Crowdsourcing as a preprocessing for complex semantic annotation tasks

Héctor Martínez Alonso, Lauren Romeo
University of Copenhagen, Pompeu Fabra University
Copenhagen (Denmark), Barcelona (Spain)
alonso@hum.ku.dk, lauren.romeo@upf.edu

Abstract

This article outlines a methodology that uses crowdsourcing to reduce the workload of experts for complex semantic tasks. We split turker-annotated datasets into a high-agreement block, which is not modified, and a low-agreement block, which is re-annotated by experts. The resulting annotations have higher observed agreement. We identify different biases in the annotation for both turkers and experts.

Keywords: Crowdsourcing, annotation, semantics

1. Introduction

In this article we outline a methodology to reduce the effort of expert annotators by crowdsourcing the annotation with Amazon Mechanical Turk (AMT) in an initial step, and taking only expert annotators in consideration for those cases where turkers have low agreement.We refer to annotators from AMT as *turkers*, and to traditional annotators as *experts*. Turkers have no guarantee of being natives—although they may be subject to language competence tests—, whereas experts are normally native or highly proficient non-natives with a background in linguistics.

Along this line, we propose using turkers to identify the items that are easiest to annotate—i.e. have higher agreement—, and only provide the experts with the hard cases, thus reducing annotation workload for the experts.

The assessment of the reliability of human annotation is a relevant topic in Natural Language Processing (NLP). Researchers over the last years have dealt with the metrics for the reliability of annotation schemes (as a proxy for the reliability of the annotation themselves) (Artstein and Poesio, 2008; Krippendorff, 2011), how to assess the usefulness of crowdsourced annotations (Snow et al., 2008; Callison-Burch, 2009), but also how to define the behavior of annotators (Hovy et al., 2013; Passonneau and Carpenter, 2013), as well as how to obtain aggregate annotations from all the annotations that an item has received (Jurgens, 2013; Graham et al., 2013).

In this article we evaluate the annotations from turkers and compare them with a re-annotated subset of the turker data that has been annotated by experts. We have chosen the regular-polysemy sense-annotated corpus from Martínez Alonso et al. (2013) to evaluate this method. We chose this corpus because it portrays a *complex semantic phenomenon* where it is reasonable to expect expert annotators to provide much more reliable data than turkers.

In Section 2. we describe the chosen semantic phenomenon to annotate. In Section 3. we describe how the data has been annotated by turkers, and Section 4. describes the re-annotation by experts. Conclusions are listed in Section 6.

2. Annotation phenomenon

Very often a word that belongs to a semantic type, like LOCATION, can behave as a member of another semantic type, like ORGANIZATION, as shown by the following examples from the American National Corpus (Ide and Macleod, 2001): The ability of certain words to switch between semantic types in a predictable manner is named by different authors as *logical metonymy* (Lapata and Lascarides, 2003), *sense extension* (Copestake and Briscoe, 1995), *transfer of meaning* (Nunberg, 1995), *logical* or *complementary polysemy* (Pustejovsky, 1995), *systematic polysemy* (Blutner, 1998) or regular polysemy.

A particularity of metonymic senses is that they can coexist with the literal sense in the same use of the word. In case a), *England* refers to the English territory (LOCATION), whereas in b) it refers to England as a political entity (ORGANIZATION). The third case refers to both the English territory and the English people.

a) Manuel died in exile in 1932 in *England*.

b) *England* was being kept busy with other concerns.

c) *England* was, after all, an important wine market.

The possibility to coordinate the two alternating senses is a key linguistic test to differentiate metaphors from metonymies, but coordinated constructions are not the only scenarios where the literal and a metonymic sense are predicated together, or copredicated. *Copredication* is a phenomenon whereby the literal and metonymic appear simultaneously. For instance, Asher (2011) describes copredication in cases of conjunctions, where each argument has a different semantic type.

d) *Lunch* was delicious but took forever.

e) *Shakespeare* has been dead for centuries and people still read him.

In example d) , we have "but" coordinating the statements "lunch was delicious"—in which *lunch* refers to food—and "lunch took forever"—in which *lunch* refers to the mealtime, which is an event. In the second example, Shakespeare means "the person William Shakespeare", as well as the metonymic sense "the works of William Shakespeare", even though the second clause has a pronoun to stand for Shakespeare. Conjunctions are one of the structures that can make both senses simultaneously active.

f) The Allies invaded *Sicily* in 1945.

g) We had a delicious leisurely *lunch*.

h) The case of *Russia* is similar.

