Predicting Style-Dependent Collocations in Russian Text Corpora

Lada Petrushenko and Olga Mitrofanova

Saint Petersburg University
Universitetskaya emb. 7-9
199034 Saint Petersburg, Russia
ladapetrushenko@gmail.com, o.mitrofanova@spbu.ru

Abstract. The paper presents the results of experiments on the development of distributional language models for predicting acceptable collocations of the ADJ+NOUN type. The models are trained on stylistically diverse Russian corpora (news, popular science, fiction, poetry). The evaluation of the models allows us to determine optimal parameters for collocation prediction and explore linguistic features of predicted collocations.

Keywords: predictive distributional models, collocations, Russian corpora.

1 Introduction

Since the emergence of Word2Vec in 2013, predictive distributional language models have become the preferred tool for dealing with semantic NLP tasks. With their help, researchers can now process huge amounts of text data at a faster rate and overcome technical limitations while working with large text collections. Moreover, predictive models have proven to be more effective at representing semantic relations between tokens compared to count-based models. Over the past ten years, a number of resources with pretrained predictive models have emerged, making it possible for any user to access the required data and investigate language phenomena both observable and unseen in corpora.

Such language models can be used to predict both paradigmatic and syntagmatic relations between tokens. To date, most Russian researchers focused on the investigation and description of paradigmatic relations: e.g., semantic similarity evaluation and taxonomy enrichment tasks were under discussion in RUSSE contests, organized within the conference on Computational Linguistics and Intellectual Technologies "Dialogue" [1]. Nevertheless, syntagmatic relations underlying lexical constructions of various types are described less thoroughly within the framework of distributional semantic models for Russian, except for a few projects, the CoCoCo database [2] and the DSM-Calculator [3] are among them. Predicting collocations of a predefined type is useful for a number of NLP tasks such as text summarization, text generation, sentiment analysis, etc., as it can improve the quality of task implementation.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 79–89, 2023. © Tribun EU 2023

The study is based on the assumption that linguistic properties of predicted collocations are determined by features of language models. Thus, the core experiments aim to define the optimal parameters that provide training of non-contextualized distributional models for predicting acceptable collocations: vector space dimensionality, context window size, corpus preprocessing, dictionary filtering, similarity measures, etc. The study focuses on predicting a particular type of collocations for Russian texts of different styles. The description of predicted style-dependent collocations, developed during our study, fills the gaps in Russian NLP. It allows us to obtain novel results that are relevant for text classification based on construction identification [4] and style transfer [5].

The paper is structured as follows: the section ″Related work″ contains the theoretical foundations of our study, the section ″Experimental design″ describes the linguistic data used in experiments and explains the experimental procedures, the section ″Results″ provides an overview of the optimal parameters for collocation prediction and the linguistic features of predicted collocations, and the section ″Conclusion and further research″ summarizes the achieved results and outlines future work.

2 Related Work

The idea of a lexical construction as a core unit of language can be traced back to the Construction Grammar (CxG) theory developed by Ch. Fillmore [6]. According to Ch. Fillmore, a lexical construction is a sequence of lexical units in which some components define the surrounding context while others serve as supplementary elements [7]. CxG emphasizes certain characteristics of lexical constructions, such as their semantic, syntactic, and pragmatic nature, as well as the potential for idiomatic meanings. Lexical constructions are usually classified as regards their idiomaticity and compositionality [8,9,10,11]. Following [11], collocations are treated as lexical constructions with partially restricted use of its components.

There are three main approaches to collocation extraction: count-based (statistical) approach involves association measures (e.g., PMI, t-score, Log Likelihood, Chi-square, etc.) and vectorization models (e.g., TF-IDF, LSA, HAL, COALS, etc.); hybrid approach, which relies on both linguistic and statistical information, is implemented in techniques for extracting lexical-grammatical patterns (e.g., keyphrase extraction algorithms like RAKE, KEA, Topia, etc.); predictive approach is implemented in distributional semantic models of dense word embeddings. Predictive models are represented by non-contextualized or static Word2Vec-type models and contextualized Transformer-based models. Among non-contextualized models, such as Word2Vec [12,13] and FastText [14,15] can be highlighted. As for contextualized models, it is worth mentioning BERT [16], ELMo [17] and contemporary developments. In this research, we focus on the first type of models due to the fact that it is more challenging to control prediction results when working with BERT-like models [18].

