An analysis of ambiguity in word sense annotations

David Jurgens
Sapienza University of Rome
jurgens@di.uniroma1.it

Abstract

Word sense annotation is a challenging task where annotators distinguish which meaning of a word is present in a given context. In some contexts, a word usage may elicit multiple interpretations, resulting either in annotators disagreeing or in allowing the usage to be annotated with multiple senses. While some works have allowed the latter, the extent to which multiple sense annotations are needed has not been assessed. The present work analyzes a dataset of instances annotated with multiple WordNet senses to assess the causes of the multiple interpretations and their relative frequencies, along with the effect of the multiple senses on the contextual interpretation. We show that contextual underspecification is the primary cause of multiple interpretations but that syllepsis still accounts for more than a third of the cases. In addition, we show that sense coarsening can only partially remove the need for labeling instances with multiple senses and we provide suggestions for how future sense annotation guidelines might be developed to account for this need.

Keywords: sense annotation, word sense disambiguation, ambiguity, syllepsis

1. Introduction

A word may take on a variety of meanings depending on the context. Word sense inventories formalize these meanings into discrete units, known as senses. Annotators later consult these senses to specify which meaning is present in a given context. In some cases, a word's senses may be related. For example, in the commonly used WordNet (Fellbaum, 1998) and OntoNotes (Hovy et al., 2006) sense inventories, *bank* may refer to the financial institution, the building that the institution occupies, or an object used to store money (e.g., a piggy bank). These related senses can cause difficulty for annotators when determining the particular sense of a word's usage. Indeed, OntoNotes even includes a special sense of *bank* to indicate when the usage is ambiguous between the institution and building senses.

While most annotation efforts restrict usages to having a single sense even when ambiguous, another possibility is to label such usages with multiple senses in order to explicitly model the multiple interpretations. Previous annotation studies have shown that annotators will use multiple senses if allowed (Véronis, 1998; Murray and Green, 2004; Erk et al., 2009; Passonneau et al., 2012; Jurgens, 2013). However, little work has assessed the underlying factors causing annotators to perceive multiple senses and what the resulting interpretation of the usage is with its multiple labels. Furthermore, no study has proposed guidelines for when annotators should use multiple senses, despite expressing the need for them (Krishnamurthy and Nicholls, 2000; Langone et al., 2004).

This paper presents an analysis of instances annotated with multiple WordNet senses in the dataset of Jurgens and Klapaftis (2013) in order to answer three open questions: (1) what is the frequency with which a lemma may have valid, differing semantic interpretations, (2) what contextual factors contribute to the perception of multiple senses and what are their relative frequencies, and (3) what is the relationship between the annotation's multiple senses and how is the usage interpreted as a result. From this analysis we propose two recommendations for sense annotation guidelines.

2. Related Work

Work on ambiguity in word sense annotation has often focused on techniques to reduce ambiguity in the sense inventory in order to improve annotator agreement. Most work has addressed the aspect of sense granularity, with many proposals for how to reduce ambiguity by adapting a sense inventory to make its senses hierarchical (Edmonds and Cotton, 2001), underspecified (Buitelaar, 2000), or more coarse-grained (Palmer et al., 2004; Palmer et al., 2007). Other work has proposed creating new sense inventories around annotators' perceived distinctions in meaning (Rumshisky and Batiukova, 2008; Biemann, 2012).

Several works have investigated allowing annotators to use multiple senses. Véronis (1998) analyzed 600 hundred contexts each for 60 French words using a common French dictionary to define their senses. The author notes that the average rate of multi-sense annotation was low, with an average of 1.02 senses per instance, but was higher for some words (e.g., an average of 1.311 senses per instance for the verb *comprendre*). Furthermore, the author notes that disagreements were not resolved by merging similar senses, with most disagreement occurring between much coarser senses. To fully capture all perceived senses, Erk et al. (2009) had three annotators label 50 contexts each for eight words, rating all senses of the word using a scale from one (inapplicable) to five (completely applicable). Their annotators readily used multiple senses per instance, with 78.8% of usages having multiple senses rated at least a three on their scale. Similar to the study of Véronis (1998), Erk et al. (2009) found that the presence of multiple sense could not be reduce to a single sense by grouping highly-correlated senses, with over 40% of contexts having two senses that were not correlated. Jurgens (2013) had crowdsourced workers rate all senses of a word according to applicability using the same dataset of Erk et al. (2009), finding that workers agreed significantly more when using multiple senses per instance than when restricted to using a single sense; this result suggests that ambiguity is a major contributing factor to annotator disagreement. Last, Alonso et al. (2013) asked crowdsourced workers and volunteers to

