Proving Basic Polysemy: Subjects Reliably Distinguish Several Senses of See
Author(s): Collin F. Baker
Proceedings of the 25th Annual Meeting of the Berkeley Linguistics Society (2000), pp. 14-25

Proving basic polysemy: subjects reliably distinguish several senses of see[1]

Collin F. Baker
UC Berkeley
International Computer Science Institute

1. Introduction

1.1 Background

Psycholinguistic experiments can be divided into those that ask the subjects to make high-level, conscious decisions about linguistic questions and those that ask subjects to make quick responses to relatively simple questions but measure reaction times, seeking clues as to the moment-to-moment processing of sentences; we can refer to these as "off-line" and "on-line" methods respectively. Off-line methods allow us to construct experiments relatively easily that seem to answer some fundamental linguistic questions directly ("Is this sentence grammatical?", "Are these words synonyms?" "What is the antecedent of this pronoun?"), but the data they provide is often hard to interpret, because so much higher-level cognition may be involved in the decision; e.g. in a grammaticality judgement, the subject may actually have time to remember some rule that she was taught in the sixth grade, rather than relying on a purely intuitive judgement. On-line tasks have a much better chance of discovering something about the semantic representation used in actual sentence understanding, but the experiments are much harder to construct and a great many factors need to be carefully controlled to produce valid results.

Among the tasks used in off-line experiments, free sorting tasks have better face validity than similarity judgments, but the resulting categories are hard to compare. For example, Jorgensen (1990) gave subjects cards containing low-polysemy nouns and high-polysemy nouns, as measured by the number of dictionary senses. In the first task, they did completely free sorting; in the second, they were told to divide the cards according to a set of dictionary definitions they had been provided with. She found that subjects basically produced about three categories for the low-polysemy words in both tasks, but created 5.6 categories on the free sorting and 9.1 on the dictionary-guided sorting. This was still less than the average number of senses given in the dictionary (14.6), but the increase was significant. Jorgensen (1990) uses measures of the number of categories produced by her subjects and the amount of agreement between them on the classification of individual items, but has nothing to say about the relation between the **semantics** of the categories produced by one subject and those of other subjects or those of the dictionary.

Various experiments have demonstrated that priming effects between related words can provide information as to the structure of semantic fields (Meyer & Schvaneveldt 1971, de Groot 1984). Other experimenters have found priming effects between the separate senses of homonyms (Swinney 1979, Simpson 1981, Seidenberg *et al.* 1982). Williams (1992) showed similar priming effects between senses of polysemous words; his experiment has important implications for those described here. In particular, Williams found that central and non-central senses had very different priming effects. Unfortunately, like many others, both Williams' sense distinctions and his decision as to which constituted the central sense were based on

a dictionary, despite the fact that commercial and other linguistically irrelevant factors influence the number and type of senses listed in dictionaries.

Given this situation, we chose to combine free sorting with two other methods, a forced classification task and a priming experiment, in the hope that the strengths of each method would complement the weaknesses of the others. In the classification task, subjects were forced to classify uses into many predetermined categories, assuming that they represent a finer breakdown than is available to most people through introspection. If subjects reliably make certain semantic distinctions, these should be representable as logical combinations of the finer ones. Of course, carrying the process to its logical conclusion, by collapsing **all** the categories together would produce complete "agreement" of the unsatisfactory sort posited in the strong version of monosemy (Ruhl 1989). As Cruse (1992) points out, claiming that a single highly abstract, undefinable "sense" accounts for all the uses of a highly polysemous word is not only *ipso facto* unprovable but also fails to distinguish such words from each other.

1.2 Predictions

On the basis of the study so far, we would predict the following:

- Since *see* seems to be highly polysemous, we would expect that speakers will be able to distinguish different senses when tested on tasks involving similarity judgements, categorization (using either predefined or their own spontaneous categories), etc. There is no reason to suppose that all speakers have exactly the same set of senses, but it is likely that there will be a great deal of overlap, which is essential to communication in general.

