Does Free Word Order Hurt? Assessing the Practical Lexical Function Model for Croatian

Zoran Medić[*] **Jan Šnajder**[*] **Sebastian Padó**[†]
[*] Faculty of Electrical Engineering and Computing, University of Zagreb
{jan.snajder, zoran.medic}@fer.hr
[†] Institut für Maschinelle Sprachverarbeitung, Stuttgart University
pado@ims.uni-stuttgart.de

Abstract

The Practical Lexical Function (PLF) model is a model of computational distributional semantics that attempts to strike a balance between expressivity and learnability in predicting phrase meaning and shows competitive results. We investigate how well the PLF carries over to free word order languages, given that it builds on observations of predicate-argument combinations that are harder to recover in free word order languages. We evaluate variants of the PLF for Croatian, using a new lexical substitution dataset. We find that the PLF works about as well for Croatian as for English, but demonstrate that its strength lies in modeling verbs, and that the free word order affects the less robust PLF variant.

1 Introduction

Compositional distributional semantic models (CDSMs) represent phrase meaning in a vector space by composing the meanings of individual words. Many CDSMs were proposed, ranging from basic ones that use element-wise operations on word vectors to compute phrase vectors (Mitchell and Lapata, 2008), to more complex models that represent predicate arguments as higher-order tensors (Baroni and Zamparelli, 2010; Guevara, 2010). The latter models assume that predicates in a phrase act as functions that act on other phrase components to yield the final representation of the phrase. For example, an adjective acts as a function on the noun in an adjective-noun phrase, while a transitive verb acts as a binary function on its subject and object. However, since the number of parameters in a tensor grows exponentially with the number of arguments of the function that it models, learning full tensors for predicates with many arguments is tedious to impractical (Grefenstette et al., 2012).

The Practical Lexical Function model (PLF, Paperno et al. (2014)) strikes a middle ground by breaking down all tensors with ranks higher than two into multiple matrices, each representing the predicate's composition with a single argument (cf. Section 2 for details). In the experiments of Paperno et al. (2014), PLF has been shown to work better than some other CDSMs in modeling semantic similarity. Particularly good results were obtained on ANVAN (adjective-noun-verb-adjective-noun) phrases, where PLF outperformed both simple CDSMs (due to its higher expressiveness) as well as the higher-order Lexical Function model (Baroni and Zamparelli, 2010).

Although the PLF shows promising results, existing work still leaves open two questions. First, it is not obvious that these results carry over to languages with free word order, such as Slavic languages, where predicates and arguments are often separated. For example, in the English sentence *'I like my dog'*, the predicate is adjacent to both the subject and the object, while in the Croatian translation *'Sviđa mi se moj pas'*, the object *'moj pas* is separated from the predicate. As corpus-derived vectors for predicate-argument combinations are a key part of the PLF, non-adjacency might make it difficult to estimate its parameters reliably for such languages. Secondly, the evaluation method reported by Paperno et al. (2014) uses a somewhat artificial setup by assuming that all phrase pairs, even ill-formed ones, can be graded for similarity.

In this work we consider both of these questions. We investigate the application of PLF to Croatian language, a Slavic language with relatively free word order. We compare PLF with other, simpler CDSMs, as well as PLF modifications proposed by Gupta et al. (2015). In contrast to Paperno et al. (2014), we adopt lexical substitution as evaluation, building a new dataset of Croatian ANVAN phrases,

*Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 115–120,
Vancouver, Canada, August 3-4, 2017. ©2017 Association for Computational Linguistics

together with word substitutes for each word. The PLF model for Croatian performs comparably well to English, outperforming simpler CDSMs in particular at the verb position.

2 The Practical Lexical Function Model

Basic model. As described above, the idea of the PLF is to represent predicates as sets of matrices for each argument slot of the predicate, plus a vector for its lexical meaning. The meaning of the predicate-argument combination is computed by multiplying all argument vectors with the predicates' slot matrices and finally adding the predicate's lexical vector. For example, the vector for the phrase *'big window'* is computed as:

$$\mathcal{P}(big\ window) = \overrightarrow{big} + \overset{\square N}{big} \times \overrightarrow{window} \quad (1)$$

This can easily be generalized to more complex ANVAN phrases, as exemplified in Figure 1.

