Measuring Word Meaning in Context

Katrin Erk*
University of Texas at Austin

Diana McCarthy**
University of Cambridge

Nicholas Gaylord*
University of Texas at Austin

Word sense disambiguation (WSD) is an old and important task in computational linguistics that still remains challenging, to machines as well as to human annotators. Recently there have been several proposals for representing word meaning in context that diverge from the traditional use of a single best sense for each occurrence. They represent word meaning in context through multiple paraphrases, as points in vector space, or as distributions over latent senses. New methods of evaluating and comparing these different representations are needed.

In this paper we propose two novel annotation schemes that characterize word meaning in context in a graded fashion. In WSsim *annotation, the applicability of each dictionary sense is rated on an ordinal scale.* Usim *annotation directly rates the similarity of pairs of usages of the same lemma, again on a scale. We find that the novel annotation schemes show good inter-annotator agreement, as well as a strong correlation with traditional single-sense annotation and with annotation of multiple lexical paraphrases. Annotators make use of the whole ordinal scale, and give very fine-grained judgments that "mix and match" senses for each individual usage. We also find that the* Usim *ratings obey the triangle inequality, justifying models that treat usage similarity as metric.*

There has recently been much work on grouping senses into coarse-grained groups. We demonstrate that graded WSsim *and* Usim *ratings can be used to analyze existing coarse-grained sense groupings to identify sense groups that may not match intuitions of untrained native speakers. In the course of the comparison, we also show that the* WSsim *ratings are not subsumed by any static sense grouping.*

* Linguistics Department. CLA Liberal Arts Building, 305 E. 23rd St. B5100, Austin, TX, USA 78712. E-mail: katrin.erk@mail.utexas.edu, nlgaylord@utexas.edu.

** Visiting Scholar, Department of Theoretical and Applied Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge, CB3 9DA, UK. E-mail: diana@dianamccarthy.co.uk.

Submission received: 3 November 2011; revised version received: 30 April 2012; accepted for publication: 25 June 2012.

doi:10.1162/COLI_a_000142

1. Introduction

Word sense disambiguation (WSD) is a task that has attracted much work in computational linguistics (see Agirre and Edmonds [2007] and Navigli [2009] for an overview), including a series of workshops, SENSEVAL (Kilgarriff and Palmer 2000; Preiss and Yarowsky 2001; Mihalcea and Edmonds 2004) and SemEval (Agirre, Màrquez, and Wicentowski 2007; Erk and Strapparava 2010), which were originally organized expressly as a forum for shared tasks in WSD. In WSD, polysemy is typically modeled through a dictionary, where the senses of a word are understood to be mutually disjoint. The meaning of an occurrence of a word is then characterized through the best-fitting among its dictionary senses.

The assumption of senses that are mutually disjoint and that have clear boundaries has been drawn into doubt by lexicographers (Kilgarriff 1997; Hanks 2000), linguists (Tuggy 1993; Cruse 1995), and psychologists (Kintsch 2007). Hanks (2000) argues that word senses have uses where they clearly fit, and borderline uses where only a few of a sense's identifying features apply. This notion matches results in psychology on human concept representation: Mental categories show "fuzzy boundaries," and category members differ in typicality and degree of membership (Rosch 1975; Rosch and Mervis 1975; Hampton 2007). This raises the question of annotation: Is it possible to collect word meaning annotation that captures degrees to which a sense applies?

Recently, there have been several proposals for modeling word meaning in context that can represent different degrees of similarity to a word sense, as well as different degrees of similarity between occurrences of a word. The SemEval Lexical Substitution task (McCarthy and Navigli 2009) represents each occurrence through multiple weighted paraphrases. Other approaches represent meaning in context through a vector space model (Erk and Pado 2008; Mitchell and Lapata 2008; Thater, Fürstenau, and Pinkal 2010) or through a distribution over latent senses (Dinu and Lapata 2010). Again, this raises the question of annotation: Can human annotators give fine-grained judgments about degrees of similarity between word occurrences, like these computational models predict?

The question that we explore in this paper is: *Can word meaning be described through annotation in the form of graded judgments?* We want to know whether annotators can provide graded meaning annotation in a consistent fashion. Also, we want to know whether annotators will use the whole graded scale, or whether they will fall back on binary ratings of either "identical" or "different." Our question, however, is not whether annotators can be *trained* to do this. Rather, our aim is to *describe word meaning as language users perceive it*. We want to tap into the annotators' intuitive notions of word meaning. As a consequence, we use untrained annotators. We view it as an important aim on its own to capture language users' intuitions on word meaning, but it is also instrumental in answering our first question, of whether word meaning can be described through graded annotator judgments: Training annotators in depth on how to distinguish predefined hand-crafted senses could influence them to assign those senses in a binary fashion.

We introduce two novel annotation tasks in which human annotators characterize word meaning in context. In the first task, they rate the applicability of dictionary senses on a graded scale. In the second task, they rate the similarity between pairs of usages of the same word, also on a graded scale. In designing the annotation tasks, we utilize techniques from psycholinguistic experimentation: Annotators give ratings on a

scale, rather than selecting a single label; we also use multiple annotators for each item, retaining all annotator judgments.[1]

The result of this graded annotation can then be used to evaluate computational models of word meaning: either to evaluate graded models of word meaning, or to evaluate traditional WSD systems in a graded fashion. They can also be used to analyze existing word sense inventories, in particular to identify sense distinctions worth revisiting—we say more on this latter use subsequently.

Our aim is not to improve inter-annotator agreement over traditional sense annotation. It is highly unlikely that ratings on a scale would ever achieve higher exact agreement than binary annotation. Our aim is also not to maximize exact agreement, as we expect to see individual differences in perceived meaning, and want to capture those differences. Still it is desirable to have an end product of the annotation that is robust against such individual differences. In order to achieve this, we average judgments over multiple annotators after first inspecting pairwise correlations between annotators to ensure that they are all doing their work diligently and with similar outcomes.

Analyzing the annotation results, we find that the annotators make use of intermediate points on the graded scale and do not treat the task as inherently binary. We find that there is good inter-annotator agreement, measured as correlation. There is also a highly significant correlation across tasks and with traditional WSD and lexical substitution tasks. This indicates that the annotators performed these tasks in a consistent fashion. It also indicates that diverse ways of representing word meaning in context—single best sense, weighted senses, multiple paraphrases, usage similarity—yield similar characterizations. We find that annotators frequently give high scores to more than one sense, in a way that is not remedied by a more coarse-grained sense inventory. In fact, the annotations are often inconsistent with disjoint sense partitions.

The work reported here is based on our earlier work reported in Erk, McCarthy, and Gaylord (2009). The current paper extends the previous work in three ways.

1. We add extensive new annotation to corroborate our findings from the previous, smaller study. In this new, second round of annotation, annotators do the two graded ratings tasks as well as traditional single-sense annotation and annotation with paraphrases (lexical substitutes), all on the same data. Each item is rated by eight annotators in parallel. This setting, with four different types of word meaning annotation on the same data, allows us to compare annotation results across tasks more directly than before.[2]

2. We test whether the similarity ratings on pairs of usages obey the triangle inequality, and find that they do. This point is interesting for psychological reasons. Tversky and Gati (Tversky 1977; Tversky and Gati 1982) found that similarity ratings on words did not obey the triangle inequality—although, unlike our study, they were dealing with words out of context. The fact that usage similarity ratings obey the triangle inequality is also important for modeling and annotation purposes.

1 We do not use as many raters per item as is usual in psycholinguistics, however, as our aim is to cover a sizeable amount of corpus data.

2 The annotation data from this second round are available at `http://www.dianamccarthy.co.uk/downloads/WordMeaningAnno2012/`.

3. We examine the extent to which our graded annotation accords with two existing coarse-grained sense groupings, and we demonstrate that our graded annotations can be used to double-check on sense groupings and find potentially problematic groupings.

2. Background

In this section, we offer an overview of previous word sense annotation efforts, and then discuss alternative approaches to the annotation and modeling of word meaning.

2.1 Word Sense Annotation

Inter-annotator agreement (also called inter-tagger agreement, or ITA) is one indicator of the difficulty of the task of manually assigning word senses (Krishnamurthy and Nicholls 2000). With WordNet, the sense inventory currently most widely used in word sense annotation, ITA ranges from 67% to 78% (Landes, Leacock, and Tengi 1998; Mihalcea, Chklovski, and Kilgarriff 2004; Snyder and Palmer 2004), depending on factors such as degree of polysemy and inter-relatedness of the senses. This issue is not specific to WordNet. Annotation efforts based on other dictionaries have achieved similar ITA levels, as shown in Table 1. The first group in that table shows two corpora in which all open-class words are annotated for word sense, in both cases using WordNet. The second group consists of two English lexical sample corpora, in which only some target words are annotated. One of them uses WordSmyth senses for verbs and WordNet for all other parts of speech, and the other uses HECTOR, with similar ITA, so the choice of dictionary does not seem to make much difference in this case.[3] Next is SALSA, a German corpus using FrameNet frames as senses, then OntoNotes, again an English lexical sample corpus. Inter-annotator agreement is listed in the last column of the table; agreement is in general relatively low for the first four corpora, which use fine-grained sense distinctions, and higher for SALSA and OntoNotes, which have more coarse-grained senses.

Sense granularity has a clear impact upon levels of inter-annotator agreement (Palmer, Dang, and Fellbaum 2007). ITA is substantially improved by using coarser-grained senses, as seen in OntoNotes (Hovy et al. 2006), which uses an ITA of 90% as the criterion for constructing coarse-grained sense distinctions. Although this strategy does improve ITA, it does not eliminate the issues seen with more fine-grained annotation efforts: For some lemmas, such as *leave*, 90% ITA is not reached even after multiple re-partitionings of the semantic space (Chen and Palmer 2009). This suggests that the meanings of at least some words may not be separable into senses distinct enough for consistent annotation.[4] Moreover, sense granularity does not appear to be the only question influencing ITA differences between lemmas. Passonneau et al. (2010) found three main factors: sense concreteness, specificity of the context in which the target word occurs, and similarity between senses. It is worth noting that of these factors, only the third can be directly addressed by a change in the dictionary.

3 HECTOR senses are described in richer detail than WordNet senses and the resource is strongly corpus-based. We use WordNet in our work due to its high popularity and free availability.

4 Examples such as this indicate that there is at times a problem with clearly defining consistently separable senses of a word. There is no clear measure of exactly how frequent such cases are, however. This is due in part to the fact that this question depends so heavily on the data being considered and the distinctions being posited.

Table 1
Word sense-annotated data, with inter-annotator agreement (ITA).

Corpus	Dictionary	Corpus reference	ITA
SemCor	WordNet	Landes, Leacock, and Tengi (1998)	78.6%
SensEval-3	WordNet	Snyder and Palmer (2004)	72.5%
SensEval-1 lex. sample	HECTOR	Kilgarriff and Rosenzweig (2000)	66.5%
SensEval-3 lex. sample	WordNet, WordSmyth	Mihalcea, Chklovski, and Kilgarriff (2004)	67.3%
SALSA	FrameNet	Burchardt et al. (2006)	86%
OntoNotes	OntoNotes	Hovy et al. (2006)	most $>$ 90%

Table 2
Best word sense disambiguation performance in SensEval/SemEval English lexical sample tasks.

Shared task	Shared task overview	Best precision	Baseline
SensEval-1	Kilgarriff and Rosenzweig (2000)	77%	69%
SensEval-2	Senseval-2 (2001)	64%	51%
SensEval-3	Mihalcea, Chklovski, and Kilgarriff (2004)	73%	55%
SemEval-1	Pradhan et al. (2007)	89%	(not given)

ITA levels in word sense annotation tasks are mirrored in the performance of WSD systems trained on the annotated data. Table 2 shows results for the best systems that participated at four English lexical sample tasks. With fine-grained sense inventories, the top-ranking WSD systems participating in the event achieved precision scores of 73% to 77% (Edmonds and Cotton 2001; Mihalcea, Chklovski, and Kilgarriff 2004). Current state-of-the-art systems have made modest improvements on this; for example, the system described by Zhong and Ng (2010) achieves 65.3% on the English lexical sample at SENSEVAL-2, though the same system obtains 72.6%, just below Mihalcea, Chklovski, and Kilgarriff (2004), on the English lexical sample at SENSEVAL-3. Nevertheless, the picture remains the same with systems getting around three out of four word occurrences correct. Under a coarse-grained approach, system performance improves considerably (Palmer, Dang, and Fellbaum 2007; Pradhan et al. 2007), with the best participating system achieving a precision close to 90%.[5] The merits of a coarser-grained approach are still a matter of debate (Stokoe 2005; Ide and Wilks 2006; Navigli, Litkowski, and Hargraves 2007; Brown 2010), however.

Although identifying the proper level of granularity for sense repositories has important implications for improving WSD, we do not focus on this question here. Rather, we propose novel annotation tasks that allow us to probe the relatedness between dictionary senses in a flexible fashion, and to explore word meaning in context without presupposing hard boundaries between usages. The resulting data sets can be used to compare different inventories, coarse or otherwise. In addition, we hope that they will prove useful for the evaluation of alternative representations of ambiguity in word

5 Zhong, Ng, and Chan (2008) report similar results (89.1%) with their state-of-the-art system when evaluating on the OntoNotes corpus, which is larger than the SENSEVAL data sets.

meaning (Erk and Pado 2008; Mitchell and Lapata 2008; Reisinger and Mooney 2010; Thater, Fürstenau, and Pinkal 2010; Reddy et al. 2011; Van de Cruys, Poibeau, and Korhonen 2011).

2.2 Representation of Word Meaning in Word Sense Inventories

One possible factor contributing to the difficulty of manual and automatic word sense assignment is the design of word sense inventories themselves. As we have seen, such difficulties are encountered across dictionaries, and it has been argued that there are problems with the characterization of word meanings as sets of discrete and mutually exclusive senses (Tuggy 1993; Cruse 1995; Kilgarriff 1997; Hanks 2000; Kintsch 2007).

2.2.1 Criticisms of Enumerative Approaches to Meaning. Dictionaries are practical resources and the nature of the finished product depends upon the needs of the target audience, as well as budgetary and related constraints (cf. Hanks 2000). Consequently, dictionaries differ in the words that they cover, and also in the word meanings that they distinguish. Dictionary senses are generalizations over the meanings that a word can take, and these generalizations themselves are abstractions over collected occurrences of the word in different contexts (Kilgarriff 1992, 1997, 2006). Regardless of a dictionary's granularity, the possibility exists for some amount of detail to be lost as a result of this process.

Kilgarriff (1997) calls into question the possibility of general, all-purpose senses of a word and argues that sense distinction only makes sense with respect to a given task. For example, in machine translation, the senses to be distinguished should be those that lead to different translations in the target language. It has since been demonstrated that this is in fact the case (Carpuat and Wu 2007a, 2007b). Hanks (2000) questions the view of senses as disjoint classes defined by necessary and sufficient conditions. He shows that even with a classic homonym like "bank," some occurrences are more typical examples of a particular sense than others. This notion of typicality is also important in theories of concept representation in psychology (Murphy 2002). Theoretical treatments of word meaning such as the Generative Lexicon (Pustejovsky 1991) also draw attention to the subtle, yet reliable, fluctuations of meaning-in-context, and work in this paradigm also provides evidence that two senses which may appear to be quite distinct can in fact be quite difficult to distinguish in certain contexts (Copestake and Briscoe 1995, page 53).

