Testing agreement between lexicographers: A case of homonymy and polysemy

Marek Maziarz,[◇] Francis Bond[♠] and Ewa Rudnicka[◇]
[♠] Nanyang Technological University, Singapore
[◇] Wrocław University of Science and Technology, Poland
bond@ieee.org, {marek.maziarz|ewa.rudnicka}@pwr.edu.pl

Abstract

In this paper we compare Oxford Lexico and Merriam Webster dictionaries with Princeton WordNet with respect to the description of semantic (dis)similarity between polysemous and homonymous senses that could be inferred from them. WordNet lacks any explicit description of polysemy or homonymy, but as a network of linked senses it may be used to compute semantic distances between word senses. To compare WordNet with the dictionaries, we transformed sample entry microstructures of the latter into graphs and cross-linked them with the equivalent senses of the former. We found that dictionaries are in high agreement with each other, if one considers polysemy and homonymy altogether, and in moderate concordance, if one focuses merely on polysemy descriptions. Measuring the shortest path lengths on WordNet gave results comparable to those on the dictionaries in predicting semantic dissimilarity between polysemous senses, but was less felicitous while recognising homonymy.

1 Introduction

We talk about *polysemy* when different word senses are semantically related. Homonymy is the opposite phenomenon in which etymologically unrelated senses are signified by the same word-form (Lyons, 1995, pp. 54-60).[1] The main source of homonyms is the diachronic process of word shortening due to their frequent use (Fenk-Oczlon and Fenk, 2008, p. 59). Though homonyms are frequent in text and speech, they remain a tough nut to crack for Natural Language Processing (Hauer and Kondrak, 2020; Klimenkov and Pokid, 2019; McCarthy, 2006; Mihalcea, 2003). One of the reasons is that wordnets lack any explicit links between the related meanings of the same word and do not discern between the two types of lexical ambiguity (Freihat et al., 2013).

The goals of this paper are two-fold: (i) we check the degree of agreement between polysemy descriptions in two general English dictionaries, namely Oxford Lexico and American English Merriam-Webster Dictionary, and in WordNet, (ii) we test the applicability of WordNet in measuring semantic similarity between senses (that is assessing polysemy vs. homonymy distinctions). For these purposes, we have created a data set of 57 nouns, noted by the three lexicons (Sec. 3.1). We represented dictionary microstructures as graphs with the equivalent WordNet synsets attached to them (Sec. 3.2). The approach resulted in 889 sense pairs in total. The set of the mapped synsets served as a common denominator for the subsequent comparisons between the three lexical resources. Measuring distances between particular sense pairs allowed us to compare the polysemy/homonymy description in the two dictionaries with the structural description in WordNet (via lexico-semantic relations, Sec. 3.3 and 4).[2] It turned out that the dictionaries are in high concordance with each other, if we consider the homonymy-polysemy distinction, and in moderate agreement, if we look at polysemy descriptions (in terms of Spearman's correlation coefficient). WordNet did not differ much from Lex-

[1] For the needs of this paper, we define *homonyms* (*homographs*) as a pair of senses which are characterised by the same part of speech, share the same lemma, but are not related semantically and etymologically (Svensén, 2009, pp. 96-7). A pair of polysemous senses (*polysemes*) – on the contrary – is constituted by the two senses of the same POS category, sharing the same lemma, semantically related and of the same etymology.

[2] The resource was published under the CC-BY 4.0 licence and is available from: https://github.com/MarekMaziarz/HomoPoly.

ico and Merriam-Webster in its capability to describe similarity between polysemous senses. It was homonymy that made the difference (Sec. 5).

