**Noname manuscript No.**
(will be inserted by the editor)

# Multiplicity and Word Sense: Evaluating and Learning from Multiply Labeled Word Sense Annotations

**Rebecca J. Passonneau · Vikas Bhardwaj · Ansaf Salleb-Aouissi · Nancy Ide**

the date of receipt and acceptance should be inserted later

**Abstract** Supervised machine learning methods to model word sense often rely on human labelers to provide a single, ground truth label for each word in its context. As part of an effort to create a community resource of sense annotations, we examine issues in establishing a single, ground truth word sense label using a fine-grained sense inventory. Our data consist of a sentence corpus for ten moderately polysemous words, and multiple sense labels (or multilabels) for 100 instances per word from trained and untrained annotators. Using a suite of asessment metrics to analyze the sets of multilabels, we conclude that the general annotation procedure is reliable, but that words differ regarding the reliability of their sense inventories, independent of the number of senses. In addition, we investigate the performance of an unsupervised machine learning method to infer ground truth labels from various combinations of labels from trained and untrained annotators. We find tentative support for the hypothesis that performance depends on the quality of the set multilabels, independent of the number of labelers or their level of training. Both sets of results indicate that whether words can be assigned ground truth sense labels depends less on the granularity of the sense inventory and more on other word-specific properties, such as their contexts of use, and the nature of the sense inventories and sense relations.

**Keywords** Word sense annotation · multilabel learning · inter-annotator reliability

## 1 Introduction

Most words have multiple meanings. In all natural languages, open class words (word classes whose membership is not fixed and where new words can be coined, borrowed, or derived), and many closed class words (such as prepositions), are more often polysemous than not. Many proposals exist for characterizing word sense in computational linguistics,

Rebecca J. Passonneau
Columbia University, New York, NY, USA
Tel.: +011-212-8701278, Fax: +011-212-8701285

Vikas Bhardwaj and Ansaf Salleb-Aouissi
Columbia University

Nancy Ide
Vassar College, Poughkeepsie, NY, USA

and there are no widely agreed upon standards for determining the number of senses for any given word. Rather, the representation one chooses for word sense is an abstraction that depends on one's theoretical or application goals. Yet resolving word sense is a prerequisite to any Natural Language Processing task that depends on utterance meaning. The fine-grained sense inventories preferred by lexicographers have been argued to lead to relatively lower annotation reliability when compared to coarse-grained inventories, in measures of agreement among two or three human labelers (annotators).

We present the results of an investigation of manual annotation of word sense for a heterogeneous corpus of present day English that relies on WordNet, a widely used lexical resource, that shows that annotators can agree well depending on the word. The work reported here is part of the Manually Annotated SubCorpus (MASC), a project to create a subset of the American National Corpus (ANC) annotated for many types of linguistic information [1]. The ANC is a large collection of present-day American English from many spoken and written genres.[1] It consists of 22 million words to date, nearly two thirds of which can be freely distributed (Open American National Corpus: OANC). MASC is a 500,000 subset of the OANC including equal portions of nineteen genres, which have been manually annotated or validated for fourteen types of annotation. One of the goals of MASC word sense annotation is to support efforts to align the sense distinctions made in WordNet [2] and FrameNet [3], as well as to facilitate investigation of alternative forms of word sense annotation and representation.

As described below, MASC word sense annotation follows best practice for creating a ground truth corpus. However, we assume that this methodology requires re-examination, in particular for word sense annotation. The key issue is how to arrive at a single ground truth label, given that different well-trained annotators and experts who often agree on a label, can also disagree, due to genuine differences in interpretation associated with specific instances. For example, of the 100 instances of adjectival *fair* in our data, there are 69 where at least one annotator selected sense 1 from WordNet. In 44 of these cases (64%), sense 1 is selected by nearly all five of the MASC annotators who worked on *fair*. In 6 of the cases (9%), annotators were split 2-to-3 between senses 1 and 2. (In the remaining 27% of cases, at least one annotator chose one of the eight other possible senses.) In WordNet, sense 1 is glossed as *free from favoritism or self-interest or bias or deception; . . .* , and one of its synonyms is *just* (an evaluative sense). Sense 2 is glossed as *not excessive or extreme* (a scalar sense), and one of its synonyms is *reasonable*. Whether a circumstance brings up issues of justice versus reasonableness is often a matter of opinion, thus leading to different interpretations, as in this example where the project annotators (A1, A2, etc.), plus one expert (E1), are split evenly between the two senses:

| | Annotators | | | | | |
|---|---|---|---|---|---|---|
| | A1 | A2 | A5 | A7 | A8 | E1 |
| Senses | s1 | s1 | s2 | s2 | s2 | s1 |

1. *And our ideas of what constitutes a fair wage or a fair return on capital are historically contingent.*

We believe the cases of near ties between these two senses of *fair* reflect an inherent open-endedness in the interpretation, rather than poor annotator performance or poor annotation methods. Data from multiple annotators reveals such instances in a way that fewer labels per instance cannot.

[1] `http://www.anc.org`

Our investigation addresses two questions about manual word sense annotation. Their combined results suggest that whether a word can be assigned a ground truth sense label depends less on the granularity of the sense inventory and more on other word-specific properties. The first question is how to collect word sense labels from trained or untrained annotators for moderately polysemous words. To investigate this question, we collected labels from approximately half a dozen trained annotators per instance, which yields a multilabel (a set of labels from different annotators) for each instance; the multilabel for example 1) is (s1, s1, s2, s2, s2, s1). Our assessment measures on the multilabels from trained annotators indicate that the annotation procedure is reliable, but that words differ regarding the ability of annotators to apply sense labels reliably. The second question is how to assign a single ground truth label for each word, given a multilabel. Recently there has been increasing interest within the NLP community in carrying out annotation efforts through crowdsourcing, which is the collective effort of a group of individuals. To examine the tradeoffs in relying on fewer trained annotators versus more untrained annotators, we collected additional labels for a subset of words, using twice as many untrained annotators as trained annotators. We then applied an unsupervised machine learning method to infer a ground truth label for each instance from several types of multilabel sets, using various combinations of labels from trained and untrained annotators. The results indicate that an expert quality labeling can be learned from a set of multilabels, but performance seems to depend in part on the quality of the multilabels, rather than solely on the number of annotators or their level of training.

The paper is structured as follows. Section 2 presents related work. Section 3 describes the ten words investigated here. Section 4 presents our assessment metrics, and section 5 assesses the sets of multilabels from trained and untrained annotators. In section 6, we present experiments using an unsupervised machine learning method to learn ground truth labels from various sets of multilabels. We conclude with a discussion (section 7) and a summary of our results and open questions for the future (section 8).

## 2 Related Work

Word meaning has been variously represented in lexicography, linguistics and computational linguistics. Approaches include providing detailed sense hierarchies for a given word (as in conventional dictionaries), WordNet's ordered inventory of sets of synonyms plus sense definitions, one or more components of a conceptual frame as in FrameNet [4], a decomposition into logical predicates and operators [5], a cluster of sentences where a word in all of them has the same meaning (as argued for in [6]), or some combination of the above. vRecent work by Erk and colleagues builds on the view that a sense can be defined as the contexts it occurs in [6], or, more specifically, as regions in a vector space model [7]. Vector space models, such as Latent Semantic Analysis [8], represent a word as an N-dimensional vector (tensor) of contextual dimensions (e.g., a 2-dimensional matrix of sentences by documents). Words with more similar contexts have similar vector representations, thus similarity of vectors captures semantic similarity. Erk and McCarty [9] rely on WordNet senses for an annotation method they refer to as graded sense assignment, in which annotators assign a score to each sense for every annotation instance. The MASC annotation task also relies on WordNet senses for sense labels. Because we collected multilabels for round 2.2, and a multilabel gives a distribution over the sense labels for a given word, this distribution is analogous to the graded sense assignment in [9]. Since all sentences for a given lemma are annotated at the same time, and the WordNet senses include glosses along with definitions (see next section), this task is similar to grouping instances by their similarity to the glosses.

There has been a decade-long community-wide effort to evaluate word sense disambiguation (WSD) systems across languages in several Senseval efforts (1998, 2001, 2004, 2007 and 2010; cf. [10–15]), with a corollary effort to investigate the issues pertaining to preparation of manually annotated gold standard corpora [13]. Differences in inter-annotator agreement and system performance across part-of-speech have been examined for two to three annotators [13, 16]. Investigations of factors that might affect human and system performance have looked at whether each annotator is allowed to assign multiple senses [17–19], the number or granularity of senses [16], merging of related senses [20], how closely related they are [21], sense perplexity [22], and entropy [22, 13]. Similarly, there have been studies of how distinguishable sense are for systems [23, 24] or humans [25, 26]. As noted below, we find a tentative part-of-speech effect for the 10 words studied here that is not borne out for the full set of MASC words. We do not find significant correlations of annotator agreement the number of senses with agreement, and only a modest correlation with the number of senses used, depending on the agreement metric. What other studies fail to consider, and that we find here, is that the general annotation procedure is reliable, but that specific words differ regarding the ability of annotators to apply the sense inventory reliably, independent of the part-of-speech or number of senses.

Previous work has suggested alternatives to pairwise agreement or the $\kappa$ family of agreement coefficients for assessing human annotations [9], automated word sense disambiguation [23], or both. In Erk & McCarthy's graded sense assignment [27], every sense in a word's inventory is assigned a grade on a 5 point scale. To evaluate graded sense assignments from human annotators, and automated word sense disambiguation (WSD) against the human data, they consider a range of metrics including Spearman's correlation coefficient, precision and recall, and Jenson Shannon Divergence (JDS), a distance metric for two probability distributions. For each annotated instance, every sense in the inventory is assigned a rating. Because individual annotators tend to be biased towards higher or lower ratings, they use JSD to provide a measure of distance that abstracts away from the absolute values assigned. They explicitly do not interpret the distribution of ratings as a probability distribution over the senses. This is in contrast to a suggestion from Resnik and Yarowsky to use cross entropy, which is related to JSD, to evaluate WSD systems that output a probability score for each available sense from the inventory [23]. They motivate their proposal in two ways: first, that even when incorrect, systems should get partial credit for assigning a relatively higher probability to the correct sense, and second, that a probabilistic result fits in well with downstream processing that relies on probabilities. Our use of JSD and similar metrics differs from both. As discussed below, we compare annotators' sense distributions on the assumption that each sense has a certain likelihood that should be roughly equivalent to its likelihood in each annotator's sense assignments.

We collected labels from multiple annotators in part to reveal differences across words with respect to annotator behavior. This has been used previously for coference phenomena: Poesio and Artstein [28] analyzed annotations from 18 annotators doing coreference annotation to detect contexts where annotators disagree because the context is ambiguous or vague. We believe the cases of disagreement described in the introduction, where annotators were split 50/50 between two word senses, are related to the cases of ambiguity or vagueness discussed by Poesio and Artstein. When there is data from many annotators, cases of disagreement can be more confidently identified as instances where no one referent (as in [28]) or no one word sense (as in our data) is significantly more probable than all others.

