2528-9705

Örgütsel Davranış Araştırmaları Dergisi
Journal Of Organizational Behavior Research
Cilt / Vol.: 4, Sayı / Is.: 1, Yıl/Year: 2019, Sayfa/Pages:162-170

A CORPUS BASED STUDY OF DEVELOPING LEXICOSEMANTIC RELATIONS AMONG THE VERBS OF PUNJABI SHAHMUKHI

Muhammad Ahmad HASHMI[1*], Muhammad Asim MAHMOOD[2], Muhammad Ilyas MAHMOOD[3]

[1] SESE English, School Education Department, Okara University, Punjab, Pakistan
[2] Chairman, Department of Applied Linguistics, Government College University, Faislabad, Punjab, Pakistan.
[3] Department of English, University of Okara, Punjab, Pakistan/Faculty of Education, Universiti Kebangsaan Malaysia.

***Corresponding Author:**
Email: muhammadahmadhashmi@gmail.com

ABSTRACT

The present study is an attempt in the development of semantic relations among the Punjabi Shahmukhi verbs. Semantic relations are taken as a key part of the WordNet. The WordNet of Punjabi Shahmukhi is not developed yet. There is no reliable and a valid online source existing to learn and understand the Punjabi Shahmukhi. WordNet is thought to have a bit complex classification, where words are classified on the basis of the word meanings rather than word forms (Miller (1998). There are four open class categories in Princeton WordNet of English including; nouns, verbs, adverbs and adjectives but the current study is limited to develop semantic nets only among the verbs. For this purpose, corpus of 2 million words of Punjabi Shahmukhi was collected from the different sources. The corpus was then tagged with Parts of Speech (POS) tagger and a generated list of verbs was analyzed manually by looking at each verb individually. Lexicosemantic relations of synonymy, antonymy, entailment and troponymy were found for the Punjabi Shahmukhi verbs. The present research is significant in the development of WordNet for Punjabi Shahmukhi for its reliable representation in the field of language and technology. Furthermore, WordNet will help to maintain an international status for Punjabi Shahmukhi.

Keywords: *Punjabi Shahmukhi, Gurmukhi Punjabi, Artificial Intelligence, Natural Language Processing, WordNet*

INTRODUCTION

The purpose of the present study is to develop semantic nets among the verbs of Punjabi Shahmukhi by following the lexicosemantic relations of English WordNet, that was developed by Miller (1990) and others at Princeton University. It is taken as a one of the most important event in the year of 1985, when linguists and psychologists developed WordNet of English. In the recent years there has been an increasing interest in the digital development of languages. The current study is an effort to promote Punjabi language by developing its WordNet in the near future. Punjabi is a rich language which is being spoken by a large majority in India and Pakistan. In Pakistan, Punjabi is considered as a major language, which is spoken by 44.15% of the population (Census of Pakistan, 2001). The status of Punjabi language is higher than Urdu, which is even the national language of Pakistan (Khokhlova, 2014). Unfortunately, Punjabi is being dealt with negligence and as a result, it is digitally very young language. As world is changing and with each day, a significant development is being recorded in the field of Natural Language Processing (NLP). Human being has made machines much intelligent to process natural languages. Artificial Intelligence has changed everything. Now machines can

Geliş tarihi/Recieved: 28.8.2018 – Kabul tarihi/Accepted: 11.03.2019 – Yayın tarihi/Published: 30.03.2019

take decisions by understanding human language. This notion of Artificial Intelligence was revolutionized with the psychological investigations of linguists. A number of WordNet have been developed in different languages and are kept in a central database (Vossen, 1998). These are linked with Princeton WordNet, while each of these WordNet shows the uniqueness of its language and its internal systems. Svensen (2009) identifies the key relations which are found among the lexical units while structuring the WordNet. According to lexicographical point of view, these lexical relations make the WordNet a unique electronic dictionary. As structuralists mention about a unique system of vocabulary in the dictionary of any language (McCarthy, 2003). Previous studies have reported that WordNet clearly distinguishes between different syntactic categories and organize them semantically as nouns are set as topical hierarchies in developing lexical memory, while adjectives are organized on the basis of antonym relations and are kept as N-dimensional hyperspaces (Gross, Fischer & Miller, 1989; Miller, 1998). Synset of verbs are organized on the basis of entailment relations and these are located towards the bottom of the trees (Vider & Orav, 2002). According to Miller (1998): WordNet is thought to have a bit complex classification, where words are classified on the basis of word meanings rather than word forms. Actually, it shows sematic relations not the semantic components. These relations are constructed within the nouns, adjectives, verbs and adverbs (Miller, Leacock, Tengi, & Bunker,1993). This development has changed the mind of human in order to utilize Artificial Intelligence in the field of NLP. Now machines can translate one language into another one. Data can be mined and speech can be recognized digitally. Speech can be converted into the text and vice versa. But a number of languages including Punjabi Shahmukhi have been ignored and the time is not far, when these languages would become the part of history. As a linguist, it is my responsibility to save and promote languages. So, the current study is significant, as it is an initiative to digitize the Punjabi Shahmukhi by developing its WordNet in a near future. It will help to save Punjabi literature digitally. WordNet will also contribute in the development of different tools to maintain the status of Punjabi Shahmukhi.

