# The MASC Word Sense Sentence Corpus

**Rebecca J. Passonneau*, Collin Baker†, Christiane Fellbaum‡, Nancy Ide¶**

*Columbia University
New York, NY, USA
becky@cs.columbia.edu

†ICSI
Berkeley, CA
collinb@icsi.berkeley.edu

‡Princeton University
Princeton, NJ, USA
fellbaum@princeton.edu

¶Vassar College
Poughkeepsie,NY, USA
ide@cs.vassar.edu

*Abstract content*

## 1. Introduction

Annotated corpora play an increasingly significant role in computational linguistics to address lexical, propositional, and discourse semantics. Each of these dimensions is important in its own right, yet also interact in the determination of meaning. As a result, there is an increased demand for high quality linguistic annotations of corpora representing a wide range of phenomena, especially at the semantic level, to support machine learning and computational linguistics research. At the same time, there is a demand for annotated corpora representing a broad range of genres, due to the increasingly apparent impact of domain on both syntactic and semantic characteristics. Finally, there is a keen awareness of the need for annotated corpora that are both easily accessible and available for use by anyone.

To address these needs, construction of the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2010) was undertaken in response to a community mandate at a 2006 workshop funded by the US National Science Foundation. MASC is a half million word corpus of American English language data drawn from the 15 million word Open American National Corpus (OANC).[1] It is annotated for an increasing number of linguistic phenomena, all of which are either manually produced or validated, and all of the data and annotations are freely available from the MASC website and the Linguistic Data Consortium (LDC).[2]

Here we give an overview of the contents of MASC and then focus on the word sense sentence corpus, describing the characteristics that differentiate it from other word sense corpora. To date, MASC has been developed and created through a combination of the efforts of the MASC collaborators and contributions from other annotation projects. However, for MASC to grow in size and in number of annotation types, we aim to encourage a variety of collective efforts. We briefly discuss a range of such efforts, including the use of Amazon Mechanical Turkers to perform word sense annotation, and ways in which the community can help grow the corpus and its annotations.

## 2. MASC Contents

MASC currently contains nineteen genres of spoken and written language data in roughly equal amounts, shown in Table 1. Roughly 15% of the corpus consists of spoken transcripts, both formal (court and debate transcripts) and informal (face-to-face, telephone conversation, etc.); the remaining 85% covers a wide range of written genres, including emerging social media genres (tweets, blogs). Because it is drawn from the OANC, all MASC data represents contemporary American English produced since 1990.

The entire MASC is annotated for logical structure, token and sentence boundaries, part of speech and lemma, shallow parse (noun and verb chunks), named entities (person, location, organization, date), and Penn Treebank syntax. Portions of MASC are also annotated for additional phenomena, including 40K of full-text FrameNet frame element annotations and PropBank, TimeML, and opinion annotations over a roughly 50K subset of the data. The list of annotation types and coverage is given in Table 2.

### 2.1. Open data, annotations, and usage

MASC differs from any existing linguistically-annotated corpus in that it represents a wide range of registers of contemporary American English, includes diverse genres, and–perhaps most notably–it is completely open: MASC data, like OANC data, is in the public domain or under a license that does not restrict redistribution of the data or its use for any purpose, including commercial (e.g., the Creative Commons Attribution license[3]). Data under li-

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

[2] urlhttp://www.ldc.upenn.edu.

[3] http://creativecommons.org/licenses/by/3.0/

| Genre | No. files | No. words | Pct corpus |
|---|---|---|---|
| Court transcript | 2 | 30052 | 6% |
| Debate transcript | 2 | 32325 | 6% |
| Email | 78 | 27642 | 6% |
| Essay | 7 | 25590 | 5% |
| Fiction | 5 | 31518 | 6% |
| Gov't documents | 5 | 24578 | 5% |
| Journal | 10 | 25635 | 5% |
| Letters | 40 | 23325 | 5% |
| Newspaper | 41 | 23545 | 5% |
| Non-fiction | 4 | 25182 | 5% |
| Spoken | 11 | 25783 | 5% |
| Technical | 8 | 27895 | 6% |
| Travel guides | 7 | 26708 | 5% |
| Twitter | 2 | 24180 | 5% |
| Blog | 21 | 28199 | 6% |
| Ficlets | 5 | 26299 | 5% |
| Movie script | 2 | 28240 | 6% |
| Spam | 110 | 23490 | 5% |
| Jokes | 16 | 26582 | 5% |
| TOTAL | 376 | 506768 | |

