Fine-grained Semantic Textual Similarity for Serbian

Vuk Batanović, Miloš Cvetanović, Boško Nikolić
School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
vuk.batanovic@student.etf.bg.ac.rs, cmilos@etf.bg.ac.rs, nbosko@etf.bg.ac.rs

Abstract

Although the task of semantic textual similarity (STS) has gained in prominence in the last few years, annotated STS datasets for model training and evaluation, particularly those with fine-grained similarity scores, remain scarce for languages other than English, and practically non-existent for minor ones. In this paper, we present the Serbian Semantic Textual Similarity News Corpus (*STS.news.sr*) – an STS dataset for Serbian that contains 1192 sentence pairs annotated with fine-grained semantic similarity scores. We describe the process of its creation and annotation, and we analyze and compare our corpus with the existing news-based STS datasets in English and other major languages. Several existing STS models are evaluated on the Serbian STS News Corpus, and a new supervised bag-of-words model that combines part-of-speech weighting with term frequency weighting is proposed and shown to outperform similar methods. Since Serbian is a morphologically rich language, the effect of various morphological normalization tools on STS model performances is considered as well. The Serbian STS News Corpus, the annotation tool and guidelines used in its creation, and the STS model framework used in the evaluation are all made publicly available.

Keywords: short-text semantic similarity, corpus annotation, morphological normalization

1. Introduction

Semantic Textual Similarity (STS), sometimes also referred to as Short-text Semantic Similarity (STSS), is the task of assigning a numerical score to a given pair of short texts based on the level of semantic equivalence between them. The minimal numerical score in a given range indicates complete semantic independence, while the maximal score indicates full semantic equality. Although STS has important implications for a whole range of other natural language processing tasks, including information retrieval, question answering, machine translation, textual entailment, etc., research on this topic started appearing only around a decade ago (Corley and Mihalcea, 2005; Mihalcea, Corley, and Strapparava, 2006; Islam and Inkpen, 2008). Semantic Textual Similarity has gained in prominence since 2012, with its inclusion in the annual *SemEval* shared tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017).

A large collection of datasets with fine-grained semantic similarity scores has been annotated in this series of shared tasks, using a standardized methodology, and has been made publicly available. The move from binary similarity scores (Dolan and Brockett, 2005) to fine-grained ones has allowed for more precise model training and evaluation. However, most of this development has been limited to English. Several other major languages have been considered recently, including Spanish (Agirre et al., 2014, 2015; Cer et al., 2017), French (Vu et al., 2014), Portuguese (Fonseca et al., 2016), Chinese and Japanese (Hayashi and Luo, 2016), Arabic (Cer et al., 2017), and Hindi (Agarwal et al., 2017). Among them, only the datasets in Spanish, Portuguese, and Arabic have been made publicly available. To the best of our knowledge, there has been no development of STS datasets with fine-grained similarity scores for minor languages so far.

In this paper, we present the Serbian Semantic Textual Similarity News Corpus (*STS.news.sr*)[1] – a publicly available STS dataset for Serbian annotated with fine-grained semantic similarity scores. Although there has been some recent work on the broader task of semantic relatedness in Polish (Wróblewska and Krasnowska-Kieraś, 2017), our dataset is, as far as we know, the first STS dataset for a Slavic language.

The remainder of this paper is structured as follows: in Section 2 we describe the creation and annotation of *STS.news.sr*, while in Section 3 we analyze and compare it with the available STS datasets in other languages. Section 4 provides some baseline model results on *STS.news.sr*, as well as an evaluation of several supervised bag-of-words STS models. Within this section, we also assess the impact of morphological normalization methods for Serbian – a language with rich morphology – on STS models. Finally, in Section 5 we present our conclusions and some potential avenues of future research.