annotate between literal and metonymic senses, allowing them to select a third option of "both" in cases of underspecification. Annotators for Danish, English, and Spanish all used the underspecified option, demonstrating that, in conjunction with the French study of Véronis (1998), the perception of multiple senses is common across languages and sense inventories.

The most similar to the proposed study is that of Passonneau et al. (2009), who analyzed factors contributing to annotator disagreement, finding that agreement was often mediated by the lemma itself and its usages. Looking across multiple lemmas, they found that the factors of sense similarity, contextual specificity, and sense concreteness most affected agreement. The present work offers a complementary analysis for understanding cases where multiple senses meaningfully apply, which potentially may have been treated as cases of annotator disagreement when usages were required to be labeled with a single sense.

3. Dataset

While many sense-annotated datasets exist, only recently have datasets included a diversity of instances annotated with multiple senses, e.g., Erk et al. (2012). We selected the dataset of Jurgens and Klapaftis (2013), which was drawn from the Open American National Corpus (OANC) (Ide and Suderman, 2004) across a variety of genres and from both the spoken and written portions of the corpus. Context is restricted to a single sentence containing the target lemma. The context size matches common real-word settings such as microtext domains like Twitter, where limited context for disambiguation is available (Gella et al., 2014). The dataset contains instances of 20 nouns, 20 verbs, and 10 adjectives for a total of 4664 usages, which were annotated using WordNet 3.1 (Fellbaum, 1998). Of those contexts, 542 (12%) were annotated with multiple senses, which is consistent with the rates other annotation studies that allowed multiple sense ratings (Véronis, 1998; Navigli et al., 2013). We restrict our analysis to the subset of instances that were labeled with multiple senses.

4. Annotation Schema

The process of determining the appropriate sense for a usage is a function of both the context and the options available in the sense inventory. Therefore, we analyze each usage with a two-way classification schema that incorporates both factors. The first classification axis assesses the type of semantic constraints put upon the usage by the context. This axis captures cases of ambiguity from too few contextual cues and cases of syllepsis where multiple dependent clauses or modifiers refer to different meanings of the same usage. The second classification axis assesses the types of senses that are elicited by the usage and how those senses together may be interpreted within in the context. For example, this axis captures cases where the context evokes highly-similar senses that do not change the overall meaning of the context. We refer to these classification types as Context and Sense Assignment, respectively. Annotation guidelines were developed for both types iteratively through a close analysis of 50 randomly-selected instances and then applied to annotate all instances. After a first pass of annotation, each instance was then re-annotated to correct errors and ensure the guidelines were clear. Following, we formalize the classifications and present examples from the corpus. For clarity, the sense descriptions with each example also include the corresponding WordNet sense keys used in the dataset.

4.1. Context Classification

Contexts were divided into two classes: (1) those containing too few semantic cues to constrain the usage to a single interpretation and (2) those exhibiting semantic syllepsis where the dependent clauses or modifiers of a usage require it to be interpreted with different meanings. We refer to these classes as Underconstrained and Sylleptic, respectively.

Underconstrained contexts may occur due to contexts having too little content as a whole and due to contexts omitting specific information needed to distinguish between related senses. We illustrate these through two examples. In the following context for *warm*:

> Rooms are classically decorated and *warm*.

the cues enable interpretations of a comfortable level of heat (warm%3:00:01::) and being colored in such a way to evoke warmth (warm%3:00:03::), despite the large difference in the senses' meanings. In contrast, consider the context for *find*.