Research Question

What are the semantic relations found among the verbs of Punjabi Shahmukhi language?

Limitations

The present study is limited to the verbs of collected corpus. It is beyond the scope of this study in order to examine other forms of words, so the present research does not deal with the semantic relations of other syntactic categories.

Significance of Study

WordNet is a modern digital form of any thesaurus or dictionary, which is quiet identical to the psychological representation of a language in human mind. Development of lexicosemantic nets are taken as a core in the development of WordNet. So, this study is very important in this regard. Furthermore, this study will show its significance to its native speaker in learning Punjabi Shahmukhi language systematically. The present research is also significant in terms of improving the standards of Punjabi Shahmukhi digitally. It will enhance the confidence of its native speakers to recognize their language in the field of language and technology.

METHODOLOGY

Data Collection

In this research, a corpus of Shahmukhi Punjabi was collected from various resources including: Newspapers, News items, Novels, Published Books, Poetry, Short stories and Articles. Corpus was taken in soft form with UTF-8 encoding in Notepad file. The next step was to tag the corpus. Due to the unavailability of Shahmukhi Punjabi tagger the corpus was translitrated into Gurmukhi Punjabi using the source developed by a South Asian organization of NLP and another software named as Akhar. After it, the corpus was tagged using Stanford Log-linear Parts of Speech (POS) tagger. The corpus was again translitrated into Shahmukhi script of Punjabi. In the next step, the list verbs was sorted using antconc software, which was freely available online http://www.laurenceanthony.net/software/antconc/. Finally, the tags were checked manually to remove machine error.

Data Analysis

After the collection of data, it was analyzed manually, looking each verb in online available sources of Shahmukhi Punjabi. The verbs were analyzed by exploring the possible sematic relations including: synonymy, antonymy, entailment and troponymy. The online sources were used to investigate each verb. An online Punjabi dictionary https://www.ijunoon.com/punjabi_dic/ which takes input in English and gives output in Punjabi was used. The second source, which has been used was https://pnb.wiktionary.org/wiki/. It is *Punjabi wiktionary* which is also known as Punjabi Wikipedia, consists of more than 9000 Punjabi words. Another online source, which has been used for data analysis was https://pnb.wikipedia.org/wiki/ . It is another Punjabi Wikipedia پنجابی وکیپیڈیا consists of 46546 articles. To get the correct ontology of a word the online dictionary https://ur.oxforddictionaries.com/ has been used known as Urdu Oxford Living Dictionaries. Semantic relations were developed by analyzing each verb individually and then verified by a Punjabi Expert from 'The Department of Punjabi' in Government College University Faisalabad.

A reference scale has been developed to verify the existence of every verb in the Punjabi Shahmukhi language using above mentioned online sources. These sources do not explain any particular lexical and semantic relations but only show the existence of particular target word in Punjabi Shahmukhi. The particular representation of above sources in analyzing Punjabi Shahmukhi nouns and verbs is given below.

1= https://pnb.wikipedia.org/wiki/
2= https://pnb.wiktionary.org/wiki/
3= https://www.ijunoon.com/punjabi_dic/
4= P. E: Punjabi Expert

After the analysis of verbs, the data has been checked by a Punjabi Expert from Government College University, Faisalabad, who checked all the verbs, their respective semantic relations and made the necessary amendments. All those amendments have been shown with a tag P.E in the reference column.