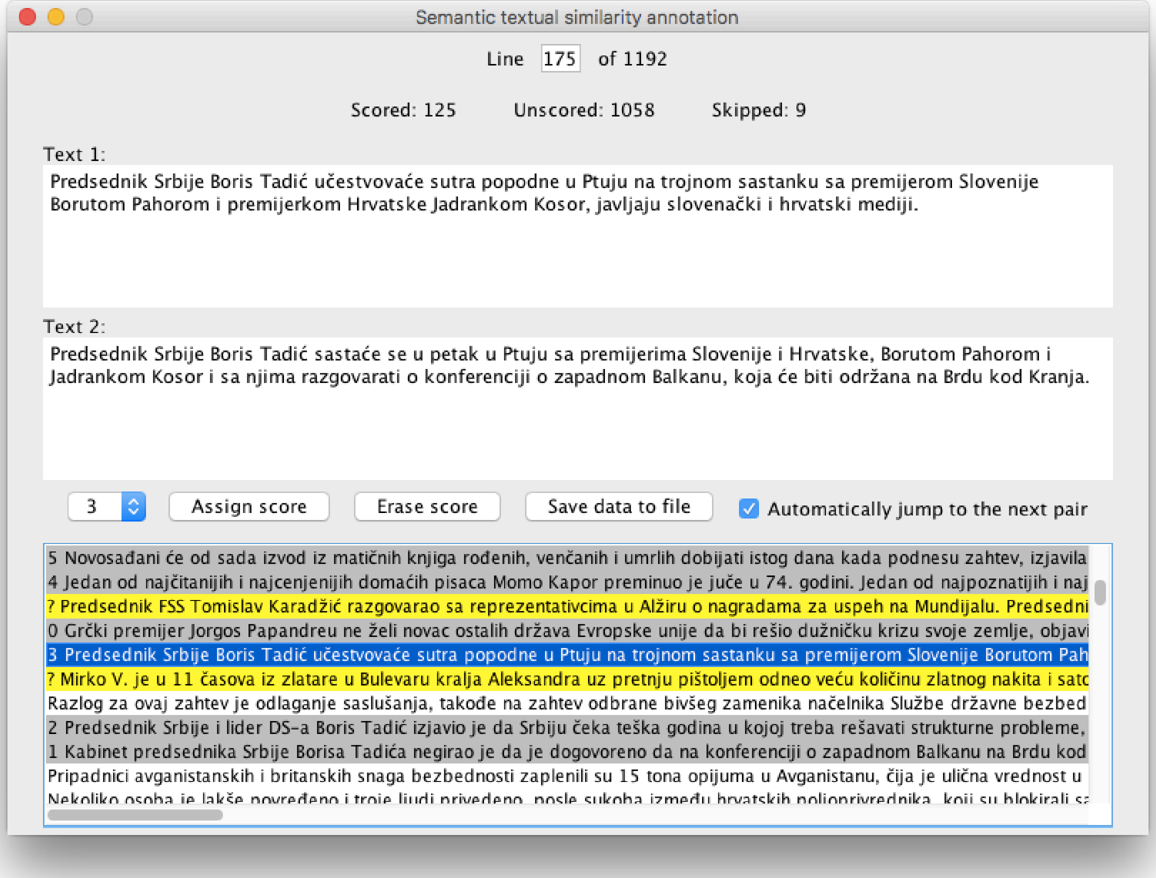
2. Dataset Creation and Annotation

The initial step in STS dataset creation is the acquisition of a suitable collection of short-text pairs. We deemed the existing Serbian Paraphrase Corpus (*paraphrase.sr*)[2] (Batanović, Furlan, and Nikolić, 2011; Furlan, Batanović, and Nikolić, 2013), a set of 1194 sentence pairs gathered from the news domain, to be a suitable source for this purpose. Firstly, we went through the corpus and manually corrected any typographical errors and restored any missing diacritical marks. Two sentence pairs were removed from the dataset since one was found to be a duplicate and the other included a text longer than one sentence. The remaining 1192 sentence pairs were then given to five annotators who independently assigned a semantic similarity score to each pair.

For the sake of standardization, we chose to follow the annotation methodology established in the *SemEval* STS tasks, and we adopted the scoring scheme (a 0 – 5 Likert scale) and the general annotation guidelines used therein (Agirre et al., 2013). However, our initial consultations with the annotators showed that the sentence pair examples for each score that are included in the *SemEval* annotation instructions can be somewhat unclear, particularly those for scores 2 – 4. This issue had an effect on lowering task comprehension and annotation quality. To rectify this, we replaced all examples with new ones, and we increased the

[1] http://vukbatanovic.github.io/STS.news.sr/

[2] http://vukbatanovic.github.io/paraphrase.sr/



Line 175 of 1192

Scored: 125 Unscored: 1058 Skipped: 9

Text 1:

Predsednik Srbije Boris Tadić učestvovaće sutra popodne u Ptuju na trojnom sastanku sa premijerom Slovenije Borutom Pahorom i premijerkom Hrvatske Jadrankom Kosor, javljaju slovenački i hrvatski mediji.

Text 2:

Predsednik Srbije Boris Tadić sastaće se u petak u Ptuju sa premijerima Slovenije i Hrvatske, Borutom Pahorom i Jadrankom Kosor i sa njima razgovarati o konferenciji o zapadnom Balkanu, koja će biti održana na Brdu kod Kranja.

3 ☑ Automatically jump to the next pair

```
5 Novosađani će od sada izvod iz matičnih knjiga rođenih, venčanih i umrlih dobijati istog dana kada podnesu zahtev, izjavila
4 Jedan od najčitanijih i najcenjenijih domaćih pisaca Momo Kapor preminuo je juče u 74. godini. Jedan od najpoznatijih i naj
? Predsednik FSS Tomislav Karadžić razgovarao sa reprezentativcima u Alžiru o nagradama za uspeh na Mundijalu. Predsedni
0 Grčki premijer Jorgos Papandreu ne želi novac ostalih država Evropske unije da bi rešio dužničku krizu svoje zemlje, objavi
3 Predsednik Srbije Boris Tadić učestvovaće sutra popodne u Ptuju na trojnom sastanku sa premijerom Slovenije Borutom Pah
? Mirko V. je u 11 časova iz zlatare u Bulevaru kralja Aleksandra uz pretnju pištoljem odneo veću količinu zlatnog nakita i satc
Razlog za ovaj zahtev je odlaganje saslušanja, takođe na zahtev odbrane bivšeg zamenika načelnika Službe državne bezbed
2 Predsednik Srbije i lider DS-a Boris Tadić izjavio je da Srbiju čeka teška godina u kojoj treba rešavati strukturne probleme,
1 Kabinet predsednika Srbije Borisa Tadića negirao je da je dogovoreno da na konferenciji o zapadnom Balkanu na Brdu kod
Pripadnici avganistanskih i britanskih snaga bezbednosti zaplenili su 15 tona opijuma u Avganistanu, čija je ulična vrednost u
Nekoliko osoba je lakše povređeno i troje ljudi privedeno, posle sukoba između hrvatskih poljoprivrednika, koji su blokirali s
```

Figure 1: The *STSAnno* annotation tool interface

number of examples from one to three per score. In order to limit our own bias in the selection of new examples, we chose suitable pairs from the 2012 *MSRPar* and the 2013-2016 *HDL SemEval* STS corpora in English (since they all belong to the news domain), and we had them professionally translated into Serbian. We considered only those pairs whose averaged scores are integers – an integer average usually means that all annotators assigned the same similarity score to a particular pair, indicating its unambiguity. The final selection was made in consultation with the annotators to ensure the representativeness of each example. Our annotation guidelines and examples, in both Serbian and English, are available on the *STS.news.sr* repository.