2.2.2 Psychological Research on Lexical and Conceptual Knowledge. Not all members of a mental category are equal. Some are perceived as more typical than others (Rosch 1975; Rosch and Mervis 1975; and many others), and even category membership itself is clearer in some cases than in others (Hampton 1979). These results are about mental concepts, however, rather than word meanings per se, which raises the question of the relation between word meanings and conceptual knowledge. Murphy (1991, 2002) argues that although not every concept is associated with a word, word meanings show many of the same phenomena as concepts in general—word meaning is "made up of pieces of conceptual structure" (Murphy 2002, page 391). A body of work in cognitive linguistics also discusses the relation between word meaning and conceptual structure (Coleman and Kay 1981; Taylor 2003).

Psycholinguistic studies on word meaning offer insight into the question of the mental representation of word senses. Unlike homonym meanings, the senses of a polysemous word are thought to be related, suggesting that the mental representations of these senses may overlap as well. The psycholinguistic literature on this question

is not wholly clear-cut, but by and large does support the position that polysemous senses are not entirely discrete in the mental lexicon. Whereas Klein and Murphy (2001, 2002) do provide evidence for discreteness of mental sense representations, it appears as though these findings may be due in part to the particular senses included in their studies (Klepousniotou, Titone, and Romero 2008).

Moreover, many psycholinguistic studies have indeed found evidence for processing differences between homonyms and polysemous words, using a variety of experimental designs, including eye movements and reading times (Frazier and Rayner 1990; Pickering and Frisson 2001) as well as response times in sensicality and lexical decision tasks (Williams 1992; Klepousniotou 2002). Brown (2008, 2010) takes the question of shared vs. separate meaning representations one step further in a semantic priming study[6] in which she shows that intuitive meaning-in-context similarity judgments have a processing correlate in on-line sentence comprehension. Response time to the target is a negative linear function of its similarity in meaning to the prime, and response accuracy is a positive linear function of this similarity. In other words, the more similar in meaning a prime–target pair was judged to be, the faster and more accurately subjects responded. This provides empirical support for a processing correlate of graded similarity-in-meaning judgments.

In our work reported here, we take inspiration from work in psychology and look at ways to model word meaning more continuously. Even though there is still some controversy, the majority of studies support the view that senses of polysemous words are linked in their mental representations. In our work we do not make an explicit distinction between homonymy and polysemy, but the data sets we have produced may be useful for a future exploration of this distinction.

2.3 Alternative Approaches to Word Meaning

Earlier we suggested that word meaning may be better described without positing disjoint senses. We now describe some alternatives to word sense inventory approaches to word meaning, most of which do not rely on disjoint senses.

2.3.1 Substitution-Based Approaches. McCarthy and Navigli (2007) explore the use of synonym or near-synonym lexical substitutions to characterize the meaning of word occurrences. In contrast to dictionary senses, substitutes are not taken to partition a word's meaning into distinct senses. McCarthy and Navigli gathered their lexical substitution data using multiple annotators. Annotators were allowed to provide up to three paraphrases for each item. Data were gathered for 10 sentences per lemma for 210 lemmas, spanning verbs, nouns, adjectives, and adverbs. The annotation took the form of each occurrence being associated with a multiset of supplied paraphrases, weighted by the frequency with which each paraphrase was supplied. We make extensive use of the LEXSUB dataset in our work reported here. An example sentence with substitutes from the LEXSUB dataset (sentence 451) is given in Table 3.

A related approach also characterizes meaning through equivalent terms, but terms in another language. Resnik and Yarowsky (2000, page 10) suggest "to restrict a word sense inventory to distinctions that are typically *lexicalized cross-linguistically*" [emphasis in original]. They argue that such an approach will avoid being too fine-grained, and that the distinctions that are made will be independently motivated by crosslinguistic

6 See McNamara (2005) for more information on priming studies.

Table 3
An example of annotation from the lexical substitution data set: sentence 451.

Sentence:	My interest in Europe's defence policy is nothing **new**.
Annotation:	original 2; recent 2; novel 2; different 1; additional 1

trends. Although substitution and translation methods are not without their own issues (Kilgarriff 1992, page 48), they constitute an approach to word meaning that avoids many of the drawbacks of more traditional sense distinction and annotation. Some cross-linguistic approaches group translations into disjoint senses (Lefever and Hoste 2010), whereas others do not (Mihalcea, Sinha, and McCarthy 2010).

2.3.2 Distributional Approaches. Recently there have been a growing number of distributional approaches to representing word meaning in context. These models offer an opportunity to model subtle distinctions in meaning between two occurrences of a word in different contexts. In particular, they allow comparisons between two occurrences of a word without having to classify them as having the same sense or different senses. Some of these approaches compute a distributional representation for a word across all its meanings, and then adapt this to a given sentence context (Landauer and Dumais 1997; Erk and Pado 2008; Mitchell and Lapata 2008; Thater, Fürstenau, and Pinkal 2010; Van de Cruys, Poibeau, and Korhonen 2011). Others group distributional contexts into senses. This can be done on the fly for a given occurrence (Erk and Pado 2010; Reddy et al. 2011), or beforehand (Dinu and Lapata 2010; Reisinger and Mooney 2010). The latter two approaches then represent an occurrence through weights over those senses. A third group of approaches is based on language models (Deschacht and Moens 2009; Washtell 2010; Moon and Erk 2012): They infer other words that could be used in the position of the target word.[7]

3. Two Novel Annotation Tasks

In this section we introduce two novel annotation schemes that draw on methods common in psycholinguistic experiments, but uncommon in corpus annotation. Traditional word sense annotation usually assumes that there is a single correct label for each markable. Annotators are trained to identify the correct labels consistently, often with highly specific a priori guidelines. Multiple annotators are often used, but despite the frequently low ITA in word sense annotation, differences between annotator responses are often treated as the result of annotator error and are not retained in the final annotation data.

In these respects, traditional word sense annotation tasks differ in design from many psycholinguistic experiments, such as the ones discussed in the previous section. Psycholinguistic experiments frequently do not make strong assumptions about how participants will respond, and in fact are designed to gather data on that very question. Participants are given general guidelines for completing the experiment but these

7 Distributional models for phrases have recently received much attention, even more so than models for word meaning in context (Baroni and Zamparelli 2010; Coecke, Sadrzadeh, and Clark 2010; Mitchell and Lapata 2010; Grefenstette and Sadrzadeh 2011; Socher et al. 2011). They are less directly relevant to the current paper, however, as we focus on eliciting judgments for individual words in sentence contexts, rather than whole phrases.

Table 4
Interpretation of the five-point scale given to the annotators. This interpretation is the same for the Usim and WSsim tasks.

1	completely different
2	mostly different
3	similar
4	very similar
5	identical

guidelines generally stop short of precise procedural detail, to avoid undue influence over participant responses. All of the psycholinguistic studies discussed earlier used participants naïve as to the purpose of the experiment, and who were minimally trained. Responses are often graded in nature, involving ratings on an ordinal scale or in some cases even a continuously valued dimension (e.g., as in Magnitude Estimation). Multiple participants respond to each stimulus, but all participant responses are typically retained, as there are often meaningful discrepancies in participant responses that are not ascribable to error. All of the psycholinguistic studies discussed previously collected data from multiple participants (up to 80 in the case of one experiment by Williams [1992]).

The annotation tasks we present subsequently draw upon these principles of experimental design. We collected responses using a scale, rather than binary judgments; we designed the annotation tasks to be accomplishable without prior training and with minimal guidelines, and we used multiple annotators (up to eight) and retained all responses in an effort to capture individual differences. In the following, we describe two different annotation tasks, one with and one without the use of dictionary senses.

Graded Ratings for Dictionary Senses. In our first annotation task, dubbed WSsim (for Word Sense Similarity), annotators rated the applicability of WordNet dictionary senses, using a five-point ordinal scale.[8] Annotators rated the applicability of every single WordNet sense for the target lemma, where a rating of 1 indicated that the sense in question did not apply at all, and a rating of 5 indicated that the sense applied completely to that occurrence of the lemma. Table 4 shows the descriptions of the five points on the scale that the annotators were given. By asking annotators to provide ratings for each individual sense, we strive to eliminate all bias toward either single-sense or multiple-sense annotation. By asking annotators to provide ratings on a scale, we allow for the fact that senses may not be perceived in a binary fashion.

Graded Ratings for Usage Similarity. In our second annotation task, dubbed Usim (for Usage Similarity), we collected annotations of word usages without recourse to dictionary senses, by asking annotators to judge the similarity in meaning of one usage of a lemma to other usages. Annotators were presented with pairs of contexts that share a word in common, and were asked to rate how similar in meaning they perceive those two occurrences to be. Ratings are again on a five-point ordinal scale; a rating of 1 indicated that the two occurrences of the target lemma were completely dissimilar in meaning, and a rating of 5 indicated that the two occurrences of the target lemma were identical in meaning. The descriptions of the five points on the scale, shown in Table 4,

8 The use of a five-point scale is a common choice when collecting ordinal ratings, as it allows more detailed responses than the "yes/no/maybe" provided by a three-point scale.

were identical to those used in the WSsim task. Annotators were able to respond "I don't know" if they were unable to gauge the similarity in meaning of the two occurrences.[9]

Annotation Procedure. All annotation for this project was conducted over the Internet in specially designed interfaces. In both tasks, all annotator responses were retained, without resolution of disagreement between annotators. We do not focus on obtaining a single "correct" annotation, but rather view all responses as valuable sources of information, even when they diverge.

For each item presented, annotators additionally were provided a comment field should they desire to include a more detailed response regarding the item in question. They could use this, for example, to comment on problems understanding the sentence. The annotators were able to revisit previous items in the task. Annotators were not able to skip forward in the task without rating the current item. If an annotator attempted to submit an incomplete annotation they were prompted to provide a complete response before proceeding. They were free to log out and resume later at any point, however, and also could access the instructions whenever they wanted.

Two Rounds of Annotation. We performed two rounds of the annotation experiments, hereafter referred to as R1 and R2.[10] Both annotation rounds included both a WSsim and a Usim task, labeled in the subsequent discussion as WSsim-1 and Usim-1 for R1, and WSsim-2 and Usim-2 for R2. An important part of the data analysis is to compare the new, graded annotation to other types of annotation. We compare it to both traditional word sense annotation, with a single best sense for each occurrence, and lexical substitution, which characterizes each occurrence through paraphrases. In R1, we chose annotation data that had previously been labeled with either traditional single sense annotation or with lexical substitutions. R2 included two additional annotation tasks, one involving traditional WSD methodology (WSbest) and a lexical substitution task (SYNbest). In the SYNbest task, annotators provided a single best lexical substitution, in contrast to the multiple substitutes annotators provided in the original LEXSUB data.[11]

Three annotators participated in each task in the R1, and eight annotators participated in R2. In R1, separate groups of annotators participated in WSsim and Usim annotation, whereas in R2 the same group of annotators was used for all annotation, so as to allow comparison across tasks for the same annotator as well as across annotators. In R2, therefore, the same annotators did both traditional word sense annotation (WSbest) and the graded word sense annotation of the WSsim task. This raises the question of whether their experience on one task will influence their annotation choice on the other task. We tested this by varying the order in which annotators did WSsim and WSbest. R2 annotators were divided into two groups of four annotators with the order of tasks as follows:

group 1: Usim-2	SYNbest	WSsim-2	WSbest
group 2: Usim-2	SYNbest	WSbest	WSsim-2

Another difference between the two rounds of annotation was that in R2 we permitted the annotators to see one more sentence of context on either side of the target

9 The "I don't know" option was present only in the Usim interface, and was not available in WSsim.

10 The annotation was conducted in two separate rounds due to funding.

11 Annotation guidelines for R1 are at `http://www.katrinerk.com/graded-sense-and-usage-annotation` and guidelines for R2 tasks are at `http://www.dianamccarthy.co.uk/downloads/WordMeaningAnno2012/`.

Table 5
Abbreviations used in the text for annotation tasks and rounds.

WSsim	Task: graded annotation of WordNet senses on a five-point scale
Usim	Task: graded annotation of usage similarity on a five-point scale
WSbest	Task: traditional single-sense annotation
SYNbest	Task: lexical substitution
R1	Annotation round 1
R2	Annotation round 2

sentence. In R1 each item was given only one sentence as context. We added more context in order to reduce the chance that the sentence would be unclear. Table 5 summarizes up the annotation tasks and annotation rounds on which we report.

Data Annotated. The data to be annotated in WSsim-1 were taken primarily from Semcor (Miller et al. 1993) and the Senseval-3 English lexical sample (SE-3) (Mihalcea, Chklovski, and Kilgarriff 2004). This experiment contained a total of 430 sentences spanning 11 lemmas (nouns, verbs, and adjectives). For eight of these lemmas, 50 sentences were included, 25 randomly sampled from Semcor and 25 randomly sampled from SE-3. The remaining three lemmas in the experiment had 10 sentences each, from the LEXSUB data. Each of the three annotators annotated each of the 430 items, providing a response for each WordNet sense for that lemma. Usim-1 used data from LEXSUB. Thirty-four lemmas were manually selected, including the three lemmas also used in WSsim-1. We selected lemmas which exhibited a range of meanings and substitutes in the LEXSUB data, with as few multiword substitutes as possible. Each lemma is the target in 10 LEXSUB sentences except there were only nine sentences for the lemma *bar.n* because of a part-of-speech tagging error in the LEXSUB trial data. For each lemma, annotators were presented with every pairwise comparison of these 10 sentences. We refer to each such pair as an SPAIR. There were 45 SPAIRs per lemma (36 for *bar.n*), adding up to 1,521 comparisons per annotator in Usim-1.

In R1, only 30 sentences were included in both WSsim and Usim. Because comparison of annotator responses on this subset of the two tasks yielded promising results, R2 used the same set of sentences for both Usim and WSsim so as to better compare these tasks. All data in the second round were taken from LEXSUB, and contained 26 lemmas with 10 sentences for each. We produced the SYNbest annotation, rather than use the existing LEXSUB annotation, so that we could ensure the same conditions as with the other annotation tasks, that is, using the same annotators and providing the extra sentence of context on either side of the original LEXSUB context. We also only required that the annotators provide one substitute. As such, there were 260 target lemma occurrences that received graded word sense applicability ratings in WSsim-2, and 1,170 SPAIRs (pairs of occurrences) to be annotated in Usim-2.

4. Analysis of the Annotation

In this section we present our analysis of the annotated data. We test inter-annotator agreement, and we test to what extent annotators make use of the added flexibility of the graded annotation. We also compare the outcome of our graded annotation to traditional word sense annotation and lexical substitutions for the same data.