The predicate matrices are estimated using ridge regression with corpus-extracted vectors for arguments ($\overrightarrow{n}$) as input and vectors for bigram phrases ($\overrightarrow{an}$) as output. For example, the predicate matrix $\overset{\square N}{a}$ for an adjective a is computed as follows:

$$\overset{\square N}{a} \triangleq \arg\min_{M} \sum_{n \in nouns(a)} \left\| M \times \overrightarrow{n} - \overrightarrow{an} \right\|^2 \quad (2)$$

PLF modifications. Gupta et al. (2015) identify an inconsistency within the PLF: there is a difference between the meaning modeled by a matrix obtained with training and its usage in phrase vector calculation. The matrix obtained using Eq. (2) directly approximates the phrase meaning for a given predicate-argument phrase, while the PLF phrase vector in Eq. (1) adds the predicate vector on top of the product of predicate matrix and argument vector. They propose two remedies, as follows.

Train phase modification changes Eq. (2) so that the predicate matrix does not learn a direct transformation from an argument vector to a phrase vector, but rather a difference between these vectors:

$$\overset{\square}{a} \triangleq \arg\min_{M} \sum_{n \in nouns(a)} \left\| M \times \overrightarrow{n} - (\overrightarrow{an} - \overrightarrow{a}) \right\|^2 \quad (3)$$

This justifies the addition of predicate vector in (1).

In contrast, *test phase modification* retains the same training process, but omits the predicate vector when computing the phrase vector:[1]

$$\mathcal{P}(big\ window) = \overset{\square N}{big} \times \overrightarrow{window} \quad (4)$$

Gupta et al. (2015) found both modifications to outperform simple baseline CDSMs for English when evaluated on ANVAN datasets, with test adaptation outperforming the original PLF.

PLF for Croatian. We implemented the basic PLF and the two above-mentioned modifications for Croatian following the procedure described by Paperno et al. (2014). As a corpus for building word and phrase lexical vectors we used fHrWaC (Šnajder et al., 2013), a filtered version of Croatian web corpus (Ljubešić and Erjavec, 2011), totaling 51M sentences and 1.2B tokens. The corpus has been parsed using the MSTParser for Croatian (Agić and Merkler, 2013).

As a first step in obtaining word vector representations, we extracted a co-occurrence matrix of 30K most frequent lemmas (nouns, verbs, and adjectives) in corpus, using a window of size 3. Next, the vectors contained in the resulting matrix were transformed using Positive Pointwise Mutual Information (PPMI) and reduced to size 300 using Singular Value Decomposition. Finally, all vectors in the matrix were normalized to unit length.

For the extraction of phrase (bigram) vectors, we consider two different approaches. The first approach considers all occurrences where the predicate and argument are adjacent in the dependency trees in fHrWaC even if they are not adjacent on the surface, sidestepping the free word order issue. The second approach extracts only those phrases in which the predicate and argument are adjacent on the surface, resulting in a smaller but potentially cleaner set of co-occurrences. The phrase vectors from both approaches use the same 30K context lemmas and window size as the unigrams.

Using the extracted lemma and bigram vectors, we train matrices for each of the predicate words from our evaluation dataset. As our dataset consists of ANVAN phrases, we train one matrix for each adjective and two matrices for each verb (one for subject and one for object). We train two versions of each matrix: one using the originally proposed training and another with modified training.

3 Experiments

Evaluation methodology. Paperno et al. (2014) evaluated the PLF on five datasets containing phrases in different forms. Two consist of free-form sentences, one of a number of differently formed phrases, and the two ANVAN datasets contain adjective-noun-verb-adjective-noun phrase

[1] For one-argument predicates, this is equivalent to the Lexical Function model (Baroni and Zamparelli, 2010).