2 Related Work

In Natural Language Processing accessing word or sense dissimilarity via measuring distances in lexical networks is a well-known procedure (Meng et al., 2013; Pedersen et al., 2004; Richardson et al., 1994). Among many measures, some are of special interest for semantic relatedness assessment, that is path-based indices (the shortest-path, Wu-Palmer's, Leakcock-Chodorow's or Li's measures) and information content-based measures (Resnik's, Lin's or Jiang's methods, see Meng et al. (2013)). In the context of recognising polysemous sense proximity, Wu-Palmer's measure was used to calculate concept similarity within the taxonomy of *The Historical Thesaurus of English* (Ramiro et al., 2018). Each sense was compared with all other word senses, then the obtained matrices of similarity were used to arrange polysemous word meanings into a chain of extended senses. In (Youn et al., 2016) polysemy networks for many world languages were compared with the use of path distances between Swadesh' concepts mapped to them. The authors found the distance distribution of polysemy structures universal across languages, despite clearly different geographical and cultural conditions. Out of various measures of semantic relatedness, we made use of one of the simplest – the shortest path length. Since our graphs were weighted, we utilised Dijkstra's distance algorithm (Dijkstra et al., 1959) which finds the geodesics for weighted networks. We applied it to measuring semantic distances in both English dictionaries and in WordNet.

Many traditional dictionaries depict word senses in the form of nested clusters of definitions. Starting from the basic sense (Atkins, 2008, p. 41), (Svensén, 2009, pp. 363-4), they unfold a network of inter-dependencies in the form of a sense hierarchy. In such structured polysemy nets main senses are linked into meaning chains with groups of subsenses attached to them (Svensén, 2009, p. 211-2, 350-1, 363). Hierarchical sense differentiation is more intuitive for dictionary users than a flat arrangement. In such a set-up senses are ordered according to their semantic "closeness" (Atkins, 2008, p. 41). Lexico and Merriam-Webster both represent this type of polysemy structuring.

The problem of consistency of lexicographic entries is widely acknowledged (Stock, 2008). The same word may be differently treated in different dictionaries (Svensén, 2009, pp. 205-6). Splitting or merging senses is not an easy task even for a specialist. The issue seems highly intuitive and decisions are supposed to be highly arbitrary. This is not entirely true. In distinguishing senses lexicographers rely on specific rules, like observing usage restrictions (e.g., for specialised vocabulary), differences in syntactic frames (cf. transitive - intransitive frame) or other grammatical properties, like grammatical number (cf. *pluralia* and *singularia tantum*) (Svensén, 2009; Jackson, 2002). Yet another way to tame lexicographers' intuitions is to rely on taxonomic and other sense relationships, in such a way *genus proximum* (a hypernym) and *differentia specifica* (a meronym/holonym, antonym etc.) might be captured (Stock, 2008, p. 153).

The lexicographic process of splitting and clustering senses was widely studied in the context of Word Sense Disambiguation (e.g. Passonneau et al. (2010)). We relate to several research studies which are most relevant to our approach. Resnik and Yarowsky (1999) proposed a method of measuring sense distances on *Hector* – a hierarchical dictionary (Atkins, 1992, cf. Tab. 3). They postulated that the penalty applied to a homonymous pair should be much higher than the cost of a polysemy step. In some aspects our methodology resembles this approach to the construction of an adjacency matrix.[3] Chugur et al. (2002) counterargued against the possibility of an honest measure based on hierarchical dictionaries. The argument is as follows: since in metaphorical shifts extended senses completely change their semantic domain, dictionary provided sense relations do not mirror mental lexicon sense proximities. To this plea we answer that polysemy topologies are often multicentred (Brugman and Lakoff, 2006) and are governed by their own rules (naming $\neq$ knowing, see (Malt et al., 1999)).

Véronis (1998) executed experiments in the assessment of the number of word senses obtained from a tagged corpus which was collated with dictionary data (*Petit Larousse*). The Spearman's

[3] We give homonymy links the distance of infinity, transformed later into the value of maximum distance of the whole polysemy network plus one. The crucial difference lies in the fact that we attach subsenses directly to the main sense, while Resnick and Yarowsky chained them. However, the idea to derive semantic dissimilarity measure out of the existing dictionaries and their hierachies remains the same.

rank correlation equal to 0.5 was reported for nouns. In various SENSEVAL editions research teams also reported rather mediocre agreement values between annotators (Artstein and Poesio, 2008, p. 587), e.g., Mihalcea et al. (2004) in SENSEVAL-3 observed ca. 70% ITA (percentage agreement) and $\kappa = 0.58$; similar results were obtained by Palmer et al. (2007). According to Artstein and Poesio (2008), "[w]ord sense tagging is one of the hardest annotation tasks."