4.1 Evaluation Measures

Because both graded annotation tasks, WSsim and Usim, use ratings on five-point scales rather than binary ratings, we measure agreement in terms of correlation. Because ratings were not normally distributed, we choose a non-parametric test which uses ranks rather than absolute values: We use Spearmans rank correlation coefficient (rho), following Mitchell and Lapata (2008). For assessing inter-tagger agreement on the R2 WSbest task we adopt the standard WSD measure of average pairwise agreement, and for R2 SYNbest, we use the same pairwise agreement calculation used in LEXSUB.

When comparing graded ratings with single-sense or lexical substitution annotation, we use the mean of all annotator ratings in the WSsim or Usim annotation. This is justified because the inter-annotator agreement is highly significant, with respectable rho compared with previous work (Mitchell and Lapata 2008).

As the annotation schemes differ between R1 and R2 (as mentioned previously, the number of annotators and the amount of visible context are different, and R2 annotators did traditional word sense annotation in the WSbest task in addition to the graded tasks) we report the results of R1 and R2 separately.[12]

4.2 WSsim: Graded Ratings for WordNet Senses

In the WSsim task, annotators rated the applicability of each sense of the target word on a five-point scale. We first do a qualitative analysis, then turn to a quantitative analysis of annotation results.

4.2.1 Qualitative Analysis. Table 6 shows an example of WSsim annotation. The target is the verb *dismiss*, which was annotated in R2. The first column gives the WordNet sense number (sn).[13] Note that in the task, the annotators were given the synonyms and full description but in this figure we only supply part of the description for the sake of space. As can be seen, three of the annotators chose a single-sense annotation by giving a rating of 5 to one sense and ratings of 1 to all others. Two annotators gave ratings of 1 and 2 to all but one sense. The other three annotators gave positive ratings (ratings of at least 3 [similar], see Table 4) to at least two of the senses. All annotators agree that the first sense fits the usage perfectly, and all annotators agree that senses 3 and 5 do not apply. The second sense, on the other hand, has an interestingly wide distribution of judgments, ranging from 1 to 4. This is the judicial sense of the verb, as in 'this case is dismissed.' Some annotators consider this sense to be completely distinct from sense 1, whereas others see a connection. There is disagreement among annotators, about sense 6. This is the sense 'dismiss, dissolve,' as in 'the president dissolved the parliament.' Six of the annotators consider this sense completely unrelated to 'dismiss our actions as irrelevant,' whereas two annotators view it as highly related (though not completely identical). It is noteworthy that each of the two opinions, a rating of 1

12 It is known that when responses are collected on an ordinal scale, the possibility exists for different individuals to use the scale differently. As such, it is common practice to standardize responses using a z-score, which maps a response X to $z = \frac{X-\mu}{\sigma}$. The calculation of z-scores makes reference to the mean and the standard deviation of an annotator's responses. Because responses were not normally distributed in our task, a transformation that relies on measures of central tendency is not appropriate. So we do not use z-scores in this paper. We repeated all analyses with z-score transform anyway, and found the results to be basically the same as those we report here with the raw values. Overall, using z-scores slightly strengthened most findings, but there were no differences in statistical significance anywhere.

13 We use WordNet 3.0 for our annotation.

Table 6
WSsim example, R2: Annotator judgments for the different senses of *dismiss*.

If we see ourselves as separate from the world, it is easy to **dismiss** our actions as irrelevant or unlikely to make any difference. (902)

sn	Description	Ratings By Annotator								Mean
1	bar from attention or consideration	5	5	5	5	5	5	5	5	5
2	cease to consider	1	4	1	3	2	2	1	3	2.125
3	stop associating with	1	2	1	1	1	2	1	1	1.25
4	terminate the employment of	1	4	1	2	1	1	1	1	1.5
5	cause or permit a person to leave	1	2	1	1	1	1	1	2	1.25
6	declare void	1	1	1	4	1	1	1	4	1.75

and a rating of 4, was chosen by multiple annotators. Because multiple annotators give each judgment, these data seem to reflect a genuine difference in perceived sense. We discuss inter-annotator agreement, both overall and considering individual annotators, subsequently.

Table 7 gives an example sentence from R1, where the annotated target is the noun *paper*. All annotators agree that sense 5, 'scholarly article,' applies fully. Sense 2 ('essay') also gets ratings of ≥ 3 from all annotators. The first annotator seems also to have perceived the 'physical object' connotation to apply strongly to this example, and has expressed this quite consistently by giving high marks to sense 1 as well as 7.

Table 8 shows a sample annotated sentence with an adjective target, *neat*, annotated in R2. In this case, only one annotator chose single-sense annotation by marking exclusively sense 4. One annotator gave ratings ≥ 3 (similar) to *all* senses of the lemma. All other annotators saw at least two senses as applying (with ratings ≥ 3) and at least one sense as not applying at all (with a rating of 1). Sense 4 has received positive ratings (that is, ratings ≥ 3) throughout. Senses 1, 2, and 6 have mixed ratings, and senses 3 and 5 have positive ratings only from the one annotator who marked everything as applying. Interestingly, ratings for senses 1, 2, and 6 diverge sharply, with some annotators seeing them as not applying at all, and some giving them ratings in the 3–5 range. Note that the

Table 7
WSsim example, R1: Annotator judgments for the different senses of *paper*.

This can be justified thermodynamically in this case, and this will be done in a separate **paper** which is being prepared. (br-j03, sent. 4)

sn	Description	Ratings			Mean
1	a material made of cellulose pulp	4	1	1	1.3
2	an essay (especially one written as an assignment)	3	3	5	3.7
3	a daily or weekly publication on folded sheets; contains news and articles and advertisements	2	1	3	2
4	a medium for written communication	5	3	1	3
5	a scholarly article describing the results of observations or stating hypotheses	5	5	5	5
6	a business firm that publishes newspapers	2	1	1	1.3
7	the physical object that is the product of a newspaper publisher	4	1	1	1.7

Table 8
WSsim example, R2: Annotator judgments for the different senses of *neat*.

Over the course of the 20th century scholars have learned that such images tried to make messy reality **neater** than it really is (103)

sn	Description	Ratings By Annotator								Mean
1	free from clumsiness; precisely or deftly executed	1	5	1	4	5	5	5	5	3.375
2	refined and tasteful in appearance or behavior or style	3	4	1	4	4	3	1	3	2.875
3	having desirable or positive qualities especially those suitable for a thing specified	1	3	1	1	1	1	1	1	1.25
4	marked by order and cleanliness in appearance or habits	4	5	5	3	4	5	5	5	4.5
5	not diluted	1	4	1	1	1	1	1	1	1.375
6	showing care in execution	1	4	1	3	4	1	3	3	2.5

Table 9
Correlation matrix for pairwise correlation agreement for WSsim-1. The last row provides the agreement of the annotator in that column against the average from the other annotators.

	A	B	C
A	1.00	0.47	0.51
B	0.47	1.00	0.54
C	0.51	0.54	1.00
against avg	0.56	0.58	0.61

annotators who give ratings of 1 are not the same for these three ratings, pointing to different, but quite nuanced, judgments of the 'make reality neater' usage in this sentence.

4.2.2 Inter-annotator Agreement. We now turn to a quantitative analysis, starting with inter-annotator agreement. For the graded WSsim annotation, it does not make sense to compute the percentage of perfect agreement. As discussed earlier, we report inter-annotator agreement in terms of correlation, using Spearman's rho. We calculate pairwise agreements and report the average over all pairs. The pairwise correlations are shown in the matrix in Table 9. We have used capital letters to represent the individuals, preserving the same letter for the same person across tasks. In the last row we show agreement of each annotator's judgments against the average judgment from the other annotators. The pairwise correlations range from 0.47 to 0.54 and all pairwise correlations were highly significant ($p \ll 0.001$), with an average of rho = 0.504. This is a very reasonable result given that Mitchell and Lapata (2008) report a rho of 0.40 on a graded semantic similarity task.[14] The lowest correlation against the average

14 Direct comparison across tasks is not appropriate, but we wish to point out that for graded semantic judgments this level of correlation is perfectly reasonable. The Mitchell and Lapata (2008) data set has been used in an evaluation exercise (GEMS-2011, `https://sites.google.com/site/geometricalmodels/shared-evaluation`). Mitchell and Lapata point out that Spearman's rho tends to yield lower coefficients compared with parametric alternatives such as Pearson's.

Table 10
Correlation matrix for pairwise correlation agreement for WSsim-2. The last row provides the agreement of the annotator in that column against the average from the other annotators.

	A	C	D	F	G	H	I	J
A	1.00	0.55	0.58	0.60	0.61	0.63	0.61	0.59
C	0.55	1.00	0.54	0.66	0.57	0.55	0.65	0.52
D	0.58	0.54	1.00	0.55	0.58	0.52	0.56	0.54
F	0.60	0.66	0.55	1.00	0.62	0.62	0.72	0.59
G	0.61	0.57	0.58	0.62	1.00	0.63	0.62	0.62
H	0.63	0.55	0.52	0.62	0.63	1.00	0.64	0.64
I	0.61	0.65	0.56	0.72	0.62	0.64	1.00	0.58
J	0.59	0.52	0.54	0.59	0.62	0.64	0.58	1.00
against avg	0.70	0.58	0.62	0.64	0.70	0.71	0.66	0.71

from the other annotators was 0.56. We discuss the annotations of individuals in Section 4.6, including our decision to retain the judgments of all annotators for our gold standard.

From the correlation matrix in Table 10 we see that for WSsim-2, pairwise correlations ranged from 0.52 to 0.72. The average value of the pairwise correlations was rho $= 0.60$, and again every pair was highly significant ($p \ll 0.001$). The lowest correlation against the average from all the other annotators was 0.58.

4.2.3 Choice of Single Sense Versus Multiple Senses. In traditional word sense annotation, annotators can mark more than one sense as applicable, but annotation guidelines often encourage them to view the choice of a single sense as the norm. In WSsim, annotators gave ratings for all senses of the target. So we would expect that in WSsim, there would be a higher proportion of senses selected as applicable. Indeed we find this to be the case: Table 11 shows the proportion of sentences where some annotator has assigned more than one sense with a judgment of 5, the highest value. Both WSsim-1 and WSsim-2 have a much higher proportion of sentences with multiple senses chosen than the traditional sense-annotated data sets SemCor and SE-3. Interestingly, we notice that the percentage for WSsim-1 is considerably higher than for WSsim-2. In principle, this could be due to differences in the lemmas that were annotated, or differences in the sense perception of the annotators between R1 and R2. Another potential influencing

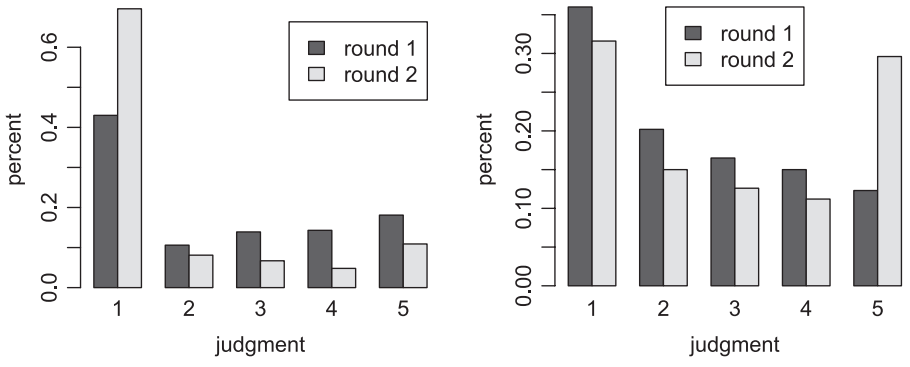
Table 11
WSsim annotation: Proportion of sentences where multiple senses received a rating of 5 (highest judgment) from the same annotator.

	Proportion
WSsim-1	46%
WSsim-2	30%
WSsim-2, WSsim first	36%
WSsim-2, WSbest first	23%
SemCor	0.3%
SE-3	8%

factor is the order of annotation experiments: As described earlier, half of the R2 annotators did WSbest annotation before doing WSsim-2, and half did the two experiments in the opposite order. As Table 11 shows, those doing the graded task WSsim-2 before the binary task WSbest had a greater proportion of multiple senses annotated with the highest response. This demonstrates that annotators in a word meaning task can be influenced by factors outside of the current annotation task, in this case another annotation task that they have done previously. We take this as an argument in favor of using as many annotators as possible in order to counteract factors that contribute noise. In our case, we counter the influence of previous annotation tasks somewhat by using multiple annotators and altering the order of the WSsim and WSbest tasks. Another option would have been to use different annotators for different tasks; by using the same set of annotators for all four tasks, however, we can better control for individual variation.

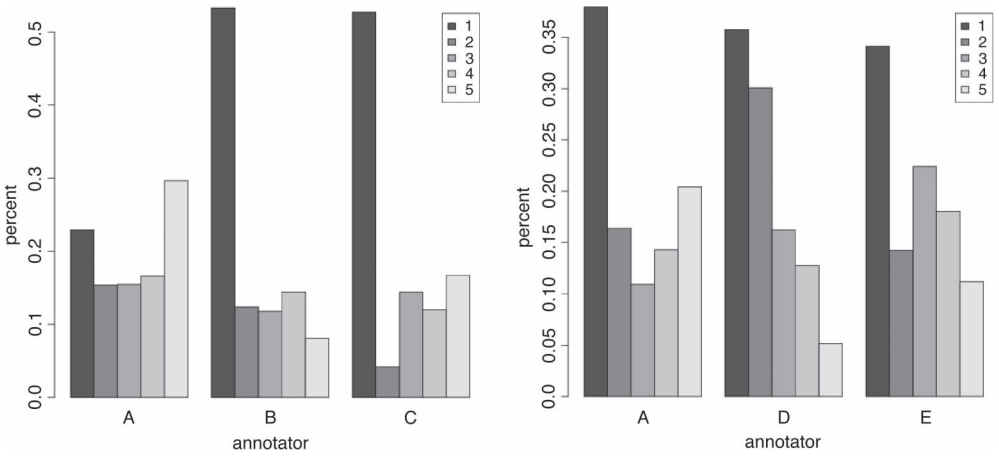
4.2.4 Use of the Graded Scale. We next ask whether annotators in WSsim made use of the whole five-point scale, or whether they mostly chose the extreme ratings of 1 and 5. If the latter were the case, this could indicate that they viewed the task of word sense assignment as binary. Figure 1a shows the relative frequency distribution of responses from all annotators over the five scores for both R1 and R2. Figures 2a and 3a show the same but for each individual annotator. In both rounds the annotators chose the rating of 1 ('completely different,' see Table 4) most often. This is understandable because each item is a sentence and sense combination and there will typically be several irrelevant senses for a given sentence. The second most frequent choice was 5 ('identical'). Both rounds had plenty of judgments somewhere between the two poles, so the annotators do not seem to view the task of assigning word sense as completely binary. Although the annotators vary, they all use the intermediate categories to some extent and certainly the intermediate category judgments do not originate from a minority of annotators.