Once the instructions were finalized, all annotators first scored a subset of 60 randomly selected pairs from the corpus (~5% of the total), before proceeding to annotate the entire dataset. This initial batch was subsequently used to calculate the annotator self-agreement scores. The annotation process was completed within approximately two months.

In order to make the annotation quicker and easier, we created *STSAnno*[3], a simple offline annotation tool. *STSAnno* allows an annotator to view in parallel the texts in a pair, assign a semantic similarity score to them, and change or erase existing scores. Annotators can also assign a special symbol to a pair to temporarily skip it, which can be useful when faced with difficult examples. Scored, unscored, and skipped pairs are highlighted in different colors to easily distinguish between them. Sentence pairs can be scored in the order in which they are given, or in any other order chosen by the annotator. At all times, a statistical overview of the annotation progress is displayed. A screenshot of *STSAnno* during the annotation of *STS.news.sr* is shown in Figure 1.

3. Dataset Analysis

The annotator self-agreements and the inter-annotator agreements were calculated using the Pearson correlation coefficient *r*. Table 1 contains the self-agreement scores and Table 2 the inter-annotator agreements. In addition to the pairwise inter-annotator scores, we also measure the agreement of each annotator with the average of the scores of all other annotators.

The agreement scores are generally very high. Even though the annotators had different backgrounds (annotator #1 is a computational linguist, annotators #2 and #3 are linguists, while annotators #4 and #5 are non-linguists) there is no major difference in correlation values due to this. This indicates that with well-chosen example pairs and clear guidelines, even non-experts can achieve very high levels of annotation quality on STS corpora. Our average inter-rater agreement between an annotator and the average of the scores of all other annotators is 0.92, which is therefore

[3] http://vukbatanovic.github.io/STSAnno/

Annotator	#1	#2	#3	#4	#5	Average
Self-agreement	0.95	0.97	0.97	0.85	0.92	0.93

Table 1: Annotator self-agreement scores

Annotator	#1	#2	#3	#4	#5	Average
#1	/					
#2	0.90	/				
#3	0.89	0.87	/			
#4	0.88	0.84	0.85	/		
#5	0.90	0.88	0.86	0.86	/	
Average of other annotators	0.94	0.92	0.91	0.90	0.92	0.92

Table 2: Inter-annotator agreement scores

the upper bound for STS model performance on this dataset. This agreement is higher than the ones reported for *SemEval* datasets from the news domain (Agirre et al., 2013, 2014, 2015) by around 0.05 – 0.1, most likely due to the increased number and quality of the examples in our annotation instructions.

The final similarity score for each sentence pair was obtained by averaging the scores of all five annotators. Figure 2 shows the distribution of sentence pairs within the Serbian STS News Corpus across the range of similarity score values. It is moderately balanced, with the exception of a large peak regarding the pairs with the score 3.0. However, similar distributional irregularities are also present in other news-based STS datasets.

A comparison between our dataset and other publicly available STS corpora created from the news domain is shown in Table 3. We consider the following corpora:

- In English: the 2012 *SemEval MSRPar* corpus (Agirre et al., 2012), the combined 2013-2016 collection of *SemEval HDL* corpora (Agirre et al., 2013, 2014, 2015, 2016), and the 2014 *SemEval Deft-news* corpus (Agirre et al., 2014).
- In Spanish: the combined 2014-2015 *SemEval News* corpora (Agirre et al., 2014, 2015).
- In Portuguese: the 2016 ASSIN corpus, divided into European and Brazilian Portuguese portions (Fonseca et al., 2016).
- In Arabic: the 2017 *SemEval* translation of a part of the *MSRPar* corpus into Arabic (Cer et al., 2017).

The size of the Serbian STS News Corpus is average when compared to the other available STS corpora, both in terms of the number of sentence pairs and in terms of token count (we counted only alphanumerical tokens). The average length of a sentence in *STS.news.sr* is greater than in most other STS datasets, while the average similarity score is 2.51 – almost ideal given the 0 – 5 score scale. In fact, *STS.news.sr* is much more balanced than the English *SemEval MSRPar* corpus, which is the one most similar to it in terms of source material, type, and size.

However, nearly all of the considered STS corpora exhibit strong distribution peaks around score values 3 and 4, in case of the 0 – 5 score scale, and scores 2 and 3 in case of the 0 – 4 scale. The ASSIN corpora score distribution is heavily skewed toward the central 2 – 4 values. The only corpus with a more uniform score distribution is the English *SemEval HDL*. This is probably at least in part a natural effect of the shortness of the texts (news headlines) in this corpus. With longer texts, the likelihood of coming across sentence pairs with near-identical semantics (score values close to 5 on a 0 – 5 scale) is lower. Similarly, in longer text pairs, there is a greater chance of encountering at least some semantic links between the sentences, lowering the probability of minimally scored pairs. We leave for further work the consideration of how and to what extent these distribution irregularities in most corpora affect the training and evaluation of STS models.

4. Evaluation

We evaluate several STS models on the Serbian STS News Corpus. As a performance metric, we utilize the Pearson correlation coefficient between the model output and the averaged annotated similarity scores, which we consider the gold standard. We first consider unsupervised models, and evaluate them on the entire *STS.news.sr*. Then, we move on to supervised algorithms, which are evaluated using 10-fold cross-validation with sorted stratification.

4.1 Unsupervised Models

The first unsupervised model we consider is the one used as a standard baseline in all *SemEval* STS shared tasks – a simple word overlap technique in which sentences are split into tokens using white space and then represented as bag-of-words vectors in the multidimensional token space (Agirre et al., 2012). Token counts in a sentence are binarized, so that each vector dimension has a value of one if that token appeared in the sentence, and zero otherwise. Cosine similarity is used to compute the similarity between such sentence vectors. We also improve upon this baseline by lowercasing the text, removing punctuation, and using the tokenizer for Serbian included in the ReLDI (*Regional Linguistic Data Initiative*) project repository[4] (Samardžić, Ljubešić, and Miličević, 2015; Ljubešić, Erjavec, et al., 2016). Since it proves to be highly beneficial, this improved tokenization approach is utilized for all subsequent models.