Many verbs of Punjabi Shahmukhi have more than one senses, as other natural languages have. To remove this ambiguity 'diacritical marks' have been used to signify the particular target sense. As Punjabi Shahmukhi is written in Urdu script, so an online Urdu dictionary has

been used named as Urdu Oxford Living Dictionaries with its webpage link as https://ur.oxforddictionaries.com/ to understand the particular sense with proper diacritical marks.

RESULTS AND DISCUSSIONS

Interpretations of Lecxico-semantic Relations of Verbs

WordNet exhibits four semantic relations for verbs, including the relations of synonymy, antonymy, entailment and troponymy. The Lecxico-semantic relations which are found among the 275 verbs are as follows.

Table 1:Comapritive Occurance of Lecxico-semantic Relations

Type of Lecxico-semantic Relation	Frequency	% of occurrence
Antonymy	200	72.73
Synonymy	270	98.19
Entailment	153	55.63
Troponymy	165	60.00

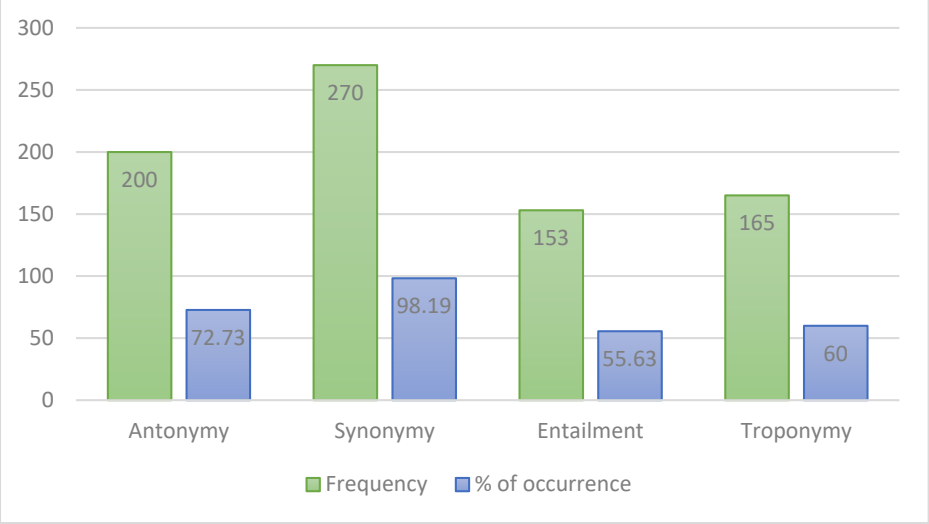

Figure 1: Semantic Relations of Shahmukhi Punjabi Verbs

Relation Synonymy

Two or more verbs are synonyms of one another, if one verb is changed with other verb as a result the changed verb does not change the meaning of context (Piasecki, 2009). Synonymous relations are found for 272 verbs out of 275 verbs as above in the table. This makes the percentage of 98.90% of the verb corpus. Only 3 verbs are left to show any synonym relation including ہونا, رنگنا and اُڈنا. The analysis of each of these 3 verbs highlights that ہونا is a very basic form of all the processes so, there might be no need for speakers to replace it. The word رنگنا means *dye krna* and it is popular now in Punjabi speakers even *dye* is an English verb. The 3rd word is اُڈنا which means *fly* as birds do. Sources of Shahmukhi Punjabi did not have the replacement of this process as happened in the case of رنگنا.

Relation of Antonymy

Relation of antonymy plays an important role in defining the sense and nature of the target word and a lexicographer takes it as a major lexical relation while designing and planning

dictionaries (Strapparava & Valitutti, 2004). 190 words of the total 275 verbs have shown the relation of antonymy which make the 69.09% of targeted corpus of verbs. It also means that 85 verbs do not have their antonym relations, for example, وجنا, رُلنا, چُگنا, مُڑنا, اپڑنا, رڑکنا, لٹنا, لٹکنا, مُکنا, سوچنا, بلانا, ورتنا, سڑنا and so on. The analysis of these verbs shows that many of the verbs like رڑکنا, مُکنا, سوچنا, سہنا, سڑنا, ڈرنا, بُجھنا, پھٹنا, تپنا, ترسنا, شرمانا, گلنا …etc. are automatic processes of natural happening, which happen without any prior attention so, they might not show antonym relation. There can be another reason of not the suitability of antonym relations in defining verbs because this relation better works in organizing adjectives and adverbs in WordNet (Miller, Beckwith, Fellbaum, Gross & Miller, 1990). On the other hand, verbs like (وسنا, وکانا,نکلنا, کرنا, دینا , رہنا, ونڈنا, آنا, کھانا …etc.) have their opposites as (اجڑنا, مُل لینا, وڑنا, نا کرنا, لینا, ونجنا, دبنا, جانا and بگنا … etc.) respectively. These verbs are intentionally doing processes and show their opposites.