Predictive models have proved to be more effective at detecting word similarity compared to count-based models in a number of tasks such as synonym detection, measuring semantic relatedness, concept categorization, etc. [19]. Research has also shown that these models can be applied to predicting specific types of collocations, such as constructions consisting of a verb and a noun [20], constructions with an attributive meaning [21], collocations expressing lexical functions [22]. In our study, we focus on the task of predicting style-dependent collocations of ADJ+NOUN type.

3 Experimental Design

In our research, we used segments of Taiga [23] and Lib.ru.sec [24] corpora consisting of stylistically diverse Russian texts: news, popular science, fiction, poetry. We conducted two sets of experiments: the first aimed at detecting the optimal parameters for collocation prediction, and the second aimed at describing the linguistic features of the predicted collocations.

For the first experiment, the subcorpora of the following size were used:

- **Fontanka** (the news subcorpus of Taiga) comprises 73,140,388 tokens and 3,885,119 sentences (the entire subcorpus was taken for both the experiments);
- **Nplus1** (the Taiga subcorpus of non-fictional (popular-science) texts) comprises 1,667,938 tokens and 72,002 sentences (the entire subcorpus was taken for both the experiments);
- **Stihi_ru** (the Taiga subcorpus of poems) comprises 5,986,693 tokens and 421,956 sentences (the first 50,000 texts were taken for the first experiment);
- **Lib.ru.sec** (the Taiga subcorpus of finction texts) comprises 9,669,140 tokens and 677,134 sentences (the first 100 texts were taken for the first experiment).

Due to technical limitations, it was not possible to process the complete versions of Stihi_ru and Lib.ru.sec, but the representative subcorpora of both of them were taken.

For working with the data of Fontanka and Nplus1 in the first experiment, we relied on the predefined annotation provided by the authors of the dataset. This implies morphosyntactic annotation performed in terms of Universal Dependencies (UD) [25]. Stihi_ru and Lib.ru.sec were downloaded in *.txt format and then annotated by us with spacy_udpipe [26]. After preprocessing, all the data was presented in the CoNLL-U format [27].

For the second experiment, we annotated all the subcorpora with the pymorphy2 morphological tagger [28] to test whether this type of annotation could increase the quality of predictions. After analyzing the predictions of the first set of the "best" models, we added the following restrictions to the algorithm of handling the data:

- token length should be more than 2 characters;

- tokenized sentence length should be more than 2 tokens;
- annotation in terms of the pymorphy2 tagset should be transformed into UD annotation.

As a result, for the second experiment we worked with the following dataset:

- **Fontanka** (the news subcorpus of Taiga) comprises 41,234,011 tokens and 3,611,338 sentences (the entire subcorpus was taken for both the experiments);
- **Nplus1** (the Taiga subcorpus of non-fictional (popular-science) texts) comprises 1,328,657 tokens and 90,313 sentences (the entire subcorpus was taken for both the experiments);
- **Stihi_ru** (the Taiga subcorpus of poems) comprises 5,961,406 tokens and 703,358 sentences (the first 100,000 texts were taken for the second experiment);
- **Lib.ru.sec** (the Taiga subcorpus of finction texts) comprises 31,591,065 tokens and 3,791,616 sentences (the first 1000 texts were taken for the second experiment).

In both experiments, we trained a set of Word2Vec and FastText models and validated the results with the help of pseudo-disambiguation procedure, a common approach to testing the quality of predictions [29,30]. Under this approach, the test data comprises combinations of three tokens:

- target word: e.g., *день* (*day*);
- candidate word that can form a lexical construction together with the target word due to their co-occurrence in the corpus: e.g., *летний день* (*summer day*);
- candidate word that can form a lexical construction but does not occur with the target unit in the specific corpus or a candidate word that cannot form a lexical construction with the target token at all: e.g., **железный день* (**iron day*).

