# Word Sense Disambiguation by Human Subjects: Computational and Psycholinguistic Applications

Thomas E. Ahlswede

*Computer Science Dept.*
*Central Michigan University*
*Mt. Pleasant, MI 48859*
*ahlswede@cps201.cps.cmich.edu*

David Lorand

*Earlham College*
*Richmond, IN 47374*
*davel@yang.earlham.bitnet*

### Abstract

Although automated word sense disambiguation has become a popular activity within computational lexicology, evaluation of the accuracy of disambiguation systems is still mostly limited to manual checking by the developer. This paper describes our work in collecting data on the disambiguation behavior of human subjects, with the intention of providing (1) a norm against which dictionary-based systems (and perhaps others) can be evaluated, and (2) a source of psycholinguistic information about previously unobserved aspects of human disambiguation, for the use of both psycholinguists and computational researchers. We also describe two of our most important tools: a questionnaire of ambiguous test words in various contexts, and a hypertext user interface for efficient and powerful collection of data from human subjects.

## 1 The need for a metric of disambiguation

Research in automatic lexical disambiguation has been going on for decades, and in recent years experimental disambiguation systems have proliferated. The problem of determining the accuracy of these systems has been little recognized: the usual check for correctness is a comparison of the test results against the experimenter's own judgment. Even less considered has been the question of what constitutes correctness in disambiguation, beyond the intuitive recognition that some disambiguations are better ("correct") and others worse ("incorrect").

A common approach to disambiguation is to select among the homographs and senses provided by a machine-readable dictionary (e.g. Lesk [1986], Byrd [1989], Krovetz [1989], Slator [1989], Guthrie et al. [1990], Ide and Veronis [1990], and Veronis and Ide [1990]. Dictionaries deal with the ambiguity of words by providing multiple definitions for sufficiently ambiguous words. These multiple definitions may be homographs (distinct words of unrelated meaning, whose written forms coincide) or senses (related but nonidentical meanings of a single word).

The inadequacy of a finite, discrete set of sense definitions to resolve all ambiguities has been pointed out by Boguraev and Pustejovsky [1990], Kilgarriff [1991], and Ahlswede [forthcoming]. For the practical task of disambiguation in natural language processing, however, the dictionary is a valuable and convenient source of sense distinctions; in our view, the best single source.

## 2 Evaluations of Human and Automatic Disambiguation

Many previous studies of human disambiguation have been from a psycholinguistic point of view. Simpson and Burgess [1988], surveying some of these studies, identify three basic models of ambiguity processing: (1) restriction by context, (2) ordered access, and (3) multiple access. Prather and Swinney [1988] consider whether the lexical component of human language processing is modular, i.e., acts independently of other components, or whether it interacts with other components.

Computationally oriented evaluations of human disambiguation began as incidental adjuncts to computational projects. Amsler and White [1979], with the help of assistants, manually (i.e., by human judgment) disambiguated the nouns and verbs used in definitions in the Merriam-Webster Pocket Dictionary. In an informal study, they found that their disambiguators' self-consistency on repeat performance was high (84%) but their consistency with respect to each other was lower.

The need for some means of evaluating automatic disambiguation methods, more rigorous than the experimenter's personal judgment, has become more obvious with the recent growing interest in the topic. Gale, Church and Yarowsky [1992], for instance, have followed the approach of estimating upper and lower bounds on the performance of a system.

## 3 Preliminary experiments

The project described in this paper began when one of us (Ahlswede) wrote disambiguation programs based on those of Lesk [1986] and Ide and Veronis [1990] for application in dictionary and corpus research. Lesk claimed 50-70% accuracy on short samples of literary and journalistic input. Ide and Veronis claimed a 90% accuracy rate for their program, although they explained that they had tested it against strongly distinct definitions – mainly homographs rather than senses.

After running the programs on test data containing ambiguities at both homograph and sense level, and evaluating the results, Ahlswede doubted whether, given this subtler mix of ambiguities, even a single human judge would achieve 90% consistency on successive evaluations of the same output; moreover, the consistency among multiple judges might well be much lower. Ahlswede recruited seven colleagues and friends to evaluate the test data, then compared their disambiguations of the test data against each other. The level of agreement averaged only 66% among the various human informants, ranging from 31% to 88% between pairs of informants [Ahlswede, forthcoming].