Some verbs take arguments that require both senses to be active at the same time and are known as *gating predicates*, following the claim by (Rumshisky et al., 2007) that "there also seem to exist gating predicates whose selectional specification may specify a transition between two simple types". Thus, a verb like *invade* is a geophysically delimited military action, which requires both the LOCATION and the ORGANIZATION sense. Compare for instance with "Mongolia declared war against Japan", where only the ORGANIZATION sense is active for both *Mongolia* and *Japan*.

Also, there are contexts in which different elements affect the sense of the predicated noun towards being literal and metonymic at the same type without being coordinated. In g), the adjective *delicious* selects for the FOOD sense of lunch, while *leisurely* activates the sense of lunch as an EVENT, as only things that happen can be leisurely (Cooper, 2005).

Example e) also shows that metonymic senses can be propagated through anaphora, as the literal referent of the metonymy is maintained in the comprehension of the metonymic predication. For more on the ability of metonymies to preserve referent (Nunberg, 1995; Stallard, 1993; Asher, 2011).

The last example has the word *Russia* placed in a context that does not indicate a strong preference for either sense. The copula has little pull towards either sense, just as the rest of the lexical environment (*case, similar*). Without more context, the sense of *Russia* cannot be adscribed to a strictly literal or metonymic reading.

Whenever a predication of a noun is potentially figurative and literal at the same time, we will refer to it as *underspecified*, regardless of the cause of such underspecification. In this way, we group cases like copredication, gating predicates, vague contexts and the presence of multiple selectors, such as illustrated in c) and g).

We consider the annotation of noun senses for regular polysemy including underspecified senses a *complex semantic task*, because it requires the analysis of figurative senses, and includes a category which is not immediate, namely the underspecified sense tag.

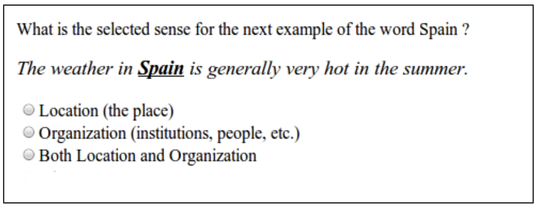
3. Crowdsourced data

The data in Martínez Alonso et al. (2013) is made up of nine datasets: five for English, two for Danish and two for Spanish. Each dataset provides five hundred sentences, each with a chosen headword belonging to a dot type. The *dot type* is the Generative Lexicon (Pustejovsky, 1995) term to account for a noun and its most common metonymic sense as a single semantic class. In this article we focus on the crowdsourced English data.

1. Animal/Meat (ANIMEAT): *"The chicken ran away"* vs. *"the chicken was delicious"*.
2. Artifact/Information (ARTINFO): *"The book fell"* vs. *"the book was boring"*.
3. Container/Content (CONTCONT): *"The box was red*" vs. *"I ate the whole box"*.
4. Location/Organization (LOCORG): *"England is far*" vs. *"England starts a tax reform"*.
5. Process/Result (PROCRES): *"The building took months to finish"* vs. *"the building is sturdy"*.

We call the first sense in the pair of metonyms that make up the dot type the *literal* sense, and the second sense the *metonymic* sense, e.g. LOCATION is the literal sense in LOCATION/ORGANIZATION.

Each block of 500 sentences belonging to a dot type was an independent annotation subtask with an isolated description. The annotator was shown an example and had to determine whether the headword in the example had the literal, metonymic or the underspecified sense. Figure 1 shows an instance of the crowdsourcing process.



What is the selected sense for the next example of the word Spain ?

*The weather in **Spain** is generally very hot in the summer.*

○ Location (the place)
○ Organization (institutions, people, etc.)
○ Both Location and Organization

Figure 1: Screen capture for a Mechanical Turk annotation instance, known as Human Intelligence Task or HIT

This annotation scheme is designed with the intention of capturing literal, metonymic and underspecified senses, and we use an inventory of three possible answers, instead of using Markert and Nissim's (Markert and Nissim, 2002; Nissim and Markert, 2005) approach with fine-grained sense distinctions, which are potentially more difficult to annotate and resolve automatically. Markert and Nissim acknowledge a *mixed* sense that they define as being literal and metonymic at the same time.