Correct collocations for pseudo-disambiguation were chosen on the basis of whether they occurred in all 4 subcorpora at least 8 times. The incorrect collocations were chosen according to relative frequencies: if the relative frequency of occurrence for an incorrect collocate was lower than the relative frequency of occurrence for a correct collocate in all the subcorpora, such incorrect collocate was taken into account in our data. As a result, we obtained a set of 155 combinations of target word-correct collocate-incorrect collocate, which we used to evaluate the performance of the models. The evaluation was conducted on the same data regardless of the genre of the text on which the model was trained.

The models were trained in gensim [31] with all possible combinations of the following 6 parameters:

- a metric to measure the degree of similarly of vectors: Euclidean distance, squared Euclidean distance, cosine similarity, correlation of vectors;
- vector size: 100, 150, 200, 250, 300;

- context window size: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10;
- a metric that determines the threshold for word frequency, indicating which words should be considered by a model (min_count): 5, 10, 15;
- a metric that indicates the approach for computing a vector of contextual features when CBOW is chosen (cbow_mean): 0 (sum), 1 (mean);
- a metric that indicates the approach for sorting the vocabulary: 0 (none), 1 (by descending frequency).
- a metric responsible for limiting the usage of RAM when building the vocabulary (max_vocab_size): None, 30000, 60000; if the number of unique tokens is larger than the defined threshold, the model neglects low-frequency tokens.

The results of collocation prediction were evaluated with such metrics as precision, recall, and F1-score. The components for these metrics were calculated as follows:

- TP is the number of times when the model predicted a collocation correctly, and the tokens within the collocation semantically combine with each other;
- TN is the number of times when the model predicted a collocation incorrectly, and the tokens within the collocation do not semantically combine with each other;
- FN is the number of times when the model did not predict a collocation and could not differentiate between incorrect and correct collocates: the similarity measure between both pairs is similar up to two decimal places.

4 Results

4.1 Optimal parameters for collocation prediction

For investigation, we took the best 500 results for each model and analyzed the training parameters. The best results consist of combinations of model training metrics and were selected by sorting all the outcomes of a specific model based on the precision score metric in solving the pseudo-disambiguation task. Based on the first experiment and comparison of precision scores for all the results (cf. major best scores for Word2Vec models in Table 1), we made the following conclusions.

The parameters that have the most influence on prediction results are similarity measure, vector size, minimum word frequency for consideration by the model, and dictionary sorting parameter.

These parameters are organized in some sort of clusters where certain combinations of them have unchanging stable values while others can fluctuate. For example, the context window size and the approach to vector calculation did not affect the overall results within these clusters. The cbow_mean parameter could be either 0 or 1, without influencing the predictions. This observation applies to both Word2Vec models and FastText models.

It can be concluded that the correlation coefficient or cosine similarity measure are the most effective in determining the semantic relationships between

Table 1: Major training parameters for the best Word2Vec models.

Corpus	Precision %	metric	vector size	min_count	sorted_vocab	n_comb
Fontanka	62.58	cosine	250	15	0	20
NPlus1	89.03	cosine	250	15	1	20
		correlation	150	10	0	20
Stihi_ru	87.1	correlation	100	15	1	20
Lib.ru.sec	70.32	correlation	150	15	0	20
		cosine	250	15	1	20

collocation elements. The best 500 results of each model did not include those that used the Euclidean distance or squared Euclidean distance as the similarity measure. This finding supports the hypothesis of other researchers that these metrics are not as effective in capturing semantic properties compared to cosine similarity measure [32].

FastText models are better at predicting synatgmatic relations compared to paradigmatic relations. To prove this point, we took two nouns: *год* (*year*) and *исследование* (*study*), and examined how frequently adjectives appeared in the best 1000 predicted words for them. As a result, it turned out that there were no adjectives at all. For comparison, Word2Vec predicted 192 and 199 adjectives for these words respectively. Because of this, we decided not to use FastText models in our second experiment.

