Analysis of a Hand-Tagging Task

Christiane Fellbaum*†, Joachim Grabowski‡, Shari Landes*
*Cognitive Science Laboratory
Princeton University
†Rider University
‡Department of Psychology
University of Mannheim, Germany

Abstract

We analyze the results of a semantic annotation task performed by novice taggers as part of the WordNet SemCor project (Landes et al., in press). Each polysemous content word in a text was matched to a sense from WordNet. Comparing the performance of the novice taggers with that of experienced lexicographers, we find that the degree of polysemy, part of speech, and the position within the WordNet entry of the target words played a role in the taggers' choices. The taggers agreed on a sense choice more often than they agreed with two lexicographers, suggesting an effect of experience on sense distinction. Evidence indicates that taggers selecting senses from a list ordered by frequency of occurrence, where salient, core senses are found at the beginning of the entry, use a different strategy than taggers working with a randomly ordered list of senses.

1 Introduction

Our present understanding of how the meanings of polysemous words are represented in speakers' minds and accessed during language use is poor. One model of the mental lexicon, implicit in much of computational linguistics, likens it to a dictionary, with a discrete entry for each word form and each sense of a polysemous word form. Language production and comprehension then would simply require "looking up" the appropriate entry and selecting the intended meaning. If this model of the mental lexicon, with its discrete and non-overlapping sense representations, were correct, both the creation and the use of dictionaries would be straightforward.

Lexicographers collect large numbers of occurrences of words from a corpus. Interpreting the different meanings of polysemous words from the corpus presents no difficulty, since lexicographers simply do what they do as competent speakers of the language. The step that is particular to lexicography is transforming the corpus occurrences of a given word form into a number of discrete senses in the format of dictionary entries. Cross-dictionary comparisons show that carving up the different meanings of a polysemous word into discrete dictionary senses is difficult. The number of senses for a polysemous word often differs, reflecting "lumping" versus "splitting" strategies; some senses are absent from one but not another dictionary. Yet postulating different mental lexicons seems unwarranted, given our rapid and successful communication. Rather, the mapping process from occurrence to dictionary entry may give rise to difficulties and discrepancies across dictionaries because speakers' meaning representations may not resemble those of dictionaries with their flat and discrete senses, thus making lexicography an artificial and therefore challenging task.

Semantic tagging is the inverse of lexicography, in that taggers identify and interpret dictionary entries with respect to words occurring in a text. Taggers, like lexicographers, first interpret the target word in the text, and then match the meaning they have identified for a given occurrence of a polysemous word with one of several dictionary senses. Our goal was to examine the difficulties associated with semantic tagging. Because taggers are faced with the same task as lexicographers–although the the former select, rather than create, dictionary senses to match word occurrences in text–we expected to see discrepancies among the results of the semantic annotation task across taggers. Moreover, we guessed that those polysemous words that receive very different treatments across dictionaries would also be tagged differently by the annotators.

2 Sources of difficulties in a semantic annotation task

We predicted that three properties of the words that were to be matched with specific WordNet senses would result in differences among the individual taggers' annotations and between those of the taggers and the more experienced lexicographers. These variables are: the degree of polysemy, the part of speech, and the position within the dictionary entry of the words.

3 Polysemy

Arguably, the degree of polysemy of a word is related to the degree of difficulty of the tagging process. The fact that dictionaries differ frequently with respect to the number of senses for polysemous words points to the difficulty of representing different meanings of a word as discrete and non-overlapping sense distinctions. In some cases (homonymy), the division between different senses seems fairly clear and agreed upon among different lexicographers, while for others, it is not at all obvious how many senses should be distinguished.