This figure was based on a simple pairwise comparison strategy. The informants rated each sense definition of a test word with a "1" indicating that it correctly represented the meaning of the word as used in the test text; "-1" if the definition did not correctly represent the meaning; and "0" if for any reason the informant could not decide one way or the other.

Pairs of informants were then compared by matching their ratings of the sense definitions of each word. The pair were considered to agree on a test word if at least one sense received a "1" from both informants and if no sense receiving a "1" from either informant was given a "-1" by the other.

This scoring method had the advantage of simplicity, but it did not reflect the agreement implicit in the rejection as well as the selection of senses by both informants. But the relative weight of common rejections and common selections among the senses of a given test word depends on the total number of senses, which varies widely. No discrete-valued scoring mechanism seems able to solve this problem.

A pairwise scoring procedure that gives much more plausible results is the coefficient of correlation, applied to the parallel evaluations by the informants being compared. It clearly distinguishes the relatively high agreement expected from human subjects from the relatively low agreement predicted for primitive automatic disambiguation systems, and from the more or less random behavior of a control series of random "disambiguations."

Table 1. Pairwise correlations of performance of human, machine and control disambiguations of test texts

| | h1 | h2 | h3 | h4 | h5 | h6 | h6a | h7 | m1 | m1a | m2 | a1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| . | | | | | | | | | | | | |
| h2 | .737 | | | | | | | | | | | |
| h3 | .463 | .497 | | | | | | | | | | |
| h4 | .743 | .669 | .428 | | | | | | | | | |
| h5 | .719 | .679 | .477 | .643 | | | | | | | | |
| h6 | .793 | .747 | .531 | .674 | .817 | | | | | | | |
| h6a | .831 | .797 | .528 | .755 | .825 | .926 | | | | | | |
| h7 | .524 | .523 | .455 | .524 | .506 | .565 | .543 | | | | | |
| . | | | | | | | | | | | | |
| m1 | .196 | .136 | .080 | .185 | .204 | .180 | .163 | .148 | | | | |
| m1a | .281 | .220 | .154 | .235 | .274 | .274 | .252 | .240 | .723 | | | |
| m2 | .097 | .083 | .016 | .033 | .100 | .104 | .085 | .033 | .027 | .060 | | |
| . | | | | | | | | | | | | |
| a1 | .013 | .014 | -.061 | .011 | .014 | .016 | .016 | .011 | .011 | .010 | .012 | |
| rand | .017 | -.008 | .047 | .010 | -.008 | .005 | -.000 | .008 | -.000 | .014 | .024 | -.016 |

Notes:

1. h1 through h7 are human informants; h6 took the test twice.
2. m1 and m1a are implementations of Lesk's algorithm. In m1a, the test texts were previously disambiguated for part of speech; senses of inappropriate parts of speech were assumed incorrect, and left out of the test data.
3. m2 is a spreading activation algorithm related to the Ide-Veronis algorithm.
4. a1 is a control in which all senses of all test words received a "1". In our first scoring strategy, a1 achieved absurdly high scores.
5. rand is a control created by randomly scrambling the sequence of answers in one of the human samples.

These results suggested that a very high accuracy rate is not so much unrealistic as meaningless: which of the human informants should the computer agree with, if the humans cannot agree among themselves?

For this reason, the informal experiment has led to the development of a larger and more formal test of human disambiguation performance. The main areas of innovation are (1) a much more systematically designed questionnaire, to be administered to hundreds of subjects rather than only seven, and (2) a user interface to facilitate both the completion of the questionnaire by this large number of human subjects, and our analysis of their performance. The biggest advantage of a computerized interface is that we can study the timing of subjects' responses: valuable information that could not be recorded in the original written test.

Combined with the user interface, the questionnaire is adapted for administration to human informants, but it can be adapted with little effort for use with dictionary-based disambiguation programs, as was done with its written (but also machine-readable) predecessor.