There are also some processes which are inborn in humans and animals and they rarely have their opposition relations, for example, ٹھگنا, ڈسنا, شرمانا, سُونا, تکنا, سجنا, تاڑنا, چبنا, چرنا, مسکرانا. In contrast many inborn processes also exhibit opposite relations as بگنا/کھانا , جینا/مرنا , رونا/ہسنا , سونا/جاگنا . It means these are selective processes and there are no hard and fast rules to categorize these processes but rules can be developed by analyzing these processes deeply.

Now come to the type of opposition these verbs exhibit. Here, in verbs the prominent kind of opposition takes place including: gradable opposites, complementary opposites, directional opposites, reversive opposites and near-opposites (Miller, 1998).

A table and bar graph has been made to show the comparative occurrence of these opposites in the present corpus of verbs. Here, we will discuss these kinds in details in relation with the verbs of present study.

Table 2: Comparitive Occurance of Relation of Opposition

Kind of Opposition	Frequency	% of occurrence
Gradable Opposites	27	13.50
Complementary Opposites	88	44.00
Reversive Opposites	75	37.50
Near-Opposites	10	5.00

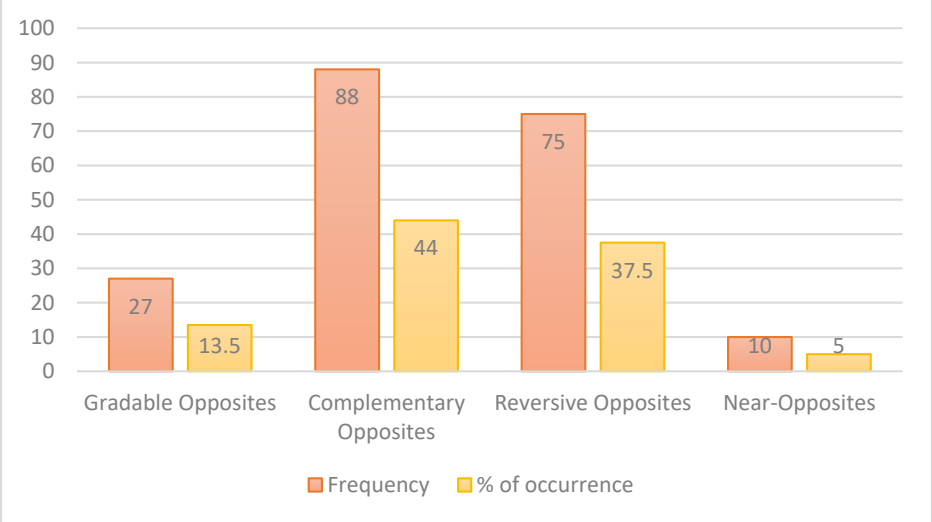

Figure 2: Graphical Representation of Kinds of Opposites

Gradable Opposites

The kind of verbs whose intensity can be defined in grades are called gradable opposites (Gross & Miller, 1990). From the total of 190 verbs 26 verbs are found gradable including دبنا/ونڈنا , تھاپڑنا/ڈانٹنا , سُکنا/ودنا , سُکنا/پھیلنا , مننا/لڑنا , ہسنا/رونا , ڈولنا/بھرنا , رونا/ہسنا , ودنا/سُکنا , بگنا/کھانا ...etc. These processes can be defined in terms of more and less, for example بہتا ہنسنا, تھوڑا لڑنا, بہتا ڈانٹنا and so on.

Complementary Opposites

These type of opposites cannot be described in "more or less". The comparative existence of complementary opposites in the corpus is 42.63%. 81 verbs out of 190 verbs have shown complementary opposition. These have only two extremes or end points. There exists no middle form exists for this type of opposition (Maziarz, Piasecki, Szpakowicz & Rabiega-Wiśniewska, 2011), for example: اجڑنا/وسنا , نا ہون/ہونا , ونجنا/رہنا , ٹُٹنا/لگنا , ڈُبنا/ترنا , جینا/مرنا , اُترنا/چڑھنا .