3 Method

3.1 Lexico and Merriam-Webster Graphs

Two dictionaries were used to obtain distances between PWN senses: Oxford Lexico[4] and American English dictionary – Merriam-Webster[5].[6] 25 lemmas representing polysemy/homonymy distinction in English, according to these dictionaries, were chosen (the set S_{HP}), as well as 31 solely polysemous noun lemmas (the set S_P).[7] For each lemma two distinct graph structures were constructed out of the dictionaries, taking into account sense orderings. Both dictionaries apply similar lexicographic rules. Senses of different etymology are split into distinct entries. Then, main senses are ordered into a chain, according to their semantic closeness (cf. (Atkins, 2008, p. 41)), starting from the primal sense. Subsenses, if they exist, are attached to their superordinate meanings. The whole sense arrangement reflects semantic relationships, sense proximity and dissimilarity, being the result of the evolutionary sense extending process (as seen by each dictionary lexicographer team).

For instance, for the noun *sink* we found in Lexico the following microstructure[8]:

sink^2 noun

- 1. 'A fixed basin with a water supply and outflow pipe';
- 2. 'A pool or marsh in which a river's water disappears by evaporation or percolation;
- 2.1. technical 'A body or process which acts to absorb or remove energy or a particular component from a system';
- 3. short for *sinkhole*;
- 4. 'A place of vice or corruption';
- 4.1. British usually as modifier 'A school or estate situated in a socially deprived area'.

We transformed it into the set of bidirectional relations in such a manner that main meanings were linked into chains of consecutive senses ($1 \leftrightarrow 2 \leftrightarrow 3 \leftrightarrow 4$), and subsenses were joint to their superordinates ($2.1 \leftrightarrow 2$ and $4.1 \leftrightarrow 4$). Subsenses were dealt differently. In a polysemy graph they were given equal distances from their superordinate sense.[9]

3.2 Mapping PWN onto Dictionaries

PWN nominal senses representing the same lemma were mapped on the Lexico graph by two professional linguists in three steps (the set S_{HP}). (1) In the first phase, the mapping of the homonymous lemmas was done independently by the two annotators, then (2) disagreement cases were again independently annotated for the second time. (3) Finally, in the 3^{rd} phase the remaining discrepancies were resolved in discussion. Cohen's κ was not worse than 0.8 in the task. Figure 1 presents the growth of kappa from the stage (1) to (2). Having assumed high agreement between lexicographers, polysemous senses from the set S_P were mapped by one of the annotators.

Thus, PWN sense *sink*-n-2 ('*technology* a process that acts to absorb or remove energy or a substance from a system') was linked to the sense sink^2-n-2-1, while PWN *sink*-n-1 ('plumbing fixture consisting of a water basin fixed to a wall...') was mapped onto the Lexico sense sink^2-n-1 resulting in the following graph structure (0s and 1s in superscripts represent relation weights) and Dijkstra's distance of two steps between the WordNet senses.

[4] `https://www.lexico.com/`

[5] `https://www.merriam-webster.com/`

[6] The dictionary entries were manually copy-pasted from the sites and then transformed into relation triples using regular expressions.

[7] The full list of the chosen words is as follows: *angle*, *band*, *bank*, *bark*, *bat*, *board*, *can*, *chapter*, *chop*, *clip*, *concealment*, *crest*, *cylinder*, *date*, *degree*, *duck*, *fall*, *fame*, *file*, *fly*, *gloss*, *intellect*, *lump*, *master*, *match*, *palm*, *pasturage*, *plant*, *ring*, *rock*, *rose*, *saw*, *scale*, *score*, *sentence*, *shilling*, *sink*, *skimmer*, *spring*, *stage*, *stalk*, *table*, *term*, *tie*, *tongue*, *trepan*, *trip*, *tune*, *veneer*, *vermin*, *victim*, *voucher*, *well*, *whirl*, *wrapping* and *wreck*.