Recent work has examined how to leverage word sense data from multiple untrained annotators, using words with very few senses [29] [30]. Snow et al. included a word sense disambiguation task among several annotation tasks presented to Amazon Mechanical Turkers

in which annotators were required to select one of three senses of the word *president* for 177 sentences taken from the SemEval Word Sense Disambiguation Lexical Sample task [31]. They show that majority voting among three annotators reaches 100% accuracy in comparison to the SemEval gold standard, after correcting a single apparent disagreement where the expert annotation turned out to be incorrect. Many approaches to learning from crowds apply a probabilistic framework, and incorporate differences in annotator expertise [32], item difficulty, or both directly into the model [33]. Rayker et al. [34] propose a Bayesian framework to estimate the ground truth and learn a classifier. One of their contributions is the extension of the approach from binary to categorical, ordinal and continuous labels. None of this work has combined learning from multilabels with assessments of them. We use the method in [33], because it models both annotator expertise and instance difficulty, factors that affect the quality of the sets of multilabels used here. To our knowledge, no one has attempted to compare trained annotators with crowdsourcing for word sense annotation.

## 3 Word Sense Annotation Data: Multiple Annotators

The most common components of best practice to create an annotated resource in NLP are development of annotation guidelines; training the annotators; documenting inter-annotator reliability for a representative subset to demonstrate that the annotation can be applied consistently, or to verify that specific annotators are reliable, or both. For the full word sense corpus, trained MASC annotators have participated in ten annotation rounds to date, with approximately ten words per round, and 1000 sentences per word. Each round began with a small sample of 50 to 100 sentences used for training annotators on the labels for new words, and for re-consideration of the word sense labels in case they needed revision; the pre-annotation samples are not included in the 1000 sentences per word. For most rounds, annotator reliability was assessed using two to four annotators on 100 sentences per word, randomly selected from the 1000 sentences. The data described here consists of annotations for ten words for one round from half a dozen trained annotators from the MASC project, plus annotations of three of these words collected from fourteen untrained annotators recruited through Amazon Mechanical Turk (AMT), and from expert annotators.

### 3.1 Availability of the Data

The MASC corpus and word sense data are available from the MASC downloads page.[2] The round 2.2 data from MASC annotators investigated here is already available for download as part of the WordNet sense annotations and interannotator agreement data. It includes MASC word sense rounds 2 through 5. The annotations from turkers and experts will be included in future MASC releases, along with data from the remaining rounds of word sense annotation.

### 3.2 MASC Data: Trained Annotators

The MASC annotators for the data presented here were six undergraduate students: three from Vassar College majoring in cognitive science or computer science, and three linguistics majors from Columbia University. They were trained using guidelines written by Christiane Fellbaum, based on her experience with previous WordNet annotation efforts. The

[2] See downloads link at `http://www.anc.org/MASC/Home.html`.

| Word | POS | Count | WN Senses |
|---|---|---|---|
| fair | Adj | 1,204 | 10 |
| long | Adj | 7,095 | 9 |
| quiet | Adj | 720 | 6 |
| land | Noun | 1,942 | 11 |
| time | Noun | 38,861 | 10 |
| work | Noun | 12,325 | 7 |
| know | Verb | 81,201 | 11 |
| say | Verb | 78,345 | 11 |
| show | Verb | 16,659 | 12 |
| tell | Verb | 14,436 | 8 |

Table 1: Round 2 words, absolute frequency in OANC, and number of WordNet 3.0 senses

annotation tool is described below. For each new word, annotators applied the same general procedures, but learned a new set of sense labels. For example, it is part of the general procedure that annotators are told to become familiar with the full set of WordNet senses for a word prior to any annotation, and to consider the WordNet sense relations (e.g., synonymy, hypernymy) during annotation. It is also part of the general procedure that each sentence exemplifies a single word to be annotated; note that all tokens of that word in a given sentence are annotated. Annotators typically completed all instances for a single word before doing the next word.

### 3.3 Annotation Materials and Tools

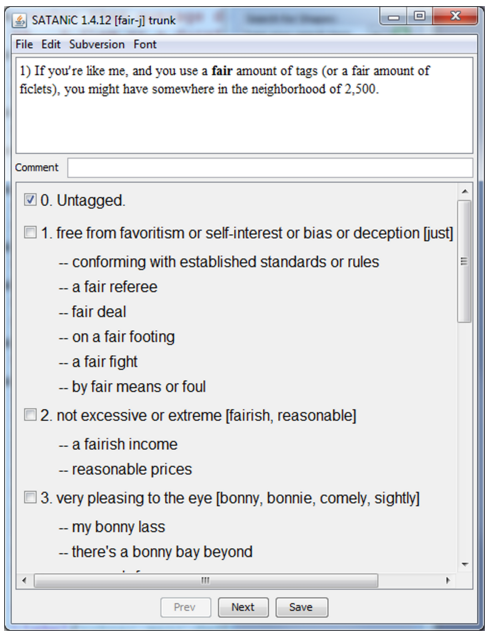
The ten words investigated in this study (round 2.2 of MASC) are fairly frequent, moderately polysemous words, balanced for part-of-speech. The ten words are shown in Table 1 with the total number of occurrences in the OANC and the number of WordNet 3.0 senses. For rounds 3 through 10, each annotation round of approximately 10 words began with a pre-annotation sample of 50 sentences per word annotated by 4 annotators for reviewing the WordNet sense inventory; any revisions to the sense inventory to support MASC annotation were included in subsequent versions of WordNet. Only then would the 1000 sentences per word be annotated, with a subset of 100 annotated by all 4 annotators for assessing reliability. Round 2.2, however, followed an initial round 2.1 where annotators used a *beta* version of the annotation tool, and where the sense inventory was reviewed, with no modifications to WordNet. For each of the ten words in the multiply annotated sample of round 2.2, 100 sentences per word were annotated by five or six trained annotators, depending on the word.[3] The resulting 1,000 sentences came from 578 texts representing eight written genres: fiction, journalism, letters, generic non-fiction, technical reports, government reports, government proceedings and travel guides. Average sentence length was 27.26 words.

Figure (1a) shows WordNet 3.0 senses for adjectival *fair* in the form displayed to all (trained and untrained) annotators. The sense number appears in the first column, followed by the glosses in italics, then sample phrases in plain font. When annotating a word for its sense, an annotator minimally considers this combination of an index (the sense number), an intensional definition (gloss), and an example. The examples for each sense can be considered a sentence cluster, where the annotator's job is to determine which cluster to assign the new sentence to. Annotators are also instructed to consider WordNet sense relations,

[3] One annotator dropped out during the round.

1 *free from favoritism or self-interest or bias or deception; conforming with established standards or rules*: a fair deal; on a fair footing; a fair fight; by fair means or foul
2 *not excessive or extreme*: a fairish income; reasonable prices
3 *very pleasing to the eye*: my bonny lass; there's a bonny bay beyond; a comely face; young fair maidens
4 *(of a baseball) hit between the foul lines:* he hit a fair ball over the third bases bag
5 *lacking exceptional quality or ability*: a novel of average merit; only a fair performance of the sonata; in fair health; the caliber of the students has gone from mediocre to above average; the performance was middling at best
6 *attractively feminine*: the fair sex
7 *(of a manuscript) having few alterations or corrections*: fair copy; a clean manuscript
8 *gained or earned without cheating or stealing*: an honest wage; a fair penny
9 *free of clouds or rain*: today will be fair and warm
10 *(used of hair or skin) pale or light-colored*: a fair complexion

(a) WordNet senses for *fair*




(b) SATANiC Annotation Tool

Fig. 1: MASC word sense annotation

such as synsets, hypernyms, troponyms, and antonyms. An example in the general guidelines discusses two similar senses of the noun *center* whose immediate hypernyms are *area* and *point*, thus further discriminating the senses into a central area versus a center point. When creating the MASC annotation tool, it was decided to achieve a balance between ease of use and richness of information, thus the annotation tool displays the sense number, gloss and example for each sense. Annotators used the browser interface to WordNet to view the remaining WordNet lexical information directly from this resource.

Figure (1b) is a screenshot of the SATANiC annotation tool developed to facilitate centralized management of annotation assignments and data collection. It connects directly to the ANC subversion (SVN) repository, allowing annotators to retrieve new assignments (SVN *check out*) and save results (SVN *commit*). The top frame displays the current sentence with the sample word in bold face. Annotators can enter free-form comments in the next frame. Below that is a scrollable window showing each WordNet sense number and its associated gloss, followed by a list of examples for the sense. Three additional labels are for uses of the word in a collocation, for sentences where the word is not the desired part-of-speech, or where no WordNet sense applies. Note that the annotation tool did not display the WordNet synsets (sets of synonymous senses). For example, the synset for sense 1 of *fair* also contains sense 3 of the adjective *just*. As noted above, however, annotators were encouraged to consult WordNet directly to view sense relations and other types of WordNet information, such as the synsets.

### 3.4 Amazon Mechanical Turk Data: Untrained Annotators

Amazon's Mechanical Turk (AMT) is a crowdsourcing marketplace where Human Intelligence Tasks (HITs; such as sense annotation for words in a sentence) can be offered, and where results from a large number of annotators (or turkers) can be obtained quickly. We used AMT to obtain annotations from turkers on the three adjectives. The task was designed to acquire annotations for 150 occurrences of the three adjectives: *fair*, *long* and *quiet*. We collected annotations from 14 turkers per word. Of the 150 occurrences, 100 were the same as those done by the trained annotators.[4] The 150 instances per word were divided into 15 HITs of 10 instances each.

Previous work has discussed some of the considerations in using AMT for language data [35] or word sense annotation [36], such as using a qualification test as a quality filter. We found that using a preliminary annotation round as a qualification test discouraged turkers from signing up for our HITs. As it would have been impractical to include all 150 sentences in a single HIT, we divided the task into 15 HITs of 10 occurrences each. To make the turker annotations parallel to the MASC data, we aimed to have each turker complete all HITs, rather than mix-and-match turkers across HITs. As a result, we had to discard or reject HITs for turkers who did not complete them all. This generated two types of protests: 1) some turkers wanted payment for the partial tasks, despite the fact that our instructions indicated that payment would be conditional on completion of all HITs; 2) rejected HITs result in lower AMT ratings for the turkers, a factor that affects whether a turker will be selected for future work. We handled the second case by creating pseudo-tasks for those turkers whose HITs we had rejected, and accepting all the pseudo-HITs. This ensured that turkers' ratings would not go down.

### 3.5 Expert Labels

We collected expert labels for evaluating the unsupervised learning approach. One of the co-authors assigned labels to two adjectives, *fair* and *long*, and worked together with an undergraduate research assistant to assign expert labels to the third (*quiet*). The sets of expert labels were reviewed twice: A first independent pass was followed by a second pass that led to a few corrections (2-3%) after comparison with the MASC annotators' results, or after comparison between the co-author and the undergraduate.

## 4 Assessment Methods

Agreement among annotators is typically measured as the proportion of pairs of agreements that occur overall (pairwise agreement), or by an agreement coefficient that calculates the proportion of observed agreement that is above chance expectation, meaning the agreements that could be expected if annotators applied labels randomly at the same rate as observed. We know *a priori* that a word's senses are not all equally likely, thus another obvious way to compare annotations is to look at the relative probabilities of each sense for each annotator. This can tell us whether annotators differ markedly with respect to the likelihood of specific senses, or with respect to the distribution of likelihoods over the set of senses.