- Since *see* is such a highly-polysemous, high-frequency word, we expect to find that our subjects will produce more senses than Jorgensen's subjects.

- Since the senses appear to have a complex structure, some senses being more central than others, we expect to find prototype effects, with more central senses more likely to be spontaneously produced and more quickly recognized. We would expect to find broad agreement among speakers as to which are the central senses.

- In a cross modal priming experiment, in accord with Williams' (1992) findings, we would predict that sentences which provide a context for one sense of *see* would facilitate responses to a probe consisting of the keyword for that sense of *see*, more than probes consisting of keywords for other senses.

2. Experimental Methodology

2.1 Stimuli

2.1.1 Experiment 1

The stimuli for this experiment consisted of two blocks of 100 sentences selected at random from the Brown corpus, combined with 43 constructed example sentences, representing a total of 19 senses. For the sorting task, each block of 100 sentences was printed on 3x5 inch cards, forming two sets. A set of 43 cards was also prepared for the constructed example sentences (the "target" set), to see how they would be sorted.

2.1.2 Experiment 2

After completing Experiment 1, the list of senses was revised and expanded from 19 to 24 total senses. At the same time it was decided that some of the senses would not be used in further experiments, as they involve collocations with other specific words; among these are: SEE-X-THROUGH, SEE-THROUGH-X, and LET'S-SEE. On the other hand, senses such as ACCOMPANY, although "idiomatic", place semantic restrictions on their complements, but do not require any specific syntax, e.g. *I'll see you as far as the bus stop, I'll see you home, I'll see you to your door.*

In evaluating the results of Experiment 1, it became apparent that much of the subjects' difficulty was due to the large number of senses involved. It was therefore decided to construct a new experiment which would contain only examples of seven clear-cut senses, which would be relatively easy to distinguish from each other. The senses chosen were: EYE, FACULTY, RECOGNIZE, DETERMINE, ENSURE, EXPERIENCE, and SETTING. Example sentences were constructed for each of these senses, systematically varying other factors such as tense and aspect, question vs. statement, negation, voice, and domain of discourse. (The three domains of discourse were academic, personal, and entertainment, broadly construed.) In practice not all of combinations of these factors produced reasonable sentences, but as many as possible were created.

2.1.3 Experiment 3

After reviewing the results of Experiment 2, seven more senses were added to the stimuli for Experiment 3: VISIT, CONSULT, PROCESS, CONDITION, ENVISION, HALLUCINATE, and ACCOMPANY. In order to keep the total set of stimuli small enough, only the clearest examples of the seven senses used in Experiment 2 were retained in Experiment 3. As before, not all combinations of the manipulated factors with the senses produced good sentences.

2.2 Tasks

In Experiment 1, only Sentence Sorting and Classification were performed. Because both of these tasks are metalinguistic, two online tasks were added in Experiments 2 and 3, Lexical Decision and Categorial Judgement.

2.2.1 Task 1: Sentence Sorting

In this task, subjects were given cards containing the examples of *see* and asked to sort them into piles according to the sense of *see* used in the sentence. No directions were given as to how many senses there should be or how the distinctions should be made. At the end of one hour, subjects were asked to write a brief definition or characterization of each group and to choose the sentence which best exemplified it.

Nine subjects were randomly given one of the two sets of 100, and instructed to group them by sense. Then subjects were given the target set and asked to add these to the groups, then (if time permitted) they were given the second set of cards and asked to continue the task. All subjects completed one set of 100 sentences and the target set; if time permitted they continued to the second set of 100.

2.2.2 Task 2: Sentence Classification

For the Classification task, the example senses and the senses to be chosen (defined above) were presented on a computer screen using HTML and a web browser. Both the responses and the response latency were recorded.

