A High-Quality, Human Annotated News Corpus on Sentence Specificity

Bridget M. O'Daniel
University of Pennsylvania
Philadelphia, PA 19104
odanielb@berea.edu

Wenli Zhao
University of Pennsylvania
Philadelphia, PA 19104
wenliz@seas.upenn.edu

Yi Di Wu
University of Pennsylvania
Philadelphia, PA 19104
wuyd@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
Philadelphia, PA 19104
nenkova@seas.upenn.edu

***Abstract*—The specificity of a sentence has yet to be fully examined as an innate semantic property necessary for human imitating automated writing. This corpus is presented as a means to explore the complexities of the specific-general scale and created through the use of a highly trained group of annotators to produce thoughtful, meaningful results. The resulting corpus is presented with an analysis of the agreement between annotators and a brief overview of potential sources of exploration using the corpus. It is shown that agreement between annotators on the subject of specificity can be reached and describes new methods of analyzing specificity of language that have yet to be explored.**

***Index Terms*—General-specific, News corpus**

I. Introduction

SENTENCES vary in how specific or general they are about the subject matter they discuss. Often specific sentences will discuss particular entities or incidents and will include a large amount of details. General sentences, on the other hand, are more prone to be descriptive, summarizing, or introductory. An example of each sentence can be found below, where the first sentence is highly specific while the second is highly general.

> *1) Together, Robert, 92, and Peter, 90, will produce a small barrel of wine -- 60 magnums to be exact -- to be sold next summer at the Napa Valley Auction, the California wine world's premier social event.*
> *2) Singapore is, by tradition, a hard-power country, though its stature is not military but economic.*

This definition of specificity based on the amount of detail present in a sentence draws from previous works [1]. The relationship between the amount of detail in a sentence and its specificity is strong, and thus for this corpus, there was a focus on researching the amount of detail that is found in a sentence and what details may be deliberately withheld from the audience in order to create a more general atmosphere in the sentence. This may be a deliberate choice in order to give contextual or otherwise necessary information to a reader that cannot be given through writing composed solely of specifics. For example, in an essay a writer is expected to begin with a general topic sentence before delving into the evidence and rationale he or she has for this claim. Similarly, an introduction will generally discuss the contents of the rest of the essay without giving too much information that will be discussed later. General sentences are enticing and understandable and when paired with more detailed, specific sentences, lends to more human, understandable pieces of writing.

Indeed, automated summaries are more likely to contain mostly specific sentences, while human written summaries tend to be have a balance between the two and may be indicative of higher quality writing [2]. In this way, a better understanding of specificity will have a variety of applications. The identification of specific and general sentences will lend to increasing the quality of automated summaries and of automatic essay grading, for example, as well as assisting in locating particularly information dense sentences in text for extraction.

Few corpora have been gathered to address specificity, and this corpus was created to add a new angle. The most direct predecessor is that of the news specificity corpus presented by Louis and Nenkova [3]. This corpus demonstrated the ease with which annotators can distinguish between general and specific sentences, although one third of all sentences disagreed greatly on the category of the sentence. Due to the fact that these annotators were working through intuition, it indicated that specificity may be a part of a spectrum as opposed to the trinary scale used in the study. The disconnect between the randomly selected sentences and their missing context was also caused a sentence to appear more general than it may have been. For example, a sentence using a pronoun rather than a specific person's name appears more general and was often categorized as such.

This new corpus is aimed to tackle these problems and attempt to improve upon the agreement in the previous study. While the previous study utilized five random annotators for each sentence, this brought about inconsistences and the inability to fully test the agreement between annotators. Instead, we aimed to create a corpus of selected annotators trained on our definition of specificity as opposed to their intuition. In addition, we aimed to provide the context for the sentences to reduce misinterpretation along with providing clear instruction on working with the context through asking questions that would be present in the corpus. Additionally,

with the added context of full articles being included, specificity beyond the sentence level could be analyzed.