The English dataset considered in the work presented here has been annotated using Amazon Mechanical Turk (AMT) with five annotations per example by turkers certified as Categorization Masters (i.e. turkers with more than 1000 validated categorization HITS). In (Snow et al., 2008), performance of sense-annotated datasets stabilizes after four turkers. For this reason we set the number of crowdsourced annotations for this task to five in order to ensure stability among the obtained annotations. The turkers were paid 0.05$ per HIT.

Turkers were asked to respond to five synthetic examples where the answer was expected to be unambiguous such as the one shown in Figure 1 to assess whether it was feasible to annotate this phenomenon with AMT at all. The feasibility test yielded a $\overline{A_o}$ of 0.85 for LOCORG and a 0.66 for CONTCONT, which we considered sufficient to conduct the study on actual corpus data. .

Table 1 shows the average observed agreement ($\overline{A_o}$) and Krippendorff's α (Artstein and Poesio, 2008) for all the English datasets. All of these datasets are made up of 500 items. Each item is a sentence with a highlighted headword belonging to a dot type, much like the examples in Section 2.. As previously mentioned, each item has been annotated by five turkers.

Dot type	$\overline{A_o}$	α
ANIMEAT	0.86	0.69
ARTINFO	0.48	0.12
CONTCONT	0.65	0.31
LOCORG	0.72	0.46
PROCRES	0.5	0.10

Table 1: Averaged observed agreement and its standard deviation and alpha

The agreement measures vary across datasets. The dataset for the ANIMEAT alternation has the highest α, whereas the ARTINFO and the PROCRES alternations have the lowest α scores. Moreover, examining these five datasets, we note that observed agreement (A_o) is not evenly distributed across examples. Figure 2 shows the frequency of the different values of A_o for the English datasets.

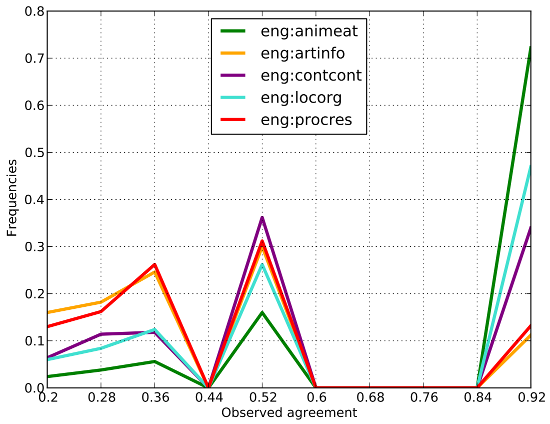


Figure 2: Distribution of agreement

In Figure 2 there are three bands of item-wise agreement values (there is a theoretical continuum of possible A_0 values, but only so many actual possible combinations for a certain number of annotators and senses). There is a peak of high-agreement items for ANIMEAT, followed by CONTCONT and LOCORG. On the low-agreement side, the peaks correspond to the datasets with lower α scores, namely ARTINFO and PROCRES.

After annotating the examples for their literal, metonymic or underspecified reading, we have determined that this scheme can provide reliable (α over 0.60) annotations for one dot type and *moderate* ($\alpha > 0.41$) for four. Not all the dot types are equally easy to annotate. The main source of variation in agreement, and thus annotation reliability, is the difficulty to identify the senses for each particular dot type. While ENG:ANIMEAT and LOCORG appear to be the easiest, ARTINFO and ENG:PROCRES obtain very low α scores.
However, we observed that agreement is unevenly distributed; while some examples have perfect agreement, others have very low A_o agreement. In view of this, we conduct a reannotation task on those low-agreement examples in an attempt to reduce the proportion of difference in agreement that is a consequence of the bias of turkers.

4. Re-annotation task

In this section we describe the re-annotation task, where experts are provided with items that received low agreement by turkers, and annotate them. The *experts* in this task are seven native or very fluent English speakers. The experts all have a background in linguistics. Each item received four annotations.

From these five datasets, we choose the CONTCONT and LOCORG datasets for this experiment because they have a great deal of high-agreement (0.67-1) items, but also enough low-agreement items to justify an additional annotation task. For instance, CONTCONT contains 171 items with agreement > 0.67, yet there are still 149 items with agreement < 0.41 while LOCORG has 238 items with agreement > 0.67 and still 135 items with agreement ¡ 0.41. This is something that ANIMEAT would not allow, as 361 of its items have an agreement > 0.67 while only 59 items have an agreement of < 0.41

Moreover, these two dot types are fully annotated by experts in the Danish and Spanish datasets, where each example has 3-4 or 6-8 expert annotations respectively.

