A Preliminary Study of Croatian Lexical Substitution

Domagoj Alagić and **Jan Šnajder**
Text Analysis and Knowledge Engineering Lab
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
{domagoj.alagic,jan.snajder}@fer.hr

Abstract

Lexical substitution is a task of determining a meaning-preserving replacement for a word in context. We report on a preliminary study of this task for the Croatian language on a small-scale lexical sample dataset, manually annotated using three different annotation schemes. We compare the annotations, analyze the inter-annotator agreement, and observe a number of interesting language-specific details in the obtained lexical substitutes. Furthermore, we apply a recently-proposed, dependency-based lexical substitution model to our dataset. The model achieves a P@3 score of 0.35, which indicates the difficulty of the task.

1 Introduction

Modeling word meaning is one of the most rewarding challenges of many natural language processing (NLP) applications, including information retrieval (Stokoe et al., 2003), information extraction (Ciaramita and Altun, 2006), and machine translation (Carpuat and Wu, 2007), to name a few. Perhaps the most straightforward task concerned with word senses is word sense disambiguation (WSD), a task of determining the correct sense of a polysemous word in its context (Navigli, 2009). Despite being a straightforward task, WSD has several drawbacks. Most often, it is criticized for relying on a fixed set of senses for each of the words (sense inventory), which – although meticulously compiled by experts – is often of inappropriate coverage or granularity (Edmonds and Kilgarriff, 2002; Snyder and Palmer, 2004). This requirement makes evaluation of WSD models across different applications rather difficult.

An alternative perspective on modeling word senses is the one of *lexical substitution* (McCarthy and Navigli, 2007), a task of finding a meaning-preserving replacement of a polysemous target word in context. For instance, in the sentence "*It took me around two hours to reach Nagoya from Kyoto by coach*", suitable substitutes for the word *coach* may be *van* or *bus*, whereas the substitute *trainer* represents a different sense of the word. Note that such a setup circumvents the need of having a fixed sense inventory, as annotators do not require any kind of resources to come up with a plausible set of substitutes for a word. This seems both more intuitive and far less restrictive than the traditional WSD task. However, the lexical substitution task is still determined by a number of parameters that need to be taken into consideration, as they affect the obtained substitutes in various ways (e.g., variety, count, etc.).

In this paper, we report on a preliminary study of the lexical substitution task for the Croatian language, a first such study so far. We compile a small-scale lexical sample dataset and annotate it using three annotation schemes to gain insights into how they affect the annotations. We analyze the obtained substitutes and report on interesting language-specific details, hoping to facilitate research on this topic for other Slavic languages. Finally, we re-implement one of the best-performing models for English lexical substitution (Melamud et al., 2015b) and evaluate it on our dataset.

2 Related Work

Most work on lexical substitution was done for English (McCarthy and Navigli, 2007; Sinha and Mihalcea, 2014; Biemann, 2012; Kremer et al., 2014). A few notable exceptions include German within the GERMEVAL-2015 (Miller et al., 2015), Italian within the EVALITA-2009 (Toral, 2009), and Spanish within a cross-lingual setup at SEMEVAL-2012 (Mihalcea et al., 2010). Recently, most research on lexical substitution closely relates

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 14–19,
Valencia, Spain, 4 April 2017.

to the task of learning meaning representations that are able to account for multiple senses of polysemous words (Melamud et al., 2015a; Melamud et al., 2016; Roller and Erk, 2016; Erk et al., 2013).

For the experiments, we adopt the work of Melamud et al. (2015b), who proposed a lexical substitution model based on dependency-based embeddings. Their model is easy to implement, yet it performs nearly at the state-of-the-art level.

