# Thesaurus Verification Based on Distributional Similarities

**Natalia Loukachevitch**
Lomonosov Moscow State University
Moscow, Russia
Louk_nat@mail.ru

**Ekaterina Parkhomenko**
Lomonosov Moscow State University
Moscow, Russia
parkat13@yandex.ru

## Abstract

In this paper we consider an approach to verification of large lexical-semantic resources as WordNet. The method of verification procedure is based on the analysis of discrepancies of corpus-based and thesaurus-based word similarities. We calculated such word similarities on the basis of a Russian news collection and Russian wordnet (RuWordNet). We applied the procedure to more than 30 thousand words and found some serious errors in word sense description, including incorrect or absent relations or missed main senses of ambiguous words.

## 1 Introduction

Large lexical-semantic resources such as Princeton WordNet (Fellbaum, 1998) and wordnets created for other languages (Bond and Foster, 2013) are important instruments for natural language processing. Developing and maintaining such resources requires special efforts, because it is difficult to find errors or gaps in structures consisting of thousands lexical units and relations between them.

In previous works, various methods on lexical enrichment of thesauri have been studied (Snow et al., 2006; Navigli and Ponzetto, 2012). But another issue was not practically discussed: how to find mistakes in existing thesaurus descriptions: incorrect relations or missed significant senses of ambiguous words, which were not included accidentally or appeared recently.

In fact, it is much more difficult to reveal missed and novel senses or wrong relations, if compared to novel words (Frermann and Lapata, 2016; Lau et al., 2014). So it is known that such missed senses are often found during semantic annotation of a corpus and this is an additional problem for such annotation (Snyder, Palmer, 2004; Bond, Wang, 2014).

In this paper, we consider an approach how to use embedding models to reveal problems in a thesaurus. Previously, distributional and embedding methods were evaluated in comparison with manual data (Baroni and Lenci, 2011; Panchenko et al., 2016). But we can use them in the opposite way: to utilize embedding-based similarities and try to detect some problems in a thesaurus.

We study such similarities for more than 30 thousand words presented in Russian wordnet RuWordNet (Loukachevitch et al., 2018)[1].

The structure of the paper is as follows. Section 2 is devoted to related work. In Section 3 we briefly present RuWordNet. Section 4 describes the procedure of calculating two types of word similarities based on thesaurus and a corpus. In Section 5 we analyze discrepancies between thesaurus-based and corpus-based word similarities, which can appear because of different reasons. In Section 6 we study groupings of distributionally similar words to an initial word using the thesaurus.

## 2 Related Work

In (Lau et al. 2014), the task of finding unattested senses in a dictionary is studied. At first, they apply the method of word sense induction based on LDA topic modeling. Each extracted sense is represented to top-N words in the constructed topics. To compute the similarity between a sense and a topic, the words in the definition are converted into the probability distribution. Then two probability distributions (gloss-based and topic-based) are compared using the Jensen-Shannon divergence. It was found that the proposed novelty measure could identify target lemmas with high- and medium-frequency novel senses. But the authors evaluated their method using word sense definitions in the Macmillan

[1] http://ruwordnet.ru/en/

dictionary and did not check the quality of relations presented in a thesaurus.

A series of works was devoted to studies of semantic changes in word senses (Gulordava and Baroni, 2011; Mitra et al., 2015; Frermann, Lapata, 2016) . Gulordava and Baroni (2011) study semantic change of words using Google n-gram corpus. They compared frequencies and distributional models based on word bigrams in 60s and 90s. They found that significant growth in frequency often reveals the appearance of a novel sense. Also it was found that sometimes the senses of words do not change but the context of their use changed significantly. For example, the context of word *parent* considerably change in 90s because of the most frequent collocation *single parent family*.

In (Mitra et al., 2015), the authors study the detection of word sense changes by analyzing digitized books archives. They constructed networks based on a distributional thesaurus over eight different time windows, clustered these networks and compared these clusters to identify the emergence of novel senses. The performance of the method has been evaluated manually as well as by comparison with WordNet and a list of slang words. But Mitra et al. did not check if WordNet misses some senses.