4.2 Linguistic features of predicted collocations

For the second experiment, we trained 4 Word2Vec models with the following best combinations of parameters:

- **Fontanka**: metric='cosine', size=200, min_count=15, sorted_vocab=0, window=any (default: 5), cbow_mean=any (default: 1);
- **Nplus1**: metric='cosine', size=150, min_count=15, sorted_vocab=0, window=any (default: 5), cbow_mean=any (default: 1);
- **Stihi_ru**: metric='cosine', size=150, min_count=10, sorted_vocab=1; window=any (default: 5), cbow_mean=any (default: 1);
- **Lib.ru.sec**: metric='cosine', size=100, min_count=15, sorted_vocab=0; window=any (default: 5), cbow_mean=any (default: 1).

We experimented with several words that were chosen from all the corpora randomly: *сайт* (*website*), *человек* (*man or human*), *научно-исследовательский* (*research or scientific-research*), *красивый* (*beautiful*), *день* (*day*), *система* (*system*).

We evaluated the results based on consistency coefficient A. The first evaluation procedure consists in the evaluation of consistency across our research models. For each of the mentioned words, we obtained 10 collocates from each model. Thus, having a total of 40 collocates for each word from the 4 models, except for *научно-исследовательский* (*scientific-research*) - for this word, we obtained 30 collocates as it was absent in the dictionary of the model trained on poetry. We performed pairwise comparisons of the results from Nplus1, Fontanka, Lib.ru.sec, and Stihi_ru. The coefficient A was calculated as the number of overlapping predictions relative to all predictions (cf. Table 2). The predictions are considered overlapping if they appear in the predictions of at least two models.

Table 2: Evaluation of consistency among the models.

Target word	**Repeating collocates**	**Consistensy of predictions A**
сайт (*website*)	*новостной, электронный, подробный* (*news, electronic, detailed*)	0,075 (3 repetitions per 40 collocates)
человек (*man or human*)	*верующий, больной, чужой, нищий* (*religious, sick, alien, poor*)	0,1 (4 repetitions per 40 collocates)
красивая (*beautiful*)	*блондинка, прелесть* (*blonde, charm*)	0,05 (2 repetitions per 40 collocates)
система (*system*)	*дистанционная, автоматическая* (*remote, automatic*)	0,05 (2 repetitions per 40 collocates)
день (*day*)	*выходной, летний, июньский, десятый, сегодняшний, бессонный* (*weekend, summer, June, tenth, today, sleepless*)	0,15 (6 repetitions per 40 collocates)
научно-исследовательский (*research or scientific-research*)	*машиностроение* (*mechanical engineering*)	0,033 (1 repetition per 30 collocates)

In some cases, the similarity value of a collocation predicted by one model could be twice as large compared to that of another one: cf. *(электронный) сайт* (*website*), cosine = 0.31 vs. 0.61. This could be due to the fact that models show differences between the strength of connections within matching collocations. At the same time, the low number of overlapping predictions can be explained by topical differences of the corpora.

In the second experiment, we compared the results of predictions with the results from the Word Portrait project of The Russian National Corpus (RNC) [33]. Additionally, we made the same requests to two models from DSM-Calculator [3,34]: the model trained on the Russian Wikipedia dump in 2017 (referred to as DSM-Wiki), and the model trained on the Lib.ru corpus in 2017 with a context window size of 5 (referred to as DSM-Lib). Coefficient A is calculated based on the total number of predictions from the Nplus1, Fontanka, and Lib.ru.sec models for 6 target words, which amounts to 60 collocates, and the Stihi_ru model, which predicted 50 collocates.

