



whitepaper

# microlanguage™ thematix™

intelligent language processing vs.  
conventional search technology

## abstract

In the search for relevant information, up to 80% of the total effort can be wasted wading through irrelevant information returned by conventional search technologies that can not overcome difficulties caused by the complexities of language in context. What is needed is a next-generation technology that can process language intelligently, and thus return only the most relevant information to the searcher. We tested Microlanguage's Thematix technology against a conventional search technology using a publicly available corpus. The results: Thematix far outperforms conventional search technologies; up to 99% of information returned is relevant, and up to 99% of the searcher's time (and the cost of finding the information) is conserved.

## table of contents

the challenge.....	3
the test .....	4
microlanguage thematix technology.....	4
conventional search technology .....	4
the corpus .....	5
the test set .....	6
the computing platform .....	6
the metrics .....	7
the results .....	7
conclusions .....	8

## table of figures

table 1: WordNet senses for <i>board</i> .....	5
table 2: examples from the DSO corpus .....	6
poor vs idealized performance .....	7
table 3: boolean search vs thematix.....	8

## about microlanguage

Microlanguage develops and markets second generation concept extraction software products and related services to OEMs, systems integrators, and portal information providers that enhance the value of their products and services to global 2000 technology, finance, healthcare and pharmaceutical, manufacturing firms, and agencies involved in national security. For more information on Microlanguage and Thematix, visit us on the web at [www.microlanguage.com](http://www.microlanguage.com), or call us at 610.995.1017.

## reclaiming lost productivity and opportunity costs: a \$2 billion opportunity

### the challenge

According to a study on knowledge worker productivity, market research firm IDC<sup>1</sup> concluded that enterprise employees may be losing as much as three hours per day on fruitless information searches. In the context of the top 1,000 U.S. companies, each with an average of 1,000 knowledge workers, of the \$80 billion budgeted annually for knowledge worker activities, IDC estimates that \$2.6 billion is being spent on the search for indexed information. An average of 80% of that time, which translates to \$2 billion, is wasted searching through irrelevant information returned by conventional search technology.

How can this be? This paper answers that question, as well as an even better question: what is the solution to this problem?

The surprisingly high productivity penalty of knowledge management and information retrieval can be attributed to two key issues:

- Human languages have inherent complexities that make the process of finding relevant information a technological challenge.
- Conventional search technologies are not up to this challenge.

One of the primary problems facing search technologies is a direct result of language's inherent complexity: a single word can have different meanings, or senses. The problem is referred to as WORD SENSE AMBIGUITY, and the challenge for search technologies is WORD SENSE DISAMBIGUATION, or being able to distinguish between the different senses that a single word can have.

The word *board*, for example, can mean 'a committee,' 'a piece of lumber,' 'a meal' (as in the phrase "room and board"), or several other things. A person searching for information on a particular board meeting knows exactly what information he or she wants, but since conventional search technologies can not distinguish between the possible senses, the search will likely return a plethora of documents that have nothing to do with board meetings, even though all of the documents contain the word *board*. In short, conventional search technologies don't know what the searcher means, which explains why the average knowledge worker spends so much time sifting through unwanted information.

---

<sup>1</sup> Infoworld Magazine, January 7, 2002 issue.

Now, the important question: if conventional search technologies are part of the problem, what is the solution? The short answer is a next-generation technology that deals intelligently with the inherent complexities in language: Microlanguage's Thematix.

Thematix, as the name suggests, is based on themes or concepts. This powerful technology is used to find relevant information by overcoming the inherent problems of language, such as a single word having different meanings. To prove that a Thematix-powered search is many times more effective and efficient than one powered by conventional search technologies, and that this type of technology is the answer to the \$2.5 billion question, Microlanguage benchmarked Thematix against the most popular search technology, the Boolean search.

## the test

The test facilities consisted of the following elements:

- Microlanguage Thematix technology
- a Boolean search
- a corpus
- the test set
- the computing platform

This section discusses each of these items in turn, which all describe how the test was set up and run. The last two sections describe the metrics used to score the test results, and the actual test results, with discussion of their relevance.

