



whitepaper

# microlanguage™ thematix™

overcoming the challenges inherent in language with  
intelligent language processing

## abstract

Language presents a set of challenges to the search for relevant information that surpass the capabilities of conventional search technology. Microlanguage's next-generation technology, Thematix, is specifically designed to overcome the challenges that language presents by processing language intelligently. This paper discusses the challenges inherent in language, explains why conventional search technology does not overcome them, and describes how Thematix not only meets the challenges but goes far beyond them in returning relevant, usable information.

## table of contents

finding relevant information .....	3
challenges to finding relevant information .....	5
one form, many meanings.....	5
one meaning, many forms.....	6
one meaning, no forms .....	6
conventional search technology .....	7
term-based technology .....	7
term weighting.....	8
term expansion .....	8
thematix: turning information into knowledge.....	9
overcoming the MULTIPLE MEANINGS CHALLENGE: access to surrounding context .....	10
overcoming the MULTIPLE FORMS CHALLENGE: concept extraction .....	10
overcoming the NO FORMS CHALLENGE: deriving information from unstructured text .....	11
conclusion .....	12

## table of figures

poor vs. ideal search performance.....	4
--	---

## about microlanguage

Microlanguage develops and markets next-generation intelligent language processing technology and services to OEMs, systems integrators, and portal information providers. Our technology enhances the value of our partners' products and services to global 2000 technology, finance, healthcare, pharmaceutical, manufacturing, and national security domains. 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.

## overcoming the challenges inherent in language

### finding relevant information

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.

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

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

In this paper we identify the inherent challenges presented by language, and show why conventional search technology can not overcome them. We then answer the more compelling question of how lost knowledge-worker productivity can be reclaimed by using a next-generation technology that can process language intelligently: Microlanguage's Thematix.

**where conventional search technology fails** The most important measure of the performance of any search technology is the difference between the information desired and the information returned. To put it plainly, it is the difference between what you want and what you get. Conventional search technology fails in both of these areas:

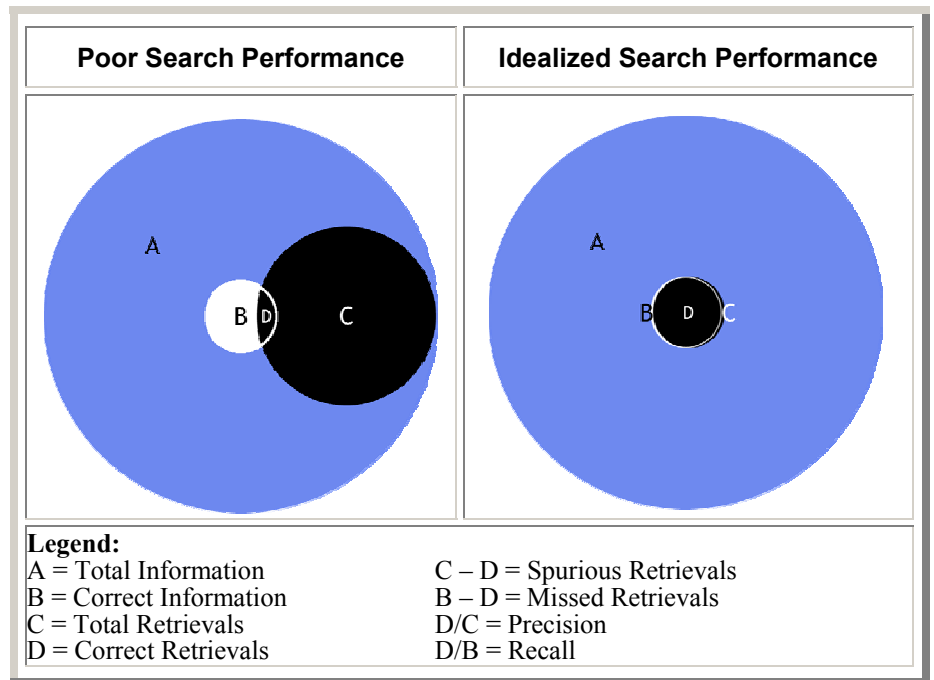
- MISSED information: It fails to return relevant information.
- SPURIOUS information: It returns irrelevant information.

These two factors account for productivity lost by knowledge workers using conventional search technology. The following figure illustrates more clearly the impact of missed and spurious information.