The annotations are complex and many different aspects of specificity and generality can be explored with the amount of data provided. Due to the training the annotators received and the dedication these tasks required, the annotations are of a high quality not yet found in previous studies of specificity. This corpus was created with the intention of unearthing the various facets of specificity in a way that previous studies could not and hopes to shed light onto the manner in which it is not only a useful attribute to have in writing, but also to provide a means of classifying this semantic property accurately and precisely for future use in automated writing.

II. Methods

The corpus was collected from annotations by three undergraduate research students. All annotators are native English speakers and are not professional linguists. Instead, each annotator was asked to provide her opinion on each sentence after an initial two week training period to improve consensus on the goals of the annotation.

The participants were asked to complete tasks that consisted of sets of eight to ten sentences. These sentences are sequential selections of political and business articles in The New York Times in January 2005. The sentences could be selected from the start, middle, or end of an article. In the latter two cases, the previous sections of the article were provided to the annotators at the start of the task for comprehension, but participants were not asked to annotate them. Each of the three annotators completed each task.

For each sentence of the set, the annotator was asked to rate its specificity on a scale from 0 to 6, with 0 being the most specific and 6 being the most general. She was also asked to consider the sentence separately from the previous context of the article for this rating and from how many questions were asked in the second aspect of the annotation as described below.

The annotator was also asked to mark which phrases, if any, added ambiguity into the sentence and to ask the question that it introduced. These marked phrases of underspecified words were to be the minimum selection of terms that brought about the question, while the questions were to be asked only about information that the annotator felt to be vital to understanding the sentence. The idea of a minimum span of words was clarified with an example. With the sentence “He sued the executive of the company.” and the question “Why did he sue?”, “sue” would be the preferred word span as opposed to “He sued” or “He sued the executive”, because the question is most closely tied to the act of suing.

Additionally, the annotator selected where the answer to that question could be found: in the immediate context (defined and identified as the preceding two sentences where applicable), in some previous context (three or more sentences earlier), not found in any previous context, or was vaguely mentioned in some previous context which indicates that some aspect of the question was answered in the previous context or was briefly touched on, but not fully defined and explained. Although in most cases annotators would ask at least one question, there were cases where she elected not to. In these instances, the annotator was prompted to select whether she asked no questions because the sentence was very specific or because it was very general.

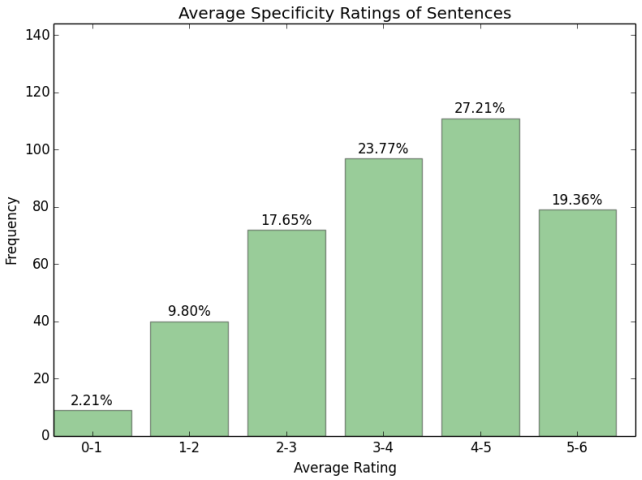

Fig. 1. The average of the three annotators’ specificity ratings for each sentence was averaged and counted as a single occurrence of that rating.

III. Corpus Summary

A. Size

The corpus is composed of 42 tasks, where a task is defined as a sequential selection of 8 to 10 sentences. The sentences in these tasks are selections from 12 articles of varying length and topic in the New York Times. A total of 408 sentences were annotated with 10,184 total words. There are three annotations per sentence, each by one of the trained annotators. They asked a total of 2,157 questions.