We notice that R2 annotators tended to give more judgments of 1 ('completely different') than the R1 annotators. One possible reason is again that half our annotators did WSbest before WSsim-2. If this were the cause for the lower judgments, we would expect more ratings of 1 for the annotators who did the traditional word sense annotation (WSbest) first. In Table 12 we list the relative frequency of each rating for the different groups of annotators. We certainly see an increase in the judgments of 1 where



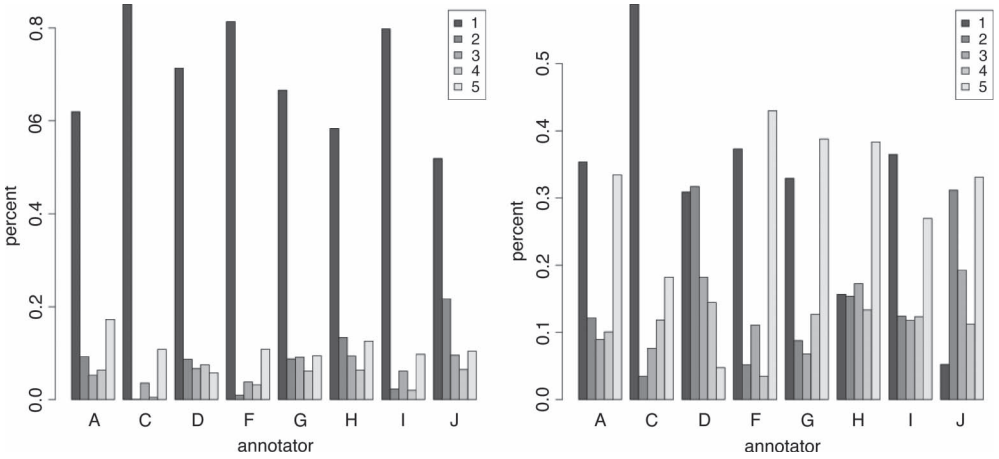
(a) WSsim (b) Usim

Figure 1
WSsim and Usim R1 and R2 ratings.



(a) WSsim (b) Usim

Figure 2
WSsim and Usim R1 individual ratings.



(a) WSsim (b) Usim

Figure 3
WSsim and Usim R2 individual ratings.

WSbest is performed before WSsim-2. Again, this may indicate that annotators were leaning more towards finding a single exact match because they were influenced by the WSbest task they had done before. Annotators in that group were also slightly less inclined to take the middle ground, but this was true of both groups of R2 annotators compared with the R1 annotators. We think that this difference between the two rounds may well be due to the lemmas and data.

In Table 18, we show the average range[15] and average variance of the judgments per item for each of the graded annotation tasks. WSsim naturally has less variation

15 As an example, the first two senses (1 and 2) in Table 6 have ranges of 0 and 3, respectively.

Table 12
The relative frequency of the annotations at each judgment from all annotators.

	Judgment				
Exp	1	2	3	4	5
WSsim-1	0.43	0.106	0.139	0.143	0.181
WSsim-2	0.696	0.081	0.067	0.048	0.109
WSsim-2, WSsim first	0.664	0.099	0.069	0.048	0.12
WSsim-2, WSbest first	0.727	0.063	0.065	0.048	0.097
Usim-1	0.360	0.202	0.165	0.150	0.123
Usim-2	0.316	0.150	0.126	0.112	0.296

compared with Usim because, for any sentence, there are inevitably many WordNet senses which are irrelevant to the context at hand and which will obtain a judgment of 1 from everyone. This is particularly the case for WSsim-2 where the annotators gave more judgments of 1, as discussed previously. The majority of items have a range of less than two for WSsim. We discuss the Usim figures further in the following section.

4.3 Usim: Graded Ratings for Usage Similarity

In Usim annotation, annotators compared pairs of usages of a target word (SPAIRs) and rated their similarity on the five-point scale given in Table 4. The annotators were also permitted a response of "don't know." Such responses were rare but were used when the annotators really could not judge usage similarity, perhaps because the meaning of one sentence was not clear. We removed any pairs where one of the annotators had given a "don't know" verdict (9 in R1, 28 in R2). For R1 this meant that we were left with a total of 1,512 SPAIRs and in R2 we had a resultant 1,142 SPAIRs.

4.3.1 Qualitative Analysis. We again start by inspecting examples of Usim annotation. Table 13 shows the annotation for an SPAIR of the verb *dismiss*. The first of the two sentences talks about "dismissing actions as irrelevant," the second is about dismissing a person. Interestingly, the second usage could be argued to carry both a connotation of 'ushering out' and a connotation of 'disregarding.' Annotator opinions on this SPAIR vary from a 1 (completely different) to a 5 (identical), but most annotators seem to view the two usages as related to an intermediate degree. This is adequately reflected in the average rating of 3.125. Table 14 compares the sentence from Table 8 to another sentence

Table 13
Usim example: Annotator judgments for a pair of usages of *dismiss*.

Sentences	Ratings
If we see ourselves as separate from the world, it is easy to **dismiss** our actions as irrelevant or unlikely to make any difference.	1, 2, 3, 3, 3, 4, 4, 5
Simply thank your Gremlin for his or her opinion, **dismiss** him or her, and ask your true inner voice to turn up its volume.	

Table 14
Usim example: Annotator judgments for a pair of usages of *neat*.

Sentences	Ratings
Over the course of the 20th century scholars have learned that such images tried to make messy reality **neater** than it really is.	3, 3, 4, 4, 4, 4, 5, 5
Strong field patterns created by hedgerows give the landscape a **neat**, well structured appearance.	

Table 15
Usim example: Annotator judgments for a pair of usages of *account*.

Sentences	Ratings
Samba-3 permits use of multiple **account** data base backends.	1, 2, 3, 3, 3, 4, 4, 4
Within a week, Scotiabank said that it had frozen some **accounts** linked to Washington's hit list.	

with the target *neat*. The first sentence is a metaphorical use (making reality neater), the second is literal (landscape with neat appearance), but still the SPAIR gets high ratings of 3–5 throughout for an average of 4.0. Note that the WordNet senses, shown in Table 8, do not distinguish the literal and metaphorical uses of the adjective, either. Table 15 shows two uses of the noun *account*. The first pertains to accounts on a software system, the second to bank accounts. The spread of annotator ratings shows that these two uses are not the same, but that some relation exists. The average rating for this SPAIR is 3.0.

4.3.2 Inter-annotator Agreement. We again calculate inter-annotator agreement as the average over pairwise Spearman's correlations. The pairwise correlations are shown in the matrix in Table 16. In the last row we show agreement of each annotator's judgments against the average judgment from the other annotators. For Usim-1 the range of correlation coefficients is between 0.50 and 0.64 with an average correlation of rho = 0.548. All the pairs are highly significantly correlated ($p \ll 0.001$). The smallest correlation for any individual against the average is 0.55. The correlation matrix for Usim-2 is provided in Table 17; the range of correlation coefficients is between 0.42 and

Table 16
Correlation matrix for pairwise correlation agreement for Usim-1. The last row provides the agreement of the annotator in that column against the average from the other annotators.

	A	D	E
A	1.00	0.50	0.64
D	0.50	1.00	0.50
E	0.64	0.50	1.00
against avg	0.67	0.55	0.67

Table 17
Correlation matrix for pairwise correlation agreement for Usim-2. The last row provides the agreement of the annotator in that column against the average from the other annotators.

	A	C	D	F	G	H	I	J
A	1.00	0.70	0.52	0.70	0.69	0.72	0.73	0.67
C	0.70	1.00	0.48	0.72	0.60	0.66	0.71	0.69
D	0.52	0.48	1.00	0.48	0.49	0.51	0.50	0.42
F	0.70	0.72	0.48	1.00	0.66	0.71	0.74	0.68
G	0.69	0.60	0.49	0.66	1.00	0.71	0.65	0.62
H	0.72	0.66	0.51	0.71	0.71	1.00	0.70	0.65
I	0.73	0.71	0.50	0.74	0.65	0.70	1.00	0.72
J	0.67	0.69	0.42	0.68	0.62	0.65	0.72	1.00
against avg	0.82	0.78	0.58	0.80	0.76	0.80	0.81	0.76

0.73. All these correlations are highly significant ($p \ll 0.001$) with an average correlation of rho = 0.62. The lowest agreement between any individual and the average judgment of the others is 0.58. Again, we note that these are all respectable values for tasks involving semantic similarity ratings.

Use of the graded scale. Figure 1b shows how annotators made use of the graded scale in Usim-1 and Usim-2. It graphs the relative frequency of each of the judgments on the five-point scale. Figures 2b and 3b show the same but for each individual annotator. In both annotation rounds, the rating 1 (completely different) was chosen most frequently. There are also in both annotation rounds many ratings in the middle points of the scale, indeed we see a larger proportion of mid-range scores for Usim than for WSsim in general, as shown in Table 12. Figures 2b and 3b show that although individuals differ, all use the mid points to some extent and it is certainly not the case that these mid-range judgments come from a minority of annotators. In Usim, annotators compared pairs of usages, whereas in WSsim, they compared usages with sense definitions. The sense definitions suggest a categorization that may bias annotators towards categorical choices. Comparing the two annotation rounds for Usim, we see that in Usim-2 there seem to be many more judgments at 5 than in Usim-1. This is similar to our findings for WSsim, where we also obtained more polar judgments for R2 than for R1.

There is a larger range on average for Usim-2 compared with the other tasks as shown earlier by Table 18. This is understandable given that there are eight annotators

Table 18
Average range and average variance of judgments for each of the graded experiments.

	avg range	avg variance
WSsim-1	1.78	1.44
WSsim-2	1.55	0.71
Usim-1	1.41	0.92
Usim-2	2.50	1.12

for R2 compared with R1,[16] and so a greater chance of a larger range per item. There is substantial variation by lemma. In Usim-2, *fire.v*, *rough.a*, and *coach.n* have an average range of 1.33, 1.76, and 1.93, respectively, whereas *suffer.v*, *neat.a*, and *function.n* have average ranges of 3.14, 3.16, and 3.58, respectively. The variation in range appears to depend on the lemma rather than POS. This variation can be viewed as a gauge of how difficult the lemma is. Although the range is larger in Usim-2, however, the average variance per item (i.e., the variance considering the eight annotators) is 1.12 and lower than that for WSsim-1.

Usim and the triangle inequality. In Euclidean space, the lengths of two sides of a triangle, taken together, must always be greater than the length of the third side. This is the *triangle inequality*:

$$\text{length(longest)} < \text{length(second\_longest)} + \text{length(shortest)}$$

We now ask whether the triangle inequality holds for Usim ratings. If Usim similarities are metric, that is, if we can view the ratings as proximity in a Euclidean "meaning space," then the triangle inequality would have to hold. This question is interesting for what it says about the psychology of usage similarity judgments. Classic results due to Tversky and colleagues (Tversky 1977; Tversky and Gati 1982) show that human judgments of similarity are not always metric. Tversky (1977), varying an example by William James, gives the following example, which involves words, but explicitly ignores context:

> Consider the similarity between countries: Jamaica is similar to Cuba (because of geographical proximity); Cuba is similar to Russia (because of their political affinity); but Jamaica and Russia are not similar at all. [...] the perceived distance of Jamaica to Russia exceeds the perceived distance of Jamaica to Cuba, plus that of Cuba to Russia—contrary to the triangle inequality.

Note, however, that Tversky was considering similarity judgments for different words, whereas we look at different usages of the same word. The question of whether the triangle inequality holds for Usim ratings is also interesting for modeling reasons. Several recent approaches model word meaning in context through points in vector space (Erk and Pado 2008; Mitchell and Lapata 2008; Dinu and Lapata 2010; Reisinger and Mooney 2010; Thater, Fürstenau, and Pinkal 2010; Washtell 2010; Van de Cruys, Poibeau, and Korhonen 2011). They work on the tacit assumption that similarity of word usages is metric—an assumption that we can directly test here. Third, the triangle inequality question is also relevant for future annotation; we will discuss this in more detail subsequently.

To test whether Usim ratings obey the triangle inequality, we first convert the *similarity* ratings that the annotators gave to *dissimilarity* ratings: Let s_{avg} be the mean similarity rating over all annotators, then we use the dissimilarity rating $d = 6 - s_{avg}$ (as 5 was the highest possible similarity score).

We examine the proportion of sentence triples where the triangle inequality holds (that is, we consider every triple of sentences that share the same target lemma). In those

16 A likely reason for the larger range in WSsim-1 compared with WSsim-2 is that in WSsim-2 half the annotators had performed WSbest before WSsim-2 and produced more judgments of 1 compared with WSsim-1.

cases where the triangle inequality is violated, we also assess the degree to which it is violated, calculated as the average distance that is missed: Let T_{miss} be the set of triples for which the triangle inequality does not hold, then we compute

$$m = \frac{1}{|T_{\text{miss}}|} \sum_{t \in T_{miss}} \text{length}(\text{longest}_t) - (\text{length}(\text{second\_longest}_t) + \text{length}(\text{shortest}_t))$$

This is the average amount by which the longest side is "too long."

For the first round of annotation, Usim-1, we found that 99.2% of the sentence triples obey the triangle inequality. For the triples that miss it, the average amount by which the longest side is too long is $m = 0.520$. This is half a point on the five-point rating scale, a low amount. In R2, all sentence triples obey the triangle inequality. One potential reason for this is that we have eight annotators for R2, and a larger sample of annotators reduces the variation from individuals. Another reason may be that the annotators in R2 could view two more sentences of context than those in R1.

Tables 19 and 20 show results of the triangle inequality analysis, but by individual annotator. Every annotator has at least 93% of sentence triples obeying the principle. For the triples that miss it, they tend to miss it by between one and two points. The results for individuals accord with the triangle inequality principle, though to a lesser extent compared with the analysis using the average, which reduces the impact of variation from individuals.

As discussed previously, this result (that the triangle inequality holds for Usim annotation triples) is interesting because it contrasts with Tversky's findings (Tversky 1977; Tversky and Gati 1982) that similarity ratings between different words are not metric. And although we consider similarity ratings for usages of the same word, not different words, we would argue that our findings point to the importance of considering the context in which a word is used. It would be interesting to test whether similarity ratings for different words, when used in context, obey the triangle inequality. To reference the Tversky example, and borrowing some terminology from Cruse, evoking the ISLAND facet of Jamaica and Cuba versus the COMMUNIST_STATE facet of Cuba and Russia would account for the non-metricality of the similarity judgments as Tversky

Table 19
Triangle inequality analysis by annotator, Usim-1.

average	A	D	E
perc obey	93.8	97.2	97.3
missed by	1.267	1.221	1.167

Table 20
Triangle inequality analysis by annotator, Usim-2.

	A	C	D	F	G	H	I	J
perc obey	94.1	97.5	98.4	97.2	93.6	97.0	97.4	97.4
missed by	1.508	1.405	1.122	1.824	1.477	1.281	1.759	1.338

Table 21
WSbest annotations.

	sense selected		Proportion with
	no	yes	multiple choice
WSbest	19,599	2,401	0.13
WSbest, WSsim-2 first	9,779	1,221	0.15
WSbest, WSbest first	9,820	1,180	0.11

points out, and moreover highlight the lack of an apt comparison between Jamaica and Russia at all. There is some motivation for this idea in the psychological literature on structural alignment and alignable differences (Gentner and Markman 1997; Gentner and Gunn 2001).

