Investigations on Word Senses and Word Usages

Katrin Erk
University of Texas at Austin
katrin.erk@mail.utexas.edu

Diana McCarthy
University of Sussex
dianam@sussex.ac.uk

Nicholas Gaylord
University of Texas at Austin
nlgaylord@mail.utexas.edu

Abstract

The vast majority of work on word senses has relied on predefined sense inventories and an annotation schema where each word instance is tagged with the best fitting sense. This paper examines the case for a graded notion of word meaning in two experiments, one which uses WordNet senses in a graded fashion, contrasted with the "winner takes all" annotation, and one which asks annotators to judge the similarity of two usages. We find that the graded responses correlate with annotations from previous datasets, but sense assignments are used in a way that weakens the case for clear cut sense boundaries. The responses from both experiments correlate with the overlap of paraphrases from the English lexical substitution task which bodes well for the use of substitutes as a proxy for word sense. This paper also provides two novel datasets which can be used for evaluating computational systems.

1 Introduction

The vast majority of work on word sense tagging has assumed that predefined word senses from a dictionary are an adequate proxy for the task, although of course there are issues with this enterprise both in terms of cognitive validity (Hanks, 2000; Kilgarriff, 1997; Kilgarriff, 2006) and adequacy for computational linguistics applications (Kilgarriff, 2006). Furthermore, given a predefined list of senses, annotation efforts and computational approaches to word sense disambiguation (WSD) have usually assumed that one best fitting sense should be selected for each usage. While there is usually some allowance made for multiple senses, this is typically not adopted by annotators or computational systems.

Research on the psychology of concepts (Murphy, 2002; Hampton, 2007) shows that categories in the human mind are not simply sets with clear-cut boundaries: Some items are perceived as more typical than others (Rosch, 1975; Rosch and Mervis, 1975), and there are borderline cases on which people disagree more often, and on whose categorization they are more likely to change their minds (Hampton, 1979; McCloskey and Glucksberg, 1978). Word meanings are certainly related to mental concepts (Murphy, 2002). This raises the question of whether there is any such thing as the one appropriate sense for a given occurrence.

In this paper we will explore using graded responses for sense tagging within a novel annotation paradigm. Modeling the annotation framework after psycholinguistic experiments, we do not train annotators to conform to sense distinctions; rather we assess individual differences by asking annotators to produce graded ratings instead of making a binary choice. We perform two annotation studies. In the first one, referred to as **WSsim** (Word Sense Similarity), annotators give graded ratings on the applicability of WordNet senses. In the second one, **Usim** (Usage Similarity), annotators rate the similarity of pairs of occurrences (usages) of a common target word. Both studies explore whether users make use of a graded scale or persist in making binary decisions even when there is the option for a graded response. The first study additionally tests to what extent the judgments on WordNet senses fall into clear-cut clusters, while the second study allows us to explore meaning similarity independently of any lexicon resource.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 10–18,
Suntec, Singapore, 2-7 August 2009.

2 Related Work

Manual word sense assignment is difficult for human annotators (Krishnamurthy and Nicholls, 2000). Reported inter-annotator agreement (ITA) for fine-grained word sense assignment tasks has ranged between 69% (Kilgarriff and Rosenzweig, 2000) for a lexical sample using the HECTOR dictionary and 78.6.% using WordNet (Landes et al., 1998) in all-words annotation. The use of more coarse-grained senses alleviates the problem: In OntoNotes (Hovy et al., 2006), an ITA of 90% is used as the criterion for the construction of coarse-grained sense distinctions. However, intriguingly, for some high-frequency lemmas such as *leave* this ITA threshold is not reached even after multiple re-partitionings of the semantic space (Chen and Palmer, 2009). Similarly, the performance of WSD systems clearly indicates that WSD is not easy unless one adopts a coarse-grained approach, and then systems tagging all words at best perform a few percentage points above the most frequent sense heuristic (Navigli et al., 2007). Good performance on coarse-grained sense distinctions may be more useful in applications than poor performance on fine-grained distinctions (Ide and Wilks, 2006) but we do not know this yet and there is some evidence to the contrary (Stokoe, 2005).