2.2.3 Timed Tasks

The stimuli in these tasks were presented by use of the PsyScope program on Macintosh computers. In each case the subjects saw one of the same example sentences as in the previous tasks, and then heard an auditory prime consisting of a single word (or sometimes in the Lexical Decision task, a single non-word). The subjects then pressed one of the keys on the keyboard to respond. The sentences were displayed for up to 4 seconds. This was a followed by the auditory probe which lasted approximately 500 ms. Subjects had 1500 ms. from the beginning of the auditory probe to respond; responses after this time period were not used in further analysis.

Blocks of 40 trials of each task (Lexical Decision and Categorial Judgement) were administered randomly across subjects. Subjects were allowed to rest after each block.

2.2.3.1 Task 3: Lexical Decision

In Lexical Decision blocks, the probe was either a keyword for the primed sense, a keyword for another sense, or a non-word. The task was a word/non-word judgement.

2.2.3.2 Task 4: Categorial Judgement

In Categorial Judgement blocks, the probe was either a keyword for the primed sense or a keyword for another sense, and the task was to decide whether the probe was an instance of the primed sense.

2.3 Subjects

The subjects were undergraduates at University of California at Berkeley, who received credit toward introductory psychology courses for their participation. The same subjects participated in all tasks within each experiment. Table 1 shows a summary of the stimuli, tasks, and number of subjects in each of the experiments.

Table 1: Summary of the Experiments

Experiment	Senses	Stimuli	Number of subjects	Tasks
1	20	Corpus	9	1 & 2
2	7	Constructed	21	all
3	14	Constructed	39	all

2.4 Statistical Measures of Agreement

Two different measures of agreement were used in the experiments reported here, omega and kappa. The omega statistic (Morey & Agresti 1984) is inherently less

powerful, since it is based on whether or not two raters classify each pair of stimuli in the same category or not, without regard to the classification of other pairs. However, omega has the advantage that it can be used in cases in which the number of categories differs from rater to rater.

The kappa statistic (Scott 1955, Cohen 1960, see also the excellent introduction in Siegel & Castellan 1988 284-91) is the standard statistic for interrater reliability used when the number of categories is fixed for all raters.

Both statistics vary from 0 for chance agreement to 1.0 for perfect agreement and are insensitive to the number of categories involved, or the distribution of instances into categories. The variance of the sampling distribution is known for both, so that the probability of a particular outcome can be calculated.

3. Results and Analysis

Because the materials and tasks used in Experiment 1 were substantially different from those used in the latter two experiments, the results for Experiment 1 will be discussed separately first, and the results for the other two experiments will be discussed together thereafter.

3.1 Experiment 1

3.1.1 Task 1: Sorting

The number of categories per subject ranged from 6 to 21, with a mean of 11. This is substantially higher than the 5.6 found by Jorgensen (1990), as we had predicted.

The sizes of the categories varied greatly, from 33% for EYE (*See the cat on the mat*), and 15% for RECOGNIZE (*See that it's red*) to 0 for some categories. The agreement between raters as to the relative sizes of the categories was high, r = .70 to .97, suggesting that there is not a large division of the population into "lumpers" and "splitters".

3.1.2 Task 2: Classification

All subjects finished 99 sentences of the first set. Some subjects continued on to other sets, but the order of the sets was randomized, so that there was little overlap beyond the first set. The overall agreement among raters, measured by the kappa statistic, was .38. This value is low, but understandable, given the large number of senses listed and the ambiguity of many of the stimuli. A more detailed analysis of the classification data will be given for Experiments 2 and 3.

3.2 Experiments 2 and 3

3.2.1 Task 1: Sorting

For Experiments 2 and 3, the median numbers of categories produced by each subject were 6 and 10 respectively, which approximate the number of senses intended by the experimenters, i.e. 7 and 14. The difference between the two medians is significant (using the median test, $\chi^2=26.09$, $p < 0.01$); this means that subjects recognized that more senses were present in Experiment 3 on the basis of the stimuli alone.