3 Dataset Construction

3.1 Data

We took a *lexical sample* approach, in which the experiments are carried out on a predefined set of words. As this is a preliminary study, we decided on using six words: two adjectives, two nouns, and two verbs. We selected these words by taking all the words that have at least three senses and that occur at least 10,000 times in hrWaC, a Croatian web corpus (Ljubešić and Erjavec, 2011). After selecting the words, we extracted 30 contexts (instances) per word from the Cro36WSD dataset (Alagić and Šnajder, 2016), a lexical sample for Croatian WSD. The words we use are: *prljav*$_A$ (dirty), *visok*$_A$ (high/tall), *težina*$_N$ (weight/difficulty), *okvir*$_N$ (frame), *oprati*$_V$ (to wash off), and *tući*$_V$ (to hit/to beat).

3.2 Annotation

Annotation schemes. One insight we wished to gain from this study is how different annotation schemes influence the lexical substitutes obtained through the annotation. We consider three different annotation schemes:

1. SINGLE – In this scheme, annotators are allowed to provide only *single-word expressions* (SWEs) as substitutes. They are also allowed to provide hypernyms if they cannot think of any other suitable substitutes;
2. MULTI – Besides SWEs, annotators can provide *multiword expressions* (MWEs) as well;
3. MULTI3 – Annotators can provide everything as in MULTI setup, but should give their best to come up with *at least three* substitutes.

The motivation for having a separate annotation scheme for single-word substitutes (SINGLE) is based upon an intuition that annotators often do not provide just every substitute they think of, but rather only a couple of those that first come to their mind. Thus, by allowing the annotators to use MWEs, they could sometimes reach for a more common MWE instead of thinking a bit harder about single-word substitutes. As an example, consider the word *preozbiljan* (too serious) in the following sentence:

(1) *On je uvijek **preozbiljan** na zabavama.*
*He is always **too serious** at parties.*

In this case, the annotators might more commonly use the idiomatic phrase *smrtno ozbiljan* (dead serious) than the single-word expression *mrk* (stern).

On the other hand, we use MULTI3 annotation scheme to investigate what substitutes the annotators provide to meet the required number of substitutes. We expect those to be less common near-synonyms or words related to the target word.

Annotation guidelines. Each annotator was presented with a sentence containing a polysemous target word and was asked to provide as many meaning-preserving substitutes as they could think of (in any order). The annotators were also instructed to give the substitutes in a lemmatized form (e.g., *kući* ⇒ *kuća*; dative case of *house*). In case of an MWE, they were asked to lemmatize the complete MWE as a single unit instead of doing it on a per-word basis (e.g., *Hrvatskoga narodnog kazališta* ⇒ *Hrvatsko narodno kazalište*, instead of *Hrvatski narodni kazalište*; genitive case of *Croatian National Theatre*). The annotators were also told not to consult any language resources during the annotation.

Annotation effort. We asked 12 native Croatian speakers to annotate our data. We split their annotation effort so that each annotator annotates all six words, but using different schemes along the way (two words for each scheme). This resulted in each instance being annotated by four annotators per annotation scheme, and each annotator completing the annotation of 180 instances in total. Each annotator spent around three person-hours on average. Lastly, to account for having only four annotators per instance, we (the authors) manually went through the annotations and corrected typos and wrong lemma forms, a step that took five person-hours.[1] We make our dataset freely-available.[2]

[1] We believe that having more annotators per instance could lessen the need of having to correct noisy annotations, as not all annotators would make slips on the same instances.

[2] http://takelab.fer.hr/data/crolexsub

Scheme	Min.	Max.	Avg.	# SWE	# MWE	# PC
SINGLE	0	10	3.92	702	4	27
MULTI	0	13	4.20	687	69	14
MULTI3	0	12	5.93	1003	64	27

Table 1: Dataset statistics. PCs have been counted only within single-word substitutes.

	PA				PAM			
Scheme	N	A	V	All	N	A	V	All
SINGLE	0.32	0.12	0.26	**0.23**	0.44	0.27	0.31	**0.35**
MULTI	0.26	0.17	0.24	0.22	0.39	0.32	0.18	0.29
MULTI3	0.20	0.09	0.29	0.20	0.18	0.16	0.16	0.17

Table 2: Inter-annotator agreement across schemes and POS tags.