[4] The remaining 50 were those used in round 2.1, and are not discussed further here.

Here we present the formulae for computing pairwise agreement and the $\alpha$ agreement coefficient [37], along with three probability-based metrics we refer to as Anveshan [38].[5]

### 4.1 Pairwise Agreement

Pairwise agreement is the ratio of the observed number of pairwise agreements among annotators to the maximum possible number. It is a descriptive statistic that provides a measure of coverage in that it answers the question, *how much of the annotated data is agreed upon*. It does not depend on any assumptions about the data, such as independence.

Computation of pairwise agreement for $c$ annotators on $i$ items from $k$ labels, where $n_{ik} \leq c$ is the number of annotators who labeled item $i$ as $k$, is given by:

$$\sum_{i=1}^{i} \sum_{k=1}^{k} \frac{\binom{n_{ik}}{2}}{\binom{c}{2}}$$

It sums the number of observed pairs of agreements on labels $k$ for the $i$ instances and divides by the total number of possible pairs of agreements.

### 4.2 Krippendorff's $\alpha$

Krippendorff's $\alpha$ is an agreement coefficient similar to $\pi$ [39], $\kappa$ [40], and related coefficients that factor out chance agreement.[6] The general formula for all of them is given by:

$$\frac{A_o - A_e}{1 - A_e}$$

where $A_o$ is the observed agreement, and $A_e$ is the agreement that would be expected by chance. For binary annotation labels, the ratio takes values in [-1,1], otherwise (-1,1], where 1 represents perfect agreement, -1 represents perfect disagreement, and 0 represents the agreement that would occur if annotators chose labels at the same rate as observed, but randomly.[7] The various agreement coefficients differ in their assumptions about and computation of $A_e$. Cohen's $\kappa$ takes each annotator's observed distribution of labels as the expected label probabilities for that annotator, whereas Krippendorff's $\alpha$ takes the distribution of labels among all annotators as the expected probability for each annotator.[8] Krippendorff's $\alpha$ evolved from measures of variance, thus casts the above ratio as a difference involving observed and expected *disagreement* (equivalent to the above agreement ratio):

$$\alpha = 1 - \frac{D_o}{D_e}$$

where $D_o$ is the observed disagreement and $D_e$ is the expected disagreement. For $i$ items, $k$ labels, and $c$ annotators, where again $n_{ik}$ is the number of annotators who assign label $k$ to item $i$, and $d_{k_j k_l}$ is the distance between a pair of label values $k_j$ and $k_l$:

[5] Anveshan is available at `http://vikas-bhardwaj.com/tools/Anveshan.zip`.

[6] To compute $\alpha$, we use Ron Artstein's perl script.Available as `http://ron.artstein.org/resources/calculate-alpha.perl`.

[7] Square brackets represent an interval that includes the endpoints; a parenthesis indicates the endpoint is not included in the interval.

[8] The values of $\kappa$ and $\alpha$ generally differ by a very small amount.

$$D_o = \frac{1}{ic(c-1)} \sum_{i=1}^{i} \sum_{j=1}^{k} \sum_{l=1}^{k} n_{ik_j} n_{ik_l} d_{k_j k_l}$$

For categorical (nominal) data such as sense labels, the distance function assigns the value 1 if $k_j \neq k_l$, and zero otherwise. All disagreements contribute to the sum $D_o$ and all agreements do not. (In other MASC rounds, where an annotator could assign multiple sense labels if they seemed equally fitting, we used a set-based distance metric to compare pairs of values $k_j k_l$.) Expected disagreement is given by:

$$D_e = \frac{1}{ic(ic-1)} \sum_{j=1}^{k} \sum_{l=1}^{k} n_{k_j} n_{k_l} d_{k_j k_l}$$

### 4.3 Comparison of Pairwise Agreement and $\alpha$

Because pairwise agreement credits all agreements between any pair of annotators, and $\alpha$ only credits agreements that would not be predicted by a random distribution, pairwise agreement is necessarily greater than or equal to $\alpha$. However, for relatively high pairwise agreement, $\alpha$ can be high or low. For example, consider the two following very simple cases for ten instances, two annotation labels, and two annotators. In the first case, the two annotators agree that five instances have label $\mathrm{L}_1$, that four instances have label $\mathrm{L}_2$, and they disagree on the tenth instance. In the second case, the annotators agree that nine instances have label $\mathrm{L}_1$, and they disagree on the tenth instance. In both cases, they have the same number of agreements: pairwise agreement is 90%. However, in the first case, $\alpha$ has the very high value of 0.81 compared with a low value of 0.00 for the second case. For the first case, it can be seen that the probability of the two labels are nearly equal (p($\mathrm{L}_1$)=$\frac{11}{20}$, p($\mathrm{L}_2$=$\frac{9}{20}$)), and that for the four combinations of labels ({$\mathrm{L}_1$,$\mathrm{L}_1$},{$\mathrm{L}_2$,$\mathrm{L}_2$},{$\mathrm{L}_1$,$\mathrm{L}_2$},{$\mathrm{L}_2$,$\mathrm{L}_1$}), the two disagreements should occur about half the time, and the two agreements should occur about half the time. At 90%, the rate of agreement is thus much higher than expected, and $\alpha$ is correspondingly high. For the second case, because the probability of label $\mathrm{L}_1$ is close to 1, the annotators can be expected to agree almost all the time on $\mathrm{L}_1$. In fact, the expected agreement equals the observed agreement, hence $\alpha$ is zero.

### 4.4 Metrics for Sense Distributions

For a dataset of $c$ annotators who label $i$ items with $k$ values, there are multiple annotations of the data that will give the same values for pairwise agreement and $\alpha$, and all the more so as $c$, $i$ or $k$ increase in size. For example, given a low $\alpha$, the disagreement might be due to a single annotator who deviates strikingly from all others (an outlier); to distinct subsets of annotators who have high agreement within but not across the subsets; or to an overall pattern of disagreement. Here, where we have relatively large values for $c$ and $k$, there are many additional facts of interest about the annotation data besides what proportion of the pairs of values are the same (pairwise agreement), or what proportion are the same after factoring out those that might arise by chance ($\alpha$). Given that sense distributions tend to be highly skewed, it is revealing to know the overall distribution of senses for each word, the distribution of senses for each annotator, and how similar these distributions are. To distinguish different sources of disagreement by comparing sense distributions within and across annotators,

we use the following metrics: Leverage [41], Kullback-Leibler Divergence (KLD) [42] and Jensen-Shannon Divergence (JSD) [43]. Each provides a measure of distance of two probability distributions. We use them in combination with pairwise agreement and $\alpha$ to provide a deeper analysis of word sense annotations.

Here we present the three metrics. In section 5, we illustrate their use in combination with $\alpha$ and pairwise agrement to identify annotators who are outliers, and subsets of annotators who are more consistent with each other than with other subsets. In section 6, we use them to create some of the subsets of annotators for the machine learning experiments.

#### *4.4.1 Leverage*

Leverage is a metric which compares two probability distributions over the same population of individuals $k$.[9] The Leverage of $P$ and $Q$ is given by:

$$Lev(P,Q) = \sum_k |P(k) - Q(k)|$$

Leverage has values in [0,1]: $Lev(P,Q) = 0$ if $Q(k) = P(k)$; $Lev(P,Q) = 1$ if $Q(k) = P(k)^{-1}$. Thus a low leverage indicates that $P$ and $Q$ are very similar, while a high score indicates the inverse. Leverage is used here to compare an individual annotator $a$'s distribution of senses ($P_a(k)$) to the distribution of the average distribution of senses $\overline{P(k)}$ for all annotators. Where $n_{k_a}$ is the number of times annotator $a$ uses sense $k$, $c$ is the number of annotators, and $i$ is the number of instances:

$$Lev(P_a(k), \overline{P(k)}) = \sum_{k=1}^{k} |\frac{n_{k_a}}{i} - \frac{\sum_{k=1}^{k}\sum_{b=1}^{c}\frac{n_{k_b}}{i}}{c}|$$

#### *4.4.2 Kullback-Leibler Divergence*

Kullback-Leibler divergence (KLD) is a non-symmetric measure of the difference between two probability distributions P and Q, where P is a reference distribution and Q is often an approximation of P. It has values in [0, $\infty$), and is given as:

$$KLD(P,Q) = \sum_{k=1}^{k} P(k) \log \frac{P(k)}{Q(k)}$$

The KLD score indicates the distance of a distribution $Q$ from $P$, with a higher score for a larger deviation. For a given annotator's distribution as the reference, we use KLD to get its comparison with the average sense distribution of all other annotators. The omission of the reference annotator from the average makes it more apparent whether this annotator differs from all the rest. (Note that if we instead take the reference distribution to be the average for other annotators, KLD becomes very large for annotators who failed to use one or more senses used by other annotators.) We compute a distance measure $KLD'$ for each annotator, by computing the KLD between each annotator's sense distribution ($P_a$) and the average of the remaining annotators (Q). Where $n_{k_b}$ is the number of times annotator $b$ uses sense $k$, $c$ is the number of annotators, and $i$ is the number of instances:

$$KLD'_a = KLD(P_a(k), Q(k)),\ where\ Q(k) = \frac{\sum_{k=1}^{k}\sum_{b\neq a}\frac{n_{k_b}}{i}}{c-1}$$

[9] Novelty [44] is another term for leverage.

*4.4.3 Jensen-Shannon Divergence*

Jensen-Shannon divergence is a modification of KLD known as *total divergence to the average*. In contrast to KLD, JSD is symmetric. It is given by:

$$JSD(P,Q) = \frac{1}{2}KLD(P,M) + \frac{1}{2}KLD(Q,M),\ \ where\ M = (P+Q)/2$$

Like KLD, JSD takes on values in [0,$\infty$), with lower scores indicating the distributions are more similar. We compute $JSD(P_{a_i}, P_{a_j})\ \forall(i,j)$, where $i,j \leq c$ and $i \neq j$.

## 5 Assessment of Label Quality

MASC is intended to cover a broad spectrum of genres, and to include accurate annotations for less frequent word senses. In the lexicographic and linguistic literature, it is taken for granted that there will be differences in judgment among language users regarding word sense, but the ramifications of preserving such differences when creating annotated resources have not been explored. Current practice in NLP word sense efforts typically assumes that appropriate annotation guidelines and training can yield a single label for word senses (cf. [45]). In our view, this achieves consensus at the expense of a more realistic view of the fluidity of sense judgments and linguistic usage. In our assessment of the labels produced by the trained annotators, we demonstrate the difficulty in the general case of producing a single ground truth sense label for each word in context, given relatively polysemous words, a very heterogeneous corpus, and half a dozen well-trained annotators. We also distinguish between the reliability of the annotators in following the general procedures, and their reliability on each sense inventory.