The omega statistic was used to compare the subjects' initial sortings with the values of the manipulated variables, including the intended sense. The results suggest that the subjects were able to follow the instructions to pay attention only to the sense of *see* occurring in each sentence and to ignore the other syntactic and semantic factors. Table 2 shows figures for a representative group of nine subjects; the agreement for the irrelevant manipulated factors is essentially zero (because of the correction for chance agreement, the value of omega can sometimes be less than zero). The agreement with the intended sense ranges from a low 0.36 for subject number 30 to a high of .82 percent for subject number 33; this variation in agreement seems to be due to individual differences.

Table 2: Agreement between subjects' sorting and manipulated variables

Subjects→ Factors↓	30	31	32	33	34	35	36	37	38
Tense/Asp.	0.03	0.04	0.02	0.02	0.03	0.05	0.03	0.02	0.02
Qn/state.	0.03	0.01	0.00	0.01	0.01	0.02	0.02	0.02	0.01
Negation	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.00
Voice	-0.03	-0.01	0.03	0.01	0.00	0.03	-0.02	0.03	0.01
Domain	0.08	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Sense	**0.36**	**0.40**	**0.75**	**0.82**	**0.62**	**0.46**	**0.50**	**0.71**	**0.61**

A calculation of the omega statistic between each pair of subjects showed that there was substantial agreement among subjects even before any instructions as to categorization were given; on Experiment 2, for example, the mean Ω =.57. There was considerable variation among subjects, but there was no cluster of subjects who agreed with each other and disagreed with the experimenters' initial categorization. This suggests that there does not exist another well-defined "dialect" for the senses of *see*, although there may be agreement among subjects and disagreement with the experimenters on individual pairs of senses.

3.2.2 Task 2: Classification

In Experiment 2, the mean kappa for agreement among all subjects on the seven categories was .74, and 84% of the items were classified as intended by the experimenters.

In Experiment 3, we found that the 10 sentences with the lowest level of agreement were causing a disproportionate amount of error, and had no more than about 50% of responses in one category, so we eliminated them from further consideration, reducing the number of stimuli from 115 to 105. There were no comparable problems in the data for Experiment 2.

In Experiment 3, with 14 senses, the mean kappa among all subjects fell only to .70; after the elimination of the 10 weakest items, it rose to .75; also, 75% of the responses agreed with the experimenters' categorization.

In addition to recording subject responses, the response latency on the Classification task was also recorded; the distribution has a strong right skew, as is typical of such measurements. The median latency was 19 seconds, with the first quartile at 13 seconds and the third quartile at 26 seconds. Latencies longer than 80 seconds were

considered errors, since it seems unlikely the subjects were actually attending to the current item for so long.

Table 3. Experiment 2: Intended senses vs. responses

Responses → Intended ↓	EYE	FACUL-TY	DETER MINE	ENSURE	RECOG NIZE	EXPERI ENCE	SETTING	Total
EYE	623	49	3	2	8	0	0	685
FACULTY	65	381	1	0	3	0	0	450
DETERMINE	46	11	394	25	2	2	2	482
ENSURE	19	7	33	444	13	3	1	520
RECOGNIZE	11	9	25	7	597	3	0	652
EXPERIENCE	10	3	4	5	4	388	131	545
SETTING	11	1	1	6	1	73	335	428
Total	785	461	461	489	628	469	469	3762

Table 3 shows the relation between intended senses and responses for all of the items on the Classification task in Experiment 2. The senses have been arranged so that those frequently confused with each other are in adjacent rows and columns. Thus, while EYE and FACULTY were correctly classified most of the time, 49 instances of intended EYE were classified as FACULTY, and 65 instances of intended FACULTY were classified as EYE. The asymmetry between the two "errors" may be due to the general bias toward the response EYE.

There is also some confusion among the three senses DETERMINE, ENSURE, and RECOGNIZE. We note that all three of these senses involved a relation between the SEER, and a proposition; in the case of ENSURE, the SEER brings the proposition about; in DETERMINE, the SEER finds out if the proposition is true; in RECOGNIZE, the SEER merely becomes aware of the proposition.