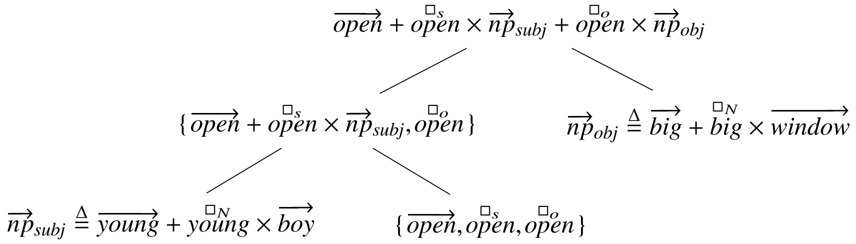
$\overrightarrow{open} + \overset{\square_s}{open} \times \overrightarrow{np}_{subj} + \overset{\square_o}{open} \times \overrightarrow{np}_{obj}$
$\{\overrightarrow{open} + \overset{\square_s}{open} \times \overrightarrow{np}_{subj}, \overset{\square_o}{open}\}$
$\overrightarrow{np}_{obj} \triangleq \overrightarrow{big} + \overset{\square_N}{big} \times \overrightarrow{window}$
$\overrightarrow{np}_{subj} \triangleq \overrightarrow{young} + \overset{\square_N}{young} \times \overrightarrow{boy}$
$\{\overrightarrow{open}, \overset{\square_s}{open}, \overset{\square_o}{open}\}$

Figure 1: Computing the vector for an ANVAN phrase (*young boy open big window*) using PLF.

ANVAN phrase (target word in bold)	Substitute words
legendaran trener voditi suparnička momčad *(**legendary** coach lead opponent team)*	cijenjen *(appreciated)*, izvanredan *(outstanding)*, poznat *(famous)*, uspješan *(successful)*, znamenit *(notable)*
dobar igrač **dati** pobjednički gol *(good player **score** winning goal)*	pogoditi *(to hit)*, postići *(to achieve)*, zabiti *(to score)*, zadati *(to give)*
sportski automobil prijeći velika **udaljenost** *(sports car travel large **distance**)*	dionica *(section)*, dužina *(length)*, put *(way)*, razdaljina *(distance)*

Table 1: Examples of ANVAN phrases with manually collected substitutes for boldfaced targets.

pairs rated for semantic similarity (Kartsaklis et al., 2013; Grefenstette, 2013). The phrases in each pair differ only in the verb. Annotators rated the similarity on a scale from 1 to 7, and CDSMs were evaluated by correlating the ratings with the similarity of the predicted phrase vectors.

The described approach is not appropriate when one or both ANVAN phrases are ungrammatical or nonsensical. Consider the following phrase pair in the ANVAN dataset by Kartsaklis et al. (2013): *'dental service file false tooth' – 'dental service register false tooth'*. While the first sentence is plausible, the second one is arguably somewhere between implausible and nonsensical. We believe that semantic similarity is not a reasonable evaluation criterion for such (relatively frequent) cases.

For our experiment, we chose a word-choice evaluation setup, which essentially builds on the idea of lexical substitution. Lexical substitution is the task of identifying a substitute for a word in a given context (McCarthy and Navigli, 2007). Typically, a system is presented with a phrase and candidate substitutes for a target word in the phrase and needs to select one or more adequate substitutes. Systems either have to rank the candidates in the appropriate order (McCarthy and Navigli, 2007; Sinha and Mihalcea, 2009), or just choose one best substitute (Melamud et al., 2016).

An additional benefit of a lexical substitution setup is that we can evaluate the predictions of the model not just globally, but at the level of individual words. We will exploit that possibility below.

Croatian ANVAN dataset. We constructed individual ANVAN phrases for Croatian like in prior English work (Kartsaklis et al., 2013; Grefenstette, 2013). We started by choosing six transitive verbs from the list of polysemous verbs on the Croatian language portal.[2] We chose verbs with high polysemy level, while avoiding those that overlap in semantic meaning. The list consist of the following verbs: *'baciti'* (to throw), *'dati'* (to give), *'izdati'* (to issue), *'prijeći'* (to cross), *'vidjeti'* (to see), and *'voditi'* (to lead). Using the distributional memory for Croatian (Šnajder et al., 2013), we selected the three most frequent subjects and objects for each verb. Finally, we chose a single adjective for each subject and object from the list of 20 most frequently co-occurring adjectives. This leaves us with 18 semantically plausible ANVAN phrases, illustrated in Table 1 (left column).