> The Random House lexicographer Jesse Sheidlower has *found* a reference to it in a passage from Varieties of Religious Experience, in which William James quotes words of Voltaire, for which he gives the date 1773: "All comes out at the end of the day, and all comes out still more even when all the days are over."

The interpretation of *find* is dependent upon the mental state of the actor and state of the passage in question. The context enables readings where the individual unintentionally encountered the reference (find%2:40:02::), was actively seeking the reference out (find%2:39:02::), or had rediscovered a lost reference (find%2:40:00::). For both example contexts, additional cues can restrict the interpretations to a single sense, e.g., adding "in winter" as a prepositional phrase modifying *warm* in the first example.

Sylleptic contexts frequently occur from two constructions. The first construction occurs when a word is associated with multiple senses, each describing different aspects of the same concept. For example, *chicken* may refer to both the animal and the food. In many cases, classes of words exhibit the same type of semantic distinctions (e.g., Animal/Food), which are known as cases of regular polysemy (Apresjan, 1974; Pustejovsky, 1995; Ravin and Leacock, 2000). These Sylleptic constructions evoke the same concept through having clauses or modifiers refer to its different aspects. We illustrate this construction with an example for *book*.

> The fat *book* here surveys the hundreds of books and articles already written about Rockefeller

	Parallel	Admissible	Conflicting	*Total*
Underconstrained	47	158	92	297
Sylleptic	122	26	15	163
Total	169	184	107	460

Table 1: Distribution of assignments to the Context and Sense Assignment classifications

The usage evokes both the physical object (book%1:06:00::) by means of the adjective fat and the book's role as a literary work (book%1:10:00::) by the discussion of its content. This object-role sense distinction is seen in other lemmas with the same type of regular polysemy, such as *magazine* or *paper*.
The second common type of Sylleptic construction comes from metaphoric usages that evoke both literal and figurative meanings of the word. We illustrate this with an example for *dark*.

> We all are relieved to lay aside our fight-or-flight reflexes and to commemorate our births from out of the *dark* centers of the women, to feel the complexity of our love and frustration with each other, to stretch our cognition to encompass the thoughts of every entity we know.

By their nature, the centers of women are devoid of light (dark%3:00:01::), but the usage also allows a reading with a secretive connotation (dark%3:00:00:concealed:00) to the role of the darkness.

4.2. Sense Assignment Classification

The fine-grained sense inventory of WordNet permits some sense interpretations to be concurrently true (e.g., describing a lemma's purpose and physical properties). However, some contexts elicit conflicting interpretations. Therefore, the second classification axis describes the consistency of the usage's interpretation according to three classifications: (1) the senses may be concurrently true (2) the senses are sufficiently related to evoke a basic interpretation that could be further refined to one sense, given more context, and (3) the senses describe conflicting interpretations of the usage in context. We refer to these classifications as Parallel, Admissible, and Conflicting, respectively, and following illustrate each with examples.
Parallel interpretations may be caused both by Underconstrained and Syllpetic contexts. Consider the usage of *severe* below.

> The shock was *severe* enough to strike her dumb, and she was committed to a London hospital.

The context allows interpretations of *severe* with respect to two aspects of the sentence, describing the shock as intensely bad or unpleasant (severe%3:00:00:intense:00) or the shock's effects on the recipient as grievous or causing harm (severe%3:00:00:critical:03). Both interpretations may be valid within the interpretation of the entire context.
Admissible interpretations reflect cases where a usage's senses express fine-grained distinctions or closely-related concepts. For example, consider the following use of *family*.

> He added that his wasn't a dysfunctional *family*.

Here, the scope of *family* could refer to both the immediate family (family%1:14:00::) or to a larger unit of blood relatives (family%1:14:01::). However, the context can still be interpreted with approximately the same meaning without resolving the ambiguity, i.e., *family* referring to people that are related in some way to the sentence's subject.
In Conflicting cases, the ambiguity in a usage's sense elicits distinct incompatible interpretations. Consider the meaning of *image* below.

> Thomason and the White House aren't talking, so I consulted my own *image* expert, Jackson Bain.

Without aid of further context, it is unclear whether the individual's expertise pertains to visual representations such as photographs (image%1:06:00::) or to managing a public persona (image%1:07:00::).