Table 1: Genre distribution in MASC

| Annotation type | No. words |
|---|---|
| Logical | 506659 |
| Token | 506659 |
| Sentence | 506659 |
| POS/lemma (GATE) | 506659 |
| POS (Penn) | 506659 |
| Noun chunks | 506659 |
| Verb chunks | 506659 |
| Named Entities | 506659 |
| FrameNet | 39160 |
| Penn Treebank | *506659 |
| PropBank | 55599 |
| Opinion | 51243 |
| TimeBank | *55599 |
| Committed Belief | 4614 |
| Event | 4614 |
| Dependency treebank | 5434 |

* under development

Table 2: Summary of MASC annotations

censes such as GNU General Public License[4] or Creative Commons Attribution-ShareAlike[5] are avoided because of the potential obstacles to use of the resource for commercial purposes imposed by the requirement to redistribute under the same terms. All MASC annotations are similarly open. All data and annotations are downloadable from http://www.anc.org/MASC.

Openness in MASC applies to not only acquisition and use, but also interoperability with diverse software and systems for searching, processing, and enhancing the corpus. All MASC annotations are represented in the ISO TC37 SC4 Linguistic Annotation Framework (LAF) GrAF format (Ide and Suderman, Submitted), with the objective to make the annotations as flexible for use with common tools and frameworks as possible. The ANC project provides a web application, called ANC2Go[6] that enables a user to choose any portion or all of MASC and the OANC together with any of their annotations to create a "customized corpus" that can be delivered in any of several widely used formats such as CONLL IOB, RDF, inline XML, etc. Modules to transduce GrAF to formats consistent with other tools and frameworks such as UIMA, GATE, and NLTK are also provided[7].

## 3. The MASC Word Sense Sentence Corpus

MASC also includes sense-tags for approximately 1000 occurrences of each of 104 words chosen by the WordNet and FrameNet teams (ca. 100,000 annotated occurrences). The sense-tagged data are distributed as a separate *sentence corpus* with links to the original documents in which they appear. Several inter-annotator agreement studies and resulting statistics have been published (Passonneau et al., 2009; Passonneau et al., 2010a), many of which are distributed with the corpus.

Word meanings are both elusive and central to many areas of NLP, such as machine translation. They are elusive because they vary in partly predictable and partly unpredictable ways due to the unique combination of the lexicogrammatical context a word occurs in, together with inferences licensed by the context of use. Although there have been many recent efforts to create corpora with sense annotations (e.g., (Hovy et al., 2006; Burchardt and Pennacchiotti, 2008)), the complete MASC word sense sentence corpus complements these in several ways. First, it focuses on finding a large number of instances of each word, and is thus analogous to the FrameNet corpus (Baker and Sato, 2003), with which it compares in size. Second, it includes words that are moderately polysemous, and selects instances from a very heterogenous, open source corpus. Here we briefly describe its design, quality and potential uses.

### 3.1. Methods

Each round of sense annotation consists of ten words selected by the co-authors, annotated using WordNet senses as labels. One of the criteria is to select words as test cases for aligning WordNet senses (Fellbaum, 1998) with FrameNet lexical units (Ruppenhofer et al., 2006), thus Christiane Fellbaum, one of the architects of WordNet, and Collin Baker, the FrameNet project manager, provide the most input. The other criteria are to achieve a rough balance between the number of nouns and verbs, with somewhat fewer adjectives; to include words with more than three or four senses but fewer than twenty or so; to include words with at least one thousand instances in MASC; and to select words used in many of the MASC genres. For some words with somewhat lower frequencies, MASC sentences have been supplemented with sentences drawn from the Open American National Corpus.

Twelve undergraduates, four at Columbia University and eight at Vassar College, performed the word sense annotation. Most performed several rounds, with 4.3 on average.

