69	0.696
GBZ u sbz5	49	0.204	GBZ v sbz5	66	0.879
GBZ0 in/ali GBZ0	47	0.213	GBZ z sbz6	53	0.962
PBZ0 in/ali PBZ0	47	0.277	zveze s predlogi	42	1.000
GBZ na sbz4	46	0.283	sbz1 Vez-gbz PBZ1	42	0.976
SBZ0 in/ali SBZ0	45	0.0	PBZ0 in/ali PBZ0	41	1.000
gbz SBZ4	42	0.714	SBZ0 in/ali SBZ0	41	0.707
GBZ sbz4	42	0.214	SBZ0 v sbz5	40	0.975
rbz PBZ0	42	0.357	gbz PBZ1	38	0.447
sbz0 SBZ2	42	0.667	gbz SBZ2	38	0.158
GBZ u sbz4	41	0.829	SBZ0 za sbz4	37	0.784
SBZ0 sbz2	32	0.656	GBZ na sbz5	36	0.972
RBZ Vez-gbz pbz1	27	0.704	GBZ o sbz5	34	0.971
gbz Inf-GBZ	25	0.64	gbz za SBZ4	34	0.941

Table 4: Precision scores for 20 most frequent grammatical relations in each evaluated language

5 Lexicon refinement

At this point we produced a recall-high resource with satisfactory precision, just over 50% for Croatian and 85% for Slovene, 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. In this section we report on the initial experiments on Croatian 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 head-word 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:

- 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)
- for the head *selo* ("village"), the most distant MWE candidate is *špansko selo* ("Spanish village", refers to something absolutely unknown to someone, like *it's all Greek to me*)
- for the head *stan* ("flat") the most distant MWE is *tkalački stan* ("sewing machine")
- for the head *ured* ("office"), the most distant MWE is *ovalni ured* (the Oval office)
- for the head *zlato* ("gold"), among the most distant MWEs is *crno zlato* ("black gold", 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

Since the grammatical relations have the same names in grammars of all the languages used in the experiment, we can use (grammatical relation, dependents) pairs as features for our context vectors, thereby obtaining a more detailed and selective formalization of the context of a lexeme than in the standard distributional approach as implemented in the previous subsection. This leads to 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 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, Serbian and Slovene 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 the languages used in this experiment as well as many 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 Croatian, Serbian and Slovene and its MWEs. Foreign language learners, as well as professional translators translating into Croatian, Serbian or Slovene as their non-mother tongue, are still lacking such a resource.

Since the recall-high approach was taken in producing the resource, the overall precision of the candidates lies slightly above 50% for Croatian, whereas it is 85% for Slovene. Nevertheless, there are big differences in accuracies of specific grammatical relation, so a lexicon with precision of $\sim 80\%$ for Croatian and $\sim 95\%$ for Slovene 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 was 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 was shown as well.

This work presents only 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 by increasing the size of the underlying corpora for the lexicons of Slovene and Serbian and publishing a three-lingual 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.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran) and the Slovenian-Croatian bilateral project "Bilingual Lexicon Construction for Closely Related Languages from Existing Language Resources" (BI-HR/14-15-047).

References

[Agić and Ljubešić, 2014] Agić, Ž. and Ljubešić, N. (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).

[Arhar Holdt, 2011] Arhar Holdt, Š. (2011). *Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev*. Trojina, zavod za uporabno slovenistiko.

[Baldwin and Kim, 2010] Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In Indurkhya, N. and

Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

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

[Bejček et al., 2013] Bejček, E., Stranak, P., and Pecina, P. (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. Association for Computational Linguistics.

[Church et al., 1991] Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum.

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

[Delač et al., 2009] Delač, D., Krleža, Z., Šnajder, J., Bašić, B. D., and Šarić, F. (2009). Termex: A tool for collocation extraction. In Gelbukh, A. F., editor, *CICLing*, volume 5449 of *Lecture Notes in Computer Science*, pages 149–157. Springer.

[Erjavec and Logar, 2012] Erjavec, T. and Logar, N. (2012). Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. In *Zbornik Osme konference Jezikovne tehnologije*.

[Gantar and Peterlin, 2006] Gantar, P. and Peterlin, A. P. (2006). Korpusni pristop v frazeologiji in slovarske aplikacije. *Slavistična revija*.

[Kilgarriff et al., 2004] Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105:116.

[Kosem et al., 2013a] Kosem, I., Gantar, P., and Krek, S. (2013a). Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0*.

[Kosem et al., 2013b] Kosem, I., Krek, S., and Gantar, P. (2013b). Automatic extraction of data: Slovenian case revisited. In *SKEW-4: 4th International Sketch Engine Workshop*, Talinn, Estonia.

[Krek and Dobrovoljc, 2014] Krek, S. and Dobrovoljc, K. (2014). Sketch grammar or parser – a comparison of two extraction methods. Poster.

[Ljubešić and Fišer, 2011] Ljubešić, N. and Fišer, D. (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.

[Ljubešić and Klubička, 2014] Ljubešić, N. and Klubička, F. (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.

[Martens and Vandeghinste, 2010] Martens, S. and Vandeghinste, V. (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. Coling 2010 Organizing Committee.

[Pecina and Schlesinger, 2006] Pecina, P. and Schlesinger, P. (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.

[Pinnis et al., 2012] Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., and Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the Terminology and Knowledge Engineering (TKE2012) Conference*, Madrid, Spain.

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

[Sag et al., 2001] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (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.

[Seretan et al., 2003] Seretan, V., Nerima, L., and Wehrli, E. (2003). Extraction of multi-word collocations using syntactic bigram composition. In *In Proceedings of the International Conference RANLP'03*, pages 424–431.

[Šnajder et al., 2013] Šnajder, J., Padó, S., and Agić, Ž. (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.

[Tadić and Šojat, 2003] Tadić, M. and Šojat, K. (2003). Finding multiword term candidates in croatian. In *Proceedings of Information Extraction for Slavic Languages 2003 Workshop*, pages 102–107.