Directional Opposites

According to Lyons (1977) directional opposites are related to adverbs and prepositions. These are rarely associated to verbs and not related to nouns. For example, اندر/باہر, اوپر/نیچے and so on. In case of present study only one verb is found having directional opposite پیجنا/بُلانا.

Reversive Opposites

According to Egan (1968) reversive opposites consist of adjectives and adverbs, which sometimes qualify nouns or a particular state of an act. This class also contains a number of different verbs, for examples کھولنا/بند کرنا, نکاح کرنا/طلاق دینا, آنا/جانا, ظاہرہونا/غائب ہونا and so on. The analysis of 190 verbs shows the occurrence of 73 reversive verbs which make the 38.42% of the total opposites. For example: رُکنا/چلنا, کڈھنا/پانا, کُھرنا/جمنا, پُچھنا/دسنا, جانا/آنا, وڑنا/نکلنا, چڑھنا/اُترنا, چٹنا/ٹُھکنا, جمنا/کُھرنا, چُکنا/رکھنا and so on. The opposites of these verbs are reverse of their original processes.

Near –Opposites

Near-opposites are not real opposites. According to Cruse (1986) these are impure opposites. Analysis of antonyms have shown that the presence of 9 near-opposite verbs out of 190 verbs showing their antonyms which make 4.73% of the total found antonyms e.g. لِکھنا/پڑھنا, کھانا/پینا, پُچھنا/دسنا, سُننا/بولنا, اٹکنا/پجنا and so on.

Relation of Entailment

Entailment is a kind of semantic relation which is rare in common dictionaries but a significant part of WordNet (Miller, Beckwith, Fellbaum, Gross & Miller, 1990). It is an association that is found in defining the verbs. Here, in the given corpus 153 verbs include other processes or the other found processes include these target verbs. In the entailment relation one process temporally includes another process. It can be a bi-directional or unidirectional relation. The verb which include other process is called *entailing* verb while the included verb is called *entailed* verb (Chklovski & Pantel, 2004). As in case of رہنا/وسنا , چبنا/کھانا , مرنا , سا رُکنا/پینا ...etc. In these examples verbs وسنا,کھانا , مرنا,پینا and پڑھنا are including other verbs رہنا, چبنا,سا رُکنا , چکھنا and سمجھنا respectively. In this way the including verbs are called entailing verbs and the included verbs are known as entailed verbs. Moreover, I have drawn a table and represented a bar-graph to show the comparative occurrence of entailing and entailed verbs as follows.

Table 3: Occurance of Entailment Relation

Kind of Entailment	Frequency	% of occurrence
Entailing Verbs	139	90.84
Entailed Verbs	14	9.16

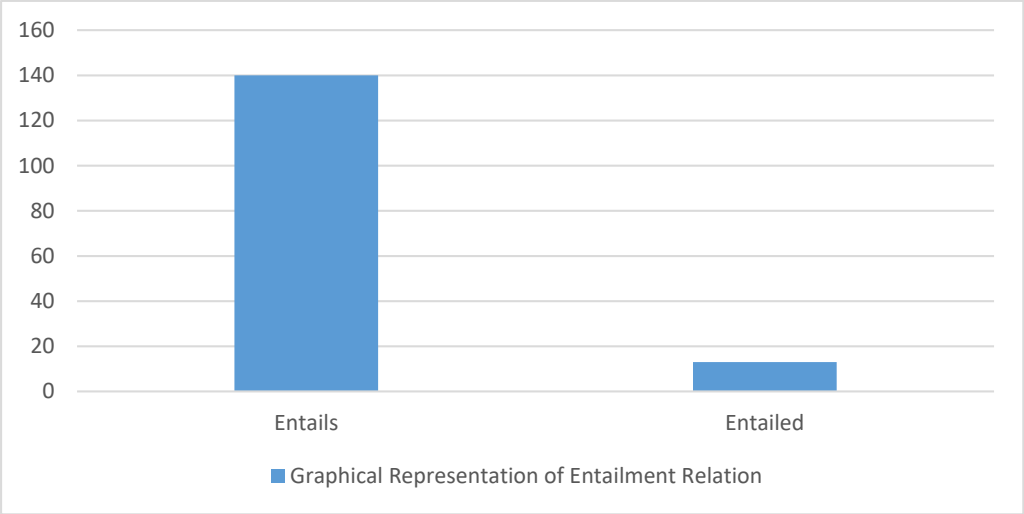

Figure 3: Graphical Representation of Entailment Relation

The above graph has been drawn by analyzing the 153 verbs which have the relation of entailment. Bar-graph shows that 90.84% verbs are entailing the other process or can say that including the other processes. The major reason behind the high frequency of entailing verbs is the focus of researcher to analyze the entailed verbs by the target verbs in the corpus. Furthermore, where entailing verbs are not found the entailed verbs are focused by the researcher.