[4] http://www.gnu.org/licenses/gpl.html

[5] http://creativecommons.org/licenses/by-sa/2.5/

[6] http://www.anc.org:8080/ANC2Go/

[7] http://www.anc.org/tools/

| | | |
|---|---|---|
| 2 | Gloss | *come into sight or view* |
| | Example | He suddenly appeared at the wedding |
| 5 | Gloss | *come into being or existence, or appear on the scene* |
| | Example | Homo sapiens appeared millions of years ago |

Figure 1: 2 of 7 WordNet senses for *appear-v*

All were trained prior to performing any annotation using guidelines created by Christiane Fellbaum. Each round had an initial step that constituted training on the sense inventory. Annotation of each round was carried out in two additional steps, using the SATANiC (Sense Annotation Tool for the ANC) graphical user interface. SATANiC connects directly to the ANC repository, so annotators can check out or commit their work.

Three to four annotators participated in most rounds. The purpose of the initial step was to familiarize the subset of annotators assigned to a given round of approximately ten words with the WordNet sense inventory for each word in the round, and to review that inventory in consultation with Christiane Fellbaum, followed by a possible revision as described below. In this step, 50 instances of each word were annotated by all the annotators for the round. In step 2, 900 instances were annotated by one annotator each. Step 3 consisted of 100 additional instances annotated by all the annotators in order to document interannotator agreement (see next section). The combined sentences from step 2 and step 3 constitute approximately 1000 sentences for the MASC sentence corpus of word senses.

When the sense annotation began, the current WordNet version was 3.0. If revisions to any of the sense inventories were required, as determined by Christiane Fellbaum and the annotators, the revisions were made and added to a working copy of WordNet, pending new releases. Many of the revisions became part of WordNet 3.1, and correspondingly, subsequent revisions will be included in later releases of WordNet.

For each new word, annotators apply the same general procedures, but learn a new set of sense labels. Examples of the general procedures are 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 relations among these senses during annotation. Figure 1 shows what annotators see for two of the seven WordNet senses for *appear-v*, giving the WordNet sense number, its gloss (or definition) in italics, and phrases exemplifying its usage. Like one of the examples in the MASC annotation guidelines, the two senses of *appear-v* shown here are similar, but their sense relations (troponyms, antonyms, etc.) further discriminate the senses. For example, senses 2 and 5 have different antonyms: sense 1 of *disappear* (*get lost, as without warning or explanation*), versus sense 3 of *disappear* (*cease to exist*), respectively.

In sum, annotators were trained with the same guidelines, had a trial annotation round for each word, used the same annotation tool, and on average, acquired experience over the course of four rounds.

### 3.2. Interannotator Agreement

The purpose of measuring interannotator agreement is to confirm that the annotation guidelines and procedures are reliable, meaning that different people can perform the same annotation and produce roughly equivalent results. An agreement coefficient such as Krippendorff's $\alpha$, a commonly used agreement coefficient (Artstein and Poesio, 2008), is a descriptive statistic that reports the proportion of observed agreement that exceeds the agreement that would be expected if annotators assigned labels randomly, given some estimate of the probability of each label (e.g., its rate over all annotators). Here, the labels for a given word are its WordNet senses, plus an additional label (Other) for cases where the sentence is not a true example for some reason, or where no WordNet sense applies. The $\alpha$ metric takes on values in [-1,1] for binary data, or in (-1,1] for data with more than two labels: 1 represents perfect agreement, 0 represents no difference from agreement predicted by chance, and -1 represents perfect opposition of two labels. Our previous work (for its use in an alternative sense annotation task, see (Passonneau et al., 2006)).

As noted in (Artstein and Poesio, 2008), when agreement coeffcients are used in the medical literature, values above 0.40 indicate good agreement. Although Arstein and Poesio recommend values of at least 0.80 on many tasks, they note for tasks like word sense annotation, where labels can be more or less similar (cf. senses 2 and 5 of *appear*, a weighted coefficient as in (Passonneau et al., 2006) would be more appropriate. In later rounds, annotators could select more than one label. Because these well-trained annotators often achieve excellent agreement, we take values above 0.50 with unweighted $\alpha$ to represent good agreement.