---

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

## poor vs. ideal search performance



For each side of the figure, the outer circle A represents all of the information assets available. The inner circle B represents all of the relevant information that the knowledge worker is after. The circle C represents the information returned by a search. The relationship between C and D is where performance is measured. For poor performance, the search misses much of the relevant information, and returns lots of irrelevant (i.e. SPURIOUS) information. With such results, the knowledge worker must spend time wading through all of the information returned in search of the relevant information. Moreover, a new search must be initiated in an attempt to gather the missed information, or worse yet, the knowledge worker has no idea how much relevant information was missed.

**the truth, the whole truth, and nothing but the truth** To clarify the relationship of these sectors to each other, we relate them to a common phrase: “the truth, the whole truth, and nothing but the truth.”

- Since circle B represents the total relevant information, we refer to B as the truth.
- The whole truth is the totality of the desired information, i.e. all the relevant information that the knowledge worker has in mind when instigating a search. The standard term for this relationship between the desired information available and the desired information returned is recall. If the search returns all of the desired information, then the searcher has the whole truth.

- Furthermore, the knowledge worker wants the search to return only that information that is relevant and not undesired information that must be waded through to find the relevant information, i.e. NOTHING BUT THE TRUTH. The standard metric that expresses the relationship between desired information available and the undesired information returned is PRECISION. If the search returns only relevant information, then the searcher has nothing but the truth.

Ideal performance is when the retrieved information is both “the whole truth” and “nothing but the truth,” or in standard terms, good precision and good recall. Even if circle B were entirely included in circle C, so that all of the relevant information was returned, circle C would still contain a large amount of irrelevant information that must be waded through to find the relevant.

Clearly what is needed is a next generation technology that delivers high performance in both of these areas.

## challenges to finding relevant information

There are two inherent features of language that virtually guarantee that conventional search technology will always deliver poor performance. If we assume a high-level perspective, language is the intersection of form and meaning. A “form” in language is nothing more than a word or group of words, and forms in language have meaning. While this fact may seem trivial, the difference between form and meaning is the basis of the inherent challenges presented by language:

- A single form can have many different meanings.
- A single meaning can be expressed by many forms.

We describe these challenges in the following sections and the productivity penalty for not meeting each challenge in terms of their impact on precision and recall.

## one form, many meanings

One word (or group of words together) can have many different meanings. For example, the word *will* has over twenty different meanings listed in the typical dictionary, including:

- volition, as in *He overcame all obstacles by an act of **will**.*
- the future-tense verb, as in *I **will** see you later.*
- a legal document, as in *They read his last **will** and testament.*
- a short form of the name *William*.

The challenge presented by multiple meanings being attached to one form is the MULTIPLE MEANINGS CHALLENGE. For example, a search for last wills and testaments will return many documents unrelated to last wills unless the search technology can differentiate the many meanings

that this one word can have. The penalty for not overcoming this challenge is a loss of precision, that is, a return of irrelevant information due to the fact that the searcher has only one meaning of a particular word in mind, but the search returns all instances of the word with all its meanings. Overcoming this challenge requires a technology that can differentiate the different meanings of a single word, e.g. that can recognize the difference between *will* as a future-tense verb and *will* as a legal document, so that a search for wills does not return all documents with the word *will* in them.

## one meaning, many forms

On the other hand, one meaning or concept can take many forms. Consider the number of different ways there are to refer to a single date in history:

*Independence Day, Seventeen Seventy Six*

*July 4th 1776*

*7-4-1776*

There is only one date in history being referred to in these examples, but there are many ways to express this date. The challenge presented by multiple forms having the same meaning is the MULTIPLE FORMS CHALLENGE. The multiple forms challenge is potentially the more difficult one to overcome because there are typically many ways to express a given meaning, whereas with the multiple meanings challenge, a single word typically has no more than 2-5 meanings on average.

The penalty for not meeting the multiple forms challenge is a loss of recall, that is, relevant information not being returned. Relevant information is missed in this case because all the ways to say something can not be anticipated. Think of how many ways there are to refer to a stock losing value: the stock *fell*, *tanked*, *plummeted*, *dropped off*, etc. The searcher interested in information on all the stocks that lost value at a particular time should not have to anticipate every way there is to say that a stock lost value. If this is not done, though, the searcher runs the risk of not returning all of the relevant information.