4 Annotation Analysis

4.1 Dataset Statistics

After correction, we measure the minimum, maximum, and average number of substitutes across annotation schemes, number of single-word (SWE) and multiword (MWE) substitutes, and number of substitutes where a POS change (PC) occurred, i.e., where substitute's and target word's POS tags are different. We report the numbers in Table 1.

4.2 Inter-Annotator Agreement

We measure the inter-annotator agreement (IAA) using the *pairwise agreement* (PA) and *pairwise agreement with modes* (PAM), following McCarthy and Navigli (2007). PA essentially measures the average overlap of substitutes between all possible annotator pairings across instances. On the other hand, PAM measures the agreement by counting the times a gold substitute mode[3] was included in the annotator substitute set. We report the IAA scores in Table 2. Even though the absolute agreement scores are generally low, we note that they are in line with those of Kremer et al. (2014). From a POS perspective, annotators agreed the most on nouns and disagreed the most on adjectives. Moreover, we note that the MULTI3 scheme has the lowest IAA, possibly because the "coerced" substitutes (especially the multiword ones) have a greater variability. We leave a more detailed analysis of the IAA for future work.

[3] A *mode* is a single substitute that received the most annotator votes, if such exists.

4.3 Observations

We present some preliminary insights into the obtained substitutes, which we think warrant further investigation. Some of the insights are language-specific, while others might be relevant for other languages as well.

Lemmatization. Even though we asked the annotators to provide substitutes in a lemmatized form, it is not obvious whether this is the best approach. Obviously, not lemmatizing the substitutes will inflate the number of proposed substitutes with inflected variants of the same word (across contexts in which the word occurs). On the other hand, lemmatizing each and every substitute may lead to information loss (for example, when lemmatizing adjectives from a superlative into a positive form).

Reflexive pronouns. It is unclear whether the verbs with obligatory reflexive pronouns, e.g., *smijati se* (to laugh) should be treated as MWEs. Currently, we prefer to treat them as SWEs.

Coreference. If a sentence contains the same target more than once, it is often possible to replace one of them with a coreferring pronoun. For example, in the sentence:[4]

(2) *Kako vam se težina nakon dijete ne bi ubrzo vratila na **težinu** prije dijete...*
*To prevent your weight after a diet from quickly reverting to **weight** before a diet...*

one could provide the pronoun substitute *onu* (one), which would perfectly preserve the sentence meaning (and in fact improve coherence of the text).

Ungrammaticality. Some substitutes may effectively break the sentence grammaticality due to the fact that they replace a multiword expression of which the target word is a part of, rather than merely the target word. As an example, consider:

(3) *... koja su započela 22. prosinca u **okviru** operativne akcije...*
*... which started on December 22 in the **scope** of an operative action...*

In this sentence, one may substitute *okviru* (frame/scope) with a preposition *unutar* (within), thus requiring to omit the preposition *u* (in) to preserve overall sentence grammaticality.

[4] The translation is slightly ungrammatical to better illustrate the issue.

5 Experiments

5.1 Models

For our experiments, we re-implemented a simple, yet powerful model of Melamud et al. (2015b), one of the best-performing models for lexical substitution. This model posits that a good lexical substitute needs to be both semantically similar to the target word (i.e., paradigmatic similarity) and suitable for a given context (i.e., syntagmatic similarity). To that end, Melamud et al. (2015b) propose four substitutability measures that combine these two concepts in different ways (Table 3). Whereas *Add* measure employs an arithmetic mean, *Mult* measure uses a stricter, geometric mean. Furthermore, they introduce *Bal* variants that balance out the effect of context size. In addition to these models, we use an *out-of-context* (OOC) model as a baseline, which calculates the substitute score simply as a cosine between the substitute's and target word's embedding (also shown in Table 3).

Substitutability measures are calculated using dependency-based word and context embeddings (Levy and Goldberg, 2014), which the authors derived from the original skip-gram negative sampling algorithm (SGNS) (Mikolov et al., 2013). In a nutshell, instead of using models that are based solely on lexical contexts, their model can be trained on arbitrary contexts (in their case, the syntactic contexts derived from dependency parse trees). The rationale behind using dependency-based embeddings is that using only regular SGNS embeddings does not account for substitute's paradigmatic fit in its context.