We manually collected substitutes for each word in the phrases. Three annotators were given a phrase and instructed to propose up to three substitutes for each word, while preserving both grammaticality and meaning; cf. the right column in Table 1. This yielded an evaluation dataset that contains 408 words: 158 adjectives, 167 nouns, and 83 verbs, each with multiple substitutes.

[2] `http://hjp.znanje.hr`

Target phrase	**odličan** đak prijeći brza cesta *(excellent pupil cross fast road)*
Possible substitutes	dobar *(good)*, potvrdan *(affirmative)*, crtani *(drawn)*, sportski *(sportive)*

Table 2: Word-choice item example. Target word in bold; correct substitute underlined.

Word Choice Task and Evaluation. We use the substitution dataset to set up a word choice task (Melamud et al., 2016): Each CDSM is presented with an ANVAN target phrase, a position in this phrase, a correct substitute and three distractors. Its task is to recognize the substitute that fits best into the context. Distractors were chosen by randomly picking three words of the same POS (adjective, noun or verb) that were not proposed as substitutes for that component in the given phrase. Table 2 shows an example of a single word-choice item.

In concrete terms, to evaluate a candidate substitute with respect to an ANVAN target phrase, we compute the cosine similarity between the compositionally computed vector for the ANVAN phrase computed "as is", and the phrase vector for the ANVAN phrase with the word at the current position replaced by the candidate substitute. The assumption is that a meaning-preserving substitution will leave the phrase vector largely unchanged and thus lead to a high cosine value. We report accuracy as the percentage of items for which the correct substitute received a higher cosine value than the incorrect substitutes.[3]

Models. We use the PLF and the two variants described in Section 2 (PLF-train and PLF-test). We build all three PLF versions for both phrase extraction approaches described in Sec. 2. In addition, we consider two baselines, namely the simple componentwise additive (add) and multiplicative (mult) models (Mitchell and Lapata, 2008).

4 Results

Table 3 shows the overall accuracy for each model. The standard PLF with dependency-extracted bigrams obtained the highest overall accuracy. The difference to the next-best model, *add*, is however not significant (p>0.01, McNemar's test).

	Counts	
Type	Surface level	Dependency level
adj-noun	14,249,655	15,548,616
subject-verb	3,147,289	3,994,552
verb-object	2,698,654	4,931,198

Table 4: fHrWaC number of predicate-argument co-occurrences at surface and dependency level.

Our new evaluation method allows us to further analyze this result by computing results for individual phrase positions (columns in Table 3). We find that PLF significantly outperforms both baselines for verbs (p<0.01, McNemar's test). This is in line with, and can potentially explain, the good results for English (Paperno et al., 2014), since in the English evaluation setup, the ANVAN phrase pairs differ only in the verbs (cf. Section 3). In contrast, *add* performs as well as or better than the PLF an adjectives and nouns.

A potential explanation for these patterns is *valency*: The verb has the highest valency of all words in the phrase (two arguments). Arguably, verbs can profit most from the additional expressiveness of PLF over the simpler CDSMs. Apparently, for adjectives (one argument) the expressiveness-learnability tradeoff is balanced between the two models, and for nouns (no arguments, thus no functional role) the additive model's simplicity wins.

Comparing the different PLF versions, we find no benefit for the modifications proposed by Gupta et al. (2015), who also obtained a null result for PLF-train, but found PLF-test to outperform plain PLF. For Croatian, PLF-test performs comparably to PLF for nouns and adjectives, but does clearly worse for verbs. A potential explanation follows from Gupta et al.'s analysis of the difference between PLF and PLF-test as a bias-variance tradeoff: the original PLF uses the lexical vector of the predicate as a "prior" for the phrase meaning, which makes it more robust, but also less flexible. PLF-test uses only the predicate matrix to compute a phrase vector and is thus more dependent on the data quality: on good data, it can outperform PLF, but it will be outperformed on noisy data.