We show that by comparing several metrics for the annotations, we can identify annotators who are outliers, meaning those whose overal sense assignments differ markedly from other annotators. and we can also provide a more nuanced assessment of a group of annotators than is given by pairwise agreement or by $\alpha$ alone. We first review the annotations of the ten words to identify outliers. An annotator can be an outlier due, for example, to gaps in training; below we identify one annotator who overuses the label that indicates the word is part of a collocation, due to a misunderstanding of the criteria for collocations. By eliminating outliers, we can arrive at a more accurate representation of the natural variation inherent in the task.

After eliminating one to two outliers, agreement among the remaining annotators is sufficiently high on some words to indicate that MASC sense annotation can be performed reliably, depending on the word. We get the same finding across sets of four well-trained annotators (distinct subsets of a larger set of ten MASC annotators) on the full set of MASC words from all rounds [46]. Here, because the same five or six well-trained annotators were used for all ten words, differences in quality after outliers are eliminated are presumed to result from properties of the sense labels themselves, such as sense similarity or confusability, or from inherent differences in how the annotators interpret the sentences. We do not attempt to distinguish these two cases in the present paper.

The next subsections identify outliers among the trained annotators for each word, assess the labels from the remaining trained annotators, and assess the mechanical turkers. Note that alll MASC annotations, including outliers, are used in the machine learning experiments described in the following section. The learning method estimates annotator quality from the observed distribution of labels, thus learns to place less trust in some annotators.

| Word-pos | WordNet Senses | Senses Used | Ann. | Pairwise Agrt. | $\alpha$ |
|---|---|---|---|---|---|
| long-j | 9 | 4 | 6 | 0.81 | 0.67 |
| fair-j | 10 | 6 | 5 | 0.71 | 0.54 |
| quiet-j | 6 | 5 | 6 | 0.64 | 0.49 |
| time-n | 10 | 8 | 5 | 0.75 | 0.68 |
| work-n | 7 | 7 | 5 | 0.71 | 0.62 |
| land-n | 11 | 9 | 6 | 0.57 | 0.49 |
| tell-v | 8 | 8 | 6 | 0.61 | 0.46 |
| show-v | 12 | 10 | 5 | 0.53 | 0.46 |
| say-v | 11 | 10 | 6 | 0.57 | 0.37 |
| know-v | 11 | 10 | 5 | 0.52 | 0.37 |

Table 2: Pairwise agreement and $\alpha$ on ten words. (*Senses Used* indicates how many of the WordNet senses were used as sense labels; *Ann.* is the number of annotators for a given word.)

### 5.1 Outlier Identification and Reliability: Trained Annotators

Table 2 shows pairwise agreement and $\alpha$ on the ten words, prior to elimination of outliers. There is a weak negative correlation of pairwise agreement with number of senses in the inventory that is not significant, using Spearman's $\rho$ ($\rho = -0.64, p \approx 0.05$), but a non-significant correlation for $\alpha$ ($\rho = -0.47, p \approx 0.16$).[10] However, there is a highly significant negative correlation of pairwise agreement with number of senses used ($\rho = -0.84, p \approx 0.002$), and similarly for $\alpha$ ($\rho = -0.72, p \approx 0.018$). Agreement goes down as the number of senses used goes up. Further discussion of pairwise agreement and $\alpha$ is deferred until after outliers are eliminated and these metrics are recomputed.

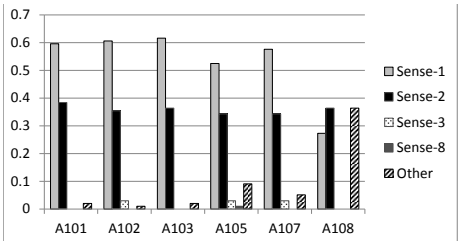
An outlier is defined as a statistical observation whose value is markedly different from others in a sample. When data fits a known distribution, outliers can be identified by measuring the distance of a data point from metrics that characterize the distribution. For example, the number of standard deviations from the mean measures indicates how far from normality a sample observation lies, given a population that follows the normal distribution. Given a heterogeneous corpus such as MASC, the distribution of senses often appears to be Zipfian, meaning a few labels occur with very high frequency, a few more occur with moderate frequency, and a long-tailed remainder occur relatively rarely. In any particular case, the observed distribution depends on factors such as the number of labels in the sense inventories, the nature of the semantic relations among the labels, and the size and constituency of the corpus of examples. Given many annotators, the rate that each sense label occurs for a given word can serve as an estimate of the true probability of the word sense, and the rate of each sense label for a given instance can serve as an estimate of the probability distribution over senses for that word in that context. We use Leverage, JSD and KLD' to identify outlier annotators.

An outlier is an annotator who uses one or more labels at a rate much higher or lower than other annotators. Outliers can result from differences in the procedures followed by the annotator, or from differences in the way annotators interpret the labels and instances; the metrics alone cannot distinguish these two cases. In this section we illustrate both cases through examples and plots that accompany the metrics to show concretely how outliers can be inferred given extreme values of one or more metrics.

[10] Due to ties in the data, the p-value computation is not exact.

| Ann | Leverage | $\overline{JSD}$ | KLD' |
|---|---|---|---|
| 108 | **0.5523** | **0.1687** | **0.8492** |
| 102 | 0.1780 | 0.0588 | 0.2792 |
| 107 | 0.1180 | 0.0428 | 0.0676 |
| 103 | 0.1787 | 0.0525 | 0.4605 |
| 101 | 0.1787 | 0.0516 | 0.4571 |
| 105 | 0.0477 | 0.0475 | 0.0271 |

(a) Leverage, $\overline{JSD}$ and KLD'




(b) Sense distributions

Fig. 2: Outlier identification for 6 annotators of *long*

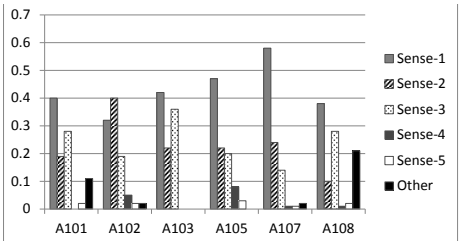
On the assumption that the observed proportion of sense labels for each annotator represents a probability distribution of senses, the three metrics of Leverage, JSD and KLD' are different measures of the similarity of a given annotator's distribution of senses to other annotators, taken as a group (Leverage, KLD') or one by one (JSD). Leverage represents how far on average the probability of annotator $a$'s sense labels $k$ are from the average probability of $k$ across all annotators. JSD($P_a(k)$,$P_b(k)$), $a \neq b$ indicates the similarity of the two sense distributions for a pair of annotators $a$ and $b$. $\overline{JSD}$ for an annotator $a$ is the average JSD for all pairings of $a$ with annotators other than $a$, and indicates the average similarity of that annotator's sense distributions to those of all other annotators. KLD' indicates how different an annotator's sense distribution is from the average of all the others. An annotator with values of Leverage, $\overline{JSD}$ and KLD' that are far above the norm is a clear outlier.

Figure 2 illustrates how annotator A108 is identified as an outlier for *long*. Table (2a) of the figure shows that A108 has much higher values of all three metrics than the five remaining annotators. The bar chart in Figure (2b) illustrates for each annotator (x-axis) the frequency of each sense (y-axis). Inspection of the sense distributions for A108 in the bar chart, compared with other annotators, shows a far greater proportion of *long* annotated as part of a collocation (*Other*; a rate of 0.36 compared with 0.09 on average). This pattern also appears in A108's annotations of other words, but exceptionally so for *long*. The marked difference in A108's sense distributions reflects a gap in training regarding the criteria for collocations. It should be noted that 108 joined the effort later than the other annotators, and received training at a different time. The remaining annotators have rather similar values of Leverage, $\overline{JSD}$ and KLD'. Figure (2b) illustrates the similarity of their sense distributions. After dropping A108, the consistency across the remaining annotators is reflected in an increase in pairwise agreement from 0.81 to 0.89 and an increase in $\alpha$ from 0.67 to 0.80. The latter value is noteworthy in that $\alpha \geq 0.80$ is taken to represent excellent annotator reliability by one of the more conservative scales of interpretation [37].

Figure 3 represents a contrasting case in which there are two annotators of *quiet* (A108 and A102) we identify as outliers. The remaining annotators fall into two subsets who are consistent within but not across the subsets. A108 and A102 have similarly high Leverage, and A108 also has very high $\overline{JSD}$ and KLD'. The plot in Figure 3b shows that A108 again has a far greater than average frequency of collocations and a compensatorily much lower than average rate of sense 2. A102 has a much greater than average rate of sense 2 and a rather lower rate of sense 1. After dropping A108 and A102 for *quiet*, the remaining annotators are not as consistent with one another as we saw above for *long*: pairwise agreement increases only from 0.64 to 0.66, and $\alpha$ does not increase. However, for two pairs, agree-

| Ann | Leverage | $\overline{JSD}$ | KLD' |
|---|---|---|---|
| A108 | **0.383** | **0.1687** | **0.8910** |
| A102 | **0.400** | 0.0588 | 0.1357 |
| A105 | 0.220 | 0.0475 | 0.1647 |
| A107 | 0.327 | 0.0428 | 0.0429 |
| A103 | 0.237 | 0.0525 | 0.1197 |
| A101 | 0.183 | 0.0516 | 0.1159 |

(a) Leverage, $\overline{JSD}$ and KLD'




(b) Sense distributions

Fig. 3: Outlier identification for 6 annotators of *quiet*

| Word-pos | Senses Used | Dropped Anns | Pairwise Agrt. | $\alpha$ | Subset | $\overline{JSD}$ | Subset $\alpha$ |
|---|---|---|---|---|---|---|---|
| long-j | 4 | A108 | 0.89 | 0.80 | NA | NA | NA |
| fair-j | 6 | A108,A102 | 0.77 | 0.63 | NA | NA | NA |
| quiet-j | 5 | A108,A102 | 0.66 | 0.49 | A101, A103 | 0.0696 | 0.61 |
| time-n | 8 | A108 | 0.77 | 0.71 | NA | NA | NA |
| work-n | 7 | A108 | 0.70 | 0.60 | NA | NA | NA |
| land-n | 9 | A108 | 0.61 | 0.54 | A101, A103 | 0.0403 | 0.60 |
| show-v | 8 | A101 | 0.55 | 0.48 | A102, A105 | 0.0132 | 0.52 |
| | | | | | A107, A108 | 0.0140 | 0.53 |
| tell-v | 12 | A103 | 0.64 | 0.50 | A101, A108 | 0.0113 | 0.57 |
| know-v | 11 | A102 | 0.62 | 0.48 | A101, A108 | 0.0492 | 0.52 |
| say-v | 11 | A101,A103 | 0.59 | 0.44 | A102, A105, A107 | 0.0302 | 0.51 |

Table 3: Pairwise agreement and $\alpha$ after dropping outlier annotators. (*Senses Used* indicates how many of the WordNet senses were used; *Ann* is the number of annotators for a given word.)

ment is very high: for (A101, A103) pairwise agreement is 0.93 and $\alpha$ is 0.86; for (A105, A107), pairwise agreement is 0.89 and $\alpha$ is 0.81. The main difference between the two pairs is that the latter use sense 1 relatively more often (52.5% versus 41% on average) and sense 3 relatively less often (17% versus 32% on average).