Rather than focus on the granularity of clusters, the approach we will take in this paper is to examine the phenomenon of word meaning both with and without recourse to predefined senses by focusing on the similarity of uses of a word. Human subjects show excellent agreement on judging word similarity out of context (Rubenstein and Goodenough, 1965; Miller and Charles, 1991), and human judgments have previously been used successfully to study synonymy and near-synonymy (Miller and Charles, 1991; Bybee and Eddington, 2006). We focus on polysemy rather than synonymy. Our aim will be to use WSsim to determine to what extent annotations form cohesive clusters. In principle, it should be possible to use existing sense-annotated data to explore this question: almost all sense annotation efforts have allowed annotators to assign multiple senses to a single occurrence, and the distribution of these sense labels should indicate whether annotators viewed the senses as disjoint or not. However, the percentage of markables that received multiple sense labels in existing corpora is small, and it varies massively between corpora: In the SemCor corpus (Landes et al., 1998), only 0.3% of all markables received multiple sense labels. In the SENSEVAL-3 English lexical task corpus (Mihalcea et al., 2004) (hereafter referred to as SE-3), the ratio is much higher at 8% of all markables[1]. This could mean annotators feel that there is usually a single applicable sense, or it could point to a bias towards single-sense assignment in the annotation guidelines and/or the annotation tool. The WSsim experiment that we report in this paper is designed to eliminate such bias as far as possible and we conduct it on data taken from SemCor and SE-3 so that we can compare the annotations. Although we use WordNet for the annotation, our study is not a study of WordNet per se. We choose WordNet because it is sufficiently fine-grained to examine subtle differences in usage, and because traditionally annotated datasets exist to which we can compare our results.

Predefined dictionaries and lexical resources are not the only possibilities for annotating lexical items with meaning. In cross-lingual settings, the actual translations of a word can be taken as the sense labels (Resnik and Yarowsky, 2000). Recently, McCarthy and Navigli (2007) proposed the English Lexical Substitution task (hereafter referred to as LEXSUB) under the auspices of SemEval-2007. It uses paraphrases for words in context as a way of annotating meaning. The task was proposed following a background of discussions in the WSD community as to the adequacy of predefined word senses. The LEXSUB dataset comprises open class words (nouns, verbs, adjectives and adverbs) with token instances of each word appearing in the context of one sentence taken from the English Internet Corpus (Sharoff, 2006). The methodology can only work where there are paraphrases, so the dataset only contains words with more than one meaning where at least two different meanings have near synonyms. For meanings without obvious substitutes the annotators were allowed to use multiword paraphrases or words with slightly more general meanings. This dataset has been used to evaluate automatic systems which can find substitutes appropriate for the context. To the best of our knowledge there has been no study of how the data collected relates to word sense annotations or judgments of semantic similarity. In this paper we examine these relation-

[1] This is even though both annotation efforts use balanced corpora, the Brown corpus in the case of SemCor, the British National Corpus for SE-3.

ships by re-using data from LEXSUB in both new annotation experiments and testing the results for correlation.

3 Annotation

We conducted two experiments through an online annotation interface. Three annotators participated in each experiment; all were native British English speakers. The first experiment, WSsim, collected annotator judgments about the applicability of dictionary senses using a 5-point rating scale. The second, Usim, also utilized a 5-point scale but collected judgments on the similarity in meaning between two uses of a word. [2] The scale was 1 – *completely different*, 2 – *mostly different*, 3 – *similar*, 4 – *very similar* and 5 – *identical*. In Usim, this scale rated the similarity of the two uses of the common target word; in WSsim it rated the similarity between the use of the target word and the sense description. In both experiments, the annotation interface allowed annotators to revisit and change previously supplied judgments, and a comment box was provided alongside each item.

WSsim. This experiment contained a total of 430 sentences spanning 11 lemmas (nouns, verbs and adjectives). For 8 of these lemmas, 50 sentences were included, 25 of them randomly sampled from SemCor [3] and 25 randomly sampled from SE-3.[4] The remaining 3 lemmas in the experiment each had 10 sentences taken from the LEXSUB data.