Indeed, there is evidence that the verb-argument matrices are noisy in Croatian: Table 4 compares co-occurrence frequencies at the surface and dependency levels for three predicate-argument combinations. It shows that >90% of A-N combinations are

[3] The annotated dataset with compiled word choice tasks is available at: `http://takelab.fer.hr/data/croanvan`

		Phrase position					
Model	Phrase vectors	A1	N1	V	A2	N2	Overall
add		73.4	**92.0**	44.6	**70.1**	89.7	74.0
mult		39.2	61.4	32.5	40.2	62.8	47.4
PLF	Dependency-based	**74.7**	85.2	**66.3** *	67.5	85.9	**76.0**
PLF-train		58.2	89.8	49.4	51.9	83.3	66.9
PLF-test		72.2	85.2	60.2	67.5	84.6	74.0
PLF	Surface-based	55.7	87.5	63.9	65.4	84.6	71.7
PLF-train		54.4	89.8	51.8	56.4	82.1	67.2
PLF-test		69.6	87.5	55.4	60.3	83.3	71.4

Table 3: Model accuracy per phrase position. Asterisk (*) indicates a statistically significant result when comparing the best PLF version with the best simple CDSM, namely *add* (McNemar's test, p<0.01).

adjacent on the surface, while this holds for less than 80% of the S-V and 55% of the V-O combinations. As it is generally true that parsing quality deteriorates for long distance dependencies, the S-V and V-O matrices are arguably built from noisier data, which can account for disadvantage for PLF-test. In this manner, the free word order of Croatian does have an effect on CDSM performance.

That being said, parsing quality is evidently good enough for syntactic analysis to pay off: the results for using surface co-occurrence based versions of the PLF model perform generally worse than the PLF using dependency-base co-occurrences, with the exception of N1 (subject) position.

5 Conclusion

We built a Practical Lexical Function (PLF) model for Croatian and evaluated it on a newly created dataset of adjective-noun-verb-adjective-noun (AN-VAN) phrases. Our evaluation differs from existing English work (Paperno et al., 2014) by using a lexical substitution setup. Crucially, this allows us to analyze performance for individual phrase components. We find that the PLF's specific strength lies in modeling verbs, while it only does as well as simple additive models for nouns and adjectives. As we use dependency parses, the free word order of Croatian does not pose a major problem of the plain PLF, although we have evidence that it does affect the less robust PLF-test by Gupta et al. (2015). For future work, we will perform similar evaluation on a wider range of models and collect more evidence on the impact of typological differences on results.

Acknowledgments

This work has been supported in part by the Croatian Science Foundation under the project UIP-2014-09-7312. The third author has been supported by the DFG (SFB 732, Project D10).

References

Željko Agić and Danijela Merkler. 2013. Three syntactic formalisms for data-driven dependency parsing of Croatian. In *Proceedings of TSD 2013, Lecture Notes in Artificial Intelligence*. Springer, pages 560–567.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*. Cambridge, MA, pages 1183–1193.

Edward Grefenstette. 2013. *Category-theoretic quantitative compositional distributional models of natural language semantics*. Ph.D. thesis, University of Oxford.

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. 2012. Multi-step regression learning for compositional distributional semantics. In *Proceedings of IWCS 2012*. Potsdam, Germany.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*. Uppsala, Sweden, pages 33–37.

Abhijeet Gupta, Jason Utt, and Sebastian Padó. 2015. Dissecting the practical lexical function model for compositional distributional semantics. In *Proceedings of the Fourth Joint Conference on Lexical and*

Computational Semantics. Denver, Colorado, pages 153–158.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *Proceedings of CoNLL*. Sofia, Bulgaria, pages 114–123.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *International Conference on Text, Speech and Dialogue*. Springer, Brno, Czech Republic, pages 395–402.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of SEMEVAL*. pages 48–53.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of CONLL*. Berlin, Germany, pages 51–61.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL*. Columbus, OH, pages 236–244.

Denis Paperno, Nghia The Pham, and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of ACL*. Baltimore, MD, pages 90–99.

Ravi Sinha and Rada Mihalcea. 2009. Combining lexical resources for contextual synonym expansion. In *Proceedings of RANLP*. Borovets, Bulgaria, pages 404–410.

Jan Šnajder, Sebastian Padó, and Željko Agic. 2013. Building and evaluating a distributional memory for Croatian. In *Proceedings of ACL*. Sofia, Bulgaria, pages 784–789.