- **Fontanka**: matches RNC in 3 predictions (A = 0.05); DSM-Wiki — in 7 (A = 0.116); DSM-lib — in 7 (A = 0.116);
- **Nplus1**: matches RNC in 3 predictions (A = 0.05); DSM-Wiki — in 5 (A = 0.083); DSM-lib — in 0 (A = 0);

- **Lib.ru.sec**: matches RNC in 2 predictions (A = 0.033); DSM-Wiki — in 2 (A = 0.033); DSM-lib — in 1 (A = 0,016);
- **Stihi_ru**: matches RNC in 5 predictions (A = 0,1); DSM-Wiki — in 1 (A = 0,02); DSM-lib — in 0 (A = 0).

The low number of matches once again shows that the model predictions strongly depend on the corpus style and main topics. At the same time, it can be unexpected that there is a low number of matches between the predictions of the Lib.ru.sec and DSM-lib models, since both had been trained on fiction texts. The differences between these models may be attributed to the fact that they were trained on different-sized datasets (around 9 million tokens and around 146 million tokens).

The predictions contain both established combinations, e.g. *официальный сайт* (*official website*), *выходной день* (*day-off*), *социальная система* (*social system*), etc.) and combinations that have represent terminological expressions, e.g. *бортовая система* (*on-board system*), etc.

The predicted constructions are mostly compositional and not idiomatic. The scientific-popular and news models perform worse in predicting constructions for the adjective *красивый* (*beautiful*) compared to fiction and poetic models. This can be explained by the fact that this adjective has a subjective interpretation that is less common in certain types of texts compared to fiction texts. There are instances of constructions where the meanings of the elements are not coordinated, for example, **красивое образование* (**beautiful education*), **красивая география* (**beautiful geography*), and **безлунный день* (**moonless day*). Such collocations are considered as anomalous.

5 Conclusion and future research

In this paper, we conducted several experiments on the prediction of noun phrases in Russian texts representing different writing styles: news, popular science, fiction, poetry. We analyzed a set of parameters and identified patterns that enable us to highlight specific parameters and approaches for predicting acceptable collocations unseen in the corpora. Multiple Word2Vec and FastText models were trained and evaluated, results leading to the conclusion that Word2Vec performs better in predicting syntagmatic relations, while FastText is better at predicting paradigmatic relations. Additionally, it is worth noting that such parameters as the association measure metric, vector size, minimum word frequency for model consideration, and dictionary sorting parameter play important roles in training the model for the prediction of noun phrases. Lastly, our experiments allowed us to observe the stylistic variation of collocations depending on the corpus type they were trained on.

The best models trained on stylistically diverse corpora are incorporated in the web-application "Construction Calculator" [35] developed on a Hugging Face platform. We plan to use the application for collocation generation in tests for studying Russian as a foreign language and for training collocationaware style-sensitive language models which are necessary in automatic style detection.

Acknowledgments. Research is performed with support of RSF grant № 21-78-10148 «Modeling the meaning of a word in individual linguistic consciousness based on distributive semantics».

References

1. RUSSE: Workshop on Russian Semantic Evaluation. https://russe.nlpub.org/, last accessed 31 Oct 2023
2. CoCoCo. https://cococo.cosyco.ru/, last accessed 31 Oct 2023
3. DSM-Calculator. https://dsm-calculator.ru/, last accessed 31 Oct 2023
4. Dubovik, A. Automatic text style identification in terms of statistical parameters. In: Computer Linguistics and Computing Ontologies. Issue 1, pp. 29-45 (2017)
5. RUSSE-2022: Detoxification. https://www.dialog-21.ru/evaluation/2022/russe/, last ac- cessed 31 Oct 2023
6. Fillmore, C.J. Syntactic Intrusion and the Notion of Grammatical Construction. In: Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society, pp. 73–86 (1985)
7. Fillmore, C.J., Kay, P., O'Connor, M.C. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. In: Linguistic Society of America, Vol. 64, No. 3 (1988)
8. Vinogradov, V.V. Selected Works: Lexicology and Lexicography (1977)
9. Gak, V.G. On the Problem of Semantic Syntagmatics. In: Language Transformations, pp. 272–297 (1998)
10. Apresyan, Ju.D. Selected Works. Vol. 1. Lexical Semantic: Synonymic Means of Language. (1995)
11. Iordanskaya, L.N. Melchuk, I.A. Sense and Compatibility in a Dictionary (2007)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient Estimation of Word Representations in Vector Space. (2013a) https://doi.org/10.48550/arXiv.1301.3781
13. Mikolov, T., Yih, W., Zweig, G. Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of NAACL-HLT 2013, pp. 746–751. https://aclanthology.org/N13-1090/, last accessed 31 Oct 2023 (2013b)
14. Joulin, A., Bojanowski, P., Mikolov, T., Jegou, H., Grave, E. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2979–2984 (2018) https://doi.org/10.18653/v1/D18-1330
15. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135–146 (2017) https://doi.org/10.1162/tacl_a_00051
16. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) – Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186 (2018) https://doi.org/10.18653/v1/N19-1423

17. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, Ch., Lee, K., Zettlemoyer, L. Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association of the Computational Linguistics: Human Language Technologies, Vol. 1, pp. 2227–2237 (2018) https://doi.org/10.18653/v1/N18-1202
18. Belyi, A.V. , Mitrofanova, O.A., Dubinina, N.A. Distributive Semantic Models in Language Learning: Automatic Generation of Lexical-Grammatical Tests for Russian as a Foreign Language. In: Proceedings of 2023 Corpus Linguistics Conference, 2023 (2023)
19. Baroni, M., Dinu, G., Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 238–247 (2014) https://doi.org/10.3115/v1/P14-1023
20. Kolesnikova, O., Gelbukh, A. A Study of Lexical Function Detection with Word2Vec and Supervised Machine Learning. In: Journal of Intelligent and Fuzzy Systems, pp. 1–8 (2020) https://doi.org/10.3233/JIFS-179866
21. Hartung, M., Kaupmann, F., Jebbara, S., Cimiano, Ph. Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases. In: 15th Meeting of the European Chapter of the Association for Computational Linguistics (EACL). `https://aclanthology.org/E17-1006/`, last accessed 31 Oct 2023 (2017).
22. Enikeeva E.V., Mitrofanova O.A. Russian Collocation Extraction based on Word Embeddings. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue», pp. 52–64 (2017)
23. Taiga. `https://tatianashavrina.github.io/taiga_site/`, last accessed 31 Oct 2023
24. Lib.ru. `http://lib.ru/`, last accessed 31 Oct 2023
25. Universal Dependencies. `https://universaldependencies.org/`, last accessed 31 Oct 2023
26. Spacy_UDPipe. `https://pypi.org/project/spacy-udpipe/`, last accessed 31 Oct 2023
27. CoNNL-U. `https://pypi.org/project/conllu/`, last accessed 31 Oct 2023
28. pymorphy2. `https://pymorphy2.readthedocs.io/en/stable/`, last accessed 31 Oct 2023
29. Gale, W.A., Church, K.W., Yarowsky, D. Work on Statistical Methods for Word Sense Disambiguation. In: AAAI Fall Symposium on Probabilistic Approaches to Natural Language: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics. `https://studylib.net/doc/13790396/work-on--statistical-methods-for--word-sense--disambiguation`, last accessed 31 Oct 2023 (1992)
30. Dagan, I., Marcus, S., Markovitch, S. Contextual Word Similarity and Estimation from Sparse Data. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 164–171 (1993) https://doi.org/10.3115/981574.981596
31. gensim. `https://radimrehurek.com/gensim/`, last accessed 31 Oct 2023
32. Rohde, D.L.T., Gonnerman, L.M., Plaut, D.C. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence (2005)
33. RNC. `https://ruscorpora.ru/`, last accessed 31 Oct 2023
34. Bukia, G., Protopopova, E., Panicheva, P., Mitrofanova, O. Estimating Syntagmatic Asso- ciation Strength Using Distributional Word Representations. Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference «Dialogue», pp. 112–122 (2016)

35. Construction Calculator. https://huggingface.co/spaces/ladapetrushenko/construction_calculator, last accessed 31 Oct 2023