Table 3 gives the three highest and three lowest $\alpha$ values across four annotators for words representing each part of speech. Annotators agree well on sense annotation of some MASC words and not others, with no obvious single explanation for the variation. For 37 words in rounds 7-10, the range is from a moderately low negative value of -0.02 on *normal-j* (3 senses) to an excellent 0.88 on *strike-n* (7 senses). Pearson's correlation coefficient shows no correlation of $\alpha$ with number of senses; $\rho$=0.07 (-0.25 for nouns; 0.48 for adjectives; 0.24 for verbs).

We find a great deal of variation not only across words, but among subsets of annotators within words. Column six of Table 3 shows the highest $\alpha$ score among the four cases where one annotator is dropped, with large increases for *poor, common* and *particular* among the adjectives, *number, control, level* and *family* among the nouns, and *ask, trace* and *fold* among the verbs.

Finally, the last column of Table 3 shows the value of $\alpha$ weighted by the MASI metric for comparing sets of labels. If both annotators assign a single sense to a given instance, $\alpha_{MASI}$ treats the comparison the same as $\alpha$, meaning every pair of annotators either agrees or not. If at least one assigns multiple labels, then $\alpha_{MASI}$ gives partial credit if there is any overlap, as described in (Passonneau, 2006). As a result, weighted $\alpha$ is never lower than unweighted. The weighted agreement is typically the same because most annotators assign only a single label. The $\alpha_{MASI}$ values for

| Rnd | Word | Pos | Senses | $\alpha$ | -1 Ann | $\alpha_{masi}$ |
|---|---|---|---|---|---|---|
| 9 | late | adj | 9 | 0.83 | 0.87 | **0.84** |
| 10 | high | adj | 9 | 0.82 | 0.85 | 0.82 |
| 7 | poor | adj | 13 | 0.54 | 0.67 | **0.59** |
| 10 | common | adj | 13 | 0.40 | 0.53 | 0.40 |
| 10 | particular | adj | 9 | 0.21 | 0.30 | 0.21 |
| 9 | normal | adj | 3 | -0.02 | 0.08 | -0.02 |
| 7 | strike | noun | 7 | 0.88 | 0.93 | 0.88 |
| 8 | state | noun | 11 | 0.72 | 0.78 | **0.73** |
| 8 | number | noun | 8 | 0.68 | 0.79 | 0.68 |
| 10 | control | noun | 15 | 0.35 | 0.44 | 0.35 |
| 9 | level | noun | 12 | 0.22 | 0.30 | **0.26** |
| 8 | family | noun | 16 | 0.14 | 0.26 | **0.35** |
| 8 | live | verb | 13 | 0.70 | 0.73 | 0.70 |
| 9 | appear | verb | 8 | 0.61 | 0.70 | 0.61 |
| 10 | book | verb | 10 | 0.63 | 0.68 | **0.65** |
| 8 | ask | verb | 8 | 0.05 | 0.48 | 0.05 |
| 7 | trace | verb | 13 | 0.10 | 0.44 | **0.11** |
| 9 | fold | verb | 10 | 0.18 | 0.29 | 0.18 |

Table 3: Agreement results for words with the three highest and three lowest agreement scores, for each part of speech

words where it made a difference (e.g., *family*) are in boldface.

In previous work (Passonneau et al., 2010b; Bhardwaj et al., 2010), we have speculated about some of the reasons for the observed differences in interannotator agreement across WordNet word sense inventories, using a small sample of the MASC word sense data. For the future, the full MASC word sense corpus will provide interannotator agreement data on over 100 words. This data can serve as a resource for comparing WordNet sense inventories that differ in interannotator agreement, and could thus lead to improvements in WordNet sense inventories, or in other approaches to sense representation.