We briefly summarize the remaining cases. *Fair* is similar to *quiet* in that again, annotator A108 uses the label *Other* much more often than other annotators. A102 uses sense 1 relatively less often than the average of other annotators (43.4% versus 54.5%). For the word *say-v*, annotators A101 and A103 have relatively high Leverage (0.40 or above versus a range of 0.12 to 0.33 for the rest), the KLD' for A103 is very high (0.88). For A101, KLD' is high relative to the others (0.39 versus a range of 0.09 to 0.31), as is $\overline{JSD}$ (above 0.12 versus below 0.10). A101 uses sense 1 56% of the time compared with an average for the rest of 34%. A103 uses sense 2 56% of the time compared with an average for the rest of 34%.

Table 3 shows pairwise agreement and $\alpha$ after dropping outliers, with an apparent pattern of higher ranges for adjectives, less high for nouns, and lowest for verbs. The last three columns show subsets of annotators for these words who have relatively low $\overline{JSD}$ (hence more similar sense distributions), and also whose $\alpha$ is relatively higher; for the last row (*say-v*) with three annotators in the Subset column, $\overline{JSD}$ for the three pairs is shown. We now see lower Spearman correlations of senses used with pairwise agreement ($\rho = -.645, p \approx 0.04$) or with $\alpha$ ($\rho = -0.581, p \approx 0.08$).

| Word-pos | WordNet Senses | Senses Used | Ann | Pairwise Agrt. | $\alpha$ |
|---|---|---|---|---|---|
| 100 instances | | | | | |
| long-j | 9 | 9 | 14 | 0.28 | 0.12 |
| fair-j | 10 | 10 | 14 | 0.48 | 0.29 |
| quiet-j | 6 | 6 | 14 | 0.25 | 0.09 |

Table 4: Pairwise agreement and $\alpha$ for labels from 14 turkers for the adjectives (*Senses Used* indicates how many of the WordNet senses were used as sense labels; *Ann* is the number of annotators for a given word.)

| Ann | Leverage | $\overline{JSD}$ | KLD' |
|---|---|---|---|
| *fair* | | | |
| T102 | 0.100 | 0.0233 | 0.0592 |
| T107 | 0.116 | 0.0171 | 0.0349 |
| T108 | 0.116 | 0.0180 | 0.0398 |
| T114 | 0.132 | **0.2655** | 0.3421 |
| T111 | **0.212** | 0.0408 | **0.7575** |
| *long* | | | |
| T104 | 0.184 | 0.0694 | **0.2932** |
| T108 | 0.250 | 0.0765 | 0.2103 |
| T119 | 0.250 | 0.0768 | 0.2120 |
| T111 | **0.294** | 0.0904 | 0.2154 |
| T107 | **0.294** | **0.0938** | 0.2485 |
| *quiet* | | | |
| T131 | 0.196 | 0.0610 | 0.1159 |
| T122 | 0.264 | 0.0597 | 0.1144 |
| T123 | 0.292 | 0.0665 | 0.1409 |
| T127 | 0.348 | **0.1121** | **0.5730** |
| T119 | **0.464** | 0.1005 | 0.2934 |

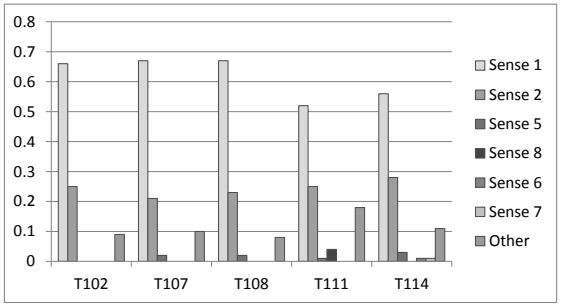
Table 5: Leverage, $\overline{JSD}$ and KLD' for the 5 best turkers for each adjective.

After dropping outliers and finding consistent subsets by means of Leverage, $\overline{JSD}$ and KLD', the values in column *Subset* $\alpha$ of Table 3–where they exist–or in column $\alpha$ otherwise, range from a low of 0.51 (moderate reliability) to 0.80 (excellent reliability). As described above for *quiet*, there are often subsets of annotators who are very consistent within but not across subsets. In this small sample, the results show greater agreement on adjectives than nouns, and on nouns than verbs. While this accords with claims for a part-of-speech effect from prior work [16, 13], it is not borne out in the full MASC data [46].

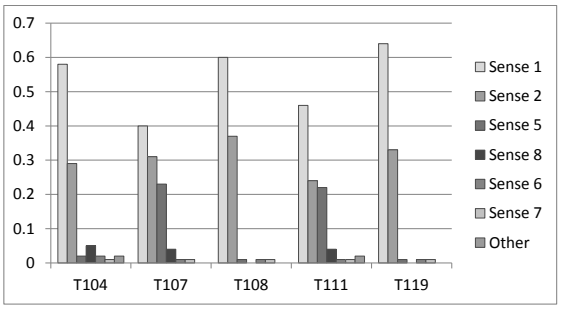
### 5.2 Untrained Annotators

As expected, when we turn to the assessment of the 14 turkers, they exhibit lower pairwise agreement and lower $\alpha$ scores than the trained annotators. This is shown in Table 4, with pairwise agreement in [0.25, 0.48], and $\alpha$ in [0.09, 0.29]. Note that the turkers use all senses in the inventory, in contrast to the trained annotators. The turkers perhaps assume the task is to find examples for all the senses. The turkers exhibit higher pairwise agreement and $\alpha$ on *fair* than on the other two adjectives, despite the fact that *fair* has the largest number of senses. The trained annotators had higher agreement on *long*; for both sets of annotators, agreement was lowest for *quiet*.

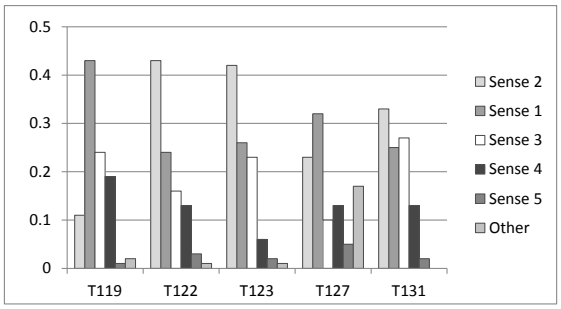
Turning to the measures based on the sense distributions, the turkers' annotations exhibit markedly higher values of Leverage, $\overline{JSD}$ and KLD' in comparison to the trained annota-




(a) Sense distributions for the 5 best turkers on *fair*




(b) Sense distributions for the 5 best turkers on *long*




(c) Sense distributions for the 5 best turkers on *quiet*

Fig. 4: Sense distributions on three adjectives for the best turkers

tors, which is also to be expected. For example, Leverage ranges from 0.047 to 0.553 for the trained annotators on the three adjectives (0.047 to 0.580 for all words), and from 0.217 to 1.433 for the turkers.[11] $\overline{JSD}$ ranges from 0.000 to 0.216 for the trained annotators on the three adjectives (0.000 to 0.220 on all words), compared with 0.003 to 0.951 for the turkers. KLD' ranges from 0.023 to 0.891 for the trained annotators (0.023 to 1.345 on all words), and from 0.093 to 2.819 for the turkers.

Despite the very high disagreement among turkers overall, and the large differences in sense distributions, it is possible to identify subsets of turkers who have agreement as good

[11] In the interest of space, we presented full Leverage, $\overline{JSD}$ and KLD' across trained annotators for only two of the eight words (Tables 2a-3a).

1 *characterized by an absence or near absence of agitation or activity*: a quiet life; a quiet throng of onlookers; quiet peace-loving people; the factions remained quiet for almost 10 years
2 *free of noise or uproar; or making little if any sound*: a quiet audience at the concert; the room was dark and quiet
3 *not showy or obtrusive*: clothes in quiet good taste
4 *in a softened tone*: hushed voices; muted trumpets a subdued whisper; a quiet reprimand

(a) Four senses of *quiet*: WordNet definitions and examples

| A101 | A102 | A103 | A105 | A107 | A108 | E101 | E102 |
|---|---|---|---|---|---|---|---|
| 2 | 4 | 2 | 4 | 2 | 4 | 3 | 3 |
| 1. In this well-produced spot, the intentionally quiet images never get in the way of the message. | | | | | | | |
| 3 | 4 | 3 | 2 | 2 | 3 | 1 | 1 |
| 2. The Armenian government downplayed the incident, claiming that the city and country are quiet and the only events are taking place around the parliament building. | | | | | | | |

(b) Labels from trained and expert annotators on two sentences

Fig. 5: Sentences with high disagreement on *quiet*

as or better than the trained annotators. For *fair*, there are five annotators among the turkers who have relatively good agreement: pairwise agreement=0.86 and $\alpha$=0.74. For use in the machine learning experiments, we chose a subset size of five rather than six because some of the sets of trained annotators were size five, and because adding any additional turker significantly lowers the quality of the sets of turker multilabels for the three words. The 5 turkers with the highest agreement on *long* have pairwise agreement=0.74 and $\alpha$=0.57. For *quiet*, the best subset of turkers has lower agreement than for the other two words: pairwise agreement=0.54 and $\alpha$=0.392.

Table 5 shows the Leverage, $\overline{JSD}$ and KLD' for the 5 best turkers for the three adjectives. *Fair*, which has the highest agreement, also has a range of values for the probability distribution metrics that is closer to the trained annotators. Not all the same turkers did all the words, but we see that certain turkers who perform well on *fair* also perform well on *long*: T107, T108 and T111. The two annotators least similar to the rest are T111 and T114. From Figure (4a), showing the sense distributions for each of the 5 best turkers on *fair*, we can see that T111 and T114 differ most in having a lower probability for sense 1. For *long*, T108 and T119 have the most similar values of the three metrics; T104 has particularly high KLD', T107 has particularly high $\overline{JSD}$, and T111 and T107 have the highest Leverage. Figure (4a) illustrates that all the annotators are similar in using sense 1 most often, followed by sense 2, and that two annotators also use sense 5 quite often. Of the three adjectives, *quiet* exhibits the least uniformity among the turkers. As shown in Table 5, T127 has the highest $\overline{JSD}$ and KLD'; T119 has the highest leverage. Figure (4c) clearly shows that there is more variation among the turkers on the senses for *quiet* than for the other two adjectives.

### 5.3 Discussion of annotator reliability

For the ten moderately polysemous words investigated here, even after eliminating outliers, there is still a wide range of agreement values (Table 3), from a high of $\alpha$ = 0.80 to a low of $\alpha$ = 0.44. Because the same annotators apply the same general procedures for all ten words, the lower agreement values cannot be explained as noise or error. Because the sense inventories

for all ten words have been carefully reviewed by the annotators and an expert member of the WordNet team (Christiane Fellbaum), the lower agreement values observed in Table 3 are also unlikely to be due to faulty application of the procedures for creating WordNet sense inventories. We believe they result instead from a natural variation across individuals regarding the meanings of certain words in context. Some contexts are more objective than others; length, for example, is a physical property that can be measured objectively, while fairness is a matter of judgment (see example 1) in section 1).