We train these word-type (lemma and POS-tag) embeddings on hrWaC, a Croatian web corpus (Ljubešić and Erjavec, 2011), using the freely available `word2vecf` tool.[5] We use default parameters: frequency threshold of 5 and negative sampling factor of 15. We did not collapse the relations including prepositions. Before training the embeddings, we discarded all lemmas that appeared fewer than 100 times in the corpus.

5.2 Evaluation

We focus on the SINGLE annotation scheme within our evaluation, as the model we use does not deal with MWEs. To compile the candidate sets for each of the instances, we follow prior work and pool candidates from all substitutes given by the annotators for a specific target word (i.e., across all target word's instances). This enables us to basically evaluate the model's ability of identifying the viable substitutes and ranking low the ones that bear a sense different of that evoked in a context. Following (Thater et al., 2010), we evaluate the models in terms of generalized average precision (GAP) (Kishida, 2005). GAP is a weighted extension of the mean average precision (MAP) measure, where weights capture how many times the annotators used a certain substitute in a goldset. In line with work of Roller and Erk (2016), we decided not to use the original lexical substitution metrics (*oot* and *best*), but standard P@3 and P@5 scores, which we find more interpretable. We report the results in Table 4.

We observe that the model based on *Add* substitutability measure consistently performs best. Usually, out of the top three substitutes predicted by the model, one of them is correct (P@3 = 0.35). Surprisingly, in terms of both GAP and P@5, the baseline *OOC* model performs comparably well.

To illustrate how the implemented model works, we show the top 10 substitute candidates predicted by *Add* model for one of the occurrences of word *prljav* (dirty) in Table 5. The top candidates perfectly capture the *filthy* sense of this word, whereas

Add	$\frac{cos(s,t) + \sum_{c \in C} cos(s,c)}{\lvert C \rvert + 1}$
BalAdd	$\frac{\lvert C \rvert \cdot cos(s,t) + \sum_{c \in C} cos(s,c)}{2 \cdot \lvert C \rvert}$
Mult	$\sqrt[\lvert C \rvert + 1]{pcos(s,t) \cdot \prod_{c \in C} pcos(s,c)}$
BalMult	$\sqrt[2 \cdot \lvert C \rvert]{pcos(s,t)^{\lvert C \rvert} \cdot \prod_{c \in C} pcos(s,c)}$
OOC	$cos(s,t)$

Table 3: The different substitutability measures for a lexical substitute s of a target word t within a context C.[6]

	Metric		
Models	GAP	P@3	P@5
Add	**0.28**	**0.35**	**0.28**
BalAdd	0.26	0.31	0.26
Mult	0.27	0.28	0.27
BalMult	**0.28**	0.31	**0.28**
OOC	0.26	0.21	0.25

Table 4: Model scores on our dataset.

[5] `https://bitbucket.org/yoavgo/word2vecf`

[6] Positive cosine is defined as $pcos(a,b) = \frac{cos(a,b)+1}{2}$.

Sentence (HR)	Sentence (EN)
"Ne diraj me tim prljavim rukama," rekla mu je s prijezirom...	*"Do not touch me with those dirty hands of yours," she told him with contempt...*
Predicted substitutes (HR)	Predicted substitutes (EN)
nečist, neopran, zmazan, uprljan, odvratan, perverzan, mutan, gadan, podmukao, zamazan	*unclean, unwashed, filthy, dirtied, disgusting, perverse, fishy, nasty, scheming, filthy*
Gold substitutes (HR)	Gold substitutes (EN)
nečist, zmazan, zamazan, neopran	*unclean, filthy, filthy, unwashed*

Table 5: Top 10 substitute candidates for instance 6086 as predicted by *Add* model.

the most of the remaining ones depict the *sordid* sense of the word, which is questionable, albeit possible within this ambiguous context.