### 3.3. Characteristics of the Corpus

For rounds 7 through 10, 56% of words have $\alpha$ values greater than 0.50. For these words with relatively good interannotator agreement (see section 3.2.), the average number of senses is 12.6, which is quite a bit higher than the average number of senses for al1 37 words. The MASC sentence corpus thus provides valuable data on sense distributions that does not exist elsewhere, namely the distribution of a word's many senses, as well as the distribution across multiple genres. Figure 2 illustrates the Zipfian distribution of senses for three words. These words were selected because all have nearly 1K instances that were annotated, all have $\alpha$ values that are reasonably high ($\alpha > 0.55$), and all have about the same number of senses used by annotators (7 for *late-j, appear-v* and 6 for *paper-n*; note that the 8th tick on the x-axis represents the label *other* for the three words). The sense, in descending order of frequency is on the x-axis; the y-axis shows the proportion of instances assigned each sense. The figure demonstrates that there are a reasonable number of instances for several less frequent senses: approximately 200 for sense 2 of *paper-n*, 100 for sense 2 of *late-j* and *appear-v*, senses 3-4 of all three words, and sense 5 of *appear-v*.

WordNet sense numbers are assigned by frequency based on corpus data, but given the size of WordNet and the lack

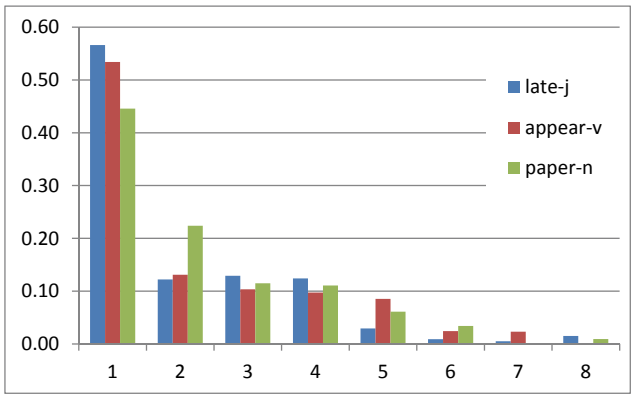


Figure 2: Sense distributions for three words

of sense-annotated corpora, sense numbers are not always reliable. The MASC data can supplement WordNet sense orderings. For example, the 7 senses of *appear-v* would be revised to swap sense numbers 2 and 5; that is, sense 5 is the second most frequent sense for *appear-v*, and sense 2 is the fifth most frequent sense (see glosses for these two senses in Figure 1).

The most common use for word sense corpora is to train automated word sense disambiguation (WSD). Besides providing training data for conventional approaches to WSD, the MASC corpus provides additional information pertaining to the agreement among multiple annotators on individual instances. Recall that 100 of the sentences for each word have been annotated by four annotators, to compute interannotator agreement. For the word *appear*, the most frequent sense is sense 1: 52 of the 100 instances have at least one annotator selecting sense 1, and in 52% of these cases, all annotators agreed on sense 1, in 37% of cases, 3 of 4 annotators agreed, 4% of the time, 2 annotators agreed, and 8% of the time, only 1 annotator selected sense 1. Only 5 instances have at least one annotator selecting sense 6, but when it does occur, 3 (40%) or 4 (40%) annotators agreed. This suggests that while sense 6 is rare, it might be possible to build an accurate classifier for this sense if the sentence

features are sufficiently discriminative.

Analysis of instances with high disagreement also has the potential to lead to improvements in WSD. Continuing with *appear-v*, we observe that there is one instance with complete disagreement among annotators (4 senses assigned), and four instances with near complete disagreement (3 senses assigned). When read in isolation, the sentence with the most disagreement has an infrequent construction with a non-animate subject, and a complement introduced by *as*: *Gum-trees or rocks (in areas of sand) appear as signs of possible water.*. The sentence appears to be about (pun intended) detecting signs of water in physical reality. However, the text is actually about the frequency of words and phrases associated with water in the writings of early explorers in arid regions, based on the hypothesis that those deprived of water will use such phrases more frequently. Knowing the full context might lead to greater agreement among annotators on such sentences, which might suggest additional features for WSD classifiers.

## 4. Growing MASC

To grow the MASC word sense sentence corpus, we are exploring the potential to use crowdsourcing to collect high quality word sense annotations. In a pilot study reported in (Passonneau et al., To Appear), we compared the quality of annotations from a half dozen trained annotators on a single round of ten words with annotations from over a dozen Amazon Mechanical Turkers (AMT). While results were inconsistent, they were sufficiently promising that we are currently collecting AMT word sense annotations for a much larger set of words. We also continue to add manual annotations, and should produce tags from multiple annotators for approximately 50 new words within the next several months.