The second baseline model that we use is one based on averaging the embeddings of words in a sentence and calculating the cosine similarity of the mean vectors. We employ the *word2vec* algorithm (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013), as implemented in the *gensim* library (Řehůřek and Sojka, 2010), since Cer et al. (2017) showed it to be superior to the other word

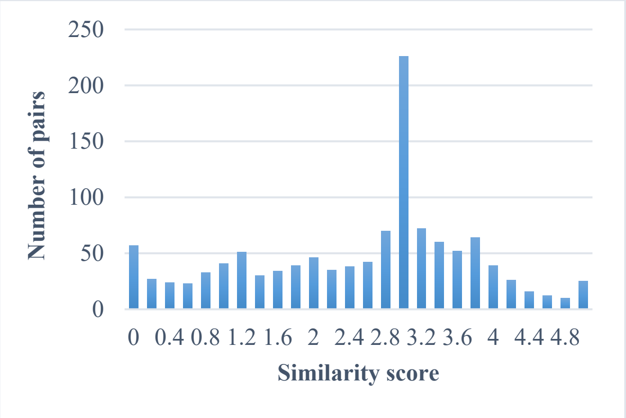


Figure 2: Sentence pair similarity score distribution in the Serbian STS News Corpus

[4] http://reldi.spur.uzh.ch/blog/tokeniser/

Corpus	Lang.	Score scale	Sentence pairs	Tokens	Average sentence length in tokens	Average similarity score	Percentage of sentence pairs with scores rounded to					
							0	1	2	3	4	5
STS.news.sr	SR	0 – 5	1192	64 K	~27	2.51	9.06%	14.93%	16.11%	39.43%	16.53%	3.94%
SemEval MSRPar	EN	0 – 5	1500	54 K	~18	3.30	0.13%	4.47%	13.87%	36.47%	36.2%	8.86%
SemEval HDL	EN	0 – 5	2499	37 K	~7	2.62	10.56%	17.89%	17.29%	19.93%	22.09%	12.24%
SemEval Deft-news	EN	0 – 5	300	9 K	~16	3.03	4.0%	11.0%	16.0%	27.33%	31.33%	10.33%
SemEval News	ES	0 – 4	980	68 K	~35	2.20	6.33%	16.33%	36.94%	29.08%	11.32%	/
ASSIN (PT)	PT-PT	1 – 5	5000	145 K	~14	3.04	/	2.72%	35.14%	24.16%	31.56%	6.42%
ASSIN (BR)	PT-BR	1 – 5	5000	130 K	~13	3.04	/	1.74%	33.94%	28.84%	30.76%	4.72%
SemEval MSRPar	AR	0 – 5	510	18 K	~18	3.36	0.2%	3.33%	13.14%	34.9%	40.39%	8.04%

Table 3: An overview of news-based STS corpora with fine-grained semantic similarity scores

embedding models when used in this context. The skip-gram *word2vec* architecture is trained on the Serbian web corpus *srWaC* (Ljubešić and Klubička, 2014), the largest publicly available text corpus in Serbian, containing 555 million tokens. The *srWaC* corpus is parsed to remove punctuation marks and words that are not in Serbian, and is then lowercased. This reduces the corpus to around 470 million tokens, with a vocabulary of around 3.8 million entries. We use 100-dimensional vectors and a window size of 10 for the skip-gram model. All other parameters are kept at the *gensim* default settings.

In both baseline models, we experiment with a simple negation-marking technique in which a single word after a negation word is marked with a special prefix in order to distinguish it from its non-negated form. Such techniques were previously found useful in simple models for other semantic tasks, such as sentiment analysis, both in English (Pang, Lee, and Vaithyanathan, 2002) and in Serbian (Batanović, Nikolić, and Milosavljević, 2016).

The results of basic unsupervised STS model evaluations are presented in Table 4. We find that the word overlap model outperforms the embedding-based one, indicating a higher level of string similarity between the sentences in *STS.news.sr*, which is to be expected given the method used to collect sentence pairs for the original *paraphrase.sr* corpus (Furlan, Batanović, and Nikolić, 2013). We therefore also consider a joint model in which the *word2vec* mean vector and the binarized bag-of-words vector of a sentence are concatenated and used in cosine similarity calculation. This joint baseline proves superior to both individual models.

Model	Pearson *r*
Word overlap (white space tokenizer)	0.6461
Word overlap (Serbian tokenizer)	0.6869
Word overlap (Serbian tokenizer + negation marking)	0.6862
word2vec averaging (Serbian tokenizer)	0.6211
word2vec averaging (Serbian tokenizer + negation marking)	0.6257
Word overlap + *word2vec* averaging (Serbian tokenizer)	**0.6949**
Word overlap + *word2vec* averaging (Serbian tokenizer + negation marking)	0.6943

Table 4: Unsupervised baseline STS model performances on the entire dataset

The proposed negation-marking technique is found beneficial on the *word2vec* baseline. However, since it has a slightly detrimental effect on the superior word overlap and joint models, we do not use it in further experiments.

4.1.1 Morphological Normalization

Next, having in mind the morphological complexity of Serbian, we evaluate the impact of morphological normalization methods on our baseline STS models. The results are presented in Table 5.

Three stemming algorithms developed for Serbian are considered – the optimal and the greedy algorithm of Kešelj and Šipka (2008), and the improvement of the greedy algorithm by Milošević (2012). We also evaluate a stemmer for Croatian, a language closely related to Serbian, by Ljubešić and Pandžić, which is a refinement of the approach presented in (Ljubešić, Boras, and Kubelka, 2007). We use the *SCStemmers* package (Batanović, Nikolić, and Milosavljević, 2016) in which all of the aforementioned algorithms are implemented.