4 Number of senses in WordNet

The dictionary that the taggers had available for tagging task is WordNet (Miller, 1990; Miller and Fellbaum, 1991). WordNet makes fairly fine-grained distinctions, roughly comparable to a collegiate dictionary. We reasoned that the greater the sense number in WordNet was, the harder the taggers' task of evaluating the different sense distinctions in terms of the target word became. We predicted that a greater degree of polysemy would lead to greater discrepancies between the taggers' matches and those of the experimenters, as well as among the taggers themselves.[1]

5 Part of speech

The semantic make-up of some words makes them more difficult to interpret, and hence harder to match to dictionary senses, than others. Some concepts are less well-defined or definable, and more abstract than others (Schwanenfluegel, 1991). Words referring to concrete and imagible entities such as objects and persons may generally be easier to interpret. If such words are polysemous, the different meanings should be relatively easy to distinguish on the grounds that each meaning has a fairly clear representation. By this reasoning, we expected nouns to present fewer difficulties to taggers. (Of course, many nouns have abstract referents, but as a class, we predicted nouns to be easier to annotate than verbs or modifiers. The nouns in the text we chose for our analysis had mostly concrete, imagible referents.)

Modifiers like adjectives and adverbs often derive much of their meanings in particular contexts from the words they modify ((Katz, 1964; Pustejovsky, 1995)). During sequential tagging, each content word in a running text is tagged, so the meanings of highly polysemous adjectives often become clear as the tagger looks to the head noun. However, adjectives in WordNet are highly polysemous and show a good deal of overlap, so that the context does not always uniquely pick out one sense. The kinds of polysemy and overlap found among the adjectives are carried over to the many derived adverbs in WordNet.

Whereas the meanings of nouns tend to be stable in the presence of different verbs, verbs can show subtle meaning variations depending on the kinds of noun arguments with which they co-occur. Moreover, the boundary between literal and metaphoric language seems particularly elusive in the case of verbs. (Gentner and France, 1988) demonstrated the "high mutability" of verbs, showing people's willingness to assign very flexible meanings to verbs while noun meanings were held constant. They argue that verb meanings are more easily altered because they are less cohesive than those of nouns. We expected the semantic flexibility of verbs to create additional difficulties for tagging. Discrete dictionary senses could be particularly ill-suited to usages where core senses have been extended beyond what the dictionary definitions cover, and where taggers must abstract from a creative usage to a more general, inclusive sense. In other cases, a usage can be assigned to several senses that have been accorded polyseme status on the basis of previously encountered usages, but may overlap with respect to other usages. We therefore expected less overall agreement for verbs tags than for nouns.

Polysemy and syntactic class membership interact: Verbs and adjectives have on average more senses than nouns in both conventional dictionaries and in WordNet. Both the number of senses and the syntactic class membership of verbs and modifiers may conspire to make these words more difficult to tag.

[1] "Polysemy" in WordNet subsumes homonymy as well as polysemy; however, the latter is far more common: in most cases, the different senses of a word are semantically related. No clearly discernible homonyms occurred in the data we analyzed for this report.

6 Sense ordering in WordNet

The order in which WordNet lists the different senses of a word corresponds to the frequency with which that sense has been tagged to words in the Brown Corpus (Landes et al., in press). Statistically, one would therefore expect the first sense to be the one that is chosen as the most appropriate one in most cases. (Gale et al., 1992) estimate that automatic sense disambiguation would be at least 75% correct if a system ignored context and assigned the most frequently occurring sense. (Miller et al., 1994) found that automatic assignment of polysemous words in the Brown Corpus to senses in WordNet was correct 58% of the time with a guessing heuristic that assumed the most frequently occurring sense to be the correct one.

The taggers whose work is analyzed here were not aware of the frequency ordering of the senses. However, other reasons led us to predict a preference for the first sense. The most frequently tagged sense also usually represents the most "central" or "core" meaning of the word in question. When it covers the largest semantic "territory," the first sense may seem like the safest choice.