To re-annotate the low-agreement items, we sort the two chosen datasets by A_o and split each turker-annotated dataset T into two blocks: a block T_h with 300 high-agreement items, and a block T_l with 200 low-agreement items.

The items in T_l have been re-annotated by experts. The experts were each asked to follow the turker annotation guidelines and determine whether the headword of each sentence is *literal*, *metonymic* or *underspecified*. The resulting annotations yield a block E with 200 expert annotations. Figure 3 illustrates the blocks in which we split the data.

5. Results

This section provides the agreement metrics for different combinations of the three blocks of annotations, namely Th, Tl, E. Table 2 shows the agreement value for the different combinations. Combinations with the same amount of items are grouped between lines.

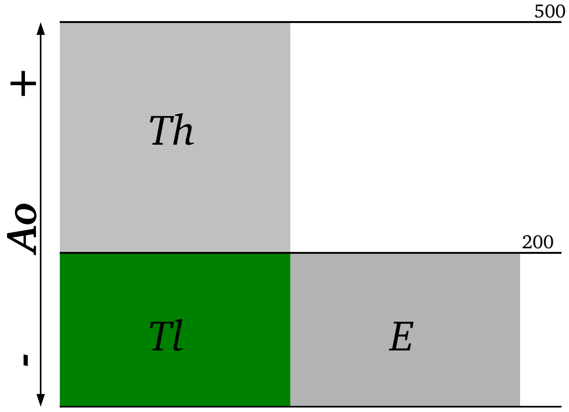


Figure 3: Splits in low- and high-agreement blocks , where T_h is the block for high-agreement items from turkers, T_l is the turker-annotated items with low agreement, and E is the expert re-annotation of T_l

	CONTCONT			LOCORG	
Section	#	$\overline{A_o}$	α	$\overline{A_o}$	α
$T_l + T_l$	500	0.65	0.30	0.72	0.46
$T_h + E$	500	**0.66**	0.30	**0.77**	0.46
$T_h + T_l + E$	500	0.65	-0.10	0.72	0.09
T_h	300	0.81	0.55	0.91	0.80
T_l	200	0.40	0.00	0.43	0.06
E	200	**0.47**	-0.07	**0.55**	0.30

Table 2: $\overline{A_o}$ and α for all blocks and combinations

$T_h + T_l$ is the original dataset with only turker annotations. $T_l + E$ is the dataset with expert annotations replacing the low-agreement items in T_l. In $T_h + T_l + E$, there are nine annotations for the low-agreement items, namely the five from T_l and the four from E.

$T_l + E$ fulfils our expectation of achieving better agreement than $T_h + T_l$, as experts agree more on the more difficult cases. However, α remains the same instead of improving, as a result of E having four annotators instead of five. $+T_l + E$. We also observe that $\overline{A_o}$ is the same as for $T_l + E$, but α drops dramatically because turkers and experts have very different biases in their annotations, and their behavior when annotating the low-agreement items is very different.Indeed, if we examine the proportions of raw annotations (before a final sense is assigned for each item) between turkers and experts in Figure 4, we observe differences in the preference towards annotating the underspecified sense, e.g. the E block in CONTCONT (CCe) has 4.45 times the proportion of underspecified sense annotations as the T_l block (CCt).

In this article we do not deal with the assignation of a final sense tag out of the pool of annotations from turkers and experts. However, turkers have a dispreference for the underspecified sense tag, and there are very few (never over 0.01%) of the examples that would receive an underspecified sense tag from simple majority.