Similarly, we consider two publicly available lemmatizers for Serbian and one for Croatian. The first lemmatizer for Serbian is *BTagger*, which is available in two variants – one that only normalizes word suffixes (Gesmundo and Samardžić, 2012b), and another that also deals with word prefixes, allowing for full lemmatization (Gesmundo and Samardžić, 2012a). In addition, we assess a lemmatization model for Croatian developed by Agić, Ljubešić, and Merkler (2013) for the CST lemmatizer (Jongejan and Dalianis, 2009). The final lemmatizer that is evaluated is the one for Serbian by Ljubešić, Klubička, et al. (2016), which relies on a large inflectional lexicon and an improved part-of-speech tagger.

Morphological normalizer	Model		
	Word overlap	*word2vec* averaging	Word overlap + *word2vec* averaging
None	0.6869	0.6211	0.6949
Stemmers			
Kešelj and Šipka (optimal)	0.7291	0.5971	0.7338
Kešelj and Šipka (greedy)	0.7218	0.5966	0.7271
Milošević	0.7210	0.5986	0.7266
Ljubešić and Pandžić	0.7287	0.6077	**0.7339**
Lemmatizers			
BTagger (suffix)	0.7031	0.5936	0.7126
BTagger (suffix + prefix)	0.7019	0.5921	0.7112
Agić et al.	0.7064	0.5915	0.7143
Ljubešić et al.	0.7225	0.5937	**0.7283**

Table 5: The effects of morphological normalization methods on unsupervised baseline STS model performances on the entire dataset

The results show that the application of morphological normalization has a consistently positive impact on the performance of word overlap and joint baseline models, and a consistently detrimental one on the purely embedding-based method. On average, stemmers tend to have a better effect on STS models than lemmatizers do. The best overall stemmer is Ljubešić and Pandžić's stemmer for Croatian, although the optimal stemmer of Kešelj and Šipka is a close second. Ljubešić and Pandžić's stemmer was also found to be the best option for sentiment classification in Serbian (Batanović and Nikolić, 2016, 2017), making it a good choice in general. The lemmatizer of Ljubešić et al. proves to be the best one in this setting, but it is still outmatched by the top two stemming algorithms.

4.2 Supervised Models

We limit the examination of supervised models to those that do not require more advanced syntactic tools, like dependency parsers, since the development of such tools for Serbian has only recently begun (Samardžić et al., 2017). We therefore evaluate the performance of several bag-of-words models. The approach proposed by Islam and Inkpen (2008) is the most basic one we consider. Within it, each word from the shorter sentence is paired to its most similar word in the longer sentence, and word pair similarities are calculated as a mixture of three string similarity metrics and one corpus-based semantic similarity measure. Supervision is used in this method to determine the optimal balance between the string and the corpus-based measures in the final score. As the corpus-based measure, we utilize the cosine similarity of the same *word2vec* 100-dimensional vectors as before.

We also evaluate three models derived from this basic approach. The first is *LInSTSS* (*Language-independent Short-text Semantic Similarity*), proposed by Furlan, Batanović, and Nikolić (2013), in which word pair similarities are weighted according to the term frequencies of the words in question. We calculate the TF values using the *srWaC* corpus.

The second one is *POST STSS* (*Part-of-speech Tag-supported Short-text Semantic Similarity*), proposed by Batanović and Bojić (2015), which utilizes similarity weighting based on the part of speech of each word in a pair. In order to obtain POS tags we use the Serbian morphosyntactic tagger developed by Ljubešić, Klubička, et al. (2016). This tagger produces morphosyntactic descriptors (MSDs) and POS tags according to the MULTEXT-East (Erjavec, 2017) version 5 standard for Croatian[5]. In this standard, a POS tag is simply the first letter of an MSD.

POST STSS relies on a two-stage optimization procedure in order to determine the best weighting settings, including the weights for each part of speech/each POS grouping, as well as the values within a POS interaction matrix that allow or disallow the pairing of words belonging to different parts of speech/POS groups. In the first phase of the *POST STSS* parameter optimization – pseudo-exhaustive search – we classify all MSDs into one of the following seven POS groups:

1. Nouns – MSDs start with *N*
2. Verbs – MSDs start with *V*
3. Adverbs – MSDs start with *R*
4. Adjectives – MSDs start with *A*
5. Pronouns – MSDs start with *P*
6. Numerals – MSDs start with *M*
7. Other – all other MSD values

In the second optimization phase – steepest ascent hill climbing – we expand the POS weights from these seven groups into 29 classes. Each class represents a different MSD category/type combination, according to the MULTEXT-East version 5 standard. For instance, the *Noun* MSD category is divided into two types – common nouns (*Nc*) and proper nouns (*Np*), while the *Numeral* category is divided into four types – cardinal numerals (*M_c*), ordinal numerals (*M_o*), multiple numerals (*M_m*), and special numerals (*M_s*). There are also MSD categories, such as adpositions (*S*), which are not divided into types – in these cases one weight is assigned to an entire category. The only exceptions to this classification scheme are the residuals (*X*), where a single weight is assigned to the entire category since only one type of residual (foreign) appears in *STS.news.sr*, and the punctuation category (*Z*), which is ignored, since punctuation is filtered out during tokenization. However, this category/type classification is applied only to those types for which actual MSD values are specified in the MULTEXT-East version 5 standard. For example, the standard allows for a separate type of copular verbs (*Vc*), but no tags are specified under this type and the utilized tagger does not employ it, so this category/type combination does not necessitate a separate weight.