## one meaning, no forms

A special problem arises when certain information is desired from the text, but there are no forms that contain that information directly. As an example, consider:

- Any information in one standard of measure but the text contains the information in another standard, e.g. information is desired in feet (English system) but all textual references are in meters (metric system). [Also, enzyme nomenclature]
- All information about a particular geographical region, when the text only contains lat-long coordinates.
- All information about buildings over a particular square footage, but the text only contained lengths and widths given in meters.
- Search for all of the board meetings that occurred in a range of dates.

The challenge of deriving information where no form exists is the NO FORMS CHALLENGE. The penalty for not being able to derive information is poor recall.

## conventional search technology

Given the three challenges outlined in the last section, this section describes conventional search technology and explains why it does not meet those challenges. The first two challenges and the penalties for not meeting them can be summarized as follows:

**multiple meanings = poor precision**

**multiple forms = poor recall**

In a nutshell, conventional search technology only operates on word forms divorced from their meaning, and so these technologies can only ever meet half of the challenge.

As for the derived meaning challenge, it is so far beyond the capability of conventional search technology that it will not be discussed in this section. We will return to this challenge when we discuss Thematix and how Thematix overcomes it.

## term-based technology

The most widespread technology behind knowledge management systems is term-based technology: the relevance of a particular document to a search is determined by the words (or “terms”) it contains. If the search is Boolean-enabled, then several words can be strung together with “or” and “and” operators. But note that the relationship between words is not considered, so that the order of words in the document has no bearing on which documents are retrieved. In fact, all the words in the document could be sorted alphabetically with no change in the retrieval results.

Because the structure of sentences is not considered by term-based technology, these approaches are often called BAG OF WORDS approaches. The obvious limitation of a term-based approach is that it can only find particular forms. This technology can not meet the

multiple forms challenge unless all the relevant forms are entered into the search every time. Nor can they meet the multiple meanings challenge because all instances of each word are returned with the search results. Microlanguage has found that the performance of term-based technology looks very much like the [left side of the figure](#), i.e. poor performance.

Because the performance of strict term-based technology is so poor, improvements have been implemented in an attempt to boost performance. These improvements take one of two forms: term weighting and term expansion.

## term weighting

With a strict term-based technology, no level of importance can be assigned to any term; all terms are treated equally, which misses the important fact that some terms are more relevant than others. Attempts to boost the precision of term-based technology take the approach of assigning different weights to every word so that the more relevant a word is to a search, the more weight it has.

The assigned weights are used to calculate the similarity of the terms in the query with those in any given document. Documents are returned based on the order of their similarity, with the most similar documents being presented first.

There are a variety of ways to give weights to terms, from an approach as simple as assigning weight based on how many times a term occurs both in a single document and across the entire collection of documents, to more complex methods of calculating the weight according to probability of relevance, for example with a Bayesian or neural net approach.

Whatever the method, the goal is the same: to make it the case that some terms are treated as more important than others according to the weight given them. While performance can be improved with term weighting, the method still relies on brute-force manipulation of the word form without its meanings and so is ultimately going to fail the multiple meanings challenge and return poor precision as a result.

## term expansion

Attempts to boost the recall of term-based technology take the approach of expanding the terms of the query. There are two ways to go about expanding the terms: (i) adding other forms of the word, i.e. form expansion, and (ii) adding other words with a related meaning, i.e. meaning expansion.

**form expansion** Also known as STEMMING, this approach recognizes that a single word can change according to the endings put on it. For example, the verb *meet* has the forms *meet*, *meets*, *meeting*, and *met*.

There is an open question as to whether the different forms of a word should be given their own weights or whether they should be combined into a single term, as in the example of *meet*. The open question is whether the different forms of a word are actually related to the term—sometimes they are, and sometimes they are not. For example, a query on a web-based search engine about stocks or stock prices will be reduced to the single term *stock*. However, for sites that contain extensive information about stockings, that term will also be reduced to *stock*, and so a query about stock prices will return many unintended hits about stockings.

While form expansion has the potential to improve recall, it must operate intelligently by differentiating the meanings of the expanded forms in order to avoid returning spurious results, as in our *stock* example. Because term-based technology works strictly on forms, it can not differentiate meanings and so the improved recall ultimately comes at the price of poorer precision.

**meaning expansion** Add terms that are considered similar to the terms in the query based on a number of correlations. The most common relationship is words having a similar meaning. Other relationships may be used as well, such as *is a kind of*, or *is a part of*, among others.