Finally we note confusion also between EXPERIENCE and SETTING, especially from intended EXPERIENCE to response SETTING. It may seem surprising that these two senses are confused, especially as the SETTING sense is unique with respect to the semantics of its subject. The similarity between the senses is that both of them allow non-animate subjects, e.g. *The house saw use as a barracks during the Revolutionary War*. In the Sorting task, several subjects created a category for non-animate SEER, and this may point to the source of the confusion between these two senses.

Table 4 shows the relationship between intended senses and responses for Experiment 3, after the 10 weakest items have been eliminated as described above. Once again, we find the general bias toward the response EYE, and some of the same confusions as noted in Experiment 2. In addition, the newly added senses create new combinations; the most striking result is that the majority of examples of intended PROCESS receive the response EYE. Although some of the subjects created a separate category in the sorting task for what we call PROCESS, the predominance of EYE responses for intended PROCESS stimuli in Experiment 3 suggests that most subjects regard perceiving a person performing an action as a simple physical perception, notwithstanding the secondary predication associated with it. The newly introduced sense CONDITION also creates considerable

Table 4. Experiment 3: Intended senses vs. responses

Responses → Intended↓	EYE	PRO-CESS	FACULTY	VISIT	CON-SULT	CONDI-TION	EXPERIENCE	SET-TING	ENVI-SION	HALLUCINATE	RECOG-NIZE	DETER-MINE	EN-SURE	AC-COM-PANY	Tot
EYE	175	4	11	0	0	2	1	1	0	0	2	1	0	0	197
PROCESS	239	174	18	4	0	7	8	2	16	0	10	1	0	1	480
FACULTY	20	0	153	0	1	0	0	0	0	0	0	0	0	0	174
VISIT	35	0	4	382	0	0	0	0	1	0	0	2	0	0	424
CONSULT	1	0	1	25	487	0	0	0	0	0	0	1	0	0	515
CONDITION	30	32	16	2	0	279	14	0	3	0	5	0	1	0	382
EXPERIENCE	0	6	1	0	0	2	122	6	6	0	1	1	0	0	145
SETTING	0	0	0	0	0	3	33	108	6	0	0	0	1	0	151
ENVISION	9	9	2	0	0	0	1	1	335	2	1	0	0	0	360
HALLUCINATE	19	0	3	0	0	1	0	0	0	397	1	0	0	0	421
RECOGNIZE	1	4	1	0	0	30	4	0	3	0	168	7	1	0	219
DETERMINE	3	2	0	5	9	4	0	0	9	0	3	210	15	2	262
ENSURE	0	0	0	0	0	1	0	0	1	0	3	2	222	2	231
ACCOMPANY	6	0	1	5	0	0	0	0	0	0	0	0	2	416	430
Total	538	231	211	423	497	329	183	118	380	399	194	225	242	421	4391

confusion, although the vast majority of cases are "correctly" recognized, and the rest of the table is remarkably low in confusion.

3.2.3 Clustering of Classification Responses

One approach to finding shared structure is to use a clustering algorithm, based on the kappa statistic. The steps are as follows: (1) For each pair of categories, compute the kappa that would result if they were combined into one. (2) Actually combine the pair which produces the greatest increase in agreement. (This represents the distinction which was hardest for the subjects to agree upon.) (3) Repeat this procedure, until combining categories produces no more improvement. Depending on the data, this may be before all categories are merged.