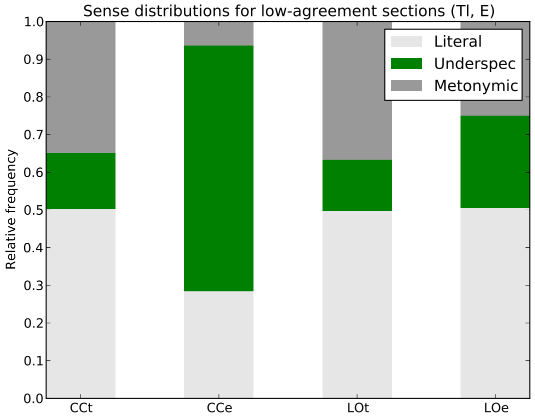


Figure 4: Raw sense distributions

5.1. On the behavior of annotators

In Section 3. we see that turkers disprefer the underspecified sense. We hypothesize that, since turkers are not always native speakers, they might lack nuances in their interpretation. But on the other hand, in NLP it is accepted that fluent non-natives can annotate data for English, as in (Biemann and Giesbrecht, 2011) or (Markert and Nissim, 2009). We have worked with this assumption ourselves when carrying out the expert annotation task with a fluent non-native as annotator.

f) Such practices are totally forbidden in *China*

Example f) shows a sentence where the expert had predicted the underspecified example because something being "forbidden in China" issued a LOCATION and ORGANIZATION reading. However, the five turkers unanimously assigned the literal sense to this example. This is an indication that turkers might have a less nuanced understanding of this phenomenon, or that they focus on very obvious features like the preposition *in*.

However, a less-nuanced understanding of the task is not the only possible explanation for the turker dispreference for the underspecified sense. Turkers are also wary of their data being rejected for invalidity and might choose options with lesser perceived risk of rejection. In fact, two of the turkers contacted us during the annotation task to inquire about their data being approved using majority rules, as they were concerned that discarding low-agreement turkers would be unfair for such a difficult task. It was explained to them that it was not the case, as our setup did not contemplate data rejection.

It is also a possibility that turkers choose, in case of doubt, the easiest sense. The easiest sense, however, is not necessarily the most literal one, and we already have seen that we do not find a general tendency to abuse first-option clicking, which would have caused a stronger bias for the literal sense. Endriss and Fernández (2013) provide a related account on the biases of turkers when annotating structured data.

We propose that turkers manifest a behavior that makes them choose the easiest option as a default in hopes of getting paid and not getting their annotations rejected. Whatever the case, turkers behave differently than the experts and show a bias against the underspecified sense that is not as strong in data annotated by volunteers.

The datasets for the dot types LOCORG and CONTCONT have been annotated by volunteers for Danish and Spanish. For these datasets we do not have an expert-annotation to compare against, like we do for English.

Still, we can contrast the general behavior of all annotators for all languages for these datasets. In Table 3 we show the raw distributions of senses chosen by the annotators before any sense assignment method was applied to the data. This is the overall proportion of times any annotator has marked an item as literal, metonymic or underspecified. We provide the LOCORG and CONTCONT datasets for all three languages. For English we provide the distribution from the turker annotation and from the expert annotation. Figure 4 reprents the information graphically.

Dot type	L	M	U
ENG:CONTCONT:TH+E	0.55	0.28	0.16
ENG:LOCORG:TH+E	0.61	0.28	0.10
ENG:CONTCONT:TH+TL	0.64	0.28	0.08
ENG:LOCORG:TH+TL	0.59	0.35	0.06
DA:CONTCONT	0.65	0.20	0.16
DA:LOCORG	0.65	0.21	0.14
SPA:CONTCONT	0.56	0.26	0.17
SPA:LOCORG	0.58	0.27	0.15

Table 3: Expert, turk and volunter sense distributions for the CONTCONT and LOCORG datasets

We can see that, for these two dot types, the literal sense is the most frequently chosen, regardless of language and type of annotator. For the underspecified sense, however, we observe two particularities: the volunteer datasets have a proportion of underspecified senses that is consistent even across Danish and Spanish, and is more similar to the English expert datasets. Experts, however, agree on the underspecified sense for 10% of the examples. This indicates that human annotators without a bias against the underspecified sense can recognize it, in spite of language differences.

Also, ENG:LOCORG:TH+E stands out as the dataset where there is the highest proportion of underspecified sense tags being assigned, twice as often as in its turker counterpart ENG:LOCORG:TH+TL.

We naively compare proportions to obtain a qualitative assessment of the behavior of the annotators. When comparing the proportion of sense tags given by the expert to the proportion of sense tags given by the turkers or volunteers we are comparing the distribution over 500 sense tags against a distribution between 1500 and 3200 sense tags. Conducting independence testing would only make sense between annotations of the same data, and for Danish and Spanish we have no alternative to the volunteer to compare against.