WSsim is a word sense annotation task using WordNet senses.[5] Unlike previous word sense annotation projects, we asked annotators to provide judgments on the applicability of *every* WordNet sense of the target lemma with the instruction: [6]

Your task is to rate, for each of these descriptions, how well they reflect the meaning of the boldfaced word in the sentence.

Applicability judgments were not binary, but were instead collected using the five-point scale given above which allowed annotators to indicate not only whether a given sense applied, but *to what degree*. Each annotator annotated each of the 430 items. By having multiple annotators per item and a graded, non-binary annotation scheme we allow for and measure differences between annotators, rather than training annotators to conform to a common sense distinction guideline. By asking annotators to provide ratings for each individual sense, we strive to eliminate all bias towards either single-sense or multiple-sense assignment. In traditional word sense annotation, such bias could be introduced directly through annotation guidelines or indirectly, through tools that make it easier to assign fewer senses. We focus not on finding the best fitting sense but collect judgments on the applicability of all senses.

Usim. This experiment used data from LEXSUB. For more information on LEXSUB, see McCarthy and Navigli (2007). 34 lemmas (nouns, verbs, adjectives and adverbs) were manually selected, including the 3 lemmas also used in WSsim. 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. For our experiment, we took every possible pairwise comparison of these 10 sentences for a lemma. We refer to each such pair of sentences as an SPAIR. The resulting dataset comprised 45 SPAIRs per lemma, adding up to 1530 comparisons per annotator overall.

In this annotation experiment, annotators saw SPAIRs with a common target word and rated the similarity in meaning between the two uses of the target word with the instruction:

Your task is to rate, for each pair of sentences, how similar in meaning the two boldfaced words are on a five-point scale.

In addition annotators had the ability to respond with "Cannot Decide", indicating that they were unable to make an effective comparison between the two contexts, for example because the meaning of one usage was unclear. This occurred in 9 paired occurrences during the course of annotation, and these items (paired occurrences) were

[2] Throughout this paper, a target word is assumed to be a word in a given PoS.

[3] The SemCor dataset was produced alongside WordNet, so it can be expected to support the WordNet sense distinctions. The same cannot be said for SE-3.

[4] Sentence fragments and sentences with 5 or fewer words were excluded from the sampling. Annotators were given the sentences, but not the original annotation from these resources.

[5] WordNet 1.7.1 was used in the annotation of both SE-3 and SemCor; we used the more current WordNet 3.0 after verifying that the lemmas included in this experiment had the same senses listed in both versions. Care was taken additionally to ensure that senses were not presented in an order that reflected their frequency of occurrence.

[6] The guidelines for both experiments are available at `http://comp.ling.utexas.edu/people/katrin_erk/graded_sense_and_usage_annotation`

excluded from further analysis.

The purpose of Usim was to collect judgments about degrees of similarity between a word's meaning in different contexts. Unlike WSsim, Usim does not rely upon any dictionary resource as a basis for the judgments.

4 Analyses

This section reports on analyses on the annotated data. In all the analyses we use Spearman's rank correlation coefficient (ρ), a nonparametric test, because the data does not seem to be normally distributed. We used two-tailed tests in all cases, rather than assume the direction of the relationship. As noted above, we have three annotators per task, and each annotator gave judgments for every sentence (WSsim) or sentence pair (Usim). Since the annotators may vary as to how they use the ordinal scale, we do not use the mean of judgments[7] but report all individual correlations. All analyses were done using the R package.[8]

4.1 WSsim analysis

In the WSsim experiment, annotators rated the applicability of each WordNet 3.0 sense for a given target word occurrence. Table 1 shows a sample annotation for the target *argument.n*. [9]

Pattern of annotation and annotator agreement. Figure 1 shows how often each of the five judgments on the scale was used, individually and summed over all annotators. (The y-axis shows raw counts of each judgment.) We can see from this figure that the extreme ratings 1 and 5 are used more often than the intermediate ones, but annotators make use of the full ordinal scale when judging the applicability of a sense. Also, the figure shows that annotator 1 used the extreme negative rating 1 much less than the other two annotators. Figure 2 shows the percentage of times each judgment was used on senses of three lemmas, *different.a*, *interest.n*, and *win.v*. In WordNet, they have 5, 7, and 4 senses, respectively. The pattern for *win.v* resembles the overall distribution of judgments, with peaks at the extreme ratings 1 and 5. The lemma *interest.n* has a single peak at rating 1, partly due to the fact that senses 5 (financial involvement) and 6 (interest group) were rarely judged to apply. For the lemma *different.a*, all judgments have been used with approximately the same frequency.