B. Overview

For each sentence in the corpus, the average of the annotators’ ratings was collected and organized into Fig. 1. The majority of sentences were found to be more general than specific, with 70.34% having an average specificity rating above 3. Very few sentences were on average rated to be “most specific”, only 2.21% averaging between 0 and 1.

Similarly, the average rating for each task was on the general side, with half of the tasks rating between 3 and 4, as shown in Fig. 2. The average specificity rating per task was calculated by averaging the raw ratings from every sentence annotated in that task. Notably there are no tasks that were, on average, very polar in specificity, as there are no tasks with average ratings less than 2 nor greater than 5.

IV. Agreement

In previous studies of sentence specificity, the annotators varied for each annotation due to the crowdsourcing methods. One major point of contention in the previous study [3] was the high rate of sentences with major disagreements on specificity and the high rate of “mixed” specificity sentences. Our goal for this corpus was to create high-quality annotations through the training of a regular set of annotators. This way the nuances of specificity may come to light more easily and reduce the number of inconsistent ratings through careful, thoughtful

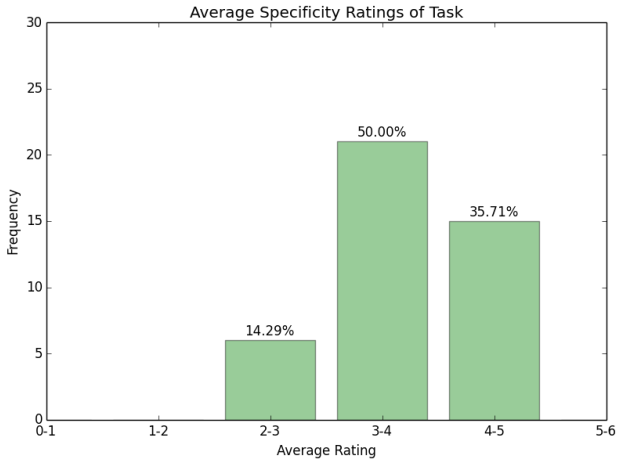

Fig. 2. The figure describes the average specificity rating for tasks in the corpus, where a task is a sequential selection of 8-10 sentences

annotation. The success of this trained regularity is measured by the agreement between annotators on various tasks, which there could be no equivalent for in previous studies due to the random nature of crowdsourcing. The three annotators will henceforth be referred to as A1, A2, and A3.

A. *Specificity Rating*

In order to give a broad overview of the agreement between annotators, the pairwise correlation between the average task specificity ratings was calculated. A1 and A2 had a correlation coefficient of 0.793, A2 and A3 had 0.728, and A1 and A3 had 0.646. While the correlation between A1 and A3 is much lower than the others, these correlations are relatively high.

In order to test each annotator's agreement with the collective, her specificity rating for a sentence was compared to the average of the ratings given by the others. A1 had the lowest correlation, 0.689, while A2 and A3 were slightly higher with 0.799 and 0.721, respectively. The increase in these numbers as compared to the pairwise correlation above suggests that as a group the annotators agreed more or less, even though they may differ slightly more individually. Though not perfect, this represents the agreement on a sentence level, in addition to the task level.

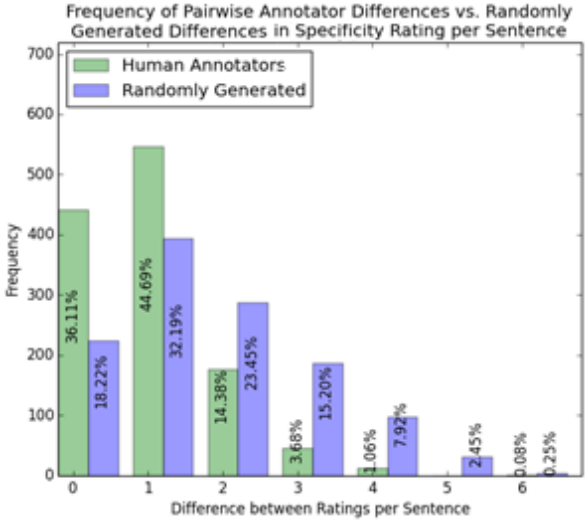

Fig. 3. The number of instances for each difference in specificity rating made by the human annotators are displayed beside the randomly generated differences based on the human distribution.