Nevertheless, we can suggest that the behavior of experts sets the standard for what to consider the output of an annotation task by well-meaning (i.e. non-spamming), native annotators. Turkers share the same kind of bias for the most frequent, literal sense, but are less willing to give the underspecified tag for the reasons we suggested in the previous section, while the experts can be overzealous when interpreting a certain usage of a dot type as underspecified.

6. Conclusions

By replacing the low-agreement items with expert judgments we improve $\overline{A_o}$ for a complex semantic annotation task. Turkers and experts have different biases, and replacing the T_l with E is more advisable than pooling them together.

We consider that turkers manifest a behavior that makes them choose the easiest option as a default in hopes of getting paid and not getting their annotations rejected. Turkers behave differently than experts and show a bias against the underspecified sense that is not as strong in data annotated by experts.

This difference in bias justifies the need to use expert annotations for complex semantic tasks where detecting very slight semantic nuances may be required. Since these difficult items are a subset of the total (cf. Figure 2), we consider that it is a viable method to crowdsource the annotation to identify difficult items and keeping the high-agreement items from turkers. We propose using crowdsourcing as a preprocessing method in order to divide high- and low-agreement examples, in order for experts to only annotate the latter. This would reduce annotation workload for experts—in our case—by 60%.

We propose that a two-step annotation setup, where (i) the dataset annotation is crowdsourced and (ii) the low-agreement examples are re-annotated by experts. This strategy can alleviate the time bottleneck caused by expert annotations while simultaneously preserving the expert intuition which is valuable for annotating low-agreement examples, which turkers are not necessarily qualified to annotate. We expect this approach to be transferrable to other complex semantic annotation tasks like those depending on figurative meaning, anaphora, etc.

The further work for this approach includes devising strategies for assigning aggregate senses from the pool of annotations for each item, and empirically adjusting the threshold for items to be considered low-agreement, as 200 out of 500 is arbitrarily set.

Using an aggregate label as a suggested sense tag could be a possible way of easing the annotation of the more difficult (i.e. low-agreement) datasets, namely ARTINFO and ANIMEAT.

The sense-annotated corpus is available at http://metashare.cst.dk/repository/search/?q=regular+polysemy

7. Acknowledgements

This work was funded with the support of the SUR of the DEC of the Generalitat de Catalunya and the European Social Fund, and by the European Commission's 7th Framework Program under grant agreement 238405 (CLARA).

8. References

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

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge University Press.

Biemann, C. and Giesbrecht, E. (2011). Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 21–28. Association for Computational Linguistics.

Blutner, R. (1998). Lexical pragmatics. *Journal of Semantics*, 15(2):115–162.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics.

Cooper, R. (2005). Do delicious lunches take a long time. In *Fourth International Workshop on Generative Approaches to the Lexicon. GL2007.*

Copestake, A. and Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of semantics*, 12(1):15–67.

Endriss, U. and Fernández, R. (2013). Collective annotation of linguistic resources: Basic principles and a formal model.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with mace. In *Proceedings of NAACL-HLT*, pages 1120–1130.

Ide, N. and Macleod, C. (2001). The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, pages 274–280. Citeseer.

Jurgens, D. (2013). Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562, Atlanta, Georgia, June. Association for Computational Linguistics.

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112.

Lapata, M. and Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.

Markert, K. and Nissim, M. (2002). Towards a corpus annotated for metonymies: the case of location names. In *LREC*. Citeseer.

Markert, K. and Nissim, M. (2009). Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138.

Martínez Alonso, H., Sandford Pedersen, B., and Bel, N. (2013). Annotation of regular polysemy and underspecification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–730, Sofia, Bulgaria. Association for Computational Linguistics.

Nissim, M. and Markert, K. (2005). Annotation scheme for metonymies. In *Technical document*.

Nunberg, G. (1995). Transfers of meaning. *Journal of semantics*, 12(2):109–132.

Passonneau, R. J. and Carpenter, B. (2013). The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 187–195, Sofia, Bulgaria, August. Association for Computational Linguistics.

Pustejovsky, J. (1995). *The generative lexicon: a theory of computational lexical semantics*. MIT Press.

Rumshisky, A., Grinberg, V., and Pustejovsky, J. (2007). Detecting selectional behavior of complex types in text. In *Fourth International Workshop on Generative Approaches to the Lexicon, Paris, France*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Stallard, D. (1993). Two kinds of metonymy. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 87–94. Association for Computational Linguistics.