[8] `https://www.lexico.com/definition/sink`

[9] In large hierarchies chaining subsenses would lead to inadequate similarity measures. Consider a hypothetical microstructure $G = (V, E)$: $V = \{1, 2, 2.1, 2.2, 2.3, 3\}$, $E = \{1 \leftrightarrow 2, 2 \leftrightarrow 3, 2.1 \leftrightarrow 2, 2.2 \leftrightarrow 2.1, 2.3 \leftrightarrow 2.2\}$. Let us measure the distance between the sense 2 and its subsense 2.3, which is $dist(2.3, 2) = 3$ steps. On the other hand, main senses 3 and 2 are only $dist(3, 2) = 1$ step ahead of each other, which seems counter-intuitive.

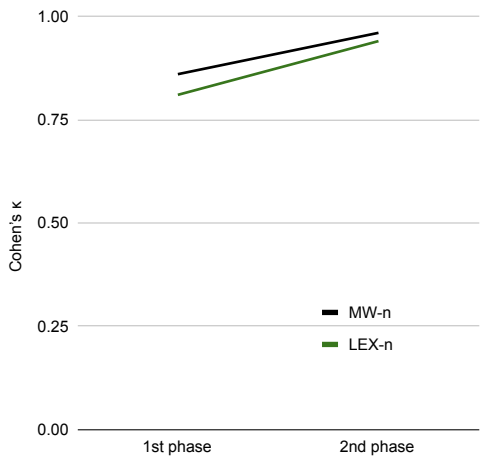

Figure 1: Cohen's κ measure of the agreement between two independent annotators for the WordNet-Lexico (LEX) and WordNet-Merriam-Webster (MW) nouns mappings, set S_{HP}.

- sink^2-n-1 $\overset{1}{\leftrightarrow}$ sink^2-n-2
- sink^2-n-2-1 $\overset{1}{\leftrightarrow}$ sink^2-n-2
- sink^2-n-2 $\overset{1}{\leftrightarrow}$ sink^2-n-3
- sink^2-n-3 $\overset{1}{\leftrightarrow}$ sink^2-n-4
- sink^2-n-4 $\overset{1}{\leftrightarrow}$ sink^2-n-4-1
- sink^2-n-1 $\overset{0}{\leftrightarrow}$ PWN-sink-n-1
- sink^2-n-2-1 $\overset{0}{\leftrightarrow}$ PWN-sink-n-2

The corresponding Merriam-Webster microstructure is the following[10]:

sink^2 noun

- 1a. 'a pool or pit for the deposit of waste or sewage: cesspool';
- 1b. 'a ditch or tunnel for carrying off sewage: sewer';
- 1c. 'a stationary basin connected with a drain and usually a water supply for washing and drainage';
- 2. 'a place where vice, corruption, or evil collects';
- 3. 'sump: the lowest part of a mine shaft into which water drains';
- 4a. 'a depression in the land surface especially : one having a saline lake with no outlet';
- 4b. 'sinkhole';
- 5. 'a body or process that acts as a storage device or disposal mechanism: such as';
- 5a. 'heat sink broadly : a device that collects or dissipates energy (such as radiation)';
- 5b 'a reactant with or absorber of a substance forests are a sink for carbon dioxide'.

From which we obtain the relational graph of polysemy instances:

- sink^2-n-2 $\overset{1}{\leftrightarrow}$ sink^2-n-1
- sink^2-n-3 $\overset{1}{\leftrightarrow}$ sink^2-n-2
- sink^2-n-4 $\overset{1}{\leftrightarrow}$ sink^2-n-3
- sink^2-n-5 $\overset{1}{\leftrightarrow}$ sink^2-n-4
- sink^2-n-1-a $\overset{1}{\leftrightarrow}$ sink^2-n-1
- sink^2-n-1-b $\overset{1}{\leftrightarrow}$ sink^2-n-1
- sink^2-n-1-c $\overset{1}{\leftrightarrow}$ sink^2-n-1
- sink^2-n-4-a $\overset{1}{\leftrightarrow}$ sink^2-n-4
- sink^2-n-4-b $\overset{1}{\leftrightarrow}$ sink^2-n-4
- sink^2-n-5-a $\overset{1}{\leftrightarrow}$ sink^2-n-5
- sink^2-n-5-b $\overset{1}{\leftrightarrow}$ sink^2-n-5
- sink^2-n-1-c $\overset{0}{\leftrightarrow}$ PWN-sink-n-1
- sink^2-n-5 $\overset{0}{\leftrightarrow}$ PWN-sink-n-2
- sink^2-n-5-a $\overset{0}{\leftrightarrow}$ PWN-sink-n-2
- sink^2-n-5-b $\overset{0}{\leftrightarrow}$ PWN-sink-n-2