The task of revising and verifying of resources is important for developers of WordNet-like resources. Some ontological tools have been proposed to check consistency of relations in WordNet (Guarino and Welty, 2004; Alvez et al., 2018).

Some authors report about revision of mistakes and inconsistencies in their wordnets in the process of linking the wordnet and English WordNet (Cristea et al., 2004; Rudnicka et al., 2012). Rambousek et al. (2018) consider a crowdsourcing tool allowing a user of Czech wordnet to report errors. Users may propose an update of any data value. These suggestions can be approved or rejected by editors. Also visualization tools can help to find problems in wordnets (Piasecki et al. 2013; Johannsen et al., 2011).

Loukachevitch (2019) proposed to use embedding-based word similarities to find possible mistakes or inconsistencies in a WordNet-like thesaurus. In the current paper we provide some additional details for the (Loukachevitch, 2019) study.

## 3 RuWordNet

RuWordNet was created on the basis of another Russian thesaurus RuThes in 2016, which was developed as a tool for natural language processing during more than 20 years (Loukachevitch and Dobrov, 2002). Currently, the published version of RuWordNet includes 110 thousand Russian words and expressions.

The important feature of RuWordNet (and its source RuThes), which is essential for this study, is that a current news collection is used as a reference collection for maintenance of RuWordNet. Periodically, a new corpus (of last year news articles) is collected, single words and phrases absent in the current version of the thesaurus are extracted and analyzed for inclusion to the thesaurus (Loukachevitch, Parkhomenko, 2018). The monitoring of news flow is important because news articles concern many topics discussed in the current society, mention new terms and phenomena recently appeared.

The current version of RuWordNet comprises the following types of relations: hyponym-hypernym, antonyms, domain relations for all parts of speech (nouns, verbs, and adjectives); part-whole relations for nouns; cause and entailment relations for verbs. Synsets of different parts of speech are connected with relations of POS-synonymy. For single words with the same roots, derivational relations are described. For phrases included in RuWordNet, relations to component synsets are given.

## 4 Comparison of Distributional and Thesaurus Similarities

To compare distributional and thesaurus similarities for Russian according to RuWordNet, we used a collection of 1 million news articles as a reference collection. The collection was lemmatized. For our study, we took thesaurus words with frequency more than 100 in the corpus. We obtained 32,596 words (nouns, adjectives, and verbs).

Now we should determine what thesaurus relations or paths are taken to determine semantically similar entries. In the current study, we consider the following entries as semantically related to the initial thesaurus entry:

- its synonyms,
- all the entries located in the 3-relation paths, consisting of hyponym-hypernyms

relations or/and part-whole relations between synsets from the initial entry;

- all the entries linked with other direct relations to the initial entry;
- for ambiguous words, all sense-related paths were considered and thesaurus entries along these paths were collected together.

In such a way, for each word, we collected the thesaurus-based "bag" of similar words (TBag).

Then we calculated embeddings according to word2vec model with the context window of 3 words, planning to study paradigmatic relations (synonyms, hypernyms, hyponyms, co-hyponyms). Using this model, we extracted twenty the most similar words $w_i$ to the initial word $w_0$. Each $w_i$ should also be from the thesaurus. In such a way, we obtained the distributional (word2vec) "bag" of similar words for $w_0$ (DBag).

Now we can calculate the intersection between TBag and DBag and sum up the similarities in the intersection. Figure 1 shows the distribution of words according to the similarity score of the TBag-DBag intersection. The axis X denotes the total similarity in the TBag-DBag intersection: it can achieve more than 17 for some words, denoting high correspondence between corpus-based and thesaurus-based similarities.

Relative adjectives corresponding to geographical names have the highest similarity values in the TBag-DBag intersection, for example, *samarskii* (related to Samara city), *vologodskii* (related to Vologda city), etc. Also nouns denoting cities, citizens, nationalities, nations have very high similarity value in the TBag-DBag intersection.