The word *quiet*, with a lower $\alpha$ (0.49) than the other adjectives, has meanings which are also a matter of judgment. Figure 5 shows the WordNet senses of *quiet* (excluding one sense specific to water, and another specific to the sun), and two sentences with labels illustrating a fair amount of disagreement. The labels are from six trained annotators, plus two expert labelers (E101, E102, one of whom is one of the co-authors).

The word *spot* in sentence 1 refers to a 1996 political advertisement in *Slate* magazine. We observe a difference in whether the annotators seem to interpret *images* as referring only to a visual dimension (sense 3) or to an audiovisual dimension (senses 2 and 4), and whether the absence of sound is the result of intentional activity (sense 4). The three senses selected by two or three annotators can be associated with the following interpretations reflecting these differences: the images are not associated with sounds in the sound track, possibly inherently (sense 2); the images are unobtrusive and backgrounded with respect to the message (sense 3); sounds associated with the depicted entities have been muted by the depicted individuals or by the filmmaker (sense 4).

Sentence 2 is from a 1999 *Slate* article reporting that gunmen killed the Armenian Prime Minister and other government leaders. It describes the city and country as *quiet*; which the annotators interpreted variously as exhibiting no activity (sense 1); being relatively free of noise (sense 2); characterized by citizens behaving in a restrained fashion (sense 3); or where people have intentionally lowered the volume of their activities (sense 4).

In both cases from Figure 5, it would be difficult to claim that there is a single correct reading; none of the readings appears to be incorrect. How one interprets each sentence presumably depends in part on the perspective one takes on the production values of political advertisements, or on the nature of claims made by a government.

## 6 Machine Learning from Multiple Labels or Features

Our next goal is to determine whether it is possible to learn expert quality labels from sets of multilabels produced by trained or untrained annotators. We present results from a series of machine learning experiments to infer true labels from multilabels. Our original hypothesis was that future annotation efforts could benefit from insight into the tradeoffs between using fewer labels from trained annotators versus more labels from untrained annotators for word sense. Ultimately, we find no consistent pattern regarding the number of annotators to use. Instead, we find that learning performance depends at least in part on the quality of a given set of multilabels, as measured by our assessment metrics.

For these experiments, we used the three adjectives *fair*, *long*, and *quiet*, because they had higher levels of agreement from the trained annotators.[12] GLAD, the unsupervised method we rely on, is an example of a family of graphical models that have been applied to NLP at least since [48], where their application to word sense disambiguation data is illustrated for nearly three dozen words, with an average of 8.5 senses each. GLAD assumes

[12] Due to lack of resources and time, we could not do all round 2.2. words.

that items vary in difficulty, and that labelers vary in accuracy [33]. It treats the true labels, labeler accuracy and instance difficulty as hidden variables to be inferred probabilistically from the observed multilabels, as illustrated in Figure 6.[13] From the distribution of observed labels $L_{ij}$ from $i$ annotators on $j$ instances, it learns the probability of true labels $Z_i$, given inferred annotator accuracies $\alpha_i$ and instance difficulties $\beta_j$:

$$p(L_{ij} = Z_j | \alpha_i, \beta_j) = \frac{1}{1 + e^{-\alpha_i \beta_j}}$$

Maximum likelihood estimates of the model parameters are obtained using Expectation-Maximization. The approach outperforms majority voting on several image datasets. The method of generating an integrated labeling quality proposed by [47] also outperformed majority voting.

Although majority voting has been proposed as a way to combine labels from multiple sources, it does not perform well in our case. When we compared the results on the five types of multilabel sources (e.g., trained annotators versus turkers) for the seven word sense tasks (thirty-five cases), we found that GLAD significantly outperformed majority voting twenty-six times out of the thirty-five on accuracy and F-measure; in the remaining nine cases GLAD results did not differ significantly from a majority vote.

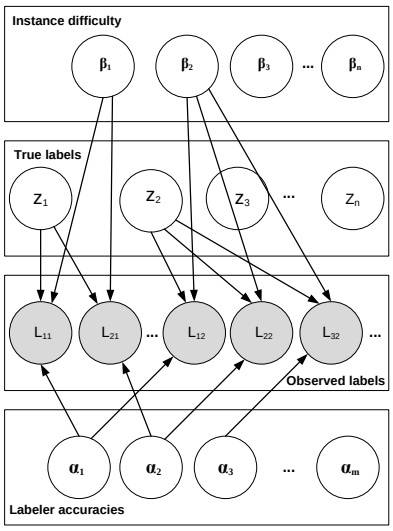



Fig. 6: Graphical model of instance difficulties ($\beta$), true labels ($Z$), observed labels ($L$), and labeler accuracies ($\alpha$; from [33]).

GLAD is designed to learn a binary classification, so we prepare seven learning tasks, using the highest frequency senses for each word: senses 1 and 2 of *fair*, senses 1 and 2 of *long*, and senses 1 through 3 of *quiet*. Column 2 of Table 6 shows the number of positive and negative instances (out of 100) assigned by the expert for each task. We run five experiments on each learning task, using different sets of labels from trained or untrained annotators. In the first experiment, GLAD is applied to the five or six labels from the trained annotators (MASC), including the outliers. In the second, learning is from the best subset of size 5 from the turkers' labels (AMT best5); these are the turkers from Table 5. In the third, learning is from subsets of size 6 from the turkers' labels: 50 random samples of size 6 are selected for each sense, and the average over the 50 samples is reported (AMT subsets$_{avg}$). In the fourth, all the turkers labels are used for learning (AMT all). In the fifth, GLAD is applied to the combination of labels from trained annotators and turkers (COMB). Evaluation uses the ground truth labels described in section 3.5. To evaluate performance, we report recall, precision, and F-measure on the positive class, and accuracy. Table 6 shows GLAD performance for the five experiments.[14] The rows with the highest recall, precision, F measure and accuracy are in boldface.

[13] GLAD is available from `http://mplab.ucsd.edu/~jake`.

[14] Note that the learning performance results for AMT subsets$_{avg}$ are averages over fifty iterations.

| WN Sense | Pos./Neg. | Exp | Rec | Pre | F | Acc |
|---|---|---|---|---|---|---|
| fair, sense 1 | 52/48 | **MASC** | **0.92** | **0.94** | **0.93** | **0.93** |
| | | AMT best5 | 0.66 | 0.70 | 0.68 | 0.78 |
| | | AMT subsets$_{avg}$ | 0.59 | 0.75 | 0.67 | 0.78 |
| | | AMT all | 1.00 | 0.71 | 0.85 | 0.79 |
| | | COMB | 1.00 | 0.74 | 0.87 | 0.82 |
| fair, sense 2 | 20/80 | MASC | 0.69 | 0.48 | 0.58 | 0.83 |
| | | AMT best5 | 0.45 | 1.00 | 0.73 | 0.97 |
| | | AMT subsets$_{avg}$ | 0.45 | 0.89 | 0.67 | 0.95 |
| | | **AMT all** | **0.81** | **0.93** | **0.87** | **0.96** |
| | | **COMB** | **0.81** | **0.93** | **0.87** | **0.96** |
| long, sense 1 | 57/43 | MASC | 0.88 | 0.84 | 0.86 | 0.84 |
| | | AMT best5 | 0.64 | 0.98 | 0.81 | 0.99 |
| | | AMT subsets$_{avg}$ | 0.36 | 0.98 | 0.55 | 0.74 |
| | | **AMT all** | **1.00** | **0.98** | **0.99** | **0.99** |
| | | **COMB** | **1.00** | **0.98** | **0.99** | **0.99** |
| long, sense 2 | 38/62 | MASC | 0.74 | 0.80 | 0.77 | 0.83 |
| | | AMT best5 | 0.69 | 0.92 | 0.81 | 0.93 |
| | | AMT subsets$_{avg}$ | 0.57 | 0.94 | 0.72 | 0.88 |
| | | AMT | 0.79 | 0.94 | 0.86 | 0.90 |
| | | **COMB** | **0.95** | **0.97** | **0.96** | **0.97** |
| quiet, sense 1 | 12/88 | **MASC** | **0.94** | **0.86** | **0.90** | **0.93** |
| | | AMT best5 | 0.50 | 0.92 | 0.71 | 0.87 |
| | | AMT subsets$_{avg}$ | 0.12 | 0.84 | 0.33 | 0.71 |
| | | AMT all | 0.00 | 0.00 | 0.00 | 0.66 |
| | | COMB | 0.50 | 0.94 | 0.72 | 0.82 |
| quiet, sense 2 | 21/79 | MASC | **0.78** | 0.64 | 0.71 | 0.88 |
| | | AMT best5 | 0.36 | 0.70 | 0.53 | 0.90 |
| | | AMT subsets$_{avg}$ | 0.19 | 0.79 | 0.41 | 0.87 |
| | | AMT | 0.10 | **1.00** | 0.55 | 0.86 |
| | | **COMB** | 0.61 | **1.00** | **0.81** | **0.93** |
| quiet, sense 3 | 13/87 | **MASC** | **0.60** | **1.00** | **0.80** | **0.82** |
| | | AMT best5 | 0.45 | 1.00 | 0.72 | 0.81 |
| | | AMT subsets$_{avg}$ | 0.14 | 0.95 | 0.54 | 0.63 |
| | | AMT | 0.05 | 1.00 | 0.53 | 0.58 |
| | | COMB | 0.42 | 1.00 | 0.71 | 0.74 |

Table 6: GLAD Results for six experiments

*Experiment 1–Half a dozen trained annotators (MASC)* This experiment, which used all the labels from the MASC annotators, addressed whether there is an advantage to a smaller set of labels from trained annotators. In three of the learning tasks, GLAD learned best from the trained annotators: sense 1 of *fair*, and senses 1 and 3 of *quiet*. For sense 2 of *quiet*, MASC labels were competitive with or better than all but COMB. For sense 1 of *long*, GLAD MASC results were better than AMT subsets$_{avg}$, about the same as AMT best5, and not as good as AMT. For sense 2 of *fair*, MASC labels yielded the poorest GLAD performance of the 5 sets of multilabels.

*Experiment 2–Best subset of five turkers (AMT best5)* This experiment addressed whether selecting high quality subsets of turkers could yield GLAD results equivalent to learning from the same number of labels from trained annotators. The answer was yes for sense 2 of *fair*, both senses of *long*, and no otherwise. AMT best5 was never the best, but was close to best on sense 2 of *fair*.

*Experiment 3–Average over random subsets of half a dozen turkers (AMT subsets$_{avg}$)* This experiment addressed the quality of learning a ground truth label by averaging over fifty iterations of random subsets of six turkers. For sense 2 of *fair*, the AMT subsets$_{avg}$ multilabels led to better performance than the MASC multilabels and nearly as good as the best (COMB). For both senses of *fair*, AMT subsets$_{avg}$ was equivalent or almost equivalent to learning from labels from the best subsets of turkers.