In addition, our finding that the triangle inequality holds for Usim annotation will be useful for future Usim annotation. Usage similarity annotation is costly (and somewhat tedious) as annotators give ratings for each pair of sentences for a given target lemma. Given that we can expect usage similarity to be metric, we can eliminate the need for some of the ratings. Once annotators have rated two usage pairs out of a triple, their ratings set an upper limit on the similarity of the third pair. In the best case, if usages s_1 and s_2 have a distance of 1 (i.e., a similarity of 5), and s_1 and s_3 have a distance of 1, then the distance of s_1 and s_3 can be at most 2. For all usage triples where two pairs have been judged highly similar, we can thus omit obtaining a rating for the third pair. A second option for obtaining more Usim annotation is to use crowdsourcing. In crowdsourcing annotation, quality control is always an issue, and again we can make use of the triangle inequality to detect spurious annotation: Ratings that grossly violate the triangle inequality can be safely discarded.

4.4 WSbest

The WSbest task reflects the traditional methodology in word sense annotation where words are annotated with the best fitting sense. The guidelines[17] allow for selecting more than one sense provided all fit the example equally well. Table 21 shows that, as one would expect given the number of senses in WordNet, there are more unselected senses than selected. We again find an influence of task order: When annotators did the graded annotation (WSsim-2) before WSbest, there were more multiple assignments (see the last column) and therefore more senses selected. This difference is statistically significant (χ^2 test, $p = 0.02$). Regardless of the order of tasks, we notice that the proportion of multiple sense choice is far lower than the equivalent for WSsim (see Table 11), as is expected due to the different annotation schemes and guidelines.

We calculated inter-annotator agreement using pairwise agreement, as is standard in WSD. There are several ways to calculate pairwise agreement in cases of multiple selection, though these details are not typically given in WSD papers. We use the size of the intersection of selections divided by the maximum number of selections from

17 See http://www.dianamccarthy.co.uk/downloads/WordMeaningAnno2012/wsbest.html.

Table 22
Inter-annotator agreement without one individual for WSbest and SYNbest R2.

	average	A	C	D	F	G	H	I	J
WSbest	0.574	0.579	0.564	0.605	0.560	0.582	0.566	0.566	0.568
SYNbest	0.261	0.261	0.259	0.285	0.254	0.256	0.245	0.260	0.267

either annotator. This is equivalent to 1 for agreement and 0 for disagreement in cases where both annotators have selected only one sense. Formally, let $i \in I$ be one annotated sentence. Let A be the set of annotators and let $P_A = \{\{a, a'\} \mid a, a' \in A\}$ be the set of annotator pairs. Let a_i be the set of senses that annotator $a \in A$ has chosen for sentence i. Then pairwise agreement between annotators is calculated as:

$$ITA \text{ WSbest} = \sum_{i \in I} \frac{\sum_{\{a,a'\} \in P_A} \frac{|a_i \cap a'_i|}{max(|a_i|, |a'_i|)}}{|P_A| \cdot |I|} \quad (1)$$

The average ITA was calculated as 0.574.[18] If we restrict the calculation to items where each annotator only selected one sense (not multiple), the average is 0.626.

For SE-3, ITA was 0.628 on the English Lexical Sample task, not including the multiword data (Mihalcea, Chklovski, and Kilgarriff 2004). This annotation exercise used volunteers from the Web (Mihalcea and Chklovski 2003). Like our study, it had taggers without lexicography background and gave a comparable ITA to our 0.626. We calculated pairwise agreement for eight annotators. To carry out the experiment under maximally similar conditions to previous studies, we also calculated ITA for items with only one response and use only the four annotators who performed WSbest first. This resulted in an average ITA of 0.638.

We also calculated the agreement for WSbest in R2 as in Equation 1 but with each individual removed to see the change in agreement. The results are in the first row of Table 22.

4.5 SYNbest

The SYNbest task is a repetition of the LEXSUB task (McCarthy and Navigli 2007, 2009) except that annotators were asked to provide one synonym at most. As in LEXSUB, agreement between a pair of annotators was counted as the proportion of all the sentences for which the two annotators had given the same response.

As in WSbest, let A be the set of annotators. I is the set of test items, but as in LEXSUB we only include those where at least two annotators have provided at least one substitute: If only one annotator can think of a substitute then it is likely to be a problematic item. As in WSbest, let a_i be the set[19] of responses (substitutes) for an item

18 Although there is a statistical difference in the number of multiple assignments depending upon whether WSsim-2 is completed before or after WSbest, ITA on the WSbest task does not significantly differ between the two sets.

19 Though, in fact, unlike LEXSUB and WSbest, we only collect one substitute per annotator for SYNbest.

$i \in I$ for annotator $a \in A$. Let P_A again be the set of pairs of annotators from A. Pairwise agreement between annotators is calculated as in LEXSUB as:

$$PA = \sum_{i \in I} \frac{\sum_{\{a,a'\} \in P_A} \frac{|a_i \cap a'_i|}{|a_i \cup a'_i|}}{|P_A| \cdot |I|} \quad (2)$$

Note that in contrast to pairwise agreement for traditional word sense annotation (WSbest), the credit for each item (the intersection of annotations from the annotator pair) is divided by the union of the responses. For traditional WSD evaluation, it is divided by the number of responses from either annotator, which is usually one. For lexical substitution this is important as the annotations are not collected over a predetermined inventory. In LEXSUB, the *PA* figure was 0.278, whereas we obtain $PA = 0.261$ on SYNbest. There were differences in the experimental set-up. We had eight annotators, compared with five, and for SYNbest each annotator only provided one substitute. Additionally, our experiment involved only a subset of the data used in LEXSUB. The figures are not directly comparable, but are reasonably in line.

In our task, out of eight annotators we had at most three people who could not find a substitute for any given item, so there were always at least five substitutes per item. In LEXSUB there were 16 items excluded from testing in the full data set of 2010 because there was only one token substitute provided by the set of annotators.

We also calculated the agreement for SYNbest as in Equation 2 but with each individual removed to see the change in agreement. The results are in the second row of Table 22.

4.6 Discussion of the Annotations of Individuals

We do not pose these annotation tasks as having "correct responses." We wish instead to obtain the annotators' opinions, and accept the fact that the judgments will vary. Nevertheless, we would not wish to conduct our analysis using annotators who were not taking the task seriously. In the analyses that follow in subsequent sections, we use the average judgment from our annotators to reduce variation from individuals. Nevertheless, before doing so, in this subsection we briefly discuss the analysis of the individual annotations provided earlier in this section in support of our decision to use all annotators for the gold standard.

Although there was variation in the profile of annotations for individuals, all of the annotators showed reasonable correlation on the graded task and at a level in excess of that achieved on other graded semantic tasks (Mitchell and Lapata 2008). There will inevitably be one annotator that has the lowest correlation with the others on any given task, but we found that this was not the same annotator on every task. For example, C on WSsim-2 has the lowest correlation with the average, yet concurs with others much more on Usim-2 and leaving C out would reduce agreement on WSbest and on SYNbest. D has lower correlation with others on several tasks, though higher than C on WSsim-2. When we redo the triangle inequality analysis in Section 4.3 individually we see from Tables 19 and 20 that annotator D is the highest performing annotator in terms of meeting the triangle inequality principle in R2 and is a close runner-up in R1. These results indicate that although annotators may use the graded scale in different ways, their annotations tally to a reasonable extent. We therefore used all annotators for the gold standard.

4.7 Agreement Between Annotations in Different Frameworks

In this paper we are considering various different annotations of the same underlying phenomenon: word meaning as it appears in context. In doing so, we contrast traditional WSD methodology (SE-3, SemCor, and WSbest) with graded judgments of sense applicability (WSsim), usage similarity (Usim), and lexical substitution as in LEXSUB and SYNbest. In this section we compare the annotations from these different paradigms where the annotations are performed on the same underlying data. For WSsim and Usim, we use average ratings as the point of comparison.

4.7.1 Agreement Between WS*sim and Traditional Sense Assignment.* To compare WSsim ratings on a five-point scale with traditional sense assignment on the same data, we convert the traditional word sense assignments to ratings on a five-point scale: Any sense that is assigned is given a score of 5, and any sense that is not assigned is given a score of 1. If multiple senses are chosen in the gold standard, then they are all given scores of 5. We then correlate the converted ratings of the traditional word sense assignment with the average WSsim ratings using Spearman's rho.

As described earlier, most of the sentences annotated in WSsim-1 were taken from either SE-3 or SemCor. The correlation of WSsim-1 and SE-3 is rho = 0.416, and the correlation of WSsim-1 with SemCor is rho = 0.425. Both are highly significant ($p \ll 0.001$).

For R2 we can directly contrast WSsim with the traditional sense annotation in WSbest on the same data. This allows a fuller comparison of traditional and graded tagging because we have a data set annotated with both methodologies, under the same conditions, and with the same set of annotators. We use the mode (most common) sense tag from our eight annotators as the traditional gold standard label for WSbest and assign a rating of 5 to that sense, and a rating of 1 elsewhere. We again used Spearman's rho to measure correlation between WSbest and WSsim and obtained rho = 0.483 ($p \ll 0.001$).

4.7.2 Agreement Between WSsim and Usim. WSsim and Usim provide two graded annotations of word usage in context. To compare the two, we convert WSsim scores to usage similarity ratings as in Usim. In WSsim, each sense has a rating (averaged over annotators), so a sentence has a vector of ratings with a "dimension" for each sense. For example, the vector of average ratings for the sentence in Table 6 is $\langle 5, 2.125, 1.25, 1.5, 1.25, 1.75\rangle$. All sentences with the same target will have vectors in the same space, as they share the same sense list. Accordingly, we can compare a pair u, u' of sentences that share a target using Euclidean distance:

$$d(\vec{u}, \vec{u'}) = \sqrt{\sum_i (\vec{u}_i - \vec{u'}_i)^2}$$

where $\vec{u}_i$ is the ith dimension of the vector $\vec{u}$ of ratings for sentence u. Note that this gives us a *dissimilarity* rating for u, u'. We can now compare these sentence pair dissimilarities to the similarity ratings of the Usim annotation.

In R1 we found a correlation of rho $= -0.596$ between WSsim and Usim ratings.[20] The basis of this comparison is small, at three lemmas with 10 sentences each, giving

20 The negative correlation is due to the comparison of dissimilarity ratings with similarity ratings.

Table 23
Spearman's correlation between lexical paraphrase overlap on the one hand, and Usim similarity or WSsim dissimilarity on the other hand.

tasks	rho
Usim-1 vs. LEXSUB	0.590
Usim-2 vs. SYNbest	0.764
WSsim-1 vs. LEXSUB	−0.495
WSsim-2 vs. SYNbest	−0.749

135 sentence pairs in total, because that is all the data available annotated in both paradigms. For R2 we can perform the analysis on the whole Usim-2 and WSsim-2 data, which gives us 26 lemmas, with 1,142 sentence pairs.[21] Correlation on R2 data is rho $= -0.816$. The degree of correlation is striking. We conclude that there is a very strong relationship between the annotations for Usim and WSsim. This bodes well for using Usim as a resource for evaluating sense inventories, an idea that we will pursue further in Section 6: It reflects word meaning but is not tied to any given sense inventory.

4.7.3 Agreement of WSsim and Usim with Lexical Substitution. Lexical paraphrases (substitutes) have been used as a means of evaluating WSD systems in a task where the inventory is not predefined (McCarthy and Navigli 2007, 2009). Because the R1 annotation was done in part on data that had previously been annotated with lexical substitutions, and R2 included lexical substitution annotation, we can compare paraphrase annotation with the results of WSsim and Usim. Again, we need to transform annotations to make the comparison feasible. We convert all annotations to a Usim-like format using sentence pair similarity or dissimilarity ratings. For WSsim, we use the transformation described previously, using Euclidean distance between sense rating vectors. We transform lexical substitution annotation using multiset intersection, as the lexical substitution annotation of a sentence is a multiset of substitutes[22] from all annotators. If sentences s_1, s_2 have substitute multisets $subs_1$ and $subs_2$, respectively, and $freq_i(w)$ is the frequency of substitute w in multiset $subs_i$, then we calculate multiset intersection as

$$\text{INTER}(s_1, s_2) = \frac{1}{max(|subs_1|, |subs_2|)} \sum_{w \in subs_1 \cap subs_2} min(freq_1(w), freq_2(w))$$

Again, as before and in LEXSUB, we only keep sentences for which at least two annotators could come up with a substitute. We also did not include any items that were tagged with the wrong POS in LEXSUB.[23]

Table 23 shows correlation, in terms of Spearman's rho, of Usim and WSsim annotation with lexical substitution annotation. The values of Usim and WSsim are based on mean scores averaged over all annotators. The INTER values computed for the

21 This is the number of pairs remaining after we exclude any pairs where one of the annotators provided a "do not know" response.

22 The frequency of a substitute in a multiset depends on the number of annotators that picked the substitute for the particular data point.

23 This was relevant only for the trial portion of LEXSUB, as the test portion was manually verified.

lexical substitution annotation yield similarity ratings for sentence pairs; accordingly, correlations of transformed lexical substitution with Usim are positive, and correlations of transformed lexical substitution with the WSsim-based sentence dissimilarity ratings are negative. All correlations are highly significant ($p \ll 0.001$). We anticipated a higher correlation of SYNbest with R2 annotation compared with that obtained using LEXSUB and R1 annotation: In R2 the set of annotators is larger, the same set of annotators do all experiments, and the SYNbest annotation focuses on obtaining one substitute per annotator (whereas LEXSUB allowed annotators to supply up to three paraphrases). This turned out to be in fact the case, as a comparison of rows 1 and 2 of Table 23 shows, and likewise a comparison of rows 3 and 4. We notice that the correlation is slightly stronger for Usim compared with WSsim, for both annotation rounds. One possible reason for this is that the comparison of lexical substitution data with Usim involves only one transformation of annotation data (the INTER calculation), whereas the comparison with WSsim involves two (INTER and also the Euclidean distance transformation). We can expect each transformation of annotation data to be "lossy" in the sense of introducing additional variance. Furthermore, WSsim relies on WordNet, which may add a layer of structure that does not reflect the overlap in semantic similarity between usages.

4.7.4 Summary. The Usim framework enables us to compare different annotation schemes for word meaning, as it is relatively straightforward to map all annotations to sentence pair (dis-)similarity ratings. We found strong relationships between WSsim and Usim annotation, and between both graded annotation frameworks on the one hand and traditional word sense annotation or lexical substitutions on the other hand. This provides some validation for the novel annotation frameworks. Also, if all labeling schemes provide comparable results, that opens up opportunities for choosing the best-fitting labeling scheme for each situation. All these tasks pursue the same endeavor, although the graded annotations and substitutions strive to capture the more subtle nuances of meaning that are not adequately represented by the winner takes all approach of traditional methodology. WSsim is closest to the traditional methodology and would suit systems needing to output WordNet sense labels, for example because they want to exploit the semantic relations in WordNet. Usim is application-independent. It allows for evaluation of systems that relate usages, whether into clusters or simply on a continuum. It could, for example, be used as a resource-independent gold standard for word sense induction. Lexical substitution tasks are particularly useful where the application being considered would benefit from lexical paraphrasing, for example, text simplification, summarization, or query expansion in information retrieval.