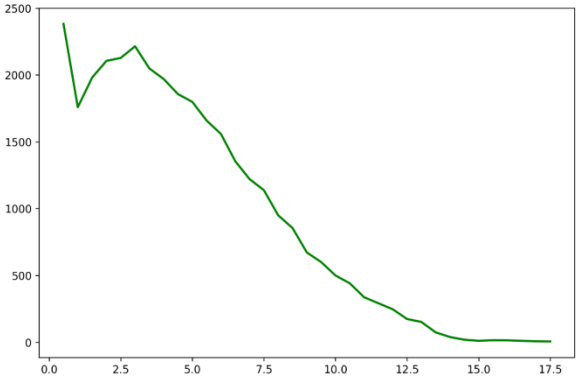


Figure 1. Distribution of numbers of thesaurus words according to total similarity in TBag-DBaf intersection

Among verbs, verbs of thinking, movement (*to drive – to fly*), informing (*to say – to inform – to warn – to assert*), value changing (*to decrease – to increase*), belonging to large semantic fields, have the highest similarity values (more than 13).

For example, according to the word2vec model, word *сказать* (*to say*) is most similar to such words as: *подчеркнуть* (*to stress*) 0.815, *заявить* (*to announce*) 0. 81, *добавить* (*to add*) 0.80, *заметить* (*to notice*) 0.79 .. And all these words are in TBag of this word in RuWordNet

On the other hand, the rise of the curve in low similarity values demonstrates the segment of problematic words.

## 5 Analyzing Discrepancies between Distributional and Thesaurus Similarities

We are interested in cases when the TBag-DBag intersection is absent or contains only 1 word with small word2vec similarity (less than the threshold (0.5)). We consider such a difference in the similarity bags as a problem, which should be explained.

For example, *троянец* (*troyanets*) is described in the thesaurus as a citizen of ancient Troya with the corresponding relations. But in the current texts, this word means a kind of malicious software (*troyan horse program*), this sense of the word was absent in the thesaurus. We can see that Dbag of word *троянец* contains:

*вредоносный* (*malicious)* 0.76, *программа* (*program*) 0.73, *троянский* (*trojan*) 0.71, *...вирус* (*virus*) 0.61,...

This means that the DBag and TBag are completely different, Dbag of word *троянец* does not contain anything related to computers and software.

We obtained 2343 such problematic "words". Table 1 shows the distribution of these words according to the part of speech.

It can be seen that verbs have a very low share in this group of problematic words. It can be explained that in Russian, most verbs have two aspect forms (Perfective and Imperfective) and also frequently have sense-related reflexive verbs. All these verb variants (perfective, imperfective, reflexive) are presented as different entries in RuWordNet.

Therefore, in most cases altogether they should easily overcome the established threshold of discrepancies. In the same time, if some verbs are

found in the list of problematic words, they have real problems of their description in the thesaurus.

| Part of speech | Number |
|---|---|
| Nouns | 1240 |
| Adjectives | 877 |
| Verbs | 226 |
| Total | 2343 |

Table 1. Distribution of parts of speech among problematic words

To classify the causes of discrepancies, we ordered the list of problematic words in decreasing similarity of their first most similar word from the thesaurus, that is in the beginning words with the most discrepancies are gathered (further, Problem List). In the subsections, we consider specific reasons, which can explain discrepancies between thesaurus and corpus-based similarities.

### 5.1 Morphological Ambiguity and Misprints

The most evident source of the discrepancies is morphological ambiguity when two different words $w_1$ and $w_2$ have the same wordform and words from DBag of $w_1$ in fact are semantically related to $w_2$ (usually $w_2$ has larger frequency). For example, in Russian there are two words *bank* (financial organization) and *banka* (a kind of container). All similar words from Dbag to *banka* are from the financial domain: *gosbank* (*state bank*), *sberbank* (*saving bank*), *bankir* (*banker*), etc. The analyzed list of problematic words includes about 90 such words.

The technical reason of some discrepancies are frequent misprints. For example, frequent Russian word *заявить* (*zayavit* – to proclaim) is often erroneously written as *завить* (*zavit* – to curl). Therefore the DBag of word *zavit* includes many words similar to *zayavit* such as *сообщить* (*to inform*), or *otmetit* (*to remark*). Another example are words *statistka (showgirl)* and *statistika (statistics)*.