This example shows that while in Lexico the relation between the senses 'plumbing fixture' and 'the absorption or removal of energy' is seen as more direct (through the Lexico sense sink^2-n-2 'a pool or marsh'), the corresponding path in Merriam-Webster is much longer due to more fine-grained sense distinctions and different conceptualisation of the sense extending path (via the senses: 5 'body/process' $\leftrightarrow$ 4 'depression/sinkhole' $\leftrightarrow$ 3 'sump' $\leftrightarrow$ 2 'place of evil' $\leftrightarrow$

[10] https://www.merriam-webster.com/dictionary/sink

1 'cesspool/ditch/basin' $\leftrightarrow$ 1-c 'drainage basin', 5 steps in total).

We assumed that senses $s_1 \in PWN$ and $s_2 \in Dict$ were to be considered equivalent iff their extensions had a non-empty and non-trivial intersection. Let $S_1 = \{x : s_1(x)\}$ and $S_2 = \{x : s_2(x)\}$ be the sets of denotata of concepts s_1 and s_2, respectively. They were mapped iff

$$S_1 \cap S_2 \notin \emptyset \Leftrightarrow$$
$$\exists x[(s_1(x) \implies s_2(x)) \wedge (s_2(x) \implies s_1(x))] \quad (1)$$

and the set of shared denotata $S_1 \cap S_2$ was intuitively not to small. The specificity of the task of linking dictionaries limited the space of choices only to different senses of the same word in WordNet (PWN) and in Lexico or Merriam-Webster ($Dict$), hence the requirement of non-triviality was easy to employ. Such an approach resulted in many-to-many mappings. An example of the process is shown in Fig. 2 (the noun *stalk*).

3.3 Semantic Distance

Having constructed semantic nets for both dictionaries and having mapped them onto PWN synsets we turned to measuring semantic distance between nodes in the graphs. For each PWN sense pair we calculated Dijkstra's distance. For 57 nominal lemmas we obtained, through combinatorics, 889 sense pairs and corresponding 889 distance values between the meanings. Homonymy groups constituted separate graphs, thus some possible paths were disjoint. Homonymy paths were given infinite lengths, while polysemy couplings obtained finite distance values. There were also cases of missed PWN meanings (a dictionary lacked any description of a given PWN sense). In such a situation we treated isolated (missed) sense exactly like homonymous ones. Table 1 jointly presents cardinalities of sets of finite ("<Inf") and infinite paths ("Inf"). As a result, we got 85% identical choices (the percentage agreement) and Cohen's $\kappa = 0.67$. Half of the remaining disagreement instances were missed senses (61 cases) and the other half were cases of real discrepancies in the homonymy/polysemy distinction (68 instances). Such a high agreement suggests that the dictionaries were pretty consistent in describing pairs of senses either as homonymous or polysemous.

Figure 3 represents a 2D histogram of actual Dijkstra's distances for the whole set of pairs.

		LEX	
		<Inf	Inf
MW	Inf	42	224
	<Inf	536	87

Table 1: Disjoint ("Inf") and finite ("<Inf") paths between PWN senses in Lexico ("LEX") and Merriam-Webster ("MW").

Homonymy couplings are posited in the top-right corner of the square. For the needs of correlation measurements, we transposed infinitives into finite values, i.e. $Inf \rightarrow max(dist) + 1$, which for Lexico was 8, and for Merriam-Webster was 9. Merriam-Webster has slightly longer sense chains than Lexico (because of deeper sense hierarchies). The concordance between Lexico and Webster was measured with Spearman's and Pearson's correlations ($\rho = .60$, $r = .60$).