5. Examining Sense Groupings Emerging from WSsim Annotation

Recently there has been considerable work on grouping fine-grained senses, often from WordNet, into more coarse-grained sense groups (Palmer, Dang, and Fellbaum 2007). The use of coarse-grained sense groups has been shown to yield considerable improvements in inter-annotator agreement in manual annotation, as well as in the accuracy of WSD systems (Palmer, Dang, and Fellbaum 2007; Pradhan et al. 2007). In our WSsim annotation, we have used fine-grained WordNet senses, but we want to check that our results are not an artifact of this fine-grained inventory. Furthermore, the annotation results might be useful for identifying senses that could be grouped or for identifying senses where grouping is not straightforward.

In WSsim, annotators gave ratings for each sense of a target word. If an annotator perceives two senses of some target word as very similar, they will probably give them similar ratings, and not just for a single sentence but across all the sentences featuring the target word in question. So by looking for pairs of senses that tended to receive similar ratings across all sentences, we can identify sense descriptions that according to our annotators describe similar senses. Conversely, we expect that unrelated senses would have dissimilar ratings. If there were many senses that the WSsim annotators implicitly "grouped" by giving them similar ratings throughout, we would have to revise our finding that WSsim annotators often perceived more than one sense to be applicable, as they would have perceived only what could be described as one implicit sense group.

If a coarse-grained sense grouping is designed with the aim of reflecting sense distinctions that would be intuitively plausible to an untrained speaker of the language, then senses in a common group should also be similar according to WSsim annotation. So when WSsim annotators give very different ratings to senses that are in the same coarse-grained group, or very similar ratings to senses that are in different groups, this can point to problems in a coarse-grained sense group.

In this section, first we describe two existing sense groupings (Hovy et al. 2006; Navigli, Litkowski, and Hargraves 2007). Then we test the extent that the annotations accord with sense groupings by:

1. comparing judgments against the existing groupings, and re-examining the question of how often WSsim annotators found multiple different WordNet senses highly applicable.
2. using the WSsim data to examine the extent that the annotations could be used to induce sense groupings.

5.1 Existing Sense Grouping Efforts

OntoNotes. The OntoNotes project (Hovy et al. 2006; Chen and Palmer 2009) annotates word sense, along with coreference and semantic roles. The senses that it uses for verbs are WordNet 2.1 and 2.0, manually grouped based on both syntactic and semantic criteria. Examples of these criteria include the causative/inchoative distinction, and semantic features of particular argument positions, like animacy. Once the sense groups for a lemma are constructed manually, they are used in trial annotation. If an inter-annotator agreement of approximately 90% is reached, the lemma's sense groups are used for annotation; otherwise they are revised. Chen and Palmer report that the sense groups used in OntoNotes have resulted in a rise in inter-annotator agreement as well as annotator productivity. The third column of Table 24 shows OntoNotes groups for the noun *account*.

5.1.1 The SemEval-2007 English All Words Task (EAW). For the English All Words task at SemEval-2007, WordNet 2.1 senses were grouped by mapping them to the more coarse-grained Oxford Dictionary of English senses. For the training data, this mapping was done automatically; for the test data, the mapping was done by hand (Navigli, Litkowski, and Hargraves 2007). For our analysis, we used only lemmas that were included in the test data where the mapping had been produced manually.

Table 24
WordNet 2.1 senses of the noun *account*, and their groups in OntoNotes (ON) and EAW.

WordNet sense	WordNet sense no.	ON group	EAW group
business relationship: "he asked to see the executive who handled his account"	3	1.1	5
report: "by all accounts they were a happy couple"	8	1.2	2
explanation: "I expected a brief account"	4	1.2	2
history, story: "he gave an inaccurate account of the plot [...]"	1	1.3	2
report, story: "the account of his speech [...] made the governor furious"	2	1.3	2
account statement: "they send me an accounting every month"	7	1.4	4
bill: "send me an account of what I owe"	9	1.4	4
score: "don't do it on my account"	5	1.5	3
importance: "a person of considerable account"	6	1.6	3
the quality of taking advantage: "she turned her writing skills to good account"	10	1.7	1

Column (4) of Table 24 shows EAW groups for the noun *account*.[24] The two resources largely agree in the groupings for *account*. But whereas EAW groups senses 1, 2, 4, and 8 together, OntoNotes splits those senses into two groups.

5.2 Does WSsim Annotation Conform to Existing Sense Groups?

In the WSsim annotation, we have used the fine-grained senses of WordNet 3.0. But annotators were free to give high ratings for a sentence to more than one sense. So it is possible that they implicitly used more coarse-grained sense distinctions. In this and the following section, we will explore the question of whether, and to what extent, WSsim annotators used implicit coarse-grained sense groups. In this section, we will first ask whether their annotation matched the sense groups of either OntoNotes or EAW. OntoNotes and EAW differ in the lemmas they cover. Also, as we saw earlier, when they both cover a lemma, they do not always agree in the sense groups that they propose. So we study the agreement of WSsim annotation with the two sense groupings separately. We only study the lemmas which are in both the WSsim data and in either OntoNotes or the EAW test data, listed in Table 25.

Tables 26 and 27 show the results. Table 26 looks at the number of sentences where two senses both had high ratings, but are in different groupings in either OntoNotes or EAW. The first row shows how many sentences there were where two senses received a judgment of ≥ 3, but the two senses were not in a common OntoNotes/EAW group. The second row shows the same for judgments ≥ 4, and the last row for judgments of 5 only. In general, the percentages are higher for EAW than for OntoNotes. This is not due to any difference in granularity between the two resources. The EAW sense groups encompass on average 2.3 fine-grained senses for the R1 lemmas and 2.6 for the

24 The table shows the EAW groups of the WordNet senses, but the group numbering is our own for ease of reference as no labels are given in EAW.

Table 25
Lemmas in R1 and R2 WSsim that have coarse-grained mappings in OntoNotes and SemEval 2007 EAW.

	R1		R2	
lemma	ON	EAW	ON	EAW
account.n			√	√
add.v	√			
ask.v	√	√		
call.v			√	√
coach.n			√	
different.a		√		
dismiss.v			√	√
fire.v			√	
fix.v			√	
hold.v			√	√
lead.n				√
new.a				√
order.v	√		√	
paper.n		√		
rich.a				√
shed.v			√	
suffer.v			√	√
win.v	√	√		

Table 26
Sentences that have positive judgments for senses in different coarse groupings: percentage, and absolute number in parentheses. J. = WSsim judgment, averaged over annotators.

	OntoNotes				EAW			
J.	Rd. 1		Rd. 2		Rd. 1		Rd. 2	
≥ 3	28%	(42)	52%	(52)	78%	(157)	62%	(50)
≥ 4	13%	(19)	16%	(16)	41%	(82)	22%	(18)
5	3%	(5)	3%	(3)	8%	(17)	6%	(5)

Table 27
Sentences that have widely different judgments for pairs of senses in the same coarse grouping: percentage, and absolute number in parentheses. J1 = WSsim judgment for the sense with the lower rating, averaged over annotators; J2 = averaged WSsim judgment for the higher-rated of the two senses.

		OntoNotes				EAW			
J1	J2	Rd. 1		Rd. 2		Rd. 1		Rd. 2	
≤ 2	≥ 4	35%	(52)	30%	(30)	20%	(39)	60%	(48)
≤ 2	5	11%	(16)	4%	(4)	2%	(4)	15%	(12)

R2 lemmas, and for OntoNotes the mean group sizes are 2.3 (R1) and 2.4 (R2). More likely it is due to the individual lemmas. We observe that ratings of "similar" or higher are frequent. In all conditions except WSsim-1/OntoNotes, we find percentages over 50%. On the other hand, there are many fewer sentences where two senses received judgments of "very similar" or "identical" but were not in the same OntoNotes or EAW group, but these cases do exist. For example, there were five sentences with the target *dismiss.v* which in WSsim received an average judgment of 4 or 5 for senses from two different OntoNotes groups, 1.1 and 1.2. As *dismiss* is an R2 lemma, for which only 10 sentences were annotated, this means that this phenomenon was found in half the sentences annotated. The two sense groups are related: One is a literal, the other a metaphorical, use of the verb. OntoNotes group 1.1 is defined as 'refuse to give consideration to something or someone,' and group 1.2 is 'discharge, let go, persuade to leave, send away.' One such sentence was the second sentence in Table 13.

Table 27 lists the number of sentences where two senses in the same OntoNotes or EAW grouping received widely different ratings in the WSsim annotation. The first row shows how many sentences there were where one sense received a rating of ≤ 2 and another sense from the same OntoNotes or EAW group had a rating of ≥ 4. The second row shows the same for sense pairs in the same coarse-grained group where one received a rating of ≤ 2 and the other the highest possible rating of 5. (Note that the table considers judgments averaged over all annotators, so this row counts only sentences where all annotators agreed on the highest rating.) An example of such a case is *Rich people manage their money well*. In WSsim the first sense in WordNet (*possessing material wealth*) received an average score of 5 (i.e. a unanimous verdict), whereas all other senses received a score of less than 2. This included the third sense (*of great worth or quality; "a rich collection of antiques"*), which had an average of 1.625, and sense 8 (*suggestive of or characterized by great expense; "a rich display"*) with an average of 1.125. Both senses 3 and 8 are in the same group as sense 1 in EAW.

These are sentences where the WSsim annotation suggests a more fine-grained analysis than the OntoNotes and EAW groups offer. The percentages are substantial: For the more inclusive analysis in the first row, the numbers are between 20% and 60% of all sentences, and between 2% and 15% of sentences even fall into the more restrictive case in the second row. There is no clear trend in whether we see more of this type of disagreement for OntoNotes or for EAW, or for the first or the second round of WSsim annotation.

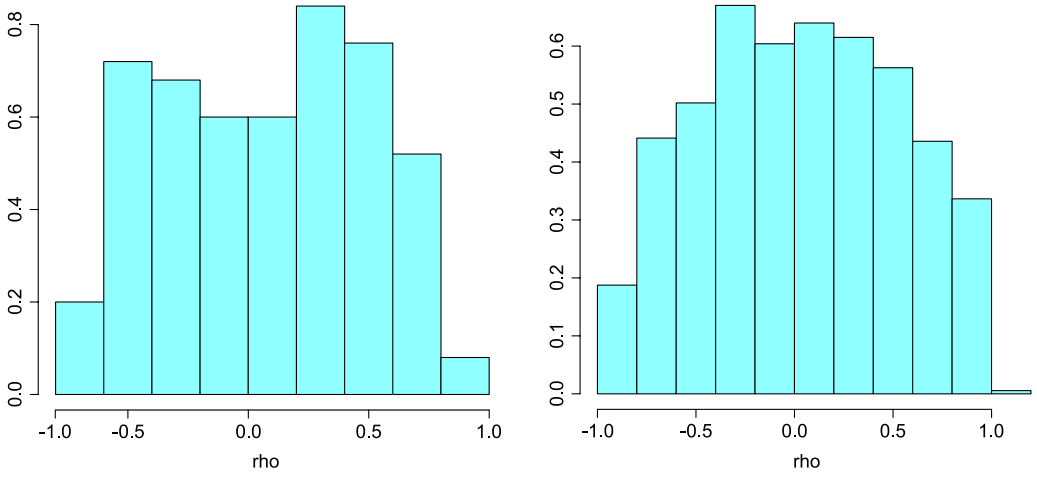
We see that there are a considerable number of sentences where either two senses from the same OntoNotes or EAW group have received diverging WSsim ratings, or two senses from different groups have received high ratings. In this way, the WSsim annotation can be used to scrutinize sense groupings: If one aim of the sense groupings is to form groups that would match intuitive sense judgments by untrained subjects, then WSsim annotation would suggest that the senses of *dismiss.v* that correspond to 'dismiss a person' and 'dismiss an idea' may be too close together to be placed in different groups.

5.3 Inducing Sense Relatedness from WSsim Annotation

In the WSsim annotation, annotators have annotated each occurrence of a target word with a rating for each of the WordNet senses for the target, as illustrated in Tables 6–8. This allows us, conversely, to chart the ratings that a WordNet sense received across all sentences. Table 28 shows this chart for two senses of the noun *account*. In the table, ratings are averaged across all annotators. In this case, the averaged ratings are

Table 28
WSsim ratings for two senses of the noun *account* for 10 annotated sentences (averaged over annotators).

WordNet	Sentence									
sense	1	2	3	4	5	6	7	8	9	10
1	1.00	2.25	1.13	4.25	1.13	1.0	1.13	1.13	1.13	4.25
4	1.50	3.00	1.25	2.88	1.50	1.50	1.63	1.00	1.38	3.88



(a) R1 (b) R2

Figure 4
Correlation between sense pairs: Distribution of rho values (Spearman's rho).

similar for the two senses: They tend to be high for the same sentences, and low for the same sentences. In general, senses that are closely related should tend to receive similar ratings: high on the same sentences, and low on the same sentences, as illustrated for the two senses in Table 28.

This then means that we can test the *correlation* on the ratings for two senses to see if the WSsim annotators perceived them to be similar. We compute correlation for any pair of senses for a common lemma, again using Spearman's rho.[25] Figure 4 shows the distribution of rho values obtained for all the sense pairs, as histograms for R1 (left) and R2 (right). When two senses are strongly positively correlated, this means that the annotators likely viewed them as similar. When two senses are strongly negatively correlated, this means they are probably so different that they tend never to be assigned high ratings for the same sentences. We see that in both rounds, there were roughly as many positive correlations as negative correlations. In R1, the rho values seem more or less equally distributed over the range from -1 to 1. In R2, there were more annotators and the distribution is closer to a normal distribution with more rho values close to 0.

25 We exclude senses that received a uniform rating of 1 on all items. For R1 there were no such cases and for R2 there were only 14 out of a total of 275 senses.