### 5.2 Named Entities and Multiword Expressions

The natural reason of discrepancies are named entities, which names coincide with ordinary words, they are not described in the thesaurus, and are frequent in the corpus under analysis. For example, *мистраль* (*mistral*) is described in RuWordNet as a specific wind, but in the current corpus French helicopter carrier Mistral is actively discussed.

Frequent examples of such named entities are names of football, hockey and other teams popular in Russia coinciding with ordinary Russian words or geographical names (*Zenith*, *Dynamo*, etc.). Some teams can have nicknames, which are written with lowercase letters in Russian and cannot be revealed as named entities, for example Russian word *ириска* (*iriska*) means a kind of candy. In the same time, it is nickname of Everton Football Club (*The Toffees*).

| Word | The most frequent phrase | Phrase Freq. (Total freq. | Most similar word according to the corpus with frequency |
|---|---|---|---|
| Топленый (adj) (toplenyi – rendered) | Топленое масло (toplenoe maslo - rendered butter) | 78 (112) | Миндальный (adj) (mindalnyi – adjective from миндаль (almond)) 180<br>Mindalnoe maslo ( almond oil) 57 |
| Размочить (verb) (razmochit' – to open (the score)) | Размочить счет (razmochit' schet – to open the score) | 183 (336) | Сравнять (verb) (sravnyat' – equalize) 6678<br>Сравнять счет (to equalize the score) 5294 |
| Капитальный (adj) (kapitalnyi – capital) | Капитальный ремонт (kapitalnii remont – major repair) | 12015 (17985) | Капремонт (noun) (kapremont – abbreviation from kapitalnii remont – major repair) 3504 |
| Заварной (adj) (zavarnoi – boiled) | Заварной крем (zavarnoi krem – custard) | 37 (126) | Тыквенный (adj) (tykvennyi – adjective from тыква (pumpkin) 175<br>Тыквенные семечки (pumpkin seeds) 15 |
| Порывистый (adj) (poryvistii )- | Порывистый ветер (poryvistii veter – rough wind) | 1176 (1512) | Метель (noun) (metel' – blizzard) 7479 |

Table. 3 Impact of multiword expressions on discrepancies between the thesaurus and corpus-based data

Some discrepancies can be based on frequent multiword expressions, which can be present or absent in the thesaurus. A component $w_1$ of multiword expression $w_2$ can be distributionally similar to other words frequently met with $w_2$ or it can be similar to words related to the whole phrase $w_1$ $w_2$.

It can be noted that if a word $w_1$ occurs in a phrase $w_1w_2$ more than half times (the order of components can be different), it can become distributionally similar to $w_2$ or $w_3$, which also often met in phrase $w_3w_2$, *even if* $w_1$ and $w_3$ are not similar in sense. Table 3 shows examples of similarity discrepancies, which seems to be explained with frequent co-occurrence in a specific phrase.

For example, word *топленый* (*toplenyi – rendered*) occurs in the phrase топленое масло (*toplenoe maslo* – rendered butter) 78 times of 112 of its total frequency. Because of this, this word is the most similar to word *миндальный* (*mindalnyi* – adjective to almond), which is met in the phrase *миндальное масло* (*mindalnoe maslo* – almond oil) 57 of 180 times. But two words *топленый* и *миндальный* cannot be considered as sense-related words.

### 5.3 Thesaurus Relations

In some cases, the idea of distributional similarity is clear, but the revision cannot be made the thesaurus. We found two types of such cases. First, such epithet as *гигант* (*giant*) in the current corpus is applied mainly to large companies (*IT-giant, cosmetics giant, technological giant*, etc.). But it can be strange to provide the relations between words *giant* and *company* in a thesaurus.

The second case can be seen on the similarity row to word *массажистка* (*women massager*), comprising such words as hairdresser, housekeeper, etc. This is a kind of specialists in specific personal services but it seems that an appropriate word does not exist in Russian to create a more detailed classification of such specialists.

Another interesting example of a similarity grouping is the group of "flaws in the appearance": word *целлюлит* (*cellulite*)[2] is most similar to words: *морщина* (*crease of the skin*), *перхоть* (*dandruff*), *кариес* (*dental caries*), *облысение* (*balding*), *веснушки* (*freckles*). It can be noted that a bald head or freckles are not necessary flaws of a specific person, but on average they are considered as flaws. On the other hand, such phrases as *недостатки внешности, недостатки внешнего вида* (*flaws in the appearance*) are quite frequent in Internet pages according to global search engines, therefore maybe it could be useful to introduce the corresponding concept for correct describing the conceptual system of the modern personality.