## 4 Design of the Questionnaire

The prototype version of the present questionnaire was a printed list of 100 test texts, each with an ambiguous word highlighted and a list of definitions following. Subjects typically took the test home, and reported needing anything from half an hour to several days to complete the questionnaire.

The test was difficult to complete for several reasons. The test texts were themselves dictionary definitions, chosen at random from the machine-readable version of the Collins English Dictionary (CED). (This was because the project grew out of an effort specifically to disambiguate definition texts in the CED.) Many of the words being defined by the test texts were highly obscure, e.g.

> **paduasoy** n. a rich strong silk fabric used for hangings, vestments, etc.

or

> **India paper** n. another name (not in technical usage) for bible paper
>
> [Ahlswede and Lorand, 1993]

Disambiguation was done (as it still is in the present questionnaire) by choosing one or more from a set of dictionary definitions of the highlighted word. This was hard work, and volunteers were hard to find. Therefore, though the present version of the questionnaire avoids "hard" words except where these are explicitly being studied, it is still tough enough that we pay our subjects a small honorarium.

Like its prototype, the present questionnaire consists of 100 test texts, each with an ambiguous test word or short phrase (e.g., ring up, go over). The number 100 was chosen, based on our experience with the prototype, as a compromise between a smaller test, easier on the subject but less informative, and a larger test which might be prohibitively difficult or time-consuming to take.

## 5 The Test Texts

Source. The test texts have been selected in part to represent a wide variety of written English, while using a minimum of different sources in order to facilitate comparison within each category as well as between categories. The distribution was:

- 24 General nonfiction (house and garden management tips, extracted from the VADisc corpus [Fox, 1988])
- 24 Fiction (selections from short stories by Mark Twain)
- 24 Journalism (The Wall Street Journal (WSJ), extracted from the ACL-DCI Corpus [Liberman, 1991];
- 20 Definitions from the CED [Collins, 1974] (selected from definitions used in the prototype questionnaire)
- 8 special texts (constructed to test specific interesting ambiguities)

One of the original criteria for both test words and test texts, neutrality between British and American usage [Ahlswede, 1992], was found virtually impossible to maintain. The CED is British, and many if not most of its multi-sense entries include definitions of idiomatic British usages. To leave these out would be to risk distorting the results as a metric for a disambiguation program that used the CED as a whole without excluding those particular definitions. The other categories are American, and in the interest of consistency, American idioms were freely permitted as well.

Several other criteria for selecting test texts were retained and followed:

1. Difficulty of resolution. This can only be estimated subjectively until the questionnaire results are in, except for the twenty dictionary definitions, where a rough measure of difficulty of resolution is provided by the "coefficient of certainty" [Ahlswede, forthcoming].

A second measure, the "coefficient of dissent", specifically measures disagreement as opposed to uncertainty. The high negative correlation between coefficient of certainty and coefficient of dissent (-0.942) indicated that, in practice, there was little difference between widespread uncertainty and widespread disagreement.

Partly because of the apparent lack of importance in this distinction, and partly for the convenience of automating the questionnaire, the "0" option in the prototype has been eliminated. The subject is forced to decide "yes" or "no" to each sense.

Size of context. The test texts are complete sentences, or (in the case of CED) complete definitions. In some cases phrases have been deleted with ellipsis, where the full text seemed unmanageably long and the deleted phrase irrelevant to the disambiguation of the test word. The net sentence length ranges from 5 to 28 with a median of 14. Results so far indicate, as did Lesk's observations, that sentence length does not significantly affect performance.

Global context was early recognized as a potential problem: human disambiguation decisions are made not only on the basis of the immediate sentence-level context, but also on an awareness of the domain: for instance, the word capital is likely (though not certain) to mean one thing in the Wall Street Journal and another thing in a political editorial about the federal government.

Since the test texts are short and have no global context whatever, we compensate by adding a small parenthetical note at the end of each text, identifying it as "WSJ", "Tips", "CED", "Twain" or "special". The meaning of these short tags is explained to the subject, and though not the same as actual global context, they provide explicitly the information the reader normally deduces during reading.