We have shown the OntoNotes and EAW sense groups for the noun *account*. We can now look at the WSsim-derived correlations for the same lemma, shown in Figure 5. The first row in each box shows the WordNet sense number, and the second row shows the OntoNotes and EAW sense groups. All three labels are those used in Table 24. Each edge represents a correlation in the WSsim annotation. To avoid clutter, only correlations with rho ≥ 0.5 are included, and a sense is only shown if it is correlated with any other sense. Edge thickness corresponds to the value of the correlation coefficient rho between each two senses; rho is also annotated on the edges. The first thing to note is that WSsim-based correlation does not give us sense groups. Correlations are of different strengths, and different cutoffs would result in different link graphs. Even for the chosen cutoff of rho $= 0.5$, the correlations do not induce cliques (in the graph-theoretic sense). For example, the sixth sense of *account* shows a correlation of rho ≥ 0.5 with the eighth sense, but not with any of the other senses to which the eighth sense is linked. The figure also shows that there are some senses that are strongly correlated in their annotation but are not grouped in one or the other of the existing groupings. For example, senses 3 (*the executive who handles his account*) and 7 (*account statement*) are strongly correlated, but are in different groups in OntoNotes as well as in EAW. There are also senses that share the same group in one of the coarse grained inventories, but have a weak or even a negative correlation based on the WSsim annotation. For example,

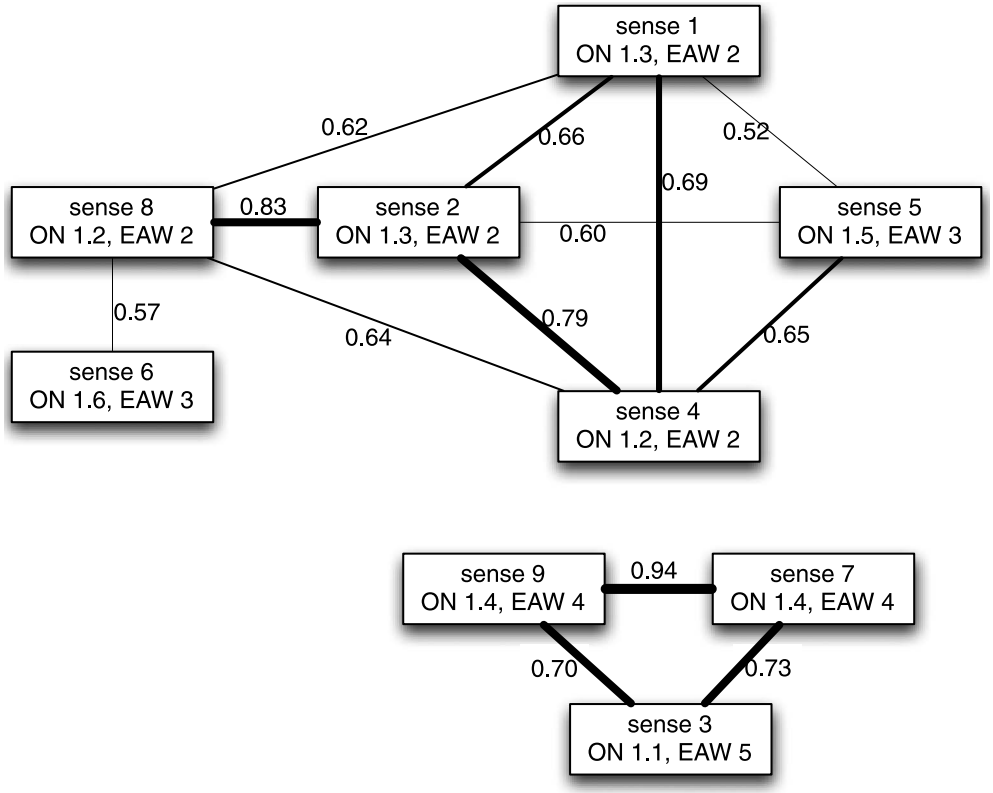
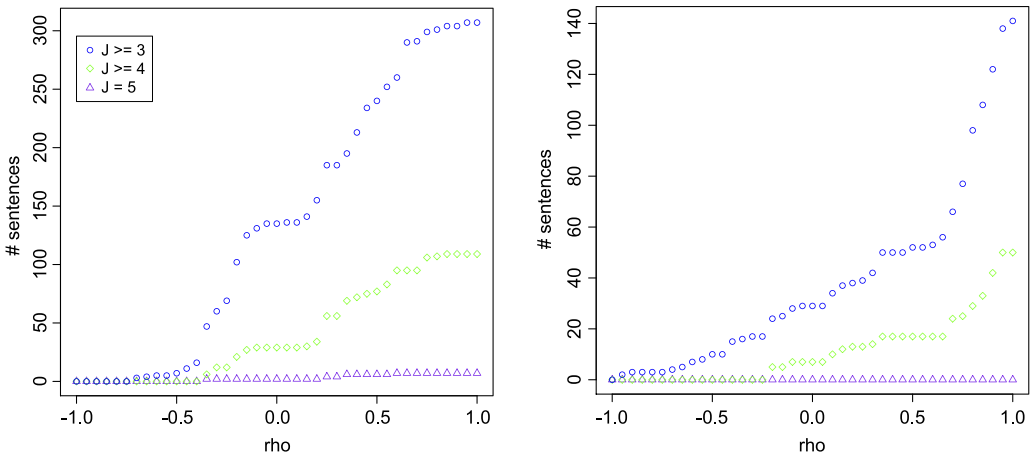


Figure 5
Sense correlation in the WSsim annotation for the noun *account*. Showing all correlations with rho ≥ 0.5. Upper row in each box: WordNet sense number. Lower row: OntoNotes and EAW sense groups. Edge thickness expresses strength of correlation.



(a) R1 (b) R2

Figure 6
Overall correlation versus annotation in single sentences: Number of sentences in which two senses with an overall correlation $\leq \rho$ have both been annotated with a judgment of $\geq j$, for $j = 3, 4, 5$. (Judgments averaged over annotators.)

for the lemma *paper*, senses 1 (*a material made from cellulose pulp*) and 4 (*a medium for written communication*) are in the same EAW group, but have a correlation in WSsim of rho $= -0.52$.

In Section 5.2 we asked whether the many cases where WSsim annotators gave high ratings to more than a single sense could possibly be explained by them implicitly using more coarse-grained senses. We answered this question by comparing the WSsim annotation to OntoNotes and EAW sense groups, finding a considerable number of sentences where two senses received a high rating but were not from the same sense group. Now we can repeat the question, but try to answer it using the WSsim sense relations obtained from correlation: Is it possible that WSsim annotators implicitly used more coarse-grained senses, but just not the OntoNotes or EAW sense groups?

We tested how often annotators gave ratings of at least *similar* (i.e., ratings ≥ 3) to senses that were related at a level $\leq$ rho, for rho ranging from -1 to 1. The question that we want to answer is: If annotators give high ratings to multiple senses on the same sentence, is it always to senses that are strongly positively correlated, or do they sometimes pick multiple senses that are not strongly correlated, or even senses that are negatively correlated? The results are shown in Figure 6. First, we can see that there is a sizeable number of sentences where two senses that are negatively correlated have both received a positive judgment. For R1, the numbers for negatively correlated senses are 135 ($j \geq 3$), 29 ($j \geq 4$), and 2 ($j = 5$). For R2, the numbers of sentences are lower absolutely and in proportion, with 29 ($j \geq 3$), 7 ($j \geq 4$), and 0 ($j = 5$). It is also interesting to look at a less stringent threshold than rho ≤ 0; we can use the significance levels $p \leq 0.05$ and $p \leq 0.01$ for this. If we look at sense pairs that were not positively correlated at $p \leq 0.05$ ($p \leq 0.01$), there were 185 (205) sentences in R1 and 54 (88) sentences in R2 where two such senses both received judgments of 3 or higher. Note that the significance levels of $p \leq 0.05$, $p \leq 0.01$ are here just arbitrary thresholds at which to inspect the data; they are not thresholds that determine the significance of

some hypothesis.[26] This brings us back to the question asked above of whether the WSsim annotators implicitly used more coarse-grained senses. If they had implicitly used more coarse-grained senses, we would have expected to see very few cases where unrelated senses got a high rating on the same sentence. What we found instead was that such cases were relatively frequent, which implies that WSsim annotators in both rounds "mix and match" senses specifically for each sentence that they evaluate. For example, the senses 1 (*she dismissed his advances*) and 5 (*I was dismissed after I gave my report*) of *dismiss* are negatively correlated (rho $= -0.61$) yet have average judgments of 3.25 and 4.125 on the second example in Table 13.

5.3.1 Summary. In this section we have analyzed the WSsim annotation in comparison with more coarse-grained sense repositories. One aim was to find out whether annotators really used the fine granularity that the WSsim task offered or whether they implicitly used more coarse-grained senses. Both by comparing the WSsim annotation to coarse-grained OntoNotes and EAW sense groups, and by comparing the WSsim annotation to the sense relations implied by WSsim, we find that annotators did make use of the ability to combine sense ratings in a way that was particular to each sentence they annotated. We also conclude that WSsim annotation can be used to evaluate OntoNotes and EAW groupings with respect to the level to which senses are intuitively distinguishable to untrained subjects. Here, WSsim annotation can uncover senses in different groups that WSsim annotators often conflate, or senses in a single coarse group that WSsim annotators treat differently.

6. Usim and Sense Groupings

One of the major motivations for the Usim task is that it allows us to examine the meanings of words without recourse to a predefined inventory. We have demonstrated in this paper that the data from this task can be compared directly to paraphrase data as well as to data annotated for word sense. In the previous section we have focused on using our WSsim data to examine existing sense groupings. WSsim is useful precisely because it has sense annotations from an existing inventory, WordNet, so we can use the graded annotations to see how these senses relate, and also relationships between coarser grained inventories with mappings to WordNet. Usim does not capture this information, nevertheless it might be useful as a resource for examining sense groupings. We can use it to examine the extent to which sense groupings keep usages together that have a high usage similarity according to Usim, and keep sentences with low usage similarity apart. In this analysis, we use the data from R2 because this has Usim judgments for sentences alongside traditional word sense annotations (WSbest). As WSbest annotation, we use the mode of the chosen senses[27] (as in the analysis in Section 4.7) for each sentence in R2, and map it to its coarse-grained sense in

26 We are performing multiple tests on the same senses, which increases the likelihood of falsely assuming two senses to be significantly correlated at some significance level (Type I errors). The significance levels are only arbitrary thresholds in our case, however. In addition, our analysis focuses on sense pairs that are *not* significantly positively correlated. For that reason, Type I errors actually reduce our estimate of the number of sentences in which two non-related senses both received high ratings. Conversely, correcting for multiple testing makes our estimate *less* conservative: If we count sentences with positive ratings for sense pairs that are not positively correlated at $p \leq 0.05$ *with Bonferroni correction*, the number of sentences rises from 185 to 207 for judgments of 3 or higher.

27 We perform this analysis only on sentences where there was one sense found as mode and where this had a coarse-grained mapping in either the EAW or OntoNotes resources.

EAW and/or OntoNotes. We then compute the average Usim similarity for all pairs of sentences with the same coarse-grained sense, and compare it with the average Usim similarity for sentence pairs with different coarse-grained senses. The results are shown in the first row of figures in Table 29. We see that the OntoNotes and EAW sense groups do indeed partition the sentences such that pairs within the same group have high usage similarity (4 or above) and those in different groups have low usage similarity (2 or below).

The second part of Table 29 performs the same analysis on the basis of individual lemmas. A dash (–) means that either there was no coarse mapping, or there were no sentence pairs in this category. For example, there were no sentence pairs identified as having different OntoNotes groups or EAW groups for the lemma *suffer.v*. For the lemmas *call.v* and *dismiss.v*, the two sense inventories give rise to the same groupings of sentence pairs.

In the table, we see many lemmas where the groupings concur with the Usim judgments. One example is *account.n*, where the sentence pairs in the same coarse group get high average Usim values, whereas sentence pairs with different coarse groups have low average Usim values. There are, however, a few lemmas where the average Usim values indicate that either the coarse groupings might benefit from another inspection, or that the lemma has meanings with subtle relationships where grouping is not a straightforward exercise. One example is *new.a*, which has the same high Usim values for both *same* and *different* categories in EAW. Another is *shed.v*, where the sentences annotated with the same OntoNotes groups actually have a lower average Usim value than those with different groups.

We can also use Usim judgments to analyze individual sense groups. This could be useful in determining specific groups that might warrant further revision, or that represent meanings which are simply difficult to distinguish. To demonstrate this, we analyzed all coarse-grained sense groups with at least one sentence pair in R2, that is, all groups that had at least two R2 sentences whose WSbest mode mapped to that coarse

Table 29
Average Usim rating for R2 where WSbest annotations suggested the same or different coarse grouping.

	OntoNotes		EAW	
	same	different	same	different
	4.0	1.9	4.1	2.0
	by lemma			
account.n	4.0	1.6	4.0	1.5
call.v	4.3	1.4	4.3	1.4
coach.n	4.6	2.3	–	–
dismiss.v	3.8	2.6	3.8	2.6
fire.v	4.6	1.2	–	–
fix.v	4.2	1.1	–	–
hold.v	4.5	2.0	3.8	1.9
lead.v	–	–	2.9	1.5
new.a	–	–	4.6	4.6
order.v	4.3	1.7	–	–
rich.a	–	–	4.6	2.0
shed.v	2.9	3.3	–	–
suffer.v	4.2	–	4.2	–

group. (Naturally, due to the skewed nature of sense distributions and the fact that we only have ten sentences for each lemma, some groups do not meet this criterion.) We find that the majority of groups that were analyzed have an average Usim rating of over 4. This is the case for 75% of the analyzed EAW groups and 76% of OntoNotes groups. There were, however, groups with very low values. One example was group 1.1 of *shed.v* in OntoNotes, with an average Usim rating of 2.9. This group includes both literal senses (*trees shed their leaves*) and metaphorical senses (*he shed his image as a pushy boss*) of the verb *shed*. Another example is group 7 of *lead.n* in EAW, also with an average Usim of 2.9. This group includes *taking the lead* as well as *lead actor*, so quite a diverse collection of usages. Two example sentences annotated with these two senses are shown here. This pair had an average Usim value of 1.25.

> My students perform a wide variety of music and they can be found singing leading roles in their high school and college musical productions, singing **lead** in rock and wedding bands, winning classical music competitions, singing at the summer conservatory of The Papermill Playhouse, and learning to sing so they can sing with local choirs.
>
> And as a result of President Bush's initiative, which he took as part of the G-8 Presidency, and also the other changes in which the US, UK has been in the **lead**, not least in Afghanistan and Iraq, you can now feel the winds of change blowing through the Arab world.

In the future we hope to obtain more Usim data. When we have more data, we will investigate whether the groupings that Usim identifies as problematic tend to be the same ones that require more iterations in inventory construction (Hovy et al. 2006). We also plan to test whether groupings with low Usim ratings tend to have lower inter-tagger agreement on traditional WSD annotation.

7. Computational Modeling

The graded meaning annotation data from Usim and WSsim annotation can be used to evaluate computational models of word meaning. In this section we summarize existing work on modeling the R1 data, which has already been made publicly available.

The WSsim data can be used to evaluate graded word meaning models as well as traditional WSD systems. Instead of evaluating only the highest-confidence sense of a WSD model, we can take a more nuanced look at a model's predictions, and give credit if it proposes multiple appropriate senses. In Erk and McCarthy (2009) we take advantage of this fact to evaluate and compare two supervised models on the WSsim data: a traditional WSD model, and a distributional model that forms one prototype vector for each sense of a given lemma. Both are trained on traditional single-sense annotation, but the prototype model does not see any negative data during training in order to avoid spurious negative data. For training, each word occurrence is represented as either a first-order or a second-order bag-of-words vector of its sentence. In an evaluation using weighted variants of precision and recall, we find that when the traditional WSD model is credited for all the senses that it proposes, rather than only the single sense with the highest confidence value, it does much better on both measures. This shows that the model does propose multiple appropriate senses, such that its performance may be underestimated in traditional evaluation. As was to be expected, the prototype models that do not see negative data during training have much higher recall at lower precision, for an overall better F-score (again using weighted variants of the evaluation measures).