## microlanguage thematix technology

Thematix is really two component technologies that work together. At the heart of Thematix is the parsing engine, which takes any symbol string as input (in this case unstructured text) and breaks the string into its constituent parts (in this case, words and punctuation). The parsing engine then interacts with the second component technology called a grammarbase™, the “brains” of Thematix. Think of a grammarbase as a repository of knowledge that instructs the parsing engine. The parsing engine uses the grammarbase instructions to disambiguate word senses so that any search tool that is integrated with Thematix will return more accurate and complete results.

## conventional search technology

The conventional search technology selected was a standard Boolean search, used by many knowledge management and information retrieval applications, including search engines of various kinds.

## the corpus

In order to achieve external validation of Thematix, a publicly-available collection of hand-marked sentences (i.e. a CORPUS) was selected to test how well each technology recognizes the correct meaning of each word. The corpus selected is the DSO Corpus of Sense-Tagged English, available from the Linguistic Data Consortium<sup>2</sup>. It consists of sense-tagged word occurrences for 121 nouns and 70 verbs in about 192,800 sentences. The different senses of each word are defined by WordNet 1.5, a publicly available dictionary of words and their different senses<sup>3</sup>. This makes the test doubly external: the corpus was tagged by a third party using word senses created by a separate third party. Table 1 shows the example definition for the word *board* as it appears in WordNet; there are eight separate senses given for this word.

table 1: WordNet senses for *board*

Sense	Definition
1	a committee having supervisory powers; "the board has seven members"
2	a plank, a stout length of sawn timber; made in a wide variety of sizes and used for many purposes
3	a flat piece of material designed for a special purpose; "he nailed boards across the windows"
4	food or meals in general; "she sets a fine table"; "room and board"
5	circuit board, circuit card, board, card -- a printed circuit that can be inserted into expansion slots in a computer to increase the computer's capabilities
6	dining table, board -- a table at which meals are served; "he helped her clear the dining table"; "a feast was spread upon the board"
7	control panel, display panel, instrument panel, control board, board, panel -- an insulated panel containing switches and dials and meters for controlling electrical devices; "he checked the instrument panel"; "suddenly the board lit up like a Christmas tree"
8	board -- a flat portable surface (usually rectangular) designed for board games; "he got out the board and set up the pieces"

Notice that these definitions are much more finely-grained than is relevant in most information applications. Conversely, one could

<sup>2</sup> <http://www ldc.upenn.edu/>

<sup>3</sup> <http://www cogsci.princeton.edu/~wn/>

reasonably argue that there are senses that are missing from the WordNet senses. However, since the test is based on third-party materials, we accepted the WordNet senses, and treated them as a problem to be addressed by the technologies.

Table 2 shows two example sentences from the corpus. Every sentence has a unique ID number, as shown in the first column. Each sentence contains an instance of the word *board* as well as the number of the WordNet sense, as shown in the third column.

table 2: examples from the DSO corpus

Sentence ID	Sense	Example Sentence
dj41.db#1579	1	Mr. Sandner, who also sits on the executive committee of the Merc's <b>board</b> , confirmed his investment in ABS but denied any conflict.
dj08.db#1664	5	"The subsidiary produces materials used in printing and in the fabrication of printed circuit <b>boards</b> ."

Thus for the first sentence, the sense of *board* is WordNet sense number one, or the 'committee' sense; for the second sentence, board has WordNet sense five, or the 'circuit board' sense.

## the test set

We chose four words from those tagged in the corpus. The words were chosen because three of them, *board*, *company*, and *state*, have at least one sense that has to do with either business or government. More importantly, they also have senses that have nothing to do with business or government, underscoring the fact that if someone needs to retrieve relevant information using these words, there is a strong likelihood that a typical information search will return irrelevant results and waste the time of the searcher. The fourth word, *lie*, is a classic example of a difficult word to use correctly, even for native speakers (recall the grammar school rules for the use of *lay* versus *lie*). It is also the only verb in the set, and verbs tend to have much more closely related meanings than nouns, making the disambiguation even more difficult.

## the computing platform

The computing platform for the test was a desktop PC featuring a 200 MHz Pentium II and 128 MB of RAM, running the Windows 2000 operating system.