TABLE I
95% CONFIDENCE INTERVAL OF RATING DIFFERENCE OCCURRENCES IN 1000 RANDOMLY GENERATED SENTENCE SETS

Difference	Mean Occurrences	Error
0	223.66	0.886
1	394.610	1.071
2	287.492	0.952
3	186.744	0.787
4	97.179	0.638
5	30.752	0.369
6	3.563	0.120

Table I. This tables describes the results of creating randomly generated specificity ratings using the distribution of differences from the three human annotators.

Besides correlation, the average difference per task between the specificity ratings was looked at. The range of these averages is from 0.40 to 1.40. This means that for any given task, the annotators were typically within 2 points on the 0-6 point rating scale, and more often than not within a 1 point, as the pairwise difference was 0.893. The difference between each point on the scale is arbitrary, and thus such a difference is expected, and in fact shows remarkable agreement between the annotators. For example a sentence where two annotators gave a "4" while the other gave a "5" is "more general than specific", as described by the rating, even though one annotator may feel that it is slightly more specific than the other. Another possibility is that the first annotator consistently selects sentences to be more specific.

The pairwise average differences per sentence were also calculated. The distribution of differences collected from these results was used to simulate human annotation through informed random generation of specificity ratings. For each sentence, three random ratings were generated based on the distribution of the annotators' average ratings. They were compared in the same pairwise manner as the three human ratings, generated one thousand times and averaged. For each category representing the frequency of ratings with a particular difference, the 95% confidence interval of the frequency was stable, ranging from ±1.032 to ±0.125, as listed in Table I. The pairwise difference frequencies for both human annotators and randomly generated annotators are shown in Fig. 3. On the specificity seven point scale, 80.80% of human pairwise ratings were within one point of each other, with 36.11% giving exactly the same rating. Only 1.14% of human ratings collected had a difference of four points or greater. Although drawn from the same distribution, the randomly generated ratings were found to have less pairwise agreement compared to that of the human annotators. Only 50.49% of randomly generated ratings were within one point of each other with 18.30% having the same rating. The percentage of ratings with differences greater than four points was increased to 10.62%.

B. *Underspecified Terms*

Annotators asked a total of 2,157 questions about phrases in the sentences that they found to underspecified and thus introduce ambiguity. These phrases could consist of any number of words, and annotators were allowed to ask any number of questions about the same phrase. Because of this, it cannot be assumed that the same question was asked in multiple annotations of an annotated phrase, nor that different

TABLE II
SHARED IDENTIFICATION OF UNDERSPECIFIED TERMS

Type of Phrasal Overlap	Percentage of Total
Equal	60.4%
Proper	39.1%
Intersecting	0.1%

Table II. The types of phrasal overlaps and their percentage of the total overlapping underspecified terms.

annotators created them. However, this method allows especially underspecified phrases to be identified with multiple relevant questions. Of the questions asked, 1,277 (59%) of the underspecified phrases marked did not overlap with other such phrases. The remaining 880 phrases were annotated two or more times. Some annotators asked more questions than others, which plays a part in this disparity.

The phrases that were annotated multiple times represent the shared identification of underspecified phrases. These could overlap in various ways. For example, one annotator might select the phrase "a rigorous test" as underspecified, while another may simply select "test". This type of overlap, where one phrase is composed of words that are a proper subset of the other, will be referred to as a proper overlap. An overlap wherein both phrases refer to exactly the same phrase will be referred to as an equal overlap. The final type, the intersecting overlap, refers to instances where the sets of words have a point of intersection but each contain words the other does not. For instance, the phrase "a rigorous test" and the phrase "test was administered" are intersecting.