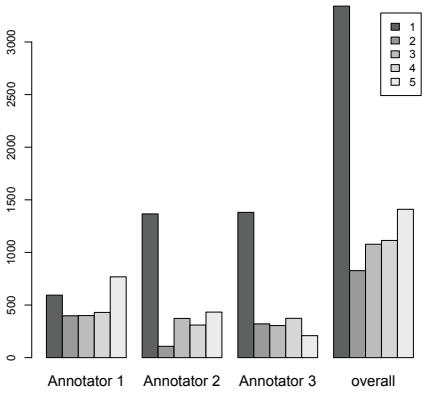

Figure 1: WSsim experiment: number of times each judgment was used, by annotator and summed over all annotators. The y-axis shows raw counts of each judgment.

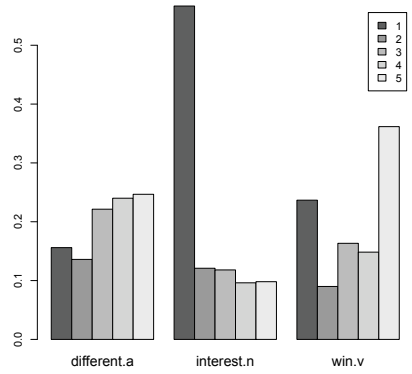

Figure 2: WSsim experiment: percentage of times each judgment was used for the lemmas *different.a*, *interest.n* and *win.v*. Judgment counts were summed over all three annotators.

We measured the level of agreement between annotators using Spearman's ρ between the judgments of every pair of annotators. The pairwise correlations were $\rho = 0.506$, $\rho = 0.466$ and $\rho = 0.540$, all highly significant with $p < 2.2\text{e-}16$.

Agreement with previous annotation in SemCor and SE-3. 200 of the items in WSsim had been previously annotated in SemCor, and 200 in SE-3. This lets us compare the annotation results across annotation efforts. Table 2 shows the percentage of items where more than one sense was assigned in the subset of WSsim from SemCor (first row), from SE-3 (second row), and

[7] We have also performed several of our calculations using the mean judgment, and they also gave highly significant results in all the cases we tested.

[8] http://www.r-project.org/

[9] We use *word*.PoS to denote a target word (lemma).

Sentence	Senses 1	2	3	4	5	6	7	Annotator
This question provoked **arguments** in America about the	1	4	4	2	1	1	3	Ann. 1
Norton Anthology of Literature by Women, some of the	4	5	4	2	1	1	4	Ann. 2
contents of which were said to have had little value as literature.	1	4	5	1	1	1	1	Ann. 3

Table 1: A sample annotation in the WSsim experiment. The senses are: 1:statement, 2:controversy, 3:debate, 4:literary argument, 5:parameter, 6:variable, 7:line of reasoning

Data	Orig.	WSsim judgment ≥ 3	≥ 4	5
WSsim/SemCor	0.0	80.2	57.5	28.3
WSsim/SE-3	24.0	78.0	58.3	27.1
All WSsim		78.8	57.4	27.7

Table 2: Percentage of items with multiple senses assigned. *Orig*: in the original SemCor/SE-3 data. WS*sim judgment*: items with judgments at or above the specified threshold. The percentages for WSsim are averaged over the three annotators.

	$p < 0.05$ pos	neg	$p < 0.01$ pos	neg
Ann. 1	30.8	11.4	23.2	5.9
Ann. 2	22.2	24.1	19.6	19.6
Ann. 3	12.7	12.0	10.0	6.0

Table 3: Percentage of sense pairs that were significantly positively (pos) or negatively (neg) correlated at $p < 0.05$ and $p < 0.01$, shown by annotator.