POST STSS requires a nested cross-validation during the first optimization phase in order to tune the model hyperparameters – the initial POS weight values, the initial POS interaction values, the initial string similarity weight, the choice of the POS weighting function, and the option of using a special weight value minimization process at the end of the first optimization phase. Here, for the sake of

[5] http://nl.ijs.si/ME/V5/msd/html/msd-hr.html

efficiency, we only optimize the initial POS weights and the initial POS interaction values in a nested three-fold CV. We consider the same options for their initial values as in (Batanović and Bojić, 2015). For the remaining hyperparameters we use the settings found optimal in previous experiments (Batanović and Bojić, 2015) – the initial string similarity weight is set to 0.5, the arithmetic mean is the chosen POS weighting function, and the value minimization process is not used.

Finally, we propose and evaluate a mixture of *LInSTSS* and *POST STSS* that uses both TF-based and POS-based weighting of word similarities. Within this model, similarities between words i and j are calculated as follows:

$$Sim(i,j) = (w_{str} \times Str(i,j) + w_{sem} \times Sem(i,j)) \\ \times TF(i,j) \times POS(i,j)$$

where w_{str} and w_{sem} are the string and the semantic similarity weights ($w_{str} + w_{sem} = 1$), $Str(i,j)$ is the mixture of three string similarity metrics as defined in (Islam and Inkpen, 2008), and $Sem(i,j)$ is the corpus-based semantic similarity measure (as noted, we use the cosine similarity of *word2vec* vectors in this paper). $TF(i,j)$ is the term frequency weighting function, as defined in (Furlan, Batanović, and Nikolić, 2013), while $POS(i,j)$ is the part-of-speech weighting function (as noted, in this paper we always use the arithmetic mean of the weights for the parts of speech of words i and j). The optimization procedure for this approach, which we name *POS-TF STSS*, is identical to the one used for *POST STSS*, since term frequencies are obtained from the *srWaC* corpus in an unsupervised way.

In all supervised models, weight values are optimized in steps of 0.1. The string similarity weight is chosen from the [0.3, 0.7] range, while the POS weights are optimized in the [0.7, 1.3] range. In order to minimize the chance of overfitting to the training set, the hill climbing part of the *POST/POS-TF STSS* optimization is stopped heuristically, when there are no hill climbing moves left whose error reduction on the training set is at least 5% of the error reduction of the first move made in the climb.

The 10-fold cross-validation results for all models are shown in Table 6. We repeat the unsupervised model evaluation using 10-fold CV to be able to make a fair comparison between the performances of the unsupervised and the supervised models. Furthermore, we measure the impact of morphological normalization on supervised models, but we limit the scope to the tools that were previously found to be the best in each category – the stemmer of Ljubešić and Pandžić, and the lemmatizer of Ljubešić et al. All of the evaluated STS models, both supervised and unsupervised, are made available as parts of *STSFineGrain*[6], a collection of STS models and a unified framework for their evaluation, implemented in Java.

Results show that supervised models perform noticeably better than unsupervised ones on non-normalized text, but the gap between the two narrows when stemming or lemmatization is applied. Stemming has a clearly positive effect on almost all models, while lemmatization only brings an improvement to word overlap methods, with mixed effects on supervised ones. The three models derived from Islam and Inkpen's approach consistently outperform the original algorithm. *LInSTSS* generally achieves results similar to *POST STSS*, but the *POS-TF STSS* mixture model performs better than both *LInSTSS* and *POST STSS* independently. In fact, when used in conjunction with the stemmer of Ljubešić and Pandžić, *POS-TF STSS* achieves the best result among all the models that we considered.

Model	Morphological normalizer		
	None	*Stemmer* Ljubešić and Pandžić	*Lemmatizer* Ljubešić et al.
Unsupervised models			
Word overlap	0.6970	0.7367	0.7278
word2vec averaging	0.6405	0.6295	0.6136
Word overlap + *word2vec* averaging	0.7050	0.7417	0.7335
Supervised models			
Islam and Inkpen	0.7387	0.7444	0.7350
LInSTSS (TF weighting)	0.7534	0.7573	0.7494
POST STSS (POS weighting)	0.7538	0.7593	0.7491
POS-TF STSS (POS and TF weighting)	0.7599	**0.7665**	0.7606

Table 6: STS model performances on 10-fold CV

4.2.1 Optimal Parameters

Naturally, the optimal parameter values for supervised models vary somewhat from one morphological normalization approach to another, but there are consistent patterns that can be observed. The optimal string similarity weight in the basic Islam and Inkpen approach tends to be 0.7, resulting in an optimal semantic similarity weight of 0.3. This is not surprising given the higher level of string similarity between the sentences in *STS.news.sr*. Nevertheless, in the *LInSTSS* model the optimal value of the string similarity weight is a bit lower (0.6) which indicates that the addition of TF weighting increases the importance of non-surface forms of similarity.

Some variation between the optimal POS weight settings of *POST STSS* and those of *POS-TF STSS* does exist. However, we did not encounter any systematic differences between the optimal parameters of these two algorithms, nor between their chosen optimal hyperparameter values. The optimal initial POS weights are most often set to the neutral value of 1.0, while the optimal initial POS interaction setting is usually to allow word pairings between all parts of speech.