Of the 880 phrases that overlapped with another, there were 1171 possible combinations, due to the possibility of three or more questions per phrase. The percentages for each category from of the total number of overlapping phrases are displayed in Table II. Only 2 of the 1171 possible instances were found to be intersecting, representing 0.1% of the overlaps. The number of proper overlaps make up a larger portion of the data, but more than half the time annotators identified the exact same underspecified phrase. The proper overlaps themselves are often composed of one annotator identifying a full clause while the other picks the head of that clause. In this way, when annotators identified aspects of a sentence underspecified, they were very likely to select the same phrase, representing an agreement in where ambiguity is found.

C. Sentences of Disagreement

A few of the sentences that annotators gave wildly different specificity ratings on were collected and the annotators were asked their choice. Occasionally an annotator would state that she would have revised her rating had she been given it again, but in the majority of cases they stuck to their choice or within one point of difference. The reasons the annotators gave for their varying choices in problematic sentences bring to light the deeper difficulties of classifying sentence specificity.

The major category of sentences with disparate specificity ratings is tied to a disparity in the importance of a specific entity to particular annotators. For example with the sentence "Christians make up about 3 percent of Iraq's total population." Annotator A1 rated it a 2 while A2 gave a 0 and A3 recorded a 6. Annotator A2 believes this sentence to be very specific as it is about the population of a specific entity, Iraq. On the other hand, A2 believes the sentence to be very general as it is a general statement about the population of a country, rather than referring to any specific incident. For A3, the inclusion of a particular country did not make the sentence specific enough to overcome the generality of a statistical fact, while it was just the opposite for A2. A1 also leaned in this direction, but commented that she did believe it to be quite mixed. The same circumstance came about with the sentence "Sikorsky is perhaps best known as maker of the Black Hawk helicopter, a military war horse that is in heavy use in Iraq." A2 reported the sentence as a 2 while A3 gave a 5, and both listed the same responses as the first sentence as to why they rated in this way: the sentence gives a specific entity but is a general description or fact about that entity.

The subjectivity of sentences to an annotator's personal specificity scale indicates that certain sentences may be impossible to correctly categorize or else that the definition given to the annotators on specificity was not clear enough. While there are many of these problematic sentences, the annotators agreed quite well in the majority of the corpus's sentences, as discussed in previous sections.

V. RESULTS

With the creation of this corpus focused on sentence specificity comes many options for analysis. As the tasks given to the annotators were quite complex, there is much that can be gained on a variety of topics that will be discussed as a preliminary overview of the potential uses and implications of this data.

A. Interrogative Analysis

The content of the questions asked by annotators was evaluated through the interrogative words found to be present. The interrogatives used for analysis were seven interrogative pronouns plus the "no interrogatives used" case, as listed in Table III. Case variants (such as "whom" as a variant of "who") and those with -ever endings ("whenever", "wherever") were considered as an instance of the matching interrogative pronoun. Questions were assumed to contain only one instance of an interrogative word if any.

As the annotators' questions are meant to ask about the details of an underspecified phrase in a sentence, the questions are unlikely to be polar in nature; rather, the questions would generally be content questions, commonly referred to as "wh-questions", and contain interrogative phrases. To this end, of

TABLE III
INTERROGATIVES IN ANNOTATOR QUESTIONS

Word	Percent of Total Questions	Average Rating of Sentences		
		With Interrogative	Without Interrogative	Difference
What	47.57%	3.682	2.969	0.713
Who	15.30%	3.948	3.322	0.626
How	12.98%	3.257	3.856	-0.599
Why	12.15%	3.417	3.721	-0.304
Which	8.30%	3.405	3.664	-0.259
Where	2.41%	3.365	3.606	-0.241
When	0.74%	2.688	3.624	-0.936
N/A	0.56%	4.444	3.561	0.883

Table III. A look at the presence of interrogatives in questions asked by annotators and their effect on the rating of the sentence they pertain to.

the total 2157 questions asked, only 0.56% were found to lack an interrogative word. Of the remaining questions, 47.57% contained the interrogative "what". For comparison, the next most used interrogative was "who", at 15.30%.