Thater, Fürstenau, and Pinkal (2010) address the WSsim data with an unsupervised model. It represents a word sense as the sum of the vectors for all synonyms in its synset, plus the vectors for all hypernyms scaled down by a factor of 10. They also use a more complex, syntax-based model to derive occurrence representations. Unfortunately their results are not directly comparable to Erk and McCarthy (2009) because they evaluate on a subset of the data (verb lemmas only).

The Usim data, which directly describes the similarity of pairs of usages, can be used to evaluate distributional models of word meaning in context. So far, only one type of model has been evaluated on this data to the best of our knowledge: the clustering-based approach of Reisinger and Mooney (2010). They use the Usim data to test to what extent their clusters correspond to human intuitions on a word's senses. Their result is negative, as a low correlation of human judgments and predictions suggests to them that the induced clusters are not a good match for human senses. The Usim data is particularly interesting for a different way of evaluating distributional and vector space approaches for word meaning in context. These have been evaluated on the tasks of lexical substitution (Erk and Pado 2008; Dinu and Lapata 2010; Thater, Fürstenau, and Pinkal 2010; Van de Cruys, Poibeau, and Korhonen 2011), information retrieval, and word sense disambiguation (Schütze 1998), but Usim, in contrast, offers a different and more direct evaluation perspective.

8. Conclusion

In this paper we have explored the question of whether word meaning can be described in a graded fashion. Our aim has been to use annotation with graded ratings to capture untrained speakers' intuitions on word meaning. Our motivation has been two-fold. On the one hand we are drawing on current theories of cognition, which hold that mental concepts have "fuzzy boundaries." On the other hand we wanted to give a basis to current computational models for word meaning in context that predicts degrees of similarity between word occurrences. We have addressed this question through two novel types of graded annotations of word meaning in context that draws on methods from psycholinguistic experimentation. WSsim obtains word sense annotations from a given sense inventory but uses graded judgments for each sense. Usim judges similarity of pairs of usages of the same lemma.

The analysis of annotation results lets us answer our main question in the affirmative. Annotators can describe word meaning through graded ratings with good inter-annotator agreement, measured through pairwise correlation. Even though no in-depth training on sense distinctions was provided, the pairwise correlations were good in every single case, indicating that all annotators did the tasks in a similar fashion. In both tasks, all annotators made use of the full graded scale, and did not treat the task as binary. The Usim annotation provides us with a means of comparing different word meaning annotation paradigms. We have used it to demonstrate that there is strong correlation of these new annotations with both traditional WSD labels, and with overlap of lexical paraphrases. This is as we anticipated, as all of these annotations are describing the same phenomenon of word meaning in context through different means.

In additional analysis of the WSsim annotation, we found a high proportion of sentences (between 23% and 46%) in which multiple senses received high positive judgments from the same annotators. At the same time, annotators used the WSsim ratings in a nuanced and fine-grained fashion, sometimes assigning high ratings on the same sentence to two senses that overall patterned very differently. Analyzing Usim annotation, we found that all annotators' ratings obey the triangle inequality in almost

all cases. This can be taken as a measure of intra-annotator consistency on the task. It also means that current distributional approaches to word meaning in context are justified in viewing usage similarity as metric. Triangle inequality can be used to check the validity of future Usim annotation.

We do not propose that either one of our annotations is a panacea for evaluation of systems that represent word meaning in context, but we argue that they provide data sets that better reflect the fluid nature of word meaning and allow us to evaluate questions related to word meaning in a new fashion. In this paper, we have used both WSsim and Usim data to analyze existing coarse-grained sense inventories. We have demonstrated that it is often not straightforward to group sentences into disjoint senses, depending on the lemma. We have also shown how both WSsim and Usim style judgments can be used to identify problematic lemmas, as well as sense groupings that may warrant another inspection to check whether they match naive speakers' intuitive judgments. The graded annotation can also be used to identify lemmas whose usages are difficult to group into clear distinct senses. This information can in the future be used to handle such lemmas differently when making sense inventories, in annotation, and in computational systems.

An important next question to consider is the use of WSsim and Usim data to evaluate computational models of word meaning. As we have shown (Erk and McCarthy 2009), WSsim data can be used to evaluate traditional WSD systems in a graded fashion. We plan to do a more large-scale evaluation to assess to what extent the performance of current WSD systems is underestimated. Also, fine-grained WSsim annotation can be used for a comparison of fine-grained and coarse-grained traditional WSD systems. We have also shown (Erk and McCarthy 2009) that WSsim can be used to evaluate graded word sense assignment systems. Although we used a supervised setting, however, we trained on traditional sense annotation. We plan to collect more WSsim annotation in order to be able to train word sense assignment systems on graded data, for example, using a regression model.

In the same vein, we will extend the available Usim data to cover many more sentences by using crowdsourcing. The use of Usim for supervised training of word meaning models is particularly interesting as all existing usage similarity models are unsupervised; given previous results in WSD, we can expect that supervision will improve the performance of models of usage meaning. One way of using Usim data in training is to learn a similarity metric. Metric learning (see, e.g., Davis et al. 2007) induces a distance measure from given constraints stating similarity or dissimilarity of items.

Our novel graded annotation frameworks, WSsim and Usim, are validated both through good agreement between those data sets themselves, as well as good agreement between those data sets and traditional word sense annotation and lexical substitutions. Because all labeling schemes provide comparable results, this allows different ways of evaluating systems providing different perspectives on system output. Furthermore, the different paradigms may suit different types of systems. Lexical substitution tasks (McCarthy 2002; McCarthy and Navigli 2009) are particularly useful where the application being considered would benefit from lexical paraphrasing, for example, text simplification, summarization, or query expansion in information retrieval. WSsim is closest to the traditional methodology (WSbest) and would suit systems needing to output WordNet sense labels, for example, because they want to exploit the semantic relations in WordNet for tasks such as inferencing or producing lexical chains. Unlike WSbest, it avoids a winner-takes-all approach and allows for more nuanced sense tagging. Usim is application-independent. It allows for evaluation of systems that relate

usages, whether into clusters or simply on a continuum. It could, for example, be used as a resource-independent gold standard for word sense induction by calculating the within and across class similarities. Aside from its use as an enabling technology within a natural language processing application, a system that performs well at the Usim task may be useful in its own right. For example, it could be used to enable lexicographers to work on groups of examples that reflect similar meanings, or find further examples close to the one being scrutinized.

Acknowledgments

The annotation was funded by a UK Royal Society Dorothy Hodgkin Fellowship to Diana McCarthy. This work was supported by National Science Foundation grant IIS-0845925 for Katrin Erk. We are grateful to Huw McCarthy for implementing the interface for round 2 of the annotation. We thank the anonymous reviewers for many helpful comments and suggestions.

References

Agirre, Eneko and Philip Edmonds, editors. 2007. *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht.

Agirre, Eneko, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague.

Baroni, Marco and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1,183–1,193, Cambridge, MA.

Brown, Susan. 2008. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*, pages 249–252, Columbus, OH.

Brown, Susan. 2010. *Finding Meaning: Sense Inventories for Improved Word Sense Disambiguation*. Ph.D. thesis, University of Colorado at Boulder.

Burchardt, Aljoscha, Katrin Erk, Annette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006. The SALSA corpus: A German resource for lexical semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 969–974, Genoa.

Carpuat, Marine and Dekai Wu. 2007a. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, pages 43–52, Skovde.

Carpuat, Marine and Dekai Wu. 2007b. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague.

Chen, Jinying and Martha Palmer. 2009. Improving English verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries. *Journal of Language Resources and Evaluation (Special Issue on SemEval-2007)*, 43:181–208.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributed model of meaning. *Lambek Festschrift, Linguistic Analysis*, 36:345–384.

Coleman, Linda and Paul Kay. 1981. Prototype semantics: The English word "lie." *Language*, 57:26–44.

Copestake, Ann and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Cruse, D. A. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Philip Saint-Dizier and Evelyne Viegas, editors, *Computational Lexical Semantics*. Cambridge University Press, pages 33–49.

Davis, Jason, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, Corvallis, OR.

Deschacht, Koen and Marie-Francine Moens. 2009. Semi-supervised semantic role labeling using the Latent Words Language Model. In *Proceedings of EMNLP*, pages 21–29, Singapore.

Dinu, Georgiana and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010*

Conference on Empirical Methods in Natural Language Processing, pages 1,162–1,172, Cambridge, MA.
Edmonds, Philip and Scott Cotton, editors. 2001. *Proceedings of the SensEval-2 Workshop*. Toulouse. See `http://www.sle.sharp.co.uk/senseval`.
Erk, Katrin and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Singapore.
Erk, Katrin, Diana McCarthy, and Nicholas Gaylord. 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, Suntec.
Erk, Katrin and Sebastian Pado. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP-08*, pages 897–906, Waikiki, HI.
Erk, Katrin and Sebastian Pado. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97, Uppsala.
Erk, Katrin and Carlo Strapparava, editors. 2010. *Proceedings of the 5th International Workshop on Semantic Evaluation(SemEval)*. Uppsala.
Frazier, Lyn and Keith Rayner. 1990. Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, 29:181–200.
Gentner, Dedre and Virginia Gunn. 2001. Structural alignment facilitates the noticing of differences. *Memory and Cognition*, 21:565–577.
Gentner, Dedre and Arthur Markman. 1997. Structural alignment in analogy and similarity. *American Psychologist*, 52:45–56.
Grefenstette, Edward and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1,394–1,404, Edinburgh.
Hampton, James A. 1979. Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18:441–461.
Hampton, James A. 2007. Typicality, graded membership, and vagueness. *Cognitive Science*, 31:355–384.
Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1–2):205–215.
Hovy, Eduard H., Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (NAACL-2006)*, pages 57–60, New York.
Ide, Nancy and Yorick Wilks. 2006. Making sense about sense. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*. Springer, Dordrecht, pages 47–73.
Kilgarriff, Adam. 1992. *Polysemy*. Ph.D. thesis, University of Sussex.
Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.
Kilgarriff, Adam. 2006. Word senses. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht, pages 29–46.
Kilgarriff, Adam and Martha Palmer, editors. 2000. *Senseval: Special Issue of the Journal Computers and the Humanities*, volume 34(1–2). Kluwer, Dordrecht.
Kilgarriff, Adam and Joseph Rosenzweig. 2000. Framework and results for English Senseval. *Computers and the Humanities*, 34(1-2):15–48.
Kintsch, Walter. 2007. Meaning in context. In T. K. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Erlbaum, Mahwah, NJ, pages 89–105.
Klein, Devorah and Gregory Murphy. 2001. The representation of polysemous words. *Journal of Memory and Language*, 45:259–282.
Klein, Devorah and Gregory Murphy. 2002. Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, 47:548–570.
Klepousniotou, Ekaterini. 2002. The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, 81:205–223.
Klepousniotou, Ekaterini, Debra Titone, and Caroline Romero. 2008. Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6):1,534–1,543.
Krishnamurthy, Ramesh and Diane Nicholls. 2000. Peeling an onion: The lexicographers' experience of manual

sense-tagging. *Computers and the Humanities*, 34(1-2):85–97.
Landauer, Thomas and Susan Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.
Landes, Shari, Claudia Leacock, and Randee Tengi. 1998. Building semantic concordances. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.
Lefever, Els and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala.
McCarthy, Diana. 2002. Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 109–115, Philadelphia, PA.
McCarthy, Diana and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague.
McCarthy, Diana and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond*, 43(2):139–159.
McNamara, Timothy P. 2005. *Semantic Priming: Perspectives from Memory and Word Recognition*. Psychology Press, New York.
Mihalcea, Rada and Timothy Chklovski. 2003. Open Mind Word Expert: Creating large annotated data collections with web users' help. In *Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora (LINC 2003)*, pages 53–60, Budapest.
Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff. 2004. The SENSEVAL-3 English lexical sample task. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings SENSEVAL-3 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 25–28, Barcelona.
Mihalcea, Rada and Phil Edmonds, editors. 2004. *Proceedings SENSEVAL-3 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Barcelona.
Mihalcea, Rada, Ravi Sinha, and Diana McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala.
Miller, George A., Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, NJ.
Mitchell, Jeff and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, OH.
Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.
Moon, Taesun and Katrin Erk. In press. An inference-based model of word meaning in context as a paraphrase distribution. *ACM Transactions on Intelligent Systems and Technology* special issue on paraphrasing.
Murphy, Gregory L. 1991. Meaning and concepts. In Paula Schwanenflugel, editor, *The Psychology of Word Meanings*. Lawrence Erlbaum Associates, Mahwah, NJ, pages 11–35.
Murphy, Gregory L. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA.
Navigli, Roberto. 2009. Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.
Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 7: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague.
Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13:137–163.
Passonneau, Rebecca, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of LREC-7*, pages 3,244–3,249, Valleta.
Pickering, Martin and Steven Frisson. 2001. Processing ambiguous verbs: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27:556–573.

Pradhan, Sameer, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task 17: English lexical sample, SRL and all words. In *4th International Workshop on Semantic Evaluations (SemEval-4) at ACL-2007*, pages 87–92, Prague.

Preiss, Judita and David Yarowsky, editors. 2001. *Proceedings of Senseval-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse.

Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Reddy, Siva, Ioannis P. Klapaftis, Diana McCarthy, and Suresh Manandhar. 2011. Dynamic and static prototype vectors for semantic composition. In *Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011)*, pages 210–218, Chiang Mai.

Reisinger, Joseph and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, CA.

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

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233.

Rosch, Eleanor and Carolyn B. Mervis. 1975. Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Senseval-2. 2001. Web page: `http://www.sle.sharp.co.uk/senseval2`.

Snyder, Benjamin and Martha Palmer. 2004. The English all-words task. In *3rd International Workshop on Semantic Evaluations (SensEval-3) at ACL-2004*, pages 41–43, Barcelona.

Socher, Richard, Eric H. Huang, Jeffrey Pennin, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems 24*, pages 801–809, Grenada.

Stokoe, Christopher. 2005. Differentiating homonymy and polysemy in information retrieval. In *Proceedings of HLT/EMNLP-05*, pages 403–410, Vancouver.

Taylor, John R. 2003. *Linguistic Categorization*. Oxford University Press, New York.

Thater, Stefan, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 948–957, Uppsala.

Tuggy, David H. 1993. Ambiguity, polysemy and vagueness. *Cognitive Linguistics*, 4(2):273–290.

Tversky, Amos. 1977. Features of similarity. *Psychological Review*, 84(4):327–352.

Tversky, Amos and Itamar Gati. 1982. Similarity, separability, and the triangle inequality. *Psychological Review*, 89(2):123–154.

Van de Cruys, Tim, Thierry Poibeau, and Anna Korhonen. 2011. Latent vector weighting for word meaning in context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1,012–1,022, Edinburgh.

Washtell, Justin. 2010. Expectation vectors: A semiotics inspired approach to geometric lexical-semantic representation. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 45–50, Uppsala.

Williams, John. 1992. Processing polysemous words in context: Evidence for interrelated meanings. *Journal of Psycholinguistic Research*, 21:193–218.

Zhong, Zhi and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala.

Zhong, Zhi, Hwee Tou Ng, and Yee Seng Chan. 2008. Word sense disambiguation using OntoNotes: An empirical study. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1,002–1,010, Honolulu, HI.