## 6 The Test Words

An factor which is probably important, but impossible to measure, is the familiarity of a test word. Two contrasting intuitions about familiarity are (1) an unfamiliar word should be harder to disambiguate because its senses are less well known to the informant; but (2) a familiar word should be harder because it is likely to have more senses and homographs. Since familiarity is not only completely subjective, but also varies widely from one individual to another, we turn to a much more measurable criterion:

Frequency. An unanswered question is whether it is more appropriate to measure word frequency based on the specialized corpora from which the texts are extracted, or based on a single average word frequency list. The texts taken from the CED, the Wall Street Journal, and the "Tips", having been extracted from multi-million-word corpora, can be measured separately. Unfortunately, we have no online corpus of Mark Twain's works, and the "special" texts are, by definition, not from any corpus at all.

Part of speech. Studies of disambiguation have focused almost exclusively on nouns, verbs and adjectives, and hardly at all on "function words" such as prepositions, conjunctions, and those adverbs not derived from adjectives. (An exception is Brugman and Lakoff [1988], who study the word over.) We are interested in in both kinds of words. Therefore the test words include 28 nouns, 22 verbs, 19 adjectives, 16 adverbs (none in -ly), and 15 assorted prepositions, conjunctions and pronouns.

Given the combination of a British dictionary with such ultra-American sources as Mark Twain, we were unable to guarantee variety neutrality in our test words as in our test texts. An alternative, however, was to include among the "special" texts two with strong variety bias: I took the tube to the repair shop, ambiguous in British but not in American, and It was a long and unpleasant fall,, ambiguous in American but not (or less so) in British. These were added in the hope that native or learned speakers of British English would handle them differently than speakers of American English.

## 7 The User Interface

An important feature of the questionnaire is its user interface. This was developed by one of us (Lorand) in Macintosh HyperCard.

The interface consists of four principal modules ("stacks" in hypertext terminology): (1) a top-level stack that drives the interface as a whole; (2) a "demographics" stack that manages a menu of demographic and identifying information that the subject fills out; (3) the "import questionnaire" stack, which allows the questionnaire to exist independently of the interface as an editable text file, and to be reinserted into the interface as desired, e.g., after changes have been made; (4) the questionnaire itself, translated automatically into MetaTalk, the MetaCard programming language.

## 8 The demographics stack

The menu of the demographics stack first solicits non-identifying portions of the subject's Social Security number and birthday, which are hashed to form a unique, confidential ID for that subject. The menu then solicits potentially relevant demographic information: age, gender, native/non-native speaker of English, number of years speaking English if non-native, and highest educational degree. This last is an extremely rough measure of literacy, but no better one is available, and the preliminary experiment showed that doctoral-level subjects agreed more closely with each other than the non-doctoral subjects did either with the doctorates or with each other [Ahlswede, forthcoming].

The ID and the demographic information are written to a text file in numerically coded form. The subject may then begin the questionnaire or cancel.

## 9 The questionnaire stack

The questionnaire is implemented as a series of windows, one for each test text and its associated definitions. The test text is displayed at the top of the window, with the test word in boldface. Below is a subwindow containing the definitions. The subject clicks on a definition to identify it as a good disambiguation; the typeface of the selected definition changes to boldface. Clicking on a selected definition will de-select it and its typeface will change back to regular. Any number of definitions may be selected. If, as sometimes happens, there are too many definitions to fit within the subwindow, it can be scrolled up and down to give access to all the definitions. Arrow buttons at the bottom right and bottom left enable the subject to go ahead to the next text or back to the previous one.

Every action by the subject is logged, as is its time, in the log file. Thus when the subject is done, we have a complete record of his or her actions, of the time at which each action took place, and thus of the interval between each pair of actions.

## 10 The Subjects

So far, most of the subjects recruited have been students, with some faculty and staff. We are presently recruiting off campus. Probably thanks to the honorarium, response has been enthusiastic: well over the 100 subjects we considered necessary for an adequate sample. Because we are still occupied with data collection, intensive analysis of the data has not begun yet.

## 11 Conclusions

As we administer the questionnaire, we are developing approaches to the analysis of the resulting data. When we have acquired a large enough collection of performances, we will begin formal analysis.