In general, however, we note that the figures are considerably lower than those obtained for the English lexical substitution task (Melamud et al., 2015b; Roller and Erk, 2016). We speculate that one of the reasons might be the morphological complexity of Croatian. Another, related reason might be the way how word embeddings are trained: we used word-type embeddings instead of word-form embeddings and we did not collapse the relations including prepositions. We leave an investigation of these issues for future work.

6 Conclusion

In this work we tackled the lexical substitution task for Croatian. We compiled a small-scale lexical sample dataset and annotated it using three different schemes. Moreover, we presented interesting insights about the annotations, some of which are specific to Croatian, while others possibly pertain to other (morphologically-rich) languages. Lastly, we re-implemented one of the best-performing models for English lexical substitution and evaluated it on our dataset. A thorough comparison of the annotation schemes, as well as the implementation of a more efficient model that also deals with MWEs are the subject of future work.

Acknowledgments

We are extremely grateful to our 12 annotators for making time to annotate our data. We would also like to thank the anonymous reviewers for their useful and insightful comments.

This work has been fully supported by the Croatian Science Foundation under the project UIP-2014-09-7312.

References

Domagoj Alagić and Jan Šnajder. 2016. Cro36WSD: A lexical sample for Croatian word sense disambiguation. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, pages 1689–1694, Portorož, Slovenia.

Chris Biemann. 2012. Turk bootstrap word sense inventory 2.0: A large-scale resource for lexical substitution. In *Proceedings of the 8th edition of the Language Resources and Evaluation Conference (LREC 2012)*, pages 4038–4042, Instanbul, Turkey.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of Conference on Computational Natural Language Learning (CoNLL 2007)*, volume 7, pages 61–72, Prague, Czech Republic.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 594–602, Sydney, Australia.

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8(04):279–291.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Kazuaki Kishida. 2005. *Property of Average Precision and Its Generalization: An Examination of Evaluation Indicator for Information Retrieval Experiments*, volume 2005. National Institute of Informatics Tokyo, Japan.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us – analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 540–549, Gothenburg, Sweden.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the*

52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), pages 302–308, Baltimore, Maryland, USA.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling web corpora for Croatian and Slovene. In *Proceedings of 14th International Conference on Text, Speech and Dialogue (TSD 2011)*, pages 395–402, Pilsen, Czech Republic.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 48–53, Prague, Czech Republic.

Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015a. Modeling word meaning in context with substitute vectors. In *The 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2015)*, pages 472–482, Denver, Colorado.

Oren Melamud, Omer Levy, and Ido Dagan. 2015b. A simple word embedding model for lexical substitution. In *Proceedings of the Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing (VSM-NLP 2015)*, pages 1–7, Denver, Colorado.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of Conference on Computational Natural Language Learning (CONLL 2016)*, pages 51–61, Vancouver, Canada.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 9–14, Uppsala, Sweden.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Neural Information Processing Systems Conference (NIPS 2013)*, pages 3111–3119, Lake Tahoe, USA.

Tristan Miller, Darina Benikova, and Sallam Abualhaija. 2015. GermEval 2015: LexSub – a shared task for German-language lexical substitution. *Proceedings of GermEval 2015*, pages 1–9.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Stephen Roller and Katrin Erk. 2016. PIC a different word: A simple model for lexical substitution in context. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, pages 1121–1126, San Diego, California.

Ravi Sinha and Rada Mihalcea. 2014. Explorations in lexical sample and all-words lexical substitution. *Natural Language Engineering*, 20(01):99–129.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of Senseval-3*, pages 41–43, Barcelona, Spain.

Christopher Stokoe, Michael P Oakes, and John Tait. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of ACM SIGIR 2013*, pages 159–166, Toronto, Canada.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010*, pages 948–957, Uppsala, Sweden.

Antonio Toral. 2009. The lexical substitution task at EVALITA 2009. In *Proceedings of EVALITA Workshop, 11th Congress of Italian Association for Artificial Intelligence*, Reggio Emilia, Italy.