5. Results and Discussion

The proposed annotation scheme was applied to all 542 instances with multiple sense assignments in the dataset of Jurgens and Klapaftis (2013). During annotation, our analysis suggested that 82 instances (15%) containing a sense rated as having low applicability should be excluded from the analysis, as these low-applicability sense would not be considered valid interpretations of the usage according to our guidelines. The remaining 460 annotated instances were used in our analysis. Instances were approximately equally distributed across part of speech classes, with 164, 151, and 145 instances for noun, verb, and adjective lemmas, respectively. Table 1 shows the distribution of instances across the two-way classification scheme.
Four analyses were performed to test (1) differences in ambiguity according to part of speech, (2) the relationship between an instance's senses' similarity and its classification type, (3) the effect of sense granularity on sense ambiguity, and (4) the presence of lemma-specific preferences towards certain types of ambiguity. Following, we discuss general observations of the instances' classifications and then describe the results of each analysis. We conclude with a discussion of the how current sense annotation guidelines might be improved.

5.1. General Observations

Three general trends appear across all instances. First, Underspecified contexts are nearly twice as common as Sylleptic constructions. Because contexts are single sentences, the high frequency of Underspecified contexts raises the possibility that the majority of ambiguous cases could potentially be resolved to a single sense using additional context outside the sentence. Second, Underspecified and

		Nouns		Verbs		Adjs.
Context	Sense Assignment	%	JCN	%	JCN	%
Underconstrained	Admissible	43.9	0.151	22.5	0.091	35.9
Underconstrained	Conflicting	24.4	0.064	16.6	0.076	18.6
Underconstrained	Parallel	4.9	0.131	5.2	0.101	21.4
Sylleptic	Admissible	1.8	0.073	11.9	0.100	3.4
Sylleptic	Conflicting	3.0	0.062	4.6	0.112	2.0
Sylleptic	Parallel	22.0	0.098	39.1	0.065	18.6

Table 2: Distribution of assignments to the two-way classifications with the percentages of total instances per part of speech and the average JCN similarity of senses assigned to each instance having that classification.

Sylleptic contexts can cause all types of sense classifications, though each context type has a clear preference. Third, only 23% of the instances have Conflicting semantic interpretations, suggesting that in the majority of multiple-sense cases, a correct interpretation of the entire context is not dependent upon refining the annotation to a single sense.

5.2. Part of Speech

The first analysis measures the differences in the classifications according to part of speech. Table 2 shows the percentage of instances assigned to each. Clear distinctions between part of speech classes emerge. Both nouns and adjectives are much more likely to be in Underconstrained contexts (73.2% and 75.9%, respectively), while verbs are slightly more likely to be in Sylleptic contexts (55.6%).

Within the observed word types, we observed a trend in adjectives and nouns where types had highly-related senses that were difficult to distinguish between with the available context. For example, in the following instance of *new*,

> By reengineering business processes in conjunction with implementing *new* technology, Owens Corning increased its ability to meet customer needs.

the context makes it unclear whether the new technology has recently been invented (new%3:00:00::) or is simply unlike the previous technology (new%3:00:00:other:00). Similarly, in the early example of *family*, distinguishing between the immediate and extended family senses is difficult in shorter contexts, though the former is a more probable interpretation. However, the number of word types in the dataset is too small to make strong generalizations about the behaviors of each part of speech class.

The frequency with which instances take on multiple senses also varies by part of speech, with 10.1% of noun instances, 8.1% of verb instance, and 15.1% of adjective instances having multiple instances. The increase in frequency for adjectives over the rates for nouns and verbs is statistically significant at $p<0.01$. We hypothesize that adjectives show an increased frequency because they may take on many different shades meanings, depending on the noun (Pustejovsky, 1995) and sparse contexts increase the difficulty in selecting only one of these related meanings.

Because the lemmas used in the dataset of Jurgens and Klapaftis (2013) were intentionally selected based on having exhibited sense ambiguity in previous annotation studies, the frequencies with which lemmas have multiple senses are likely to be higher than those for a larger sample of lemmas. However, the frequencies may still provide soft upper bounds for those expected in larger corpora and could be useful for identifying annotators who are over-zealous in using multiple senses.