Taggers may often be reluctant to examine a large number of senses when one appears quite appropriate. While reading each new WordNet entry for a given word, taggers must modify the corresponding entry in their mental lexicons. When encountering a sense that appears to match the usage, taggers do not know whether another sense, which they have not yet read, will present a still more subtle meaning difference. Since the first sense usually represents the most inclusive meaning of the word, taggers daunted by the task of examining a large number of closely related senses or unsure about certain sense distinctions may simply chose the first sense rather than continue searching for further subdifferentiations. We therefore predicted a tendency on the part of the taggers to select the first sense even when it was not the one chosen by us.

7 The experiment

We analyzed the data from the paid training session that all taggers underwent before they were assigned to work on the semantic concordance (cite landesinpress). The taggers were 17 undergraduate and graduate students (6 male, 11 female). In all cases, the taggers' sense selections were compared to those made by two of the authors, who have years of experience in lexicography. While these "expert" sense selections constituted the standard for evaluating the taggers' performance, they should not be regarded as the "right" choice, implying that all other choices are "wrong." Rather, the matches between taggers' and experts' choices reflect the extent to which the ability to match mental representations of meanings with dictionary entries overlap between untrained annotators and lexicographers practiced in drawing subtle sense distinctions and familiar with the limitations of dictionary representations.

In addition to evaluating the taggers' annotations against those of the "experts," we examined the degree of inter-tagger agreement, which would shed some light on the representation of meanings in the lexicons of novice taggers unpracticed at drawing a large number of fine-grained sense distinctions, and their ability to deal with potentially overlapping and redundant entries in WordNet. A high inter-tagger agreement rate would be indicative of the stability of naive inter-subject meaning discrimination. We expected less agreement for words that we predicted to be more difficult. Significant disagreement for highly polysemous words would be compatible with (Jorgenson, 1990), whose subjects discriminate only about three senses of highly polysemous nouns. Moreover, we expected less inter-tagger agreement for verbs and modifiers than for nouns.

The material was a 660-word section taken from a fiction passage in the Brown Corpus. We eliminated the 336 function words and proper nouns, and the 70 monosemous content words. Of the remaining 254 polysemous words, 88 were nouns, 100 were verbs, 39 were adjectives, and 27 were adverbs, a distribution similar to that found in standard prose texts. The task of the taggers was to select appropriate senses from WordNet for these 254 words.[2]

The number of alternative WordNet senses per word ranged from two to forty-one (the mean across all POS was 6.62). The mean number of WordNet senses for the verbs in the text was 8.63; for adjectives 7.95; for nouns 4.74; for adverbs 3.37.

Taggers received a specially created booklet with the typed text and a box in which they marked their sense choices.[3]

Taggers further received a dictionary booklet containing the senses for the words to be tagged as they are represented in WordNet. Word senses were provided as synonym sets along with defining glosses. For nouns and verbs, the corresponding superordinate synonym sets were presented; adjectives were

[2]We had made a few minor alterations to the text; for example, we omitted short phrases containing word senses that had previously occurred in the text.

[3]In addition, the taggers participants indicated the degree of confidence with which they made their choice; these ratings are reported in (Fellbaum et al., in press).

given with their antonyms. Two versions of the dictionary booklet were prepared, one for each training condition.

In the first condition ("frequency" condition), 8 taggers were given a dictionary booklet listing the WordNet senses in the order of frequency with which they appear in the already tagged Brown Corpus. If, in the frequency condition, there was a significant tendency to chose the first sense, which was usually also the most inclusive, general one, it would indicate that the taggers adopted a "safe" strategy in picking the core sense rather than to continue searching for more subtle distinctions. While the taggers were not told anything about the sense ordering in the dictionary booklet, we expected those taggers working in the frequency condition to realize fairly quickly in the course of their annotations that the sense listed at the top was often most inclusive or salient one.

In the second condition ("random order condition"), the remaining 9 taggers were given a dictionary booklet with the same WordNet senses arranged in random order generated by means of a random number generator. Here, the first sense was no longer necessarily the most inclusive, general one. A strong tendency towards picking the first sense in the random order would point to a reluctance to examine and evaluate all available senses, independent of whether this sense represented the most salient or core sense.