	$j \geq 3$	$j \geq 4$	$j = 5$
Ann. 1	71.9	49.1	8.1
Ann. 2	55.3	24.7	8.1
Ann. 3	42.8	24.0	4.9

Table 4: Percentage of sentences in which at least two uncorrelated ($p > 0.05$) or negatively correlated senses have been annotated with judgments at the specified threshold.

all of WSsim (third row). The *Orig.* column indicates how many items had multiple labels in the original annotation (SemCor or SE-3) [10]. Note that no item had more than one sense label in SemCor. The columns under WS*sim judgment* show the percentage of items (averaged over the three annotators) that had judgments at or above the specified threshold, starting from rating 3 – *similar*. Within WSsim, the percentage of multiple assignments in the three rows is fairly constant. WSsim avoids the bias to one sense by deliberately asking for judgments on the applicability of each sense rather than asking annotators to find the best one.

To compute the Spearman's correlation between the original sense labels and those given in the WSsim annotation, we converted SemCor and SE-3 labels to the format used within WSsim: Assigned senses were converted to a judgment of 5, and unassigned senses to a judgment of 1. For the WSsim/SemCor dataset, the correlation between original and WSsim annotation was $\rho = 0.234$, $\rho = 0.448$, and $\rho = 0.390$ for the three annotators, each highly significant with $p <$ 2.2e-16. For the WSsim/SE-3 dataset, the correlations were $\rho = 0.346$, $\rho = 0.449$ and $\rho = 0.338$, each of them again highly significant at $p <$ 2.2e-16.

Degree of sense grouping. Next we test to what extent the sense applicability judgments in the WSsim task could be explained by more coarse-grained, categorial sense assignments. We first test how many pairs of senses for a given lemma show similar patterns in the ratings that they receive. Table 3 shows the percentage of sense pairs that were significantly correlated for each annotator.[11] Significantly positively correlated senses can possibly be reduced to more coarse-grained senses. Would annotators have been able to designate a single appropriate sense given these more coarse-grained senses? Call two senses *groupable* if they are significantly positively correlated; in order not to overlook correlations that are relatively weak but existent, we use a cutoff of $p = 0.05$ for significant correlation. We tested how often annotators gave ratings of at least *similar*, i.e. ratings ≥ 3, to senses that were *not* groupable. Table 4 shows the percentages of items where at least two non-groupable senses received ratings at or above the specified threshold. The table shows that regardless of which annotator we look at, over 40% of all items had two or more non-groupable senses receive judgments of at least 3 (*similar*). There

[10] Overall, 0.3% of tokens in SemCor have multiple labels, and 8% of tokens in SE-3, so the multiple label assignment in our sample is not an underestimate.

[11] We exclude senses that received a uniform rating of 1 on all items. This concerned 4 senses for annotator 2 and 6 for annotator 3.

1) We study the methods and concepts that each writer uses to defend the cogency of legal, deliberative, or more generally political prudence against explicit or implicit charges that practical thinking is merely a knack or form of cleverness.

2) Eleven CIRA members have been convicted of criminal charges and others are awaiting trial.

Figure 3: An SPAIR for *charge.n*. Annotator judgments: 2,3,4

were even several items where two or more non-groupable senses each got a judgment of 5. The sentence in table 1 is a case where several non-groupable senses got ratings ≥ 3. This is most pronounced for Annotator 2, who along with sense 2 (controversy) assigned senses 1 (statement), 7 (line of reasoning), and 3 (debate), none of which are groupable with sense 2.

4.2 Usim analysis

In this experiment, ratings between 1 and 5 were given for every pairwise combination of sentences for each target lemma. An example of an SPAIR for *charge.n* is shown in figure 3. In this case the verdicts from the annotators were 2, 3 and 4.

Pattern of Annotations and Annotator Agreement Figure 4 gives a bar chart of the judgments for each annotator and summed over annotators. We can see from this figure that the annotators use the full ordinal scale when judging the similarity of a word's usages, rather than sticking to the extremes. There is variation across words, depending on the relatedness of each word's usages. Figure 5 shows the judgments for the words *bar.n, work.v* and *raw.a*. We see that *bar.n* has predominantly different usages with a peak for category 1, *work.v* has more similar judgments (category 5) compared to any other category and *raw.a* has a peak in the middle category (3). [12] There are other words, like for example *fresh.a*, where the spread is more uniform.