5.3. Sense Similarity

Fine-grained sense distinctions are often reflected in the high degree of similarity between two senses of a word. We hypothesize that sense similarity may reveal whether fine-grain distinctions contribute to certain types of ambiguity as expressed in the classifications. Therefore, in the second analysis, for each noun and verb instance, we calculate the similarity of its senses using the Jiang and Conrath (JCN) similarity measure (Jiang and Conrath, 1997), which was shown to most-closely approximate human similarity judgments (Budanitsky and Hirst, 2006). The JCN columns in Table 2 show the average similarity for nouns and verb instances assigned to each classification type.

Among the nouns, two trends emerge. First, the selected senses are more similar when a context is Underconstrained than when Sylleptic, which highlights the differences in the underlying mechanisms. In Underconstrained contexts, the lack of semantic cues makes distinguishing between similar senses difficult, resulting in multiple senses in the annotation, while Syllpetic contexts are more likely to evoke senses related to a single concept which are not necessarily themselves similar. Second, instances with a Conflicting annotation classification have sense with lower similarity than those of the other two classifications; however, we note that the number of Sylleptic nouns instances is too few to draw statistically significant conclusions with respect to sense similarity.

The sense similarity of verb instances did not show the same trends as noun instances. Though minor difference in verb sense similarities are present, differences between classifications are not significant at $p<0.05$.

5.4. Sense Granularity

In our third analysis, we consider whether multiple interpretations could be due to the granularity of the sense inventory. To test this hypothesis, each instance's WordNet annotation was evaluated to see whether its multiple senses were all subsumed by a single, more coarse-grained OntoNotes sense (Hovy et al., 2006), using the publicly-available sense mappings for nouns and verbs between the two inventories.

Context	Sense Assign.	Nouns	Verbs
Underconstrained	Admissible	90.3	97.1
Underconstrained	Conflicting	57.5	100.0
Underconstrained	Parallel	87.5	100.0
Sylleptic	Admissible	100.0	100.0
Sylleptic	Conflicting	60.0	100.0
Sylleptic	Parallel	69.4	100.0

Table 3: Percentage of instances with multiple WordNet senses that would receive a single OntoNotes sense label

Table 3 lists the percentage of instances for each classification and part of speech whose multiple senses would be represented by a single OntoNotes sense.

The merging percentages shown in Table 3 reveal two clear differences between parts of speech and the classifications. First, nearly all verbs (99.3%) would be annotated with a single OntoNotes sense. WordNet often contains fine-grain sense distinctions of a single action for a verb, reflecting aspects such as the circumstances, effect, and implications related to the action. For example, the verb *transfer* contains different senses for the action of moving an object depending on whether the movement implies a change of ownership. In the following instance, it not clear whether the authorities provided copies of their documents (transfer%2:40:01::) or provided the originals such that they no longer have ownership (transfer%2:40:00::):

> "The US Administration stated on January 12 that all documents on 'the Iranian case' have been *transferred* to the competent Russian agencies through diplomatic channels," Segodnya said.

Indeed, OntoNotes merges six WordNet senses of *transfer* relating to this action into one sense. However, coarsening the senses loses information. In the above example, an OntoNotes annotation would not longer convey whether documents were retained by the US Administration, which could potentially affect downstream applications using senses, such as Textual Entailment.

In the second difference between parts of speech, sense coarsening reveals that Admissible contexts were nearly always represented with a single OntoNotes sense for nouns and verbs, irrespective of the context classification, but in contrast, a large minority (23.2%) of Conflicting and Parallel noun instances would still require multiple OntoNotes senses. The latter minority is due to cases where the instance requires significantly different interpretations (e.g., the earlier example instance of *image*) or where OntoNotes does not merge cases of regular polysemy into a single sense.