Not surprisingly, the expert choice was at the top of the list in the frequency condition for most words. The mean position of the expert choice for all parts of speech in the frequency order was 2.29; in the random condition, the mean position of the expert choice was 3.55.

The taggers, who worked independently from each other, were not aware of having been assigned to one of two groups of participants. They finished the task within 4-6 hours.

8 Results

We first report the percentage of overlap between taggers' and experts' choices in terms of the three main variables: POS, degree of polysemy, and the order of senses in WordNet. We give the results in percentages here; however, calculation of the significant effects is based on analyses of variance carried out on the raw data.

In the frequency condition, taggers overall chose the same sense as the experts 75.2% of the time; in the random condition, the overall agreement was 72.8%. In both conditions, performance was significantly ($p < 0.01$) higher for nouns than for the other parts of speech. For all four parts of speech, we found more tagger-expert matches in the frequency condition than in the random condition. The difference, however, was significant ($p < 0.05$) only for nouns.

The target words were classified into four groups depending on their polysemy count. Group 1 contained words with 2 senses; Group 2 words with 3-4 senses; the words in Group 3 had 5-7, and in Group 4, 8 or more senses. The groups were created so that each contained approximately 25% of the words from each part of speech, i.e., the groups were similar in size for each syntactic category.

Tagger-expert matches decreased significantly with increasing number of senses ($p<0.01$) in both conditions. This effect was found for all parts of speech, but it was especially strong for adverbs, where performance dropped from a mean 83.3% tagger-expert agreement for adverbs with two senses to 32.5% for adverbs with 5-7 senses, and to only 29.4% for the most polysemous adverbs. Except for words with two senses, we found more tagger-expert matches in the frequency condition than in the random condition.

In both conditions, significantly more tagger-expert matches occurred for all parts of speech when the expert choice was in first position than when it occurred in a subsequent position (80.2% vs. 70.5%, $p<0.01$ for the frequency condition; 79% vs. 70%, $p<0.05$ for the random condition). This effect was also found with the same level of significance for verbs alone, in both conditions. In the frequency condition, we found the effect of the expert choice being at the top of the list of senses to be particularly strong for the most polysemous words ($p<0.05$); the overall effect of the expert choice being the first choice for all polysemy classes was significant at the $p<0.01$ level. (For words with only two senses in WordNet, the position had no significant effect on the rate of agreement between taggers and experts.)

We now turn to the sense choices that were made by most taggers. We asked, what percentage of taggers selected the most frequently chosen sense, and did the syntactic class membership of the words, their degree of polysemy, or the order of the senses in WordNet have an effect on the rate of agreement?

Taggers agreed among themselves significantly more often than they did with the experts (82.5% in the frequency condition, and 82% in the random condition). Inter-tagger agreement followed the same pattern as tagger-expert matches: agreement decreased with increasing polysemy; agreement rates were highest for nouns and lowest for verbs and adjectives in both conditions.

Inter-tagger agreement decreased significantly ($p<0.01$) with increasing polysemy for all parts of speech in both conditions. This supports our expectation that more choices render the matching task more difficult, making agreement less likely. The decrease in inter-tagger agreement with increasing polysemy was especially strong in the case of adverbs.

In the frequency order condition, the overall agreement was significantly ($p<0.01$) higher (87%) when the agreed-upon sense was the first choice rather than a subsequent one (78%) on the list of alternative senses in the dictionary. This effect was also found separately for all POS except nouns. Similarly, we found that in the random order condition, inter-tagger agreement was higher for all POS when the agreed-upon sense was the first in the dictionary (85.5% vs. 79.6%). For the different polysemy groups, the choice most often made was in first position for low and medium high polysemy words, but for high polysemy words (5 or more senses), the most frequently selected sense was less often in the first position.

9 Discussion

The rather high tagger-expert agreement indicated that the novice taggers found the annotation task feasible. We found the predicted main effects for degree of polysemy, POS, and the order in which the senses were presented in the dictionary booklet.