*Experiment 4–Fourteen turkers (AMT)* This experiment addressed whether with untrained annotators, doubling the number of labels always improves results, as reported elsewhere [29] [34]. Learning from all the turkers improved over AMT subsets$_{avg}$ for senses of *fair* and *long*. For senses 2 and 3 of *quiet*, performance on AMT was rather comparable to AMT subsets$_{avg}$, but for *quiet* it was quite a bit lower on accuracy (0.66 versus 0.71), and was zero for recall, precision and f-measure. The low performance here is due to the fact that the negative label was always assigned; the probability of the positive label is so low (0.06) that given the relatively few instances, it becomes harder to estimate its probability using EM. However, the AMT labels did produce the highest or next highest results for three senses (sense 2 of *fair*, and both senses of *long*). Overall, experiment 4 results were good but not the best on accuracy, and were often poor on F measure. As we will see next, the combination of all turkers with trained annotators never degraded results.

*Experiment 5–Combination of trained annotators and turkers (COMB)* This experiment addressed whether combining labels from trained and untrained annotators improves results. Results improved over untrained labels alone in four of the seven cases, and were roughly equivalent in the remaining cases.

*Comparison across experiments* The comparison of the five cases does not yield consistent results across the seven learning tasks. Learning from trained annotators often yields results closest to an expert's labels, but not always. Learning from many turkers' labels is as good or better than from fewer trained annotators only half the time. This suggests that the overall quality of the set of multilabels might matter when using less than the maximum set of multilabels (COMB). We next ask whether assessments of the sets of labels for each experiment sheds any light on the pattern of results.

The assessment metrics presented in section 5 were for all senses per word. Because the GLAD experiments use a modified form of the data in which all labels other than the target sense are treated as *Other*, we recompute the assessment metrics using this binary representation for each target sense. Table 7 gives the pairwise agreement and $\alpha$ scores across all annotators for a given experiment on a given binary sense label.[15] For a given sense label, such as sense 1 of *fair*, the new data representation obscures the fact that annotators who did not choose sense 1 might have disagreed with sense 1 in different ways (e.g., sense 2 versus sense 3), therefore the absolute values of the assessments no longer measure the actual agreement. Howevever, they still show the relative degrees of agreement. Thus the trained annotators (MASC) have a slighly lower pairwise agreement on *fair* sense 1 (0.82) than the best subset of five turkers (AMT best5: 0.89), but on average, the MASC annotators sense

[15] This table reports averages. The column headed $\overline{Lev}$ gives the average across all annotators of Lev(P$_a(k), \overline{P(k)}$), followed by the average JSD ($\overline{JSD}$) for all pairs of annotators $a, b, where a \neq b$ of $JSD(P_a(k), P_b(k))$, followed by the average KLD' ($\overline{KLD'}$): $KLD'(P_a(k), \overline{P_b(k)}), b \neq a$. Note that the assessment results for AMT subsets$_{avg}$ are averages of averages over fifty iterations.

| WN Sense | Pos./Neg. | Exp | Agt. | $\alpha$ | $\overline{Lev}$ | $\overline{JSD}$ | $\overline{KLD'}$ |
|---|---|---|---|---|---|---|---|
| fair, sense 1 | 52/48 | **MASC** | 0.82 | 0.65 | **0.089** | **0.004** | **0.011** |
| | | AMT best5 | 0.89 | 0.77 | 0.122 | 0.008 | 0.019 |
| | | AMT subsets$_{avg}$ | 0.65 | 0.28 | 0.354 | 0.097 | 0.213 |
| | | AMT all | 0.67 | 0.33 | 0.337 | 0.081 | 0.203 |
| | | COMB | 0.59 | 0.18 | 0.235 | 0.045 | 0.108 |
| fair, sense 2 | 20/80 | MASC | 0.79 | 0.45 | 0.108 | 0.009 | 0.022 |
| | | **AMT best5** | 0.93 | 0.82 | **0.038** | **0.001** | **0.003** |
| | | AMT subsets$_{avg}$ | 0.79 | 0.35 | 0.123 | 0.016 | 0.038 |
| | | **AMT all** | 0.77 | 0.43 | 0.090 | 0.020 | 0.070 |
| | | **COMB** | 0.70 | 0.24 | 0.093 | 0.019 | 0.050 |
| long, sense 1 | 57/43 | MASC | 0.85 | 0.69 | 0.177 | 0.026 | 0.061 |
| | | AMT best5 | 0.80 | 0.59 | **0.170** | **0.015** | **0.037** |
| | | AMT subsets$_{avg}$ | 0.68 | 0.22 | 0.388 | 0.135 | 0.289 |
| | | **AMT all** | 0.66 | 0.22 | 0.453 | 0.140 | 0.272 |
| | | **COMB** | 0.60 | 0.16 | 0.406 | 0.110 | 0.206 |
| long, sense 2 | 38/62 | MASC | 0.93 | 0.86 | **0.024** | **0.000** | **0.000** |
| | | AMT best5 | 0.89 | 0.74 | 0.069 | 0.004 | 0.010 |
| | | AMT subsets$_{avg}$ | 0.75 | 0.34 | 0.183 | 0.052 | 0.101 |
| | | AMT all | 0.76 | 0.29 | 0.180 | 0.047 | 0.081 |
| | | **COMB** | 0.68 | 0.17 | 0.171 | 0.045 | 0.077 |
| quiet, sense 1 | 12/88 | **MASC** | 0.79 | 0.57 | **0.129** | **0.012** | **0.028** |
| | | AMT best5 | 0.77 | 0.45 | **0.120** | **0.010** | **0.026** |
| | | AMT subsets$_{avg}$ | 0.72 | 0.10 | 0.209 | 0.059 | 0.126 |
| | | AMT all | 0.72 | 0.10 | 0.213 | 0.052 | 0.100 |
| | | COMB | 0.73 | 0.11 | 0.218 | 0.058 | 0.110 |
| quiet, sense 2 | 21/79 | MASC | 0.81 | 0.45 | **0.122** | **0.019** | **0.047** |
| | | AMT best5 | 0.75 | 0.42 | 0.214 | 0.035 | 0.082 |
| | | AMT subsets$_{avg}$ | 0.70 | 0.12 | 0.188 | 0.044 | 0.095 |
| | | AMT | 0.69 | 0.17 | 0.226 | 0.051 | 0.100 |
| | | **COMB** | 0.70 | 0.13 | 0.196 | 0.046 | 0.088 |
| quiet, sense 3 | 13/87 | **MASC** | 0.85 | 0.59 | 0.130 | 0.013 | 0.030 |
| | | AMT best5 | 0.82 | 0.45 | 0.112 | 0.012 | 0.028 |
| | | AMT subsets$_{avg}$ | 0.72 | 0.08 | 0.105 | 0.011 | 0.027 |
| | | AMT | 0.74 | 0.11 | **0.093** | **0.007** | **0.016** |
| | | COMB | 0.72 | 0.09 | 0.105 | 0.010 | 0.022 |

Table 7: Five assessment metrics on labels for the five learning experiments

distributions are more similar to one another than the AMT best5 annotators, as reflected by the lower average values for AMT best5 of leverage, $\overline{JSD}$ and KLD'.

For each learning task (e.g., sense 1 of *fair*), the experiment label in column 2 is in boldface for the experiment that had the best result, or the experiments that had similarly good results. Values in columns 6 through 8 (the probability-based assessment metrics) are in boldface to indicate which of the five sets of labels had the best (lowest) values for average leverage, $\overline{JSD}$ and KLD'. Here we see a possible explanation for the difference in performance shown in Table 6. There is an apparent trend for GLAD to perform well in predicting expert labels when the sense distributions across annotators are similar. In four of the seven learning tasks, GLAD results are best for the set of multilabels that had the lowest average leverage, $\overline{JSD}$ and KLD' (senses 1 and 2 of *fair*, sense 2 of *long*) or nearly tied for the lowest (sense 1 of *quiet*). In a fifth case–sense 3 of *quiet*–the probability-based metrics are rather low in all the experiments, and the two that had the highest GLAD performance (MASC, AMT best5) are the only ones that had non-chance values of $\alpha$. While there are no strong correlations of F-measure or accuracy with any of our metrics, the density in

the lower right corner of Figure 7 shows an association between accuracies above 0.80 and KLD' below 0.10. We see a similar pattern for $\overline{JSD}$ and Leverage.

The two remaining cases are somewhat anomalous. COMB had the highest performance for sense 2 of *quiet*, but nothing in the assessment data to distinguish this experiment among the five. All had relatively low $\overline{JSD}$ and $\overline{KLD'}$; three of the experiments had relatively low $\overline{Lev}$ along with a distribution of $\alpha$ scores similar to sense 3 of *quiet*. For sense 1 of *long*, we see no explanation for the unusually good GLAD performance for AMT all and COMB. $\overline{JSD}$ and $\overline{KLD'}$ are low, while $\overline{Lev}$ is rather high (0.453 and 0.406). The number of instances is high (57), but is also high for sense 1 of *fair* (52). Comparison of the average values for annotator accuracy and item difficulty produced by GLAD was also unrevealing.

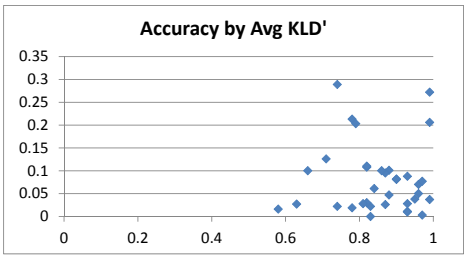



Fig. 7: Plot of accuracy (x-axis) by $\overline{KLD}$ (y-axis), all experiments.

In summary, the results presented here suggest there is no a priori best number of annotators or level of annotator training that consistently yields expert quality labels. On the other hand, it is possible to learn a single label close to expert quality. Further, it appears that crowdsourcing could substitute for trained labelers even on word sense labeling using fine-grained sense inventories given a sufficient number of labelers with sufficient consistency in sense distributions.

## 7 Discussion

Regarding our first question of how to collect word sense labels for moderately polysemous words, to assess the annotation quality, we have shown the aptness of using leverage, JSD and KLD to compare distributions of word sense in data from multiple annotators. We find that the annotation procedure we followed is reliable, and that it is possible to collect reliable labels from trained annotators for some polysemous words. For other words, the sense labels cannot be applied as reliably. Because the same annotators follow the same procedures well with some words, we assume that lower performance on other words is due to properties of the words and their sense inventories, or to the contexts of use, or both. Besides assisting in the identification of outliers, these metrics help pinpoint the source of similarities among subsets of annotators who agree well with one another, but not with other subsets. In previous work, we speculated that confusability of pairs of sense labels for a given word correlated with an inter-sense similarity measure [49], or with the relative concreteness of senses, or with specificity of contexts of use [50]. To go beyond speculation would require much more data than we have investigated here, so this is an endeavor we leave for future work.

We have also explored whether we can posit criteria to collect sets of multilabels that can be used to infer a true sense label for each instance, give our task of labeling contexts of relatively polysemous words. Results from our suite of metrics indicate that there is a trend for accuracy of the inferred true label to be higher when annotators' probability distributions over senses are more similar. Previous work on the use of noisy labels investigated

the performance of inferring a true label given multilabels from annotators with a uniform probability of being correct, and found that as long as the probability of a true label was greater than 0.5, adding new labelers eventually yielded high performance [47]. In the same spirit, we suggest that for the word sense task, it might be possible to monitor quality as noisy labels are collected from untrained annotators, and to continue acquiring new labels until a certain quality threshold is reached. We are currently pursing this possibility, using Bayesian methods to estimate annotator quality, and to learn a full set of sense categories rather than binary senses.