Common nouns typically retain a neutral POS weight value of 1.0 and are found to be more important than proper nouns, whose weight revolves around the 0.8 – 0.9 mark. The weight for main verbs is almost universally set to the 1.3 maximum, indicating the central role of a verb in conveying the meaning of a sentence. This effect was also evident when applying *POST STSS* to data in English (Batanović and Bojić, 2015), and was previously noted by other researchers as well (Wiemer-Hastings, 2004). Auxiliary verbs, on the other hand, carry far less semantic

[6] http://vukbatanovic.github.io/STSFineGrain/

content and are therefore assigned a lower weight, most often in the 0.7 – 0.9 range. With regard to this, participial adjectives are consistently found to be the most important kind of adjectives, with the maximum POS weight value. The weight of possessive adjectives[7] also tends to be augmented, but to a lesser extent, while other adjectives are most often assigned the 0.9 weight value. The weights within the adverb category follow a similar pattern – adverbial participles are assigned higher weights, around 1.2, while the weight of adverbs proper is usually 1.0 or 1.1. Numerals, particularly ordinal ones, are found to be quite important – their weight values usually approach the upper POS weight bound. The high weight values assigned to numerals, adverbs, and most adjectives probably indicate their importance in correctly measuring the exact level of semantic similarity between sentences whose main actions/verbs are the same. The POS weights of pronouns are generally lower, ranging between 0.7 and 1.0, while the weight value assigned to abbreviations tends to fluctuate between 0.8 and 1.0. The remaining parts of speech mostly consist of functional words, such as conjunctions, adpositions, interjections, etc., which do not contain salient semantic content and are, thus, assigned low weight values in the 0.7 – 0.8 range.

The optimized POS interaction matrix allows the pairing of words belonging to different parts of speech in most cases, but some nonsensical pairings are generally prohibited, like the one of pronouns and purely functional words like conjunctions. However, the fact that most pairings remain permitted shows that such strict prohibitions are only useful in a very limited number of cases, and that, in performance optimization, the *POST/POS-TF STSS* models rely primarily on the modification of POS weight values. This conclusion is further validated by the fact that the optimal string similarity weight in these models usually remains at the starting value of 0.5. Consequently, the optimal semantic similarity weight has the same value.

5. Conclusion

In this paper, we have presented the Serbian STS News Corpus, the first STS corpus with fine-grained semantic similarity scores in a Slavic language. We have compared it to similar STS corpora in other languages and have evaluated several unsupervised baseline STS models on it. A number of previously presented supervised models have also been considered. In addition, we have proposed *POS-TF STSS*, a new bag-of-words method that uses both term frequency weighting and part-of-speech weighting, and outperforms similar algorithms on *STS.news.sr*. The effects of various morphological normalization techniques on STS model performances have also been evaluated. In particular, we have found that using the stemmer for Croatian by Ljubešić and Pandžić alongside the *POS-TF STSS* approach yields the best results among the evaluated models. Finally, the optimal values of supervised model parameters have been discussed.

In the future, we plan to construct additional, topically distinct STS corpora in Serbian, and to use them to conduct a more thorough model evaluation. We also aim to examine the influence of gold score distribution irregularities on the behavior of STS models.

6. Acknowledgements

This work was partially supported by the III 44009 research grant of the Ministry of Education, Science, and Technological Development of the Republic of Serbia. The annotation of the Serbian STS News Corpus was supported by the Regional Linguistic Data Initiative (*ReLDI*) via the Swiss National Science Foundation grant no. 160501. The authors would like to thank Aleksandar Milinković for assisting in *srWaC* lemmatization, and Ana Bjelogrlić, Jelena Bošnjak, Marko Janković, and Ognjen Krešić for annotating the *STS.news.sr* corpus. The first author also served as an annotator.

7. Bibliographical References

Agarwal, D., Mujadia, V., Sharma, D. M., and Mamidi, R. (2017). A Modified Annotation Scheme for Semantic Textual Similarity. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2017)*, Budapest, Hungary.

Agić, Ž., Ljubešić, N., and Merkler, D. (2013). Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the Fourth Biennial International Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria, Association for Computational Linguistics, pp. 48–57.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2014). SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, Association for Computational Linguistics, pp. 81–91.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, USA, Association for Computational Linguistics, pp. 252–263.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., … Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, USA, Association for Computational Linguistics, pp. 497–511.

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada, Association for Computational Linguistics, pp. 385–393.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, Atlanta, Georgia, USA, Association for Computational Linguistics, pp. 32–43.

Batanović, V., and Bojić, D. (2015). Using Part-of-Speech Tags as Deep-Syntax Indicators in Determining Short-Text Semantic Similarity. *Computer Science and*

[7] NB: Possessive adjectives in Serbian correspond to possessive noun forms in English.

Information Systems, 12(1), pp. 1–31.

Batanović, V., Furlan, B., and Nikolić, B. (2011). A software system for determining the semantic similarity of short texts in Serbian. In *Proceedings of the 19th Telecommunications Forum (TELFOR 2011)*, Belgrade, Serbia, IEEE, pp. 1249–1252. [in Serbian]

Batanović, V., and Nikolić, B. (2016). Sentiment Classification of Documents in Serbian: The Effects of Morphological Normalization. In *Proceedings of the 24th Telecommunications Forum (TELFOR 2016)*, Belgrade, Serbia, IEEE.

Batanović, V., and Nikolić, B. (2017). Sentiment Classification of Documents in Serbian: The Effects of Morphological Normalization and Word Embeddings. *Telfor Journal*, 9(2), pp. 104–109.

Batanović, V., Nikolić, B., and Milosavljević, M. (2016). Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, European Language Resources Association (ELRA), pp. 2688–2696.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, Vancouver, Canada, Association for Computational Linguistics, pp. 1–14.

Corley, C., and Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment (EMSEE 2005)*, Ann Arbor, Michigan, USA, Association for Computational Linguistics, pp. 13–18.

Dolan, W. B., and Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, South Korea, Asia Federation of Natural Language Processing, pp. 9–16.

Erjavec, T. (2017). MULTEXT-East. In *Handbook of Linguistic Annotation*, Springer, Dordrecht, pp. 441–462.

Fonseca, E. R., Santos, L. B. dos, Criscuolo, M., and Aluísio, S. M. (2016). Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual. *Linguamática*, 8(2), pp. 3–13. [in Portuguese]

Furlan, B., Batanović, V., and Nikolić, B. (2013). Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, 55(3), pp. 710–719.