Annotators were urged to only annotate underspecified elements of a sentence if the question was necessary for understanding the sentence itself, either without its context or relative to its purpose. In this way, the questions themselves not only indicate which phrases are underspecified, but the importance of that phrase to the sentence's specificity. It was found that while interrogatives such as "when" and "where" were much less commonly asked than others, their presence was not necessarily less or more important as indicators of the specificity of the sentence they question. The importance was measured by averaging the ratings for each sentence that had at least one question containing the interrogative word and then for those not containing the interrogative word. These values, as well as the difference between them, are listed in Table III.

"What" and "N/A" were the only options that had an average rating difference above 0.700, indicating that the presence of these interrogatives have a larger effect on the annotator's rating and thus may be more important to answer for a sentence to be labelled specific. Questions that contain "what" are both common and relatively influential, which may indicate a connection between the importance of phrases asked about using this interrogative and the specificity of a sentence. The questions using "what" phrases asked about noun phrases 64% of the time, placing importance upon the participants, places, and objects. Based on this assessment, specified nouns are significantly more important to include than other parts of speech for increased sentence specificity. While sentences without an interrogative were found to have a larger difference in sentence specificity, the rarity of such questions and a lack of connection to parts of a sentence make a similar analysis difficult. In the case of "when" questions, it was found that sentences where such questions were on average nearly one point more specific than if a "when" question was not asked. "When" questions being very rare, appearing in 0.74% of questions, indicates that including temporal details in a sentence may not be as important in an article, but if all other details are included as in a specific sentence, an annotator may ask for clarification. Other interrogatives such as "why", "which", and "where" have no substantial difference in average sentence specificity.

Interestingly none of the differences in specificity rating are above one, which may indicate that the importance of the questions asked themselves relative to the specificity rating of the sentence is negligible. This is in addition to the idea that the number of questions asked for a sentence may not be indicative of a change in specificity. Although the differences in the number of questions each annotator asks on average varies greatly, the total number of questions remains relatively consistent and can thus be compared to the specificity ratings of the sentence. The correlation between the number of questions for a sentence and its average rating is a mere 0.108. As each question is tied to a particular underspecified phrase in the sentence, this lack of correlation implies that the inclusion of these terms in a sentence may not be a main factor in the specificity of a sentence, although the fact that the annotators do have the context of these sentences may be closely tied to this finding.

B. Term Frequency

In order to identify whether the terms identified by the annotators as adding ambiguity into a sentence through underspecification were sufficiently different in specificity from the terms that the annotators did not identify, the term frequency–inverse document frequency (tf-idf) was calculated for every word in the sentences of the corpus, both for those identified as underspecified and those that were not. The inverse document frequencies were calculated by using a New York Times corpus of all articles from 1987 through 2006, about two million articles, where each article was counted as a single document. In order to give a value to words that were not accounted for in the New York Times corpus, the variation of the tf-idf formula using add one smoothing was used.

Using the term frequency of each term where a document was each set of sentences, the 95% confidence intervals for the average tf-idf of terms that were and were not identified as ambiguous were calculated. For those that were identified by annotators, the average was 3.216 with an error of ±0.128, while unidentified terms had an average of 4.422 and an error of ±0.106. Because the tf-idf of terms that were asked about was found to be lower, those that were identified were more common, with the most asked about term being "the".

A tf-idf value is meant to represent the importance of a term based on commonness, where common terms are given less weight and unusual terms more. However, in a study on specificity, the ubiquity of a term may indicate its generality rather than its importance, as such a term would be used in many situations rather than limiting itself to very particular, specific ones. Therefore the terms that the annotators found to be underspecified were not only more commonly used terms, but also more general than those terms that were not asked out.