5.5. Distribution of Annotations

Given the differences in classification distribution per part of speech, in the fourth analysis, we assess whether the classification types were also unevenly distributed across the lemmas themselves; i.e., whether all instances of a lemmas occurred with the same classification type. Figure 1 shows the distribution of the six classification combinations across all 49 lemmas having multiple sense annotations. Lemmas had on average 3.02 different classifications

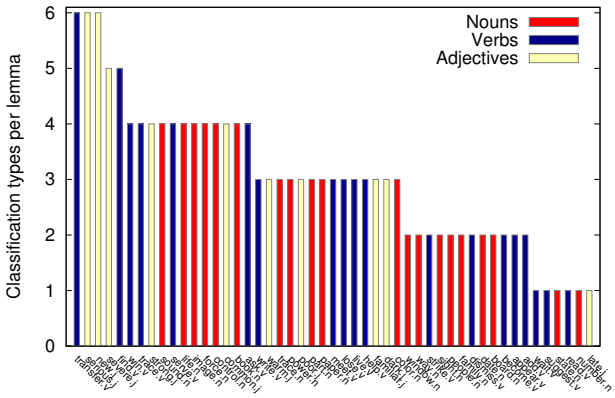

Figure 1: A comparison of the number of unique Context-Sense Assignment classifications seen per lemma

across their instances. Furthermore, the number of classes seen was moderately correlated with the total number of instances per lemma, Pearson's r=0.468, which is statistically significant at $p<0.01$. Together, these suggests that most lemmas have sense distinctions that may be ambiguous in certain circumstances and that given enough instances of the lemma, its usages expressing multiple interpretations will be seen with a variety of classification types.

5.6. Sense Annotation Recommendations

The two-way classification was designed with an additional motivation of providing guidance to sense annotators when deciding whether an instance should have multiple senses. While prior annotator efforts have allowed annotators to use multiple senses, no guidelines have been proposed for when annotators should use them, despite an expressed need (Langone et al., 2004). For example, Krishnamurthy and Nicholls (2000) note ambiguous sylleptic examples for *onion*, stating that during the annotation effort for SensEval-1, annotators chose only a single sense of these using their own *ad hoc* heuristics.

Our first recommendation is that sense annotators should be encouraged to assign multiple senses but do so only in cases where they are able to additionally classify the usage according to our proposed classification schema. This recommendation is motivated by two factors. First, the process of classifying the instance encourages annotators to scrutinize the context in order to justify why multiple senses would be present. This discourages the introduction of multi-sense annotations due to an annotator accidentally overlooking sense-specific cues. However, the classification process is sufficiently lightweight and related to the main task so as not to overburden annotators. Second, the classifications can be important for later applications needing to interpret the multiple senses. Similarly, the classifications can be of use for the evaluation of Word Sense Disambiguation (WSD) systems. In cases of an Underspecified context, a WSD system's output could be considered correct if it contains any of the multiple senses; however, in Sylleptic contexts, the WSD system would need to recognize all the intended meanings.

Our second recommendation is that in cases of Underconstrained contexts, annotators should be able to mark which sense is more likely. We motivate this with an example for *strike*.

Traditionally eight bells are *struck*.

The likely meaning of the context is that the bells were hit to make a sound (strike%2:35:00::); however, in certain domains, the eight bells could have been forged (strike%2:36:02::), perhaps in commemoration. In the absence of available cues, annotators can use their background knowledge to indicate which sense is more probable. We note that this weighting differs from the sense applicability ratings used by Erk et al. (2009) for labeling multiple senses per word; in our setting, both senses would already be applicable, with the weight specifying only their likelihoods of being the correct interpretation. The proposed weights could be used by downstream applications to select a single interpretation and by WSD evaluations to favor a system that reports the more probable sense, much like prior suggestions for evaluating based on sense similarity (Resnik and Yarowsky, 2000) or applicability (Jurgens, 2012).