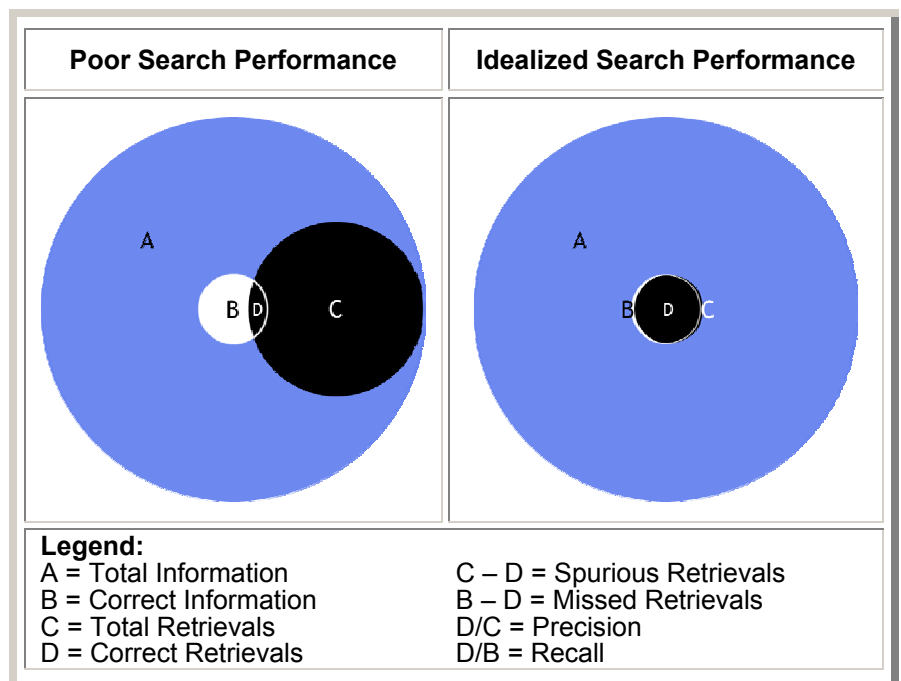
## the metrics

The task is to recognize the 24 correct senses of the four words in the test set. The standard metrics of precision and recall are used to gauge performance:

- **RECALL:** the percentage of the total number of relevant sentences in the corpus that were retrieved
- **PRECISION:** the percentage of the total number of sentences that were retrieved that were actually relevant

These are two very different metrics, and they tend to be inversely proportional. For example, the more information you retrieve (i.e. higher recall), the more likely you are to retrieve information that is not relevant (i.e. lower precision). Conversely, if you are more careful to retrieve only relevant information (higher precision), you are more likely to miss relevant information (lower recall).

## poor vs idealized performance



The concepts of precision and recall are illustrated in the figure above. On the left is an illustration of poor search performance. On the right is an illustration of idealized search performance. Note the legend and the formula for calculating precision and recall.

## the results

Table 3 provides the test results of both the Boolean and Thematix technologies on the word sense disambiguation task.

table 3: boolean search vs thematix

Word	Boolean Precision	Boolean Recall	Thematix Precision	Thematix Recall	Delta
board	13%	100%	99%	100%	762%
company	20%	100%	97%	98%	475%
lie	13%	100%	98%	98%	754%
state	33%	100%	96%	98%	291%

The Boolean search technology produced desired (precise) results 20% or 33% of the time, depending on which word was used. The effect of that performance on productivity is 67-80% of search time is wasted, which explains how it is possible that 2.5 hrs of a knowledge worker's day is spent on information retrieval. Corporations suffer both the loss of productivity and the opportunity cost of the lost time.

By contrast, as Table 3 indicates, Thematix produces results that approach complete precision and recall. For the word *board*, the results indicate an improvement of 762% over conventional Boolean search technology, as well as tremendous improvements for all words indicated in the last column of Table 3.

## conclusions

This white paper presented a test methodology and test results that suggest much of the \$2.5 billion incurred annually by Global 1000 corporations is wasted due inefficient and ineffective search technologies, and that a productivity gain of up to 80%, which translates into \$2 billion, can be realized by switching to a second generation search technology like Thematix, Microlanguage's intelligent language processing technology.