Since investigating the coverage of a dictionary in terms of the noticed senses was not the aim of this research, we linked missed senses manually to the closest PWN senses (with weights equal to 1). This supplementary set of missed sense linkages was used in consecutive experiments as the shared extension of both dictionary graphs and Princeton WordNet. Having attached the set, we obtained the correlation of $\rho = .71$ and $r = .80$, see Table 2. The calculations show that our dictionaries give a similar semantic depiction of polysemy (lower distance values) and unrelated homonymous meanings (maximal distances).

Figure 4 illustrates the relationship between Lexico and Merriam-Webster after the removal of senses with infinite paths (homonymy cases). Now, correlations decrease to moderate values ($\rho = 0.43$ and $r = 0.41$). This proves that particular paths in each dictionary for the very same sense pair must differ (as we saw in the case of the PWN noun *sink*, senses 1 and 2).

4 Comparison with WordNet

Dictionaries are in high agreement when we consider both homonymy and polysemy, and in moderate concordance when we look solely at polysemy. The moderate correlations in polysemy depiction are not surprising. If one took into account the fact that our dictionaries might have differently clustered meanings, subsenses and meaning shades; might have distinguished more or less

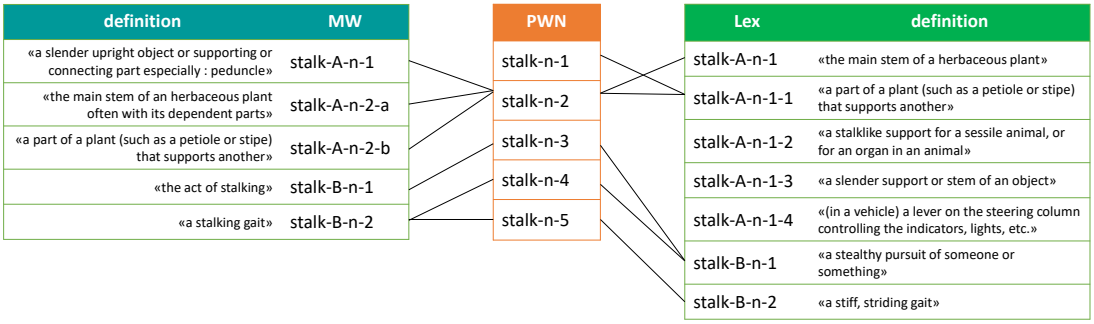

Figure 2: Mapping the equivalents of the noun *stalk* in Princeton WordNet (PWN), Lexico (Lex) and Merriam-Webster (MW). The number of possible choices was 25 for Merriam-Webster (5×5) and 35 (5×7) for Lexico, ca. $\frac{1}{5}$ of the combinatorial possibilities was real semantic equivalence, as defined by the proposition 1.

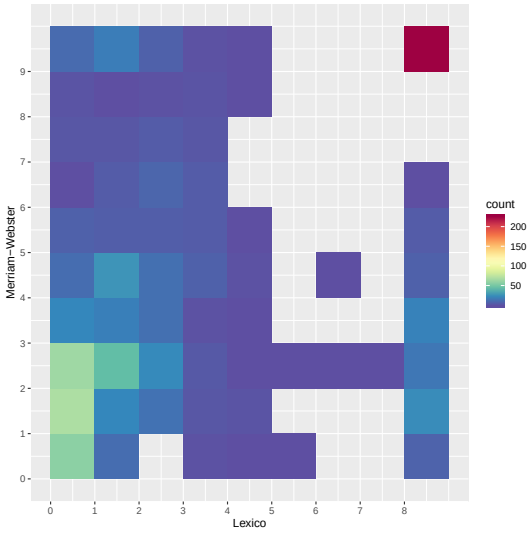

Figure 3: 2D histogram of Dijkstra's distances between PWN senses (in steps). This time the overlooked senses landed in the top-most and right-most sides of the square.