C. Parts of Speech

The parts of speech of the underspecified terms may also be important for identifying the specificity of a sentence. Using Stanford CoreNLP's part of speech tagger, the terms in all sentences were categorized into their parts of speech and then separated by whether an annotator had identified them as underspecified or not. In Table IV, each part of speech category is paired with the percent of that category that was marked as underspecified. The three categories with the highest percentages are nouns, pronouns, and adjectives with 35%, 33%, and 31% respectively.

This means that of all nouns that were not proper nouns or pronouns, 35% of them were asked about by an annotator. Pronouns were also asked about frequently, although only 13% of all proper nouns were identified as unspecified. Because proper nouns are inherently specific in that they refer to a very particular entity, this low number is expected. Many of the sentences involving questions about proper nouns may be due to the necessity of a previous context to properly understand what that entity is.

The high rate of selection for pronouns is likely due to the fact that the antecedent to a pronoun is often included only in preceding sentences, while the annotators were asked to

TABLE IV
PERCENTAGE OF PARTS OF SPEECH IDENTIFIED AS UNDERSPECIFIED

Part of Speech	Percentage of Part of Speech
Noun	35%
Pronoun	33%
Adjective	31%
Adverb	25%
Determiner	25%
Verb	18%
Particle	17%
Proper Noun	13%
Predeterminer	11%
Possessive Ending	10%
Symbol	9%
Preposition	6%
Existential There	6%
Conjunction	5%

Table IV. The percentages listed in this table describe the percent of the total number of that particular part of speech that annotators marked as underspecified in a sentence.

consider each sentence separately from its context. The high number of adjectives found to be ambiguous is also interesting, as one might expect adjectives to add specificity to a noun—a "cat" is more general than a "black cat" for example. The addition of comparative or superlative adjectival forms may lend to further ambiguity, as often the comparison or quantifying information is not present in the context. For example, the phrase "one of the worst chapters in the war" gives a very clear feeling that the war is going badly, but "worst" is ambiguous if one does not know the previous horrors of the war, nor how the author meant to define it as worse than another chapter. In the same vein, it is difficult to say if someone is tall without having a general reference to the "average" height being referred to. These sorts of ambiguities may be why adjectives, where included, are often identified as lacking in specificity. Conjunctions have the lowest frequency of identification, perhaps due to the idea that conjunctions join two related ideas, often adding information—and thus specificity—to a sentence.

The amount of questions about these different parts of speech may indicate that these aspects of a sentence are considered more important to the annotators, and by proxy, to readers. Nouns, for example, are the subjects of a sentence and without specific nouns a sentence can easily become very ambiguous. It also implies that certain aspects of a sentence simply cannot introduce ambiguity the way that others can, such as conjunctions.

VI. CONCLUSION

Even with the complex nature of the tasks given to the annotators, the training was successful in improving the agreement between annotators, even on the sentences that were of mixed generality and specificity. This corpus serves as a basis for further research into the properties of specificity and the fine-tuning of features that would be useful for accurate classification with the eventual goal of applying such findings in a variety of areas including information extraction, autosummarization, and analysis of human writing. This paper introduced a few possible ways in which this corpus can be analyzed to provide a new look at specificity in news articles and hopes to establish interest in further research delving into this complex semantic property.

REFERENCES

[1] A. Louis and A. Nenkova. (2011, Nov.). Automatic identification of general and specific sentences by leveraging discourse annotations. Presented at International Joint Conference on Natural Language Processing. Available: http://aclweb.org/anthology/I/I11/I11-1068.pdf

[2] A. Louis and A. Nenkova. (2011, June). Text specificity and impact on quality of news summaries. Annual Meeting of the Association for Computational Linguistics. Available: http:www.aclweb.org/anthology/W/W11/W11-1605.pdf

[3] A. Louis and A. Nenkova. (2012). A corpus of general and specific sentences from news. The International Conference on Language Resources and Evaluation. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/657_Paper.pdf