Increasing polysemy of the target words produced less tagger-expert and inter-tagger agreement. Besides having to weigh and compare more options, the taggers needed to adjust their own ideas of the polysemous words' meanings to the particular way these are split up and represented in WordNet. The more alternative senses there were, the less likelihood there was that the taggers' mental representations of the senses overlapped significantly with those in WordNet.

In both conditions, nouns were tagged significantly more often in agreement with the experts' choice than verbs and adjectives. For nouns, we found no significant increase in the number of agreed-upon choices when they were at the top of the list of alternative senses, indicating that the taggers were fairly sure of their choices independent of the order in which the different noun senses were listed in the dictionary. This effect could be attributed at best only partly to the relatively low polysemy of nouns. Nouns may be "easier" because they commonly denote concrete, imagible referents. Verb and adjective meanings, on the other hand, are more context-dependent, particularly on the meanings of the nouns with which they co-occur. People's mental representations of noun concepts may be more fixed and stable and less vague than those of verbs and adjectives. In fact, the larger number of dictionary sense numbers for verbs in particular may be due less to actual meaning distinctions than to the lexicographer's attempt to account for the great semantic flexibility of many verbs.

Overall, taggers chose the expert selection less frequently than they agreed on a sense among themselves. While it is possible that the expert choice did not always reflect the best match, we suspect that novice taggers annotate differently from lexicographers. The latter are necessarily highly sensitive to sense distinctions and have developed a facility to retrieve and distinguish the multiple meanings of a word more easily than naive languge users, who may have a less rich representation of word meanings at their fingertips. This possibility is supported by (Jorgenson, 1990), whose naive subjects consistently distinguished fewer senses of a word than dictionaries do, even when they were given dictionaries to consult in the course of the sense discrimination task. Jorgenson's subjects agreed substantically on discriminating the three most central, salient senses of polysemous nouns but did not distinguish subsenses. Dictionaries likewise often agree among each other on the most central, core, senses of words but differ in the number and kinds of subtle distinctions. But whereas lexicographers are trained in drawing fine distinctions, naive language users appear to be aware of large-grained sense differences only. Our results indicate, in the case of finer sense distinctions, a lack of shared mental representations among the taggers, and a decrease in agreement. This explanation is also consistent with the decrease in tagger-expert matches along with increasing polysemy.

The salience and the shared mental representation of certain word senses might further account for our third main effect. Taggers agreed with the experts and with each other significantly more often when the WordNet senses were presented in the order of frequency of occurrence. This was generally true for words from all polysemy groups and POS. We suggest that taggers recognized the most appropriate sense more easily in this condition because they did not use the same strategy as in the random order condition. In the frequency condition, the most salient, "core," senses usually occurred first, or at least fairly high, on the list of senses. These senses also had a high chance of being the appropriate ones in the text, since we had selected a fiction passage with non-technical, everyday language. Taggers working in the frequency condition proba-

bly realized that the sense ordering resembled that of most standard dictionaries and chose the first sense that seemed at all to be a good match rather than examining all senses carefully, as they would have to do in the random order condition.

When the first sense was also the one the lexicographers had chosen as the most appropriate one, the taggers' task was relatively easy. Given that they recognized that the first sense was appropriate, selecting it meant that they did not have to examine and compare the remaining senses in search of an even better choice. Weighing all available senses against each other and against the given usage can be a difficult task especially for novice taggers, and we expected a general tendency to gravitate towards the first choice for this reason. Stopping to read after one has encountered the first sense that seems appropriate resembles the dictionary look-up strategy where one stops reading the entry when one has found a sense that seems to match the given usage (Kilgarriff, 1993).

The first senses in the frequency condition, which generally express the most salient and central meanings, might be most clearly representend in both naive and expert speakers' mental lexicons and might show the greatest overlap across speakers. These senses were presumably easily understood by the taggers and increased any reluctance to examine the remaining options.