Our first concern in this effort has been to develop a useful corpus or set of "norms" of human disambiguation behavior, against which automatic disambiguation systems, at least those based on machine-readable dictionaries, can be compared. We also believe, however, that our results will be interesting to psycholinguists studying human disambiguation: since our approach has been different from previous psycholinguistic experiments, we expect that considerable new knowledge will emerge from the data we are now gathering.

## 12 References

Ahlswede, Thomas E. and David Lorand, 1993. The Ambiguity Questionnaire: A Study of Lexical Disambiguation by Human Informants. Proc. of the Fifth Annual Midwest Artificial Intelligence and Cognitive Science Society Conference, Chesterton, Indiana, pp. 21-25.

Ahlswede, Thomas E., 1992. Issues in the Design of Test Data for Lexical Disambiguation by Humans and Machines. /f2Proc. of the Fourth Annual Midwest Artificial Intelligence and Cognitive Science Society Conference, Starved Rock, Illinois, pp. 112-116.

Ahlswede, Thomas E., forthcoming. An Experiment in Human vs. Machine Disambiguation of Word Senses. (Submitted).

Ahlswede, Thomas E. and Martha Evens, 1988. Generating a Relational Lexicon from a Machine-Readable Dictionary. International Journal of Lexicography, vol. 1, no. 3, pp. 214-237.

Boguraev, Branimir, and James Pustejovsky, 1990. Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design. Proc. of COLING-90, Helsinki, vol. 2, pp. 36-41.

Brugman, Claudia, and George Lakoff, 1989. Cognitive Topology and Lexical Networks. In Small et al., eds., pp. 477-508.

Byrd, Roy, 1989. Discovering Relationships among Word Senses. In Dictionaries in the Electronic Age: Proc. of the Fifth Annual Conference of the UW Centre for the New OED, Waterloo, Ontario, pp. 67-80.

Collins English Dictionary, 1974. Collins, Birmingham.

Fox, Edward, 1988. Virginia Disc One. Virginia Polytechnic Institute, Blacksburg, Virginia.

Gale, William, Kenneth Church and David Yarowsky, 1992. Estimating Upper and Lower Bounds on the Performance of Word Sense Disambiguation Programs. Proc. of the 30th Annual Meeting of the ACL, Newark, Delaware.

Guthrie, Louise, Brian M. Slator, Yorick Wilks, and Rebecca Bruce, 1990. Is there content in empty heads? Proc. of COLING-90, Helsinki, vol. 3, pp. 138-143.

Hirst, Graeme, 1987. Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press, Cambridge.

Ide, Nancy M. and Jean Veronis, 1990. Mapping Dictionaries: A Spreading Activation Approach. Proc. of the 6th Annual Conference of the UW Centre for New OED, Waterloo, Ontario, pp. 52-64.

Krovetz, Robert, 1989. Lexical Acquisition and Information Retrieval. Proc. of the First International Lexical Acquisition Workshop, IJCAI, Detroit, 1989.

Lesk, Michael, 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. Proc. of SIGDOC, Toronto, pp. 1-9.

Liberman, Mark, 1991. ACL/DCI Corpus, University of Pennsylvania.

Prather, P. A., and David A. Swinney, 1988. Lexical Processing and Ambiguity Resolution: An Autonomous Process in an Interactive Box. In Small et al., eds., pp. 289-310.

Pustejovsky, James, 1989. Current Issues in Computational Lexical Semantics. Proc. of the Fourth Conference of the European Chapter of the ACL,, Manchester, pp. xvii-xxv.

Simpson, Greg B. and Curt Burgess, 1988. Implications of Lexical Ambiguity Resolution for Word Recognition and Comprehension. In Small et al., eds., pp. 271-288.

Slator, Brian M., 1989. Using Context for Sense Preference. Proc. of the First International Lexical Acquisition Workshop, IJCAI, Detroit, 1989.

Small, Steven L., Garrison W. Cottrell, and Michael K. Tanenhaus, eds., 1988. Lexical Ambiguity Resolution. Morgan Kaufman, San Mateo, California.

Veronis, Jean, and Nancy M. Ide, 1990. Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. Proc. of COLING-90, Helsinki, vol. 2, pp. 398-394.