To gauge the level of agreement between annotators, we calculated Spearman's ρ between the judgments of every pair of annotators as in section 4.1. The pairwise correlations are all highly significant ($p < 2.2e\text{-}16$) with Spearman's $\rho =$ 0.502, 0.641 and 0.501 giving an average correlation of 0.548. We also perform leave-one-out resampling following Lapata (2006) which gave us a Spearman's correlation of 0.630.

[12] For figure 5 we sum the judgments over annotators.

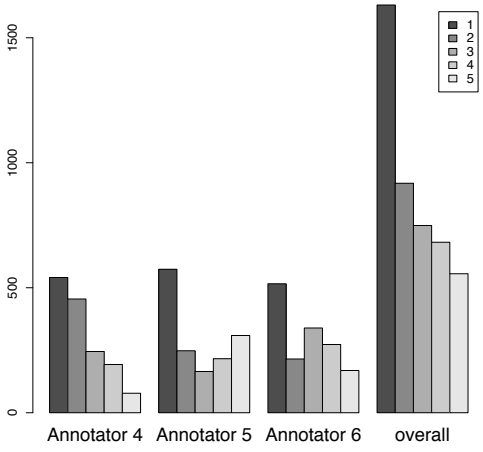

Figure 4: Usim experiment: number of times each judgment was used, by annotator and summed over all annotators

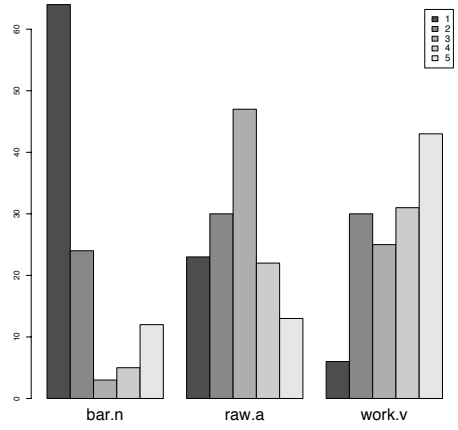

Figure 5: Usim experiment: number of times each judgment was used for *bar.n, work.v* and *raw.a*

Comparison with LEXSUB substitutions Next we look at whether the Usim judgments on sentence pairs (SPAIRs) correlate with LEXSUB substitutes. To do this we use the overlap of substitutes provided by the five LEXSUB annotators between two sentences in an SPAIR. In LEXSUB the annotators had to replace each item (a target word within the context of a sentence) with a substitute that fitted the context. Each annotator was permitted to supply up to three substitutes provided that they all fitted the context equally. There were 10 sentences per lemma. For our analyses we take every SPAIR for a given lemma and calculate the overlap (*inter*) of the substitutes provided by the annotators for the two usages under scrutiny. Let s_1 and s_2 be a pair of sentences in an SPAIR and

x_1 and x_2 be the multisets of substitutes for the respective sentences. Let $freq(w,x)$ be the frequency of a substitute w in a multiset x of substitutes for a given sentence. [13] $\text{INTER}(s_1,s_2) =$

$$\frac{\sum_{w \in x_1 \cap x_2} min(freq(w,x_1), freq(w,x_2))}{max(|x_1|,|x_2|)}$$

Using this calculation for each SPAIR we can now compute the correlation between the Usim judgments for each annotator and the INTER values, again using Spearman's. The figures are shown in the leftmost block of table 5. The average correlation for the 3 annotators was 0.488 and the p-values were all $<$ 2.2e-16. This shows a highly significant correlation of the Usim judgments and the overlap of substitutes.