But also real problems of thesaurus descriptions were found. They included word relations, which could be presented more accurate. For example, word *тамада* (*tamada – toastmaster*) was linked to more general word, not to *ведущий* (*veduschii – master of ceremonies*).

### 5.4 Senses Unattested in Thesaurus

Also significant missed senses including serious errors for verbs were found. As it was mentioned before, in Russian there are groups of related verbs: perfective, imperfective, and reflexive. These verbs usually have a set of related senses, and also can have their own separate senses. In the comparison of discrepancies between TBag and Dbag of verbs, it was found that at least for 25 verbs some of senses were unattested in the current version of the thesaurus, which can be considered as evident mistakes. For example, the imperfective sense of verb *otpravlyatsya* (*depart*) was not presented in the thesaurus.

Several dozens of novel senses, which are the most frequent senses in the current collection, were identified. Most such senses are jargon (sports or journalism) senses, i.e. *дерби* (derby as a game between main regional teams) or *навес* as a type of a pass in football (*high-cross pass*). Also several novel senses that belong to information technologies were detected: *прошивка* (*proshivka – firmware*), *соцсеть* (abbreviation from *социальная сеть* (*social network*).

The modern news discourse allows using words and expressions of the colloquial register (Patrona, 2011; Busa, 2013). In our analysis, several colloquial (but well-known) word senses absent in RuWordNet were found. For example, verb *обжечься* (*obzech'sya*) in the main sense means 'burn oneself'. In Dbag the colloquial sense 'make a mistake' is clearly seen.

For word *корректор* (*corrector*), two most frequent unattested senses were found: cosmetic corrector and correction fluid. The Dbag of this word looks as a mixture of cosmetics and stationary terms: *гуашь* (*gouache*), *кисточка* (*tassel*),

[2] https://en.wikipedia.org/wiki/Cellulite

*тональный* (*tonal*), *чернила* (*ink*), *типографский* (*typographic*), etc.

Currently, about 90 evident missed senses (different from named entities), which are most frequent senses of words in the collection, are identified from the analysis of the differences in two similarity lists .

### 5.5 Other cases

In some cases, paths longer than 3 should be used to provide better correspondence between thesaurus-based and corpus-based similar words.

Besides, the collected news corpus contains some number of Ukrainian texts, which are also written in the Cyrillic alphabet. Some Russian words coincide with Ukrainian words but have different senses and contexts in texts. Therefore, distributional similarities of such words are very different from the Russian thesaurus similarities.

## 6 Searching for regularities in Dbags

We supposed that we can group words in the corpus-based set of similar words (DBag) of problematic words using synonyms and part-of-speech synonyms of RuWordNet.

In such a way we can find more clear indications to some missed relations or novel senses. We have gathered synonyms, summed up their similarity scores to the target word, and again reordered list according to the descending order of the maximum similarity in DBag. For example, we obtained for word *рассекать (to cut* in the thesaurus sense) the maximum similarity score 3.58 with the following group of words: *мчаться, промчаться, пронестись, нестись, носиться* (*rush, race, hasten*). And this is the clear indication of the novel sense of this word absent in the thesaurus.

At the same time we obtained for word *длинноногий* (*long-legged*) the following most similar group *белокурый светловолосый блондинистый* (*blond, blonde, light-haired*). There is no semantic similarity between words *длинноногий* (*long-legged*) and *светловолосый* (*light-haired*) but there frequent co-occurrence and occurrence with the same nouns (*девушка, красавица, красотка - girl, beauty*) generate such similarity values.

It is also evident, that word *кроссворд* (*crossword*) is distributionally similar to group *разгадывание, разгадывать, отгадывание* (*guess, guessing, solve*) (score 1.51) only because of their frequent co-occurrence.