6. Conclusion

Word sense annotation is a challenging task where annotators may find multiple valid interpretations of a usage, leading to multiple senses in the same annotation. The present study has analyzed 460 instances annotated with multiple senses from the dataset of Jurgens and Klapaftis (2013). We then classified each according to a proposed two-way classification to quantify (1) the contextual features contributing to the multiple interpretations and (2) the impact of multiple senses on the interpretation of the usage in context. Our work offers three main contributions. First, we showed that contextual underspecification is responsible for instances with multiple sense annotations in nearly two-thirds of the cases, while the remaining are due to syllepsis, where multiple contextual features each select different senses. However, verbs are more likely to appear in syllleptic contexts. Second, we showed that the fine granularity of WordNet is not completely responsible for the perception of multiple senses: While nearly all verb instances with multiple senses would be annotated with a single OntoNotes sense, for nouns, conflicting interpretations or cases of regular polysemy would still require the instance to have multiple OntoNotes senses in approximately 23% of cases. Last, we proposed a two-way classification scheme, which we argue would serve sense annotators as a guideline when considering multiple senses for a usage.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant Multi-JEDI No. 259234.

7. References

Alonso, H. M., Pedersen, B. S., and Bel, N. (2013). Annotation of regular polysemy and underspecification. In *Procedings of the 51st Meeting of the Associatation for Computation Linguistics (ACL)*.

Apresjan, I. D. (1974). Regular polysemy. *Linguistics*, 142:5–32.

Biemann, C. (2012). Turk bootstrap word sense inventory 2.0: A large-scale resource for lexical substitution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 4038–4042.

Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Buitelaar, P. (2000). Reducing lexical semantic complexity with systematic polysemous classes and underspecification. In *Proeceedings of the NAACL-ANLP Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*, pages 14–19. ACL.

Edmonds, P. and Cotton, S. (2001). Senseval-2: Overview. In *Proceedings of Senseval-2*, pages 1–5. ACL.

Erk, K., McCarthy, D., and Gaylord, N. (2009). Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18. ACL.

Erk, K., McCarthy, D., and Gaylord, N. (2012). Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Fellbaum, C., editor. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

Gella, S., Cook, P., and Baldwin, T. (2014). One Sense per Tweeter... and Other Lexical Semantic Tales of Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL)*, pages 57–60. ACL.

Ide, N. and Suderman, K. (2004). The american national corpus first release. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 1681–1684.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33.

Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval), in the Second Joint Conference on Lexical and Computational Semantics*. ACL.

Jurgens, D. (2012). An Evaluation of Graded Sense Disambiguation using Word Sense Induction. In *Proceedings of *SEM, the First Joint Conference on Lexical and Computational Semantics*. ACL.

Jurgens, D. (2013). Embracing Ambiguity: A Compar-

ison of Annotation Methodologies for Crowdsourcing Word Sense Labels. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. ACL.

Krishnamurthy, R. and Nicholls, D. (2000). Peeling an onion: The lexicographer's experience ofmanual sense-tagging. *Computers and the Humanities*, 34(1-2):85–97.

Langone, H., Haskell, B. R., and Miller, G. A. (2004). Annotating WordNet. In *Proceedings of the NAACL-HLT Workshop on Frontiers in Corpus Annotation*.

Murray, G. and Green, R. (2004). Lexical knowledge and human disagreement on a wsd task. *Computer Speech & Language*, 18(3):209–222.

Navigli, R., Jurgens, D., and Vannella, D. (2013). Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval), in the Second Joint Conference on Lexical and Computational Semantics*.

Palmer, M., Babko-Malaya, O., and Dang, H. T. (2004). Different sense granularities for different applications. In *Proceedings of the Second Workshop on Scalable Natural Language Understanding Systems*.

Palmer, M., Dang, H. T., and Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

Passonneau, R., Salleb-Aouissi, A., and Ide, N. (2009). Making sense of word sense variation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

Passonneau, R., Bhardwaj, V., Salleb-Aouissi, A., and Ide, N. (2012). Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Pustejovsky, J. (1995). *The Generative Lexicon: A Theory of Computational Lexical Semantics*. Cambridge, MA: The MIT Press.

Ravin, Y. and Leacock, C. (2000). *Polysemy: Theoretical and computational approaches*. MIT Press.

Resnik, P. and Yarowsky, D. (2000). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.

Rumshisky, A. and Batiukova, O. (2008). Polysemy in verbs: systematic relations between senses and their effect on annotation. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 33–41. ACL.

Véronis, J. (1998). A study of polysemy judgments and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*.