We also compare the WSsim judgments against the LEXSUB substitutes, again using the INTER measure of substitute overlap. For this analysis, we only use those WSsim sentences that are originally from LEXSUB. In WSsim, the judgments for a sentence comprise judgments for each WordNet sense of that sentence. In order to compare against INTER, we need to transform these sentence-wise ratings in WSsim to a WSsim-based judgment of sentence similarity. To this end, we compute the Euclidean Distance[14] (ED) between two vectors J_1 and J_2 of judgments for two sentences s_1, s_2 for the same lemma ℓ. Each of the n indexes of the vector represent one of the n different WordNet senses for ℓ. The value at entry i of the vector J_1 is the judgment that the annotator in question (we do not average over annotators here) provided for sense i of ℓ for sentence s_1.

$$ED(J_1,J_2) = \sqrt{(\sum_{i=1}^{n}(J_1[i]-J_2[i])^2)} \qquad (1)$$

We correlate the Euclidean distances with INTER. We can only test correlation for the subset of WSsim that overlaps with the LEXSUB data: the 30 sentences for *investigator.n, function.n* and *order.v*, which together give 135 unique SPAIRs. We refer to this subset as W∩U. The results are given in the third block of table 5. Note that since we are measuring *distance* between SPAIRs for WSsim whereas INTER is a measure of *similarity*, the correlation is negative. The results are highly significant with individual p-values from $<$ 1.067e-10 to $<$ 1.551e-08 and a mean correlation of -0.495. The results in the first and third block of table 5 are not directly comparable, as the results in the first block are for all Usim data and not the subset of LEXSUB with WSsim annotations. We therefore repeated the analysis for Usim on the subset of data in WSsim and provide the correlation in the middle section of table 5. The mean correlation for Usim on this subset of the data is 0.532, which is a stronger relationship compared to WSsim, although there is more discrepancy between individual annotators, with the result for annotator 4 giving a p-value = 9.139e-05 while the other two annotators had p-values $<$ 2.2e-16.

Usim All		Usim W∩U	WSsim W∩U	
ann.	ρ	ρ	ann.	ρ
4	0.383	0.330	1	-0.520
5	0.498	0.635	2	-0.503
6	0.584	0.631	3	-0.463

Table 5: Annotator correlation with LEXSUB substitute overlap (*inter*)

The LEXSUB substitute overlaps between different usages correlate well with both Usim and WSsim judgments, with a slightly stronger relationship to Usim, perhaps due to the more complicated representation of word meaning in WSsim which uses the full set of WordNet senses.

4.3 Correlation between WSsim and Usim

As we showed in section 4.1, WSsim correlates with previous word sense annotations in SemCor and SE-3 while allowing the user a more graded response to sense tagging. As we saw in section 4.2, Usim and WSsim judgments both have a highly significant correlation with similarity of usages as measured using the overlap of substitutes from LEXSUB. Here, we look at the correlation of WSsim and Usim, considering again the subset of data that is common to both experiments. We again transform WSsim sense judgments for individual sentences to distances between SPAIRs using Euclidean Distance (ED). The Spearman's ρ range between -0.307 and -0.671, and all results are highly significant with p-values between 0.0003 and $<$ 2.2e-16. As above, the correlation is negative because ED is a distance measure between sentences in an SPAIR, whereas the judg-

[13] The frequency of a substitute in a multiset depends on the number of LEXSUB annotators that picked the substitute for this item.

[14] We use Euclidean Distance rather than a normalizing measure like Cosine because a sentence where all ratings are 5 should be very different from a sentence where all senses received a rating of 1.

ments for Usim are similarity judgments. We see that there is highly significant correlation for every pairing of annotators from the two experiments.

5 Discussion

Validity of annotation scheme. Annotator ratings show highly significant correlation on both tasks. This shows that the tasks are well-defined. In addition, there is a strong correlation between WSsim and Usim, which indicates that the potential bias introduced by the use of dictionary senses in WSsim is not too prominent. However, we note that WSsim only contained a small portion of 3 lemmas (30 sentences and 135 SPAIRs) in common with Usim, so more annotation is needed to be certain of this relationship. Given the differences between annotator 1 and the other annotators in Fig. 1, it would be interesting to collect judgments for additional annotators.

Graded judgments of use similarity and sense applicability. The annotators made use of the full spectrum of ratings, as shown in Figures 1 and 4. This may be because of a graded perception of the similarity of uses as well as senses, or because some uses and senses are very similar. Table 4 shows that for a large number of WSsim items, multiple senses that were not significantly positively correlated got high ratings. This seems to indicate that the ratings we obtained cannot simply be explained by more coarse-grained senses. It may hence be reasonable to pursue computational models of word meaning that are graded, maybe even models that do not rely on dictionary senses at all (Erk and Pado, 2008).