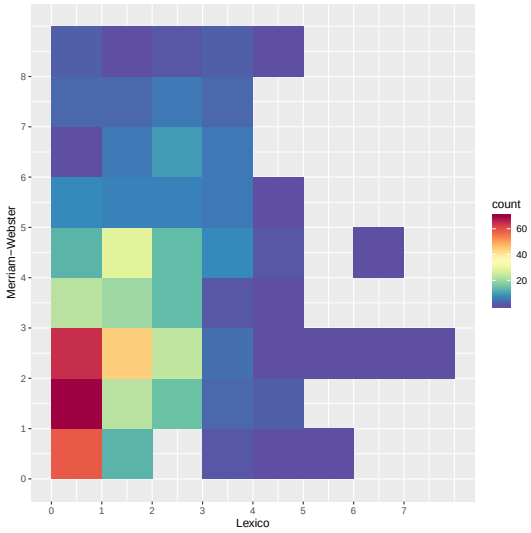

Figure 4: 2D histogram of Dijkstra's distances between PWN senses (in steps) for polysemous sense pairs noted in both dictionaries.

ρ	MW	
	HP	P
LEX	.71	.43
CI	$(0.65, 0.76)$	$(0.38, 0.53)$

Table 2: Spearman's rank correlation ρ between Lexico (LEX) and Merriam-Webster (MW) graph distances in two testing scenarios: HP – homonymy and polysemy cases, P – only polysemy cases. 'CI' signifies 99% jackknife pseudovalue intervals, $n = 57$ lemmas, cf. (Efron and Stein, 1981).

sense distinctions; might have merged and split the same semantic space in various ways – it would become obvious that they should differ. Eventually, differences do not necessarily indicate errors and may be signs of equally justified semantic descriptions.

We calculated Dijkstra's shortest path lengths between PWN senses mapped on the two dictionaries within WordNet 3.0. The undirected graph of WordNet was used. It contained 365,000 bidirectional relation instances. All relation instances were treated democratically, receiving weights of 1.

Table 3 presents the comparison between Lex-

ico, Merriam-Webster and WordNet in terms of Spearman's correlation ρ for polysemy and homonymy cases. In general, WordNet distances behaved obviously worse than dictionaries, when homonymy was considered altogether with polysemy ('HP' scenario). However, when hints from the oracle were applied (the 'OHP' case), the results became fully comparable with the Lexico-Merriam-Webster agreement. When we cut off homonymy pairs, we found the WordNet-based measure performed almost as well as both dictionary-based distances (it achieved the lower confidence limit). It seems that what dictionaries and WordNet differ in is the proper treatment of homonymy pairs. In dictionaries the information is provided by etymologists; WordNet lacks it.

	WN		
ρ	HP	OHP	P
LEX	.46	**.70**	.36
MW	.46	**.67**	.38
minML	.47	**.68**	.38
LEX-MW CI	$(0.65, 0.76)$		$(0.38, 0.53)$

Table 3: Spearman's rank correlation ρ between WordNet (WN) and dictionary graph distances in three testing scenarios. Symbols: HP – homonymy & polysemy cases; OHP – homonymy cases given by the oracle; P – homonymy cases excluded ($n = 680$ sense pairs, 57 lemmas); LEX – Lexico, MW – Merriam-Webster, minML = $min(dist'_{LEX}, dist'_{MW})$, the lowest of two distance values, where $dist'$ signifies the standardisation of distance measures. In bold we indicated results that fitted corresponding 99% confidence intervals for the LEX-MW comparison.

When one merges the information from both dictionaries (see Table 3, *minML* measure), the Spearman's correlation increases. We calculated the minimum value from standardised distances on both dictionaries, i.e.

$$minML = min(dist'_{LEX}, dist'_{MW}). \quad (2)$$

The obtained scores indicate that dictionaries might have presented rather complementary pieces of sense description than inconsistent information.

5 Conclusions

The performed experiments aimed at comparing how similarly two dictionaries described semantic distances in polysemy and homonymy. We found out that traditional English dictionaries showed traces of positive correlation between Dijkstra's path lengths on corresponding polysemy nets (0.7 for polysemy and homonymy, and $\rho = 0.4$ for sole polysemy). With regard to the homonymy/polysemy binary distinction, we obtained Cohen's $\kappa = 0.67$ and 85% percentage agreement.