For semantic and pragmatic distinctions, it is to be expected that some judgments are more difficult to make than others, and that they will give rise to less agreement among annotators working independently. In such cases, interesting patterns that would not be apparent using only two or three annotators are revealed by collecting labels from multiple annotators. In particular, multilabels provide greater evidence for instances that are more difficult for everyone to agree on. Figure (5b) illustrated two examples where there is no single high probability sense for *quiet*; there is no pattern to the disagreement. This contrast with example 1), where annotators were split evenly between two senses of *fair*, and where there is a systematic pattern of disagreement between senses 1 and 2 for many instances. For the cases of systematic disagreement between two senses, while it is difficult to assign a true sense label, it is clear that the true label is not any of the senses other than 1 or 2. Presumably, methods to distinguish among instances that lead to high agreement versus systematic disagreement versus noisy disagreement could increase our understanding of word sense, and improve the performance of automated word sense disambiguation systems.

## 8 Conclusion

We presented a dataset consisting of word sense multilabels from trained and untrained annotators for moderately polysemous words. This data, which will be included with MASC releases, shows the benefits of multilabels for discriminating between instances that yield high agreement across annotators, those associated with a split among annotators (as in example 1), and those where annotators choose many senses (as in Figure (5b)). While it is expensive for individual research groups to collect such data, incorporating it as part of a community resource provides researchers an opportunity to investigate in new ways the complex interaction among words, senses, contexts of use, and annotators.

The assessment of the multilabels demonstrates that word sense annotation based on labels from a relatively finer-grained sense inventory can achieve excellent to moderate reliability, depending on the word. We find the same range of differences in reliability across words for the entire MASC word sense corpus, using different sets of four well-trained annotators [46]. In previous work, we have suggested that the factors that differentiate words with respect to the reliability of their sense inventories might include characteristics that typify their contexts of uses, such as whether the contexts tend to be more specific; measurable properties of a word's sense inventory, such as inter-sense similarity; or other properties whose measurement remains difficult, such as relative concreteness of the senses. In the full set of reliability annotations for the MASC word sense corpus, there are roughly 90 additional words beyond those investigated here, each with four annotators per 100 instances for each word, for a total of 9,000 additional instances of multilabels of size 4. This should provide valuable data for investigating such factors more deeply.

Annotation, which has long been an investigative tool in Natural Language Processing, seems to be growing in importance, given the increasing number venues to report the results

of annotation projects. Relatively new venues include the Linguistic Annotation Workshops (LAW), and the inclusion of a *Resources/Evaluation* track for recent annual meetings of the Association for Computational Linguistics. This suggests that inexpensive methods to achieve high quality annotation will become increasingly important. To this end, our analysis of sets of multilabels from trained and untrained annotators on 1000 instances (100 for each of 10 words) includes a deeper investigation of 300 instances (for three adjectives). Our learning experiments demonstrate that expert quality labels for word sense can be learned from noisy multilabels acquired by crowdsourcing. At the same time, they also show that many questions remain to be addressed regarding the best tradeoff between the cost of adding new labelers and the quality of unsupervised learning of the *true* labels.

## 9 Acknowledgements

This work was supported by NSF award CRI-0708952, including a supplement to fund Vikas Bhardwaj as a Graduate Research Assistant for one semester. The authors thank the annotators for their excellent work and thoughtful comments on sense inventories. We thank Bob Carpenter for discussions about data from multiple annotators, and for his generous and insightful comments on drafts of the paper. Finally, we thank the anonymous reviewers who provided deep and thoughtful critiques.

## References

1. N. Ide, C. Baker, C. Fellbaum, R.J. Passonneau, The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the Association for Computational Linguistics* (2010), pp. 68-73.
2. G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: An on-line lexical database (revised). Tech. Rep. Cognitive Science Laboratory (CSL) Report 43, Princeton University, Princeton (1993). Revised March 1993.
3. J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, J. Scheffczyk. Framenet II: Extended theory and practice (2006). Available from http://framenet.icsi.berkeley.edu/index.php.
4. C.J. Fillmore, C.R. Johnson, M.R.L. Petruck, Background to FrameNet. International Journal of Lexicography **16**(3), 235-250 (2003).
5. D. Dowty, *Word Meaning and Montague Grammar* (D. Reidel, Dordrecht, 1979).
6. A. Kilgarriff, I don't believe in word senses. Computers and the Humanities **31**, 91-113 (1997).
7. K. Erk, Representing words as regions in vector space. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (2009), pp. 57-65.
8. T. Landauer, S. Dumais, A solution to Plato's problem: the latent semantic analysis theor of acquisition, induction, and representation of knowledge. Psychological Review **104**(2), 211-240 (1977).
9. K. Erk, D. Mccarthy, Graded word sense assignment. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 09)* (2009), pp. 440-449.
10. A. Kilgarriff, SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)* (Granada, 1998), pp. 581-588.
11. T. Pedersen, Assessing system agreement and instance difficulty in the lexical sample tasks of SENSEVAL-2. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (2002), pp. 40-46.
12. T. Pedersen, Evaluating the effectiveness of ensembles of decision trees in disambiguating SENSEVAL lexical samples. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (2002), pp. 81-87.
13. M. Palmer, H.T. Dang, C. Fellbaum, Making fine-grained and coarse-grained sense distinctions, both manually and automatically. Natural Language Engineering **13**(2), 137-163 (2007).
14. S. Manandhar, I. Klapaftis, D. Dligach, S. Pradhan, SemEval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)* (Association for Computational Linguistics, Uppsala, Sweden, 2010), pp. 63-68. URL `http://www.aclweb.org/anthology/S10-1011`.

15. E. Agirre, O.L. de Lacalle, C. Fellbaum, S.K. Hsieh, M. Tesconi, M. Monachini, P. Vossen, R. Segers, SemEval-2010 Task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (2010), pp. 75-80. URL `http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1283775102/publikoak/semeval`.
16. H.T. Ng, C.Y. Lim, S.K. Foo, A case study on inter-annotator agreement for word sense disambiguation. In *SIGLEX Workshop On Standardizing Lexical Resources* (1999).
17. J. Véronis, A study of polysemy judgements and inter-annotator agreement. In *SENSEVAL Workshop* (1998), pp. Sussex, England.
18. N. Ide, T. Erjavec, D. Tufis, Sense discrimination with parallel corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (2002), pp. 54-60.
19. R.J. Passonneau, N. Habash, O. Rambow, Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* (Genoa, Italy, 2006), pp. 1951-1956.
20. R. Snow, D. Jurafsky, A.Y. Ng, Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Prague, 2007), pp. 1005-1014.
21. I. Chugur, J. Gonzalo, F. Verdejo, Polysemy and sense proximity in the SENSEVAL-2 test suite. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (Philadelphia, 2002), pp. 32-39.
22. M. Diab, Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (2004), pp. 303-311.
23. P. Resnik, D. Yarowsky, Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. Natural Language Engineering **5**(2), 113-133 (1999).
24. N. Ide, Cross-lingual sense determination: Can it work? Computers and the Humanities: Special Issue on the Proceedings of the SIGLEX/SENSEVAL Workshop **34(1-2)**, 223-234 (2000).
25. D. Klein, G. Murphy, Paper has been my ruin: Conceptual relations of polysemous words. Journal of Memory and Language **47**, 548 (2002).
26. N. Ide, Y. Wilks, Making sense about sense. In *Word Sense Disambiguation: Algorithms and Applications*, ed. by E. Agirre, P. Edmonds (Springer, Dordrecht, The Netherlands, 2006), pp. 47-74.
27. K. Erk, D. McCarthy, N. Gaylord, Investigations on word senses and word usages. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing* (2009), pp. 10-18.
28. M. Poesio, R. Artstein, The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky* (2005), pp. 76-83.
29. R. Snow, B. O'Connor, D. Jurafsky, A.Y. Ng, Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)* (Honolulu, 2008), pp. 254-263.
30. C. Callison-Burch, Fast, cheap, and creative: evaluating translation quality using Amazons Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Morristown, NJ, 2009), pp. 286-295.
31. S. Pradhan, E. Loper, D. Dligach, M. Palmer, SemEval-2007 Task-17: English lexical sample, SRL and all words. In *Proceedings of Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (Prague, Czech Republic, 2007), pp. 87-92.
32. Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L.G. Moy, J. Dy, Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2010), pp. 932-939.
33. J. Whitehill, P. Ruvolo, T. fan Wu, J. Bergsma, J. Movellan, Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, A. Culotta (MIT Press, 2000), pp. 2035-2043.
34. V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds. Journal of Machine Learning Research **11**, 1297-1322 (2010).
35. C. Callison-Burch, M. Dredze, Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (2010), pp. 1-12.
36. C. Akkaya, A. Conrad, J. Wiebe, R. Mihalcea, Amazon Mechanical Turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Association for Computational Linguistics, Los Angeles, 2010), pp. 195-203. URL `http://www.aclweb.org/anthology/W/W10/W10-0731.pdf`.

37. K. Krippendorff, *Content analysis: An introduction to its methodology* (Sage Publications, Beverly Hills, CA, 1980).
38. V. Bhardwaj, R.J. Passonneau, A. Salleb-Aouissi, N. Ide, Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)* (2010).
39. W.A. Scott, Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly **17**, 321-325 (1955).
40. J. Cohen, A coeffiecient of agreement for nominal scales. Educational and Psychological Measurement **20**, 37-46 (1960).
41. G. Piatetsky-Shapiro, Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, ed. by G. Piatetsky-Shapiro, W.J. Frawley (AAAI Press, 1999), pp. 229-248.
42. S. Kullback, R.A. Leibler, On information and sufficiency. Annals of Mathematical Statistics **22**(1), 79-86 (1951).
43. J. Lin, Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory **37**(1), 145-151 (1991).
44. N. Lavrac, P.A. Flach, B. Zupan, Rule evaluation measures: a unifying view. In *Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP-99)* (1999), pp. 174-185.
45. E. Hovy, M. Marcus, M. Palmer, L. Ramsha, R. Weischedel, Ontonotes: The 90% solution. In *Proceedings of HLT-NAACL 2006* (2006), pp. 57-6.
46. R.J. Passonneau, C. Baker, C. Fellbaum, N. Ide. The MASC word sense sentence corpus (Submitted).
47. V.S. Sheng, F. Provost, P.G. Ipeirotis, Get another label? improving data quality and data mining using multiple noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, New York, NY, USA, 2008), KDD '08, pp. 614-622.
48. R.F. Bruce, J.M. Wiebe, Decomposable modeling in natural language processing. Computational Linguistics **25**(2), 195-208 (1999).
49. R.J. Passonneau, A. Salleb-Aouissi, V. Bhardwaj, N. Ide, Word sense annotation of polysemous words by multiple annotators. In *Seventh International Conference on Language Resources and Evaluation (LREC)* (2010).
50. R.J. Passonneau, A. Salleb-Aouissi, N. Ide, Making sense of word sense variation. In *Proceedings of the NAACL-HLT 2009 Workshop on Semantic Evaluations* (2009).