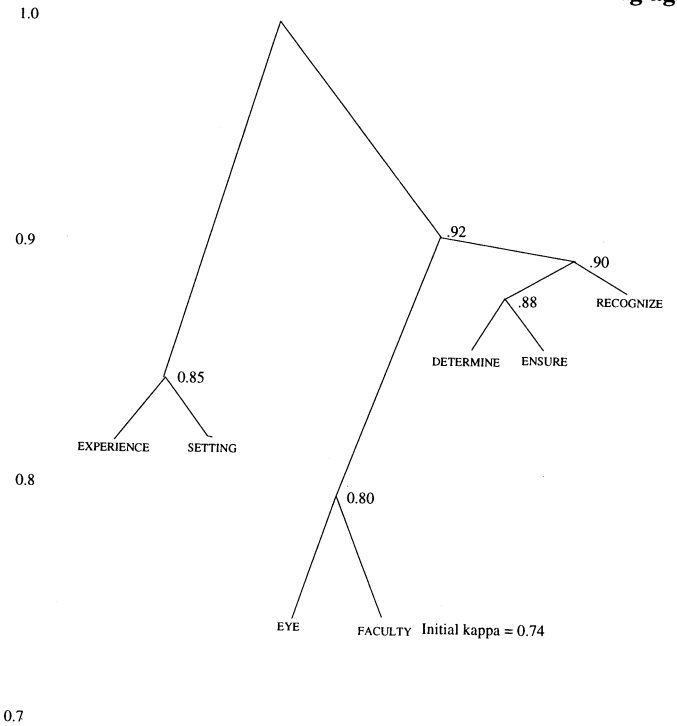
The order of combining can be represented as a tree, with the branchings at the bottom of the tree representing the categories most easily confused. The height of each branching represents the new level of agreement produced by combining the categories below. The clusters can be thought as reflecting the speakers' hierarchy of mental representations in this semantic space.

Figure 1 shows the results of clustering on the basis of agreement (i.e. kappa) for Experiment 2. This can be thought of as another way of looking at the confusion between intended senses and responses. Three clusters are noticeable (in order of decreasing confusion),EYE/FACULTY, EXPERIENCE/SETTING, and DETERMINE/ENSURE/RECOGNIZE. While these clusters were not foreseen by the experimenters, they seem reasonable *post hoc*, and also reflect the subjects' naive categorizations, as noted above. A somewhat similar, but less clear-cut tree (not shown) is produced from the results of Experiment 3. EYE and PROCESS are the first senses to merge, and ACCOMPANY lies at the top of the tree, but the other expected clusters are not apparent.

3.2.4 Timed Tasks

The data for the Lexical Decision task and the Categorial Judgement task are still being analyzed. The priming effects expected on the basis of Williams (1992) are very small (less than 50 milliseconds), and may require more accurate methods to detect them. In particular, our measurement of reaction times is subject to an error of approximately 16 ms., due to the polling frequency of the Macintosh keyboard. Preliminary results suggest that there is priming in the predicted directions in Experiment 2; it appears that the data is too sparse for good statistical analysis of the timed tasks in Experiment 3. We are continuing to work on eliminating outliers and finding appropriate groupings in this data.

Figure 1. Experiment 2: Clustering of senses based on increasing agreement



We have found quite a high percentage of "correct" responses in the Categorial Judgement task. On Experiment 2, the median is 92% correct (Q1 = 87%, Q3 = 96%). On Experiment 3 (after eliminating one subject who pressed the "yes" key on all of his responses) the median is 94% (Q1 = 89%, Q3 = 97%), even with the larger number of senses.

4. Conclusions and Future Directions

Our subjects were able to distinguish a relatively large number of senses for the highly ambiguous word *see* in Experiments 2 and 3, suggesting that the relatively low rate of agreement in Experiment 1 was due to the ambiguous and unbalanced stimuli rather than to inherent difficulty of the tasks. The level of agreement within subjects between the sorting task and the classification task suggests that the categories which we used in the classification task were fairly well matched with the categories which the subjects had at the beginning of the experiments.

The very high accuracy found on the Categorial Judgement task under timed conditions might be interpreted as proving that the subjects actually use the categories which they displayed on classification task in understanding natural language sentences. However, we cannot rule out the possibility that subjects have merely learned the categories created by the experimenters extremely well by that point. If the latter is occurring, there may be no way to get at the subjects' naive

representations except by creating several sets of a priori categories and determining which ones produce the highest agreement, presumably because of greater "naturalness",

Nothing in this experimental setup will help to resolve the vexed question of retrieval of fixed representations vs. different processing strategies, with which cognitive psychologists have been so concerned. Nor do these experiments provide any evidence as to the relative importance (or temporal precedence) of semantic vs. syntactic factors. Many senses have quite specific restrictions (syntactic and/or semantic) on their arguments, such as ACCOMPANY or DETERMINE, The subjects may be learning to distinguish at least some of senses from relatively straightforward syntactic cues, despite our best efforts to vary the syntactic patterns within senses. But the assumption that syntactic cues are more straightforward than semantic cues may itself be characteristic of linguists rather than most language users.