From this experiment, we can conclude that trying to extract some novel senses or missed relations on the basis of corpus-based embeddings, it is important to account for the diversity of contexts and co-occurrence of words predicted to be related. Low diversity of frequent contexts and significant co-occurrence can lead to erroneous conclusion on word semantic similarity.

## 7 Conclusion

In this paper we discuss the usefulness of applying a checking procedure to existing thesauri. The procedure is based on the analysis of discrepancies between corpus-based and thesaurus-based word similarities. We applied the procedure to more than 30 thousand words of Russian wordnet RuWordNet, classified sources of differences between word similarities and found several dozens of serious errors in word sense description including too general relations, missed relations or unattested main senses of ambiguous words. It is impossible to find such diverse problems in short time without automatic support.

We highly recommend to use this procedure for checking wordnets – it is possible to find a lot of unexpected knowledge about the language and the thesaurus.

In future, we plan to develop an automatic procedure of finding thesaurus regularities in DBag of problematic words, which can make more evident what kind of relations or senses are missed in the thesaurus.

## Acknowledgments

The reported study was funded by RFBR according to the research project N 18-00-01226 (18-00-01240).

## References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations* Association for Computational Linguistics: 7-12.

Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. Cross-checking WordNet and SUMO using meronymy. 2018. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).*

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In

*Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Edinburgh, Scotland, pages 1–11.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages. 1352-1362.

Francis Bond and Shan Wang. 2014. Issues in building English-Chinese parallel corpora with WordNets. In *Proceedings of the Seventh Global Wordnet Conference:* 391-399.

M.Grazia Busa. Introducing the language of the news: a student's guide. – Routledge, 2013.

Paul Cook and Graeme Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*.

Cristea, D., Mihaila, C., Forascu, C., Trandabat, D., Husarciuc, M., Haja, G., & Postolache, O. (2004). Mapping princeton WordNet synsets onto Romanian WordNet synsets. *Romanian Journal of Information Science and Technology*, *7*(1-2), 125-145.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Lea Frermann and Mirella Lapata. 2016. Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistic*s. V. 4. pages 31-45.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics: 67-71.

Nicola Guarino, and Christopher A. Welty. 2004. An overview of OntoClean. *Handbook on ontologies*. Springer, Berlin, Heidelberg: 151-171.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers):* 259-270.

Anders Johannsen, and Bolette Sandford Pedersen. "Andre ord"–a wordnet browser for the Danish wordnet, DanNet. *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*. 2011.

Natalia Loukachevitch and Boris Dobrov. 2002. Development and Use of Thesaurus of Russian Language RuThes. In *Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation*.(LREC 2002): 65-70.

Natalia Loukachevitch, German Lashevich and Boris Dobrov, Boris. 2018. Comparing Two Thesaurus Representations for Russian. In *Proceedings of Global WordNet Conference GWC-2018*, pages 35-44.

Natalia Loukachevitch and Ekaterina Parkhomenko. 2018. Recognition of Multiword Expressions Using Word Embeddings." *Russian Conference on Artificial Intelligence*. Springer, Cham, pages 112-124.

Natalia Loukachevitch. 2019. Corpus-based Check-up for Thesaurus. In Proceedings of ACL-2019: 5773-5779.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, *21*(5), 773-798.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR).* V. 41, №. 2, pages 10.

Roberto Navigli and Simone Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pages 217-250.

Alexander Panchenko, Anastasiya Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. 2018. RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In *Proceedings of Intern. conference Dialogue-2018*, pages 547--564.

Marianna Patrona. 2011. When journalists set new rules in political news discourse. Talking politics in broadcast media: *Cross-cultural perspectives on political interviewing, journalism and accountability*, 42, 157.

Maciej Piasecki, Michal Marcińczuk, Radoslaw and Marek Maziarz. 2013. WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*, *5*(3): 210-232.

Adam Rambousek, Ales Horak, and Karel Pala. 2018. Sustainable long-term WordNet development and maintenance: Case study of the Czech WordNet. *Cognitive Studies*, 18.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz 2012. A strategy of Mapping Polish WordNet onto Princeton WordNet. In Proceedings of COLING 2012, pages 1039-1048.

Rion Snow, Daniel Jurafsky, and Andrew Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics: 801-808.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. *In Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.