Troponymy

Troponymy is a kind of relation between two processes, where the process remains same but the manner of happening becomes different (Miller, 1995). For example, in cry and speak, the process is same but the manner is different. For the current corpus of 275 verbs the 164 verbs have their dual processes in different manners, for example: بجھنا/کھوجنا, گرفتار کرنا/ڈکنا, مِننا/ناپنا, کُرلانا/تڑفنا, ہکنا/چلانا, بولنا/چِلّانا and so on.

The remaining 111 verbs of doing and happening did not show this relation. The study of these verbs shows that many of the verbs like ہونا,کرنا and سِکھنا are very basic that these are the most generalized form for many actions so, these do not have their specific manner to entertain.

Many of the verbs are basic skills like چڑھنا,اُڑنا, پڑھنا , ترنا, لِکھنا and so on, which have not their representation in any alternative manners. Further, some of the verbs like کُھرنا,مُکنا, اُگنا , ٹھرنا, جلنا,جمنا , پھسنا , جننا …etc. are automatic processes which happen naturally, so these also do not have alternative manners of happening.

CONCLUSION

The current study was set out to determine the lexicosemantic relations among verbs with a future aim to develop Punjabi Shahmukhi WordNet. As WordNet is thought to be a solid lexical database to run the digital applications of any language. Punjabi Shahmukhi is digitally

very young language so, it is very important to develop its WordNet. Development of Semantic nets of different word forms is one of the most important aspect of WordNet.

The purpose of current study was to develop lexicosemantic relations among Punjabi Shahmukhi verbs. It is concluded that the verbs of Punjabi Shahmukhi are frequent in showing the lexical relation of synonymy and antonymy. The semantic relation of entailment is found rare among the Punjabi Shahmukhi verbs.

Finally, these lexicosemantic relations will be the part of Punjabi Shahmukhi WordNet, which will be developed in near future. This WordNet will help to digitize all the literature of Punjabi Shahmukhi which is available only in hard form.

References

Chklovski, T., & Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Cruse, D. A., & Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Egan, R. 1968. Survey of the History of English Synonymy. In: Gove, P. (ed.) *Webster's New Dictionary of Synonyms*. Springfield, MA: Merriam-Webster. 5a–31a.

Gross, D., & Miller, K. J. (1990). Adjectives in wordnet. *International Journal of lexicography*, *3*(4), 265-277.

Gross, D., Fischer, U., & Miller, G. A. (1989). The organization of adjectival meanings. *Journal of memory and language*, *28*(1), 92.

Khokhlova, L. V. (2014). 2. Majority language death. University of Hawai'i Press.

Lyons, J. 1977. *Semantics*. 2 vols. New York: Cambridge University Press

Maziarz, M., Piasecki, M., Szpakowicz, S., & Rabiega-Wiśniewska, J. (2011). Semantic relations among nouns in Polish WordNet grounded in lexicographic and semantic tradition. *Cognitive Studies| Études cognitives*, (11).

McCarthy, D., Keller, B., & Carroll, J. (2003, July). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18* (pp. 73-80). Association for Computational Linguistics.

Miller, G. (1998). *WordNet: An electronic lexical database*. MIT press.

Miller, G. A. (1990). Nouns in WordNet: a lexical inheritance system. *International journal of Lexicography*, *3*(4), 245-264.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, *3*(4), 235-244.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993, March). A semantic concordance. In *Proceedings of the workshop on Human Language Technology* (pp. 303-308). Association for Computational Linguistics.

Piasecki, M., Broda, B., & Szpakowicz, S. (2009). *A wordnet from the ground up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Strapparava, C., & Valitutti, A. (2004, May). Wordnet affect: an affective extension of wordnet. In *Lrec* (Vol. 4, pp. 1083-1086).

Svensen, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Vider, K., & Orav, H. (2002). *Estonian wordnet and lexicography*. na.

Vossen, P. (1998). A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, *10*, 978-94.