The technique of clustering on the basis of agreement statistics, described in Section 3.2.3 above, is useful in revealing certain aspects of the underlying structure of the senses. It is, however, naturally one-dimensional, and thus cannot reveal the complexity underlying the sense divisions. Several approaches for further, multidimensional analysis are being considered.

From a linguistic point of view, it would be possible to treat all of the syntactic factors (and perhaps some of the semantic factors) connected with each sense as features in a high dimensional space. As mentioned above, it was difficult to construct examples of particular senses with particular combinations of syntactic characteristics; the participation of certain senses in certain patterns of alternation and not others could be used as a set of features for discriminating the senses, somewhat in the manner of Levin 1993. These features need not be binary, and could even be continuous values. Such an approach would depend more on the analysts' linguistic judgments, but would not be limited by the particular alternations exemplified in a given experiment. A more experimental approach would be to consider the responses on the classification tasks as (partially) independent dimensions, and to find a method to reduce their dimensionality.

Notes

1. This is joint work with Jane A. Edwards, who has taken part in the design, running, and analysis of the experiments. This research will be discussed in more detail in my dissertation (Baker forthcoming). My colleague Chris Johnson also participated in the initial establishment of the list of senses, and I have received innumerable suggestions from the members of my committee, other UCB graduate students and faculty, and audience members at the presentation of this paper at the Berkeley Linguistics Society, February, 1999. Any errors which remain are my own responsibility.

References

Baker, Collin F. Forthcoming, 1999. Ph. D. dissertation. University of California at Berkeley.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement (E&PM)*. 20.37--46.

Cruse, D. Alan. 1992. Monosemy vs. polysemy. *Linguistics*. 3.577--599.

Durkin, Kevin, & Jocelyn Manning. 1989. Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research (JPLR)*. 18.577--612.

de Groot, A. M. B. 1984. Primed lexical decision: Combined effects of the proportion of related prime-target pairs and the stimulus-onset asynchrony of prime and target. *Quarterly Journal of Experimental Psychology*. 36A.253--280.

Jorgensen, Julia C. 1990. The psychological reality of word senses. *JPLR*. 19.167--190.

Levin, Beth. 1993. English Verb Classes and Alternations: A Preliminary Investigation. Chicago: University of Chicago Press.

Meyer, D. E., & R. W. Schvanenveldt. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* 90.227--234.

Morey, Leslie C., & Alan Agresti. 1984. The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *E&PM*. 44.33--37.

Ruhl, Charles. 1989. *On monosemy: a study in linguistic semantics*. Albany, N.Y.: State University of New York Press.

Scott, J. 1955. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*. 19.321--325.

Seidenberg, M. S., M. K. Tannenhaus, J. M. Leiman, &M. Bienkowski. 1982. Automatic access of the meanings of ambiguous words in context: Some limitations on knowledge-based processing. *Cognitive Psychology*. 14.489--532.

Siegel, Sidney, & Jr. Castellan, N. John. 1988. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 2nd edition.

Simpson, G. B. 1981. Meaning dominance and semantic context in the processing of semantic ambiguity. *Journal of Verbal Learning and Verbal Behavior (JVLVB)*. 20.120--136.

Swinney, David A. 1979. Lexical access during sentence comprehension: (re)consideration of context effects. *JVLVB*. 18.645--59.

Williams, John N. 1992. Processing polysemous words in context: Evidence for interrelated meanings. *JPLR*. 21.193--218.