MASC as a whole is intended to serve as a base for a sustained collaborative resource development effort by the community. In our view, a community-wide, collaborative effort to produce open, high quality annotated corpora is one of the very few possible ways to address the high costs of resource production and ensure that the entire community, including large teams as well as individual researchers, has access and means to use these resources in their work. The ultimate goal is to build on the MASC base to provide an ever-expanding, *open linguistic infrastructure* for the field, in which the community engages in a distributed effort to provide, enhance, and evaluate data, annotations, and other linguistic resources that are easily accessible and free for community use.

MASC is at present a half million words, substantially smaller than some other multiply-annotated corpora and therefore less ideal for training language models. The corpus will be soon increased in size to a million words, although there are currently no resources for further in-house validation; we will depend on the community to contribute annotations to fill in the gap. At present, researchers and developers are invited to contribute annotations of MASC and/or OANC data of any type, in any format, to be incorporated into MASC in a common format that makes all MASC annotations usable together. Others can contribute their annotations of MASC and/or OANC data by sending them in email to `anc-contrib@anc.org`. The MASC/OANC team transduces contributed annotations to GrAF so that all MASC annotations are usable together and can be input services such as ANC2Go.[8]

In addition to growing MASC and the sense-annotated sentence corpus, an effort to create "Multi-MASC", consisting of open corpora in other languages that are built to be comparable in genre distribution to MASC, is just getting underway (Ide, 2012). The goal is to produce a massive multi-lingual corpus including language-specific data with comparable genre distribution and annotations, ultimately with linkages among annotations across languages. Like the continued development of the American English MASC, creation of comparable Multi-MASC corpora will rely on community effort.

## 5. Conclusion

The MASC project has produced a multi-genre corpus with multiple layers of linguistic annotation, together with a "sentence" corpus containing WordNet-3.1 sense tags for 1000 occurrences of each of 100 words produced by multiple annotators, accompanied by in-depth inter-annotator agreement data. All data and annotations are completely open and free for community use. The intent is that the community will use these resources as a base upon which to build by contributing additional data and annotations of all kinds. Hopefully, a substantial community-based collaborative effort to develop the MASC resources will ultimately serve to avoid redundant work, enable substantive evaluation and replication of results, and empower all members of the community with access to high-quality resources in the years to come.

## Acknowledgments

This work was supported by National Science Foundation grants CRI-0708952 and CRI-1059312.

## References

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

Collin F. Baker and Hiroaki Sato. 2003. The FrameNet data and software. In *Companion Volume to the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 161–164.

Vikas Bhardwaj, Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*, pages 47–55.

Alioscha Burchardt and Marco Pennacchiotti. 2008. FATE: a FrameNet-annotated corpus for textual entailment. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

[8] We have developed and will soon make available software to transduce Penn Treebank and PropBank formats to GrAF.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Eduard Hovy, Mitchel Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of HLT-NAACL 2006*, pages 57–60.

Nancy Ide and Keith Suderman. Submitted. The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging. *Language Resources and Evaluation*.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca J. Passonneau. 2010. The manually annotated subcorpus: a community resource for and by the people. In *Proceedings of the Association for Computational Linguistics*, pages 68–73.

Nancy Ide. 2012. MultiMASC: An open linguistic infrastructure for language research. In *Proceedings of the Workshop on Building and Using Comparable Corpora*.

Rebecca J. Passonneau, Nizar Habash, and Owen Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1951–1956, Genoa, Italy.

Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *SEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 2–9, Morristown, NJ, USA. Association for Computational Linguistics.

Rebecca Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010a. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC)*, Paris. European Language Resources Association.

Rebecca J. Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010b. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 3244–3249.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. To Appear. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*.

Rebecca J. Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 831–836, Genoa, Italy, May.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. FrameNet II: Extended theory and practice. Available from http://framenet.icsi.berkeley.edu/index.php.