The difference between the tagger-expert matches for words in the first position and words in subsequent positions was particularly strong for verbs and (in the frequency order condition) for words with eight or more senses. These were the cases that were generally more difficult for the taggers, as reflected in lower tagger-expert agreement. The results therefore indicate that the expert choice being the first made the decision process for the taggers much easier by eliminating the need for a difficult comparison of all the available senses, and, in the frequency condition, by the fact that the first sense was generally the most salient one.

The preference for the first among the available senses was even more pronounced in the inter-tagger agreement. There was a highly significant difference for the agreed-upon choice between the first and subsequent positions in the case of verbs and adjectives and words with eight or more senses in the frequency order condition ($p<0.01$). Again, the taggers probably understood the first, most frequent and often most salient sense easily and were reluctant to consider more fine-grained sense differentiations.

In the random order condition, no bias towards the first sense existed, so the strategy of choosing the first sense or an appropriate sense near the top of the list was not available. The taggers had to examine and consider each sense in the entry, which made the task more difficult. This is reflected in lower inter-tagger and tagger-expert agreement rates. Yet the high percentages of matches in this condition show that the taggers worked well. When the expert sense was the first on the list, taggers working in the random order condition selected the expert sense less frequently than the taggers working in the frequency order condition. This result further indicates that taggers here were not biased towards the first sense, but considered all senses equally.

In sum, we found that matching word usages to word senses in a dictionary is a hard task, whose difficulty depends on the part of speech of the target word and increases with the number of senses given in the dictionary. Among the available choices, the first sense of each polysemous word was a significant attractor.

Our findings suggest that randomly ordered senses would weaken taggers' strategy of relying on the first sense being the best match and encourage more scrupulous examination of the available choices.[4] Confidence ratings reflected the degree of difficulty of the items in that they paralleled the taggers' performance as measured by tagger-expert and inter-tagger agreement. Highly polysemous words were tagged with less confidence, and taggers were more confident when tagging nouns rather than verbs and modifiers. Confidence was slightly higher for inter-tagger than expert-tagger matches, supporting the reality of a "naive" lexicon as opposed to representation of polysemous words in the mental lexicon of practiced lexicographers or linguists. In the random order condition, taggers made their decision with more confidence than in the frequency order condition, although was less agreement with the experts. We believe that this result further supports the claim that taggers in the two conditions proceeded differently: Taggers working with a randomly ordered list of senses did not rely on the first sense being the correct one. They worked more scrupulously, which is reflected in the higher confidence ratings.

References

C. Fellbaum, J. Grabowski, and S. Landes. in press. Confidence and performance. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database.* MIT Press.

[4](Fellbaum et al., in press) report the confidence ratings of the taggers for their choices.

W. Gale, K. Church, and D. Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of 30th Annual Meeting of the Association for Computational Linguistics*, pages 249–256.

D. Gentner and I. France. 1988. The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. Small, G. Cottrell, and Tanenhaus, editors, *Lexical Ambiguity Resolution.* Morgan Kaufmann.

J. Jorgenson. 1990. The psycholinguistic reality of word senses. *Journal of Psycholinguistic Research*, 19:167–190.

J. J. Katz. 1964. Semantic theory and the meaning of "good". *Journal of Philosophy*, 61:739–766.

A. Kilgarriff. 1993. Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26:365–387.

S. Landes, C. Leacock, and R. Tengi. in press. Building semantic concordances. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database.*

G. A. Miller and C. Fellbaum. 1991. Semantic networks of english. *Cognition*, 41.

G. A. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the Human Language Technology Workshop*, pages 240–243. Morgan Kaufmann.

G. A. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4).

J. Pustejovsky. 1995. *The Generative Lexicon.* MIT Press.

P. Schwanenfluegel. 1991. Why are abstract concepts hard to understand? In P. Schwanenfluegel, editor, *The Psychology of Word Meaning.* Erlbaum.