Gesmundo, A., and Samardžić, T. (2012a). Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, Association for Computational Linguistics, pp. 368–372.

Gesmundo, A., and Samardžić, T. (2012b). Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, European Language Resources Association (ELRA), pp. 2103–2106.

Hayashi, Y., and Luo, W. (2016). Extending Monolingual Semantic Textual Similarity Task to Multiple Cross-lingual Settings. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, European Language Resources Association (ELRA), pp. 1233–1239.

Islam, A., and Inkpen, D. (2008). Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2), p. Article No. 10.

Jongejan, B., and Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre- , in- and suffixes alike. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL 2009)*, Suntec, Singapore, ACL and AFNLP, pp. 145–153.

Kešelj, V., and Šipka, D. (2008). A Suffix Subsumption-Based Approach to Building Stemmers and Lemmatizers for Highly Inflectional Languages with Sparse Resources. *INFOtheca*, 9(1–2), p. 23a–33a.

Ljubešić, N., Boras, D., and Kubelka, O. (2007). Retrieving Information in Croatian: Building a Simple and Efficient Rule-Based Stemmer. In *INFuture2007: Digital Information and Heritage*, Zagreb, Croatia, Department for Information Sciences, Faculty of Humanities and Social Sciences, pp. 313–320.

Ljubešić, N., Erjavec, T., Fišer, D., Samardžić, T., Miličević, M., Klubička, F., and Petkovski, F. (2016). Easily Accessible Language Technologies for Slovene , Croatian and Serbian. In *Proceedings of the Conference on Language Technologies & Digital Humanities*, Ljubljana, Slovenia, pp. 120–124.

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)*, Gothenburg, Sweden, Association for Computational Linguistics, pp. 29–35.

Ljubešić, N., Klubička, F., Agić, Ž., and Jazbec, I.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, European Language Resources Association (ELRA), pp. 4264–4270.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, Boston, Massachusetts, USA, AAAI Press, pp. 775–780.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations Workshop (ICLR 2013)*, Scottsdale, Arizona, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013)*, Lake Tahoe, Nevada, USA, Curran Associates, Inc., pp. 3111–3119.

Milošević, N. (2012). Stemmer for Serbian language, arXiV 1209.4471.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pp. 79–86.

Řehůřek, R., and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, European Language Resources Association (ELRA), pp. 45–50.

Samardžić, T., Ljubešić, N., and Miličević, M. (2015). Regional Linguistic Data Initiative (ReLDI). In *Proceedings of the Fifth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, Hissar, Bulgaria, pp. 40–42.

Samardžić, T., Starović, M., Agić, Ž., and Ljubešić, N. (2017). Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, Valencia, Spain, Association for Computational Linguistics, pp. 39–44.

Vu, H. H., Villaneau, J., Saïd, F., and Marteau, P. F. (2014). Sentence similarity by combining explicit semantic analysis and overlapping n-grams. In *Proceedings of the 17th International Conference on Text, Speech and Dialogue (TSD 2014)*, Brno, Czech Republic, Springer International Publishing, pp. 201–208.

Wiemer-Hastings, P. (2004). All parts are not created equal: SIAM-LSA. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, Illinois, USA, Erlbaum.

Wróblewska, A., and Krasnowska-Kieraś, K. (2017). Polish evaluation dataset for compositional distributional semantics models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, Canada, Association for Computational Linguistics, pp. 784–792.

8. Language Resource References

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). MSRPar (*SemEval* 2012 STS shared task), distributed online: https://www.cs.york.ac.uk/semeval-2012/task6/index.php%3Fid=data.html, 1.0.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). HDL (*SemEval* 2013 STS shared task), distributed online: http://ixa2.si.ehu.es/sts/index.php%3Foption=com_content&view=article&id=49&Itemid=56.html, 1.0.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2014). Deft-news (*SemEval* 2014 STS shared task), distributed online: http://alt.qcri.org/semeval2014/task10/index.php?id=data-and-tools, 1.0.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2014). HDL (*SemEval* 2014 STS shared task), distributed online: http://alt.qcri.org/semeval2014/task10/index.php?id=data-and-tools, 1.0.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2014). News [Spanish] (*SemEval* 2014 STS shared task), distributed online: http://alt.qcri.org/semeval2014/task10/index.php?id=data-and-tools, 1.0.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2015). HDL (*SemEval* 2015 STS shared task), distributed online: http://alt.qcri.org/semeval2015/task2/index.php?id=data-and-tools, 1.0.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2015). News [Spanish] (*SemEval* 2015 STS shared task), distributed online: http://alt.qcri.org/semeval2015/task2/index.php?id=data-and-tools, 1.0.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., … Wiebe, J. (2016). HDL (*SemEval* 2016 STS shared task), distributed online: http://alt.qcri.org/semeval2016/task1/index.php?id=data-and-tools, 1.0.

Batanović, V., Furlan B., and Nikolić B. (2011). The Serbian Paraphrase Corpus (*paraphrase.sr*), distributed online: http://vukbatanovic.github.io/paraphrase.sr/, 1.0, ISLRN 192-200-046-033-9.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). MSRPar [Arabic] (*SemEval* 2017 STS shared task), distributed online: http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools, 1.0.

Dolan, W. B., Brockett, and C., Quirk, C. (2005). Microsoft Research Paraphrase Corpus, distributed online: http://www.microsoft.com/en-us/download/details.aspx?id=52398, 1.0.

Fonseca, E. R., Santos, L. B. dos, Criscuolo, M., and Aluísio, S. M. (2016). ASSIN (Avaliação de Similaridade Semântica e INferência textual), distributed online: http://nilc.icmc.usp.br/assin/, 1.0.

Ljubešić, N., and Klubička, F. (2014). Serbian web corpus srWaC, distributed via CLARIN.SI: http://hdl.handle.net/11356/1063, 1.1.