The agreement with WordNet was moderate in the case of homonymy and polysemy ($\rho \sim 0.46$). When the oracle was considered (hints on the status of homonymous pairs), the correlation rose to the level of $\rho = 0.7$ which value was comparable to the confidence interval calculated for dictionaries. The values calculated for the sole polysemy (i.e. excluding homonymy) were slightly smaller than those obtained from the Lexico and Merriam-Webster comparison. The achieved results resembled agreement measurements reported in the literature (see Sec. 2 above).

The performed experiments gave an insight into the debate on the quality of dictionary descriptions. It turned out that lexicographers from different publishing companies provided very similar semantic description of homonymy – senses were similarly grouped according to their shared etymology. Dictionaries comparably described also semantic distance between related senses, when measured shortest paths on entry microstructures (micro-hierarchies). WordNet proved its usefulness in capturing the strength of polysemy links, but failed in homonymy recognition.

Acknowledgments

This research was financed by the National Science Centre, Poland, grant number 2018/29/B/HS2/02919, and supported by the CLARIN-PL[11] research infrastructure, and the NTU Digital Humanities Research Cluster.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Beryl TS Atkins. 1992. Tools for computer-aided corpus lexicography: the hector project. *Acta Linguistica Hungarica*, 41(1/4):5–71.

Sue Atkins. 2008. *Practical Lexicography: A Reader*, chapter Theoretical Lexicography and

[11] http://clarin-pl.eu

Dictionary-making, pages 31–50. Oxford University Press.

Claudia Brugman and George Lakoff. 2006. *Cognitive Linguistics: Basic Readings*, chapter Radial network: Cognitive topology and lexical networks, pages 185–239. Mouton de Gruyter: Berlin.

Irina Chugur, Julio Gonzalo, and Felisa Verdejo. 2002. A study of polysemy and sense proximity in the senseval-2 test suite. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*, pages 32–39.

Edsger W Dijkstra et al. 1959. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

Bradley Efron and Charles Stein. 1981. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

Gertrud Fenk-Oczlon and August Fenk. 2008. *Language Complexity: Typology, contact, change*, chapter Complexity trade-off between subsystems of language, pages 1–15. John Benjamins Publishing.

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. 2013. Regular polysemy in wordnet and pattern based approach. *International Journal On Advances in Intelligent Systems*, 6.

Bradley Hauer and Grzegorz Kondrak. 2020. One homonym per translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7895–7902.

Howard Jackson. 2002. *Lexicography: An Introduction*. Routledge.

Sergey Klimenkov and Alexander Pokid. 2019. Designing a model of contexts for word-sense disambiguation in a semantic network.

John Lyons. 1995. *Linguistic semantics: An introduction*. Cambridge University Press.

Barbara C Malt, Steven A Sloman, Silvia Gennari, Meiyi Shi, and Yuan Wang. 1999. Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40(2):230–262.

Diana McCarthy. 2006. Relating wordnet senses for word sense disambiguation. In *Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*.

Lingling Meng, Runqing Huang, and Junzhong Gu. 2013. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1):1–12.

Rada Mihalcea. 2003. Turning wordnet into an information retrieval resource: Systematic polysemy and conversion to hierarchical codes. *International journal of pattern recognition and artificial intelligence*, 17(05):689–704.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28. Association for Computational Linguistics, Barcelona, Spain. URL `https://www.aclweb.org/anthology/W04-0807`.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Nat. Lang. Eng.*, 13(2):137–163.

Rebecca J Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *LREC*.

Ted Pedersen, Siddharth Patwardhan, Jason Michelizzi, et al. 2004. Wordnet:: Similarity-measuring the relatedness of concepts. In *AAAI*, volume 4, pages 25–29.

Christian Ramiro, Mahesh Srinivasan, Barbara C. Malt, and Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, 115(10):2323–2328.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural language engineering*, 5(2):113–133.

Ray Richardson, A Smeaton, and John Murphy. 1994. Using wordnet as a knowledge base for measuring semantic similarity between words.

Penelope F. Stock. 2008. *Practical Lexicography: A Reader*, chapter Polysemy, pages 1–15. Oxford University Press.

Bo Svensén. 2009. A handbook of lexicography. *The theory and practice of dictionary-making. Cambridge: CUP*.

Jean Véronis. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, pages 2–4. Citeseer.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.