Comparison to previous word sense annotation. Our graded WSsim annotations do correlate with traditional "best fitting sense" annotations from SemCor and SE-3; however, if annotators perceive similarity between uses and senses as graded, traditional word sense annotation runs the risk of introducing bias into the annotation.

Comparison to lexical substitutions. There is a strong correlation between both Usim and WSsim and the overlap in paraphrases that annotators generated for LEXSUB. This is very encouraging, and especially interesting because LEXSUB annotators freely generated paraphrases rather than selecting them from a list.

6 Conclusions

We have introduced a novel annotation paradigm for word sense annotation that allows for graded judgments and for some variation between annotators. We have used this annotation paradigm in two experiments, WSsim and Usim, that shed some light on the question of whether differences between word usages are perceived as categorial or graded. Both datasets will be made publicly available. There was a high correlation between annotator judgments within and across tasks, as well as with previous word sense annotation and with paraphrases proposed in the English Lexical Substitution task. Annotators made ample use of graded judgments in a way that cannot be explained through more coarse-grained senses. These results suggest that it may make sense to evaluate WSD systems on a task of graded rather than categorial meaning characterization, either through dictionary senses or similarity between uses. In that case, it would be useful to have more extensive datasets with graded annotation, even though this annotation paradigm is more time consuming and thus more expensive than traditional word sense annotation.

As a next step, we will automatically cluster the judgments we obtained in the WSsim and Usim experiments to further explore the degree to which the annotation gives rise to sense grouping. We will also use the ratings in both experiments to evaluate automatically induced models of word meaning. The SemEval-2007 word sense induction task (Agirre and Soroa, 2007) already allows for evaluation of automatic sense induction systems, but compares output to gold-standard senses from OntoNotes. We hope that the Usim dataset will be particularly useful for evaluating methods which relate usages without necessarily producing hard clusters. Also, we will extend the current dataset using more annotators and exploring additional lexicon resources.

Acknowledgments. We acknowledge support from the UK Royal Society for a Dorothy Hodkin Fellowship to the second author. We thank Sebastian Pado for many helpful discussions, and Andrew Young for help with the interface.

References

E. Agirre and A. Soroa. 2007. SemEval-2007 task 2: Evaluating word sense induction and dis-

crimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic.

J. Bybee and D. Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language*, 82(2):323–355.

J. Chen and M. 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*. in press.

K. Erk and S. Pado. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP-08*, Waikiki, Hawaii.

J. A. Hampton. 1979. Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18:441–461.

J. A. Hampton. 2007. Typicality, graded membership, and vagueness. *Cognitive Science*, 31:355–384.

P. Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1-2):205–215(11).

E. H. Hovy, M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw, and R. 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.

N. Ide and Y. Wilks. 2006. Making sense about sense. In E. Agirre and P. Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 47–73. Springer.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for English Senseval. *Computers and the Humanities*, 34(1-2):15–48.

A. Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

A. Kilgarriff. 2006. Word senses. In E. Agirre and P. Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 29–46. Springer.

R. Krishnamurthy and D. Nicholls. 2000. Peeling an onion: the lexicographers' experience of manual sense-tagging. *Computers and the Humanities*, 34(1-2).

S. Landes, C. Leacock, and R. Tengi. 1998. Building semantic concordances. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

M. Lapata. 2006. Automatic evaluation of information ordering. *Computational Linguistics*, 32(4):471–484.

D. McCarthy and R. 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, Czech Republic.

M. McCloskey and S. Glucksberg. 1978. Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, 6:462–472.

R. Mihalcea, T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *3rd International Workshop on Semantic Evaluations (SensEval-3) at ACL-2004*, Barcelona, Spain.

G. Miller and W. Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

G. L. Murphy. 2002. *The Big Book of Concepts*. MIT Press.

R. Navigli, K. C. Litkowski, and O. 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, Czech Republic.

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

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

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

H. Rubenstein and J. Goodenough. 1965. Contextual correlates of synonymy. *Computational Linguistics*, 8:627–633.

S. Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

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