Consider the following real-world example. An intelligence analyst in the national security domain searches for all recent reports of hijackings from a news source. A typical semantic expansion would include the term “piracy” in the query along with “hijacking,” because piracy is a kind of hijacking. This inclusion opens the gate for any article with the word “piracy” in it to be returned along with the search for hijacking. Recently, most articles on the Napster controversy used some form of the word “piracy” in it, and so the search for recent hijackings returns most of the articles on Napster, which must be sorted through in order to get to the ones on hijacking.

Meaning expansion has the potential to improve recall, but it has the same shortcoming as form expansion: it can not differentiate the meanings of the expanded forms and thus improved recall again comes at the price of poorer precision.

## thematix: turning information into knowledge

Thematix is an intelligent language processing technology that can do what no conventional search technology can:

- It overcomes the multiple meanings challenge by distinguishing between the different meanings of words and groups of words together.
- It overcomes the multiple forms challenge by correctly identifying concepts in a variety of forms.
- It overcomes the no forms challenge by deriving high-level meaning from unstructured text.

## overcoming the MULTIPLE MEANINGS CHALLENGE: access to surrounding context

The meaning of a word or group of words together depends on everything in the surrounding text, or context. Conventional search technologies are restricted to processing one word at a time. What is worse, some systems that use this technology discard relevant symbols such as capitalization, punctuation, and special characters in order to arrive at a “normalized” text. A normalized text can be desirable from the perspective of term-based technology, but an intelligent technology mines all of the surrounding context to arrive at the meaning in the text and therefore needs access to all of the context.

Consider a seemingly simple example like punctuation. By changing just a few punctuation marks, the meaning of a sentence can be changed:

*Woman, without her man, is nothing.*

*Woman! Without her, man is nothing.*

Similarly with capitalization, the precision of a search suffers if the search is not case-sensitive, for example *US* is an abbreviation for *United States*, but *us* is a pronoun.

It is just these sorts of textual cues that determine the meaning of a word or group of words. Thematrix not only retains all of the symbols that contribute to meaning such as punctuation, capitalization, and special characters, but it can look into the surrounding context of a word to determine its meaning so that only relevant information is returned. In this way, Thematrix is able to achieve a very high level of precision. See the Microlanguage whitepaper *Intelligent Language Processing vs. Conventional Search Technology* for further discussion of the near-complete precision returned by Thematrix as opposed to the poor precision of conventional technology.

## overcoming the MULTIPLE FORMS CHALLENGE: concept extraction

Thematrix is able to intelligently search for concepts even though they may reside in a variety of forms. Suppose the knowledge worker from

the previous example on hijacking uses Thematix instead of the standard technology. Thematix is intelligent enough to find the following.

- Forms of the verb *hijack*:  
*hijack, hijacks, hijacked, hijacking, hijackings.*
- Related concepts such as *piracy* and all of its forms, but not the meanings associated with software or computer piracy.
- Reports of hijackings that do not contain any form of the word *hijack*, e.g. the phrase *taken over by force*.

See the Microlanguage whitepaper *Intelligent Language Processing vs. Conventional Search Technology* for further discussion of the near-complete recall returned by Thematix.

## overcoming the NO FORMS CHALLENGE: deriving information from unstructured text

Thematix has an important feature that very few systems of any sort have: the ability to extract new information that does not exist in any one place in the text. That is, Thematix can derive meaning from a text in which the desired meaning does not reside in any particular form. In this way, Thematix is the bridge between unstructured text and structured, relevant information. Following are some examples of how this technology can be used.

**output database structures from free text** For example, consider the parts that make up the contact information for a person:

- name
- street number
- street name
- street identifier (street, avenue, boulevard, etc., including all possible abbreviations)
- city
- state
- zip code
- e-mail address
- one or more phone numbers.

When this information is returned, Thematix can output each part as a field in a database, all of which combine to create a database record.

**output one format from varying formats** In some domains, there are very precise ways to refer to things but the standards for doing so can vary. That is, one thing can go by several different names, and so the multiple forms challenge must be met. For any domain in which something can go by several different names, Thematix can

recognize the different names so that all of the relevant information is returned without the searcher having to keep in mind all of the ways there are to refer to the information desired. Furthermore, Thematix can return all of the information in one standard, so that the searcher need have only one standard in mind when doing the search, and sees all the returned information in that same standard.

Consider an example from bioinformatics. There are several standards for naming enzymes, i.e. one enzyme can go by several different names depending on the standards used. Thematix is able to recognize the different ways there are to refer to a particular enzyme so that the researcher does not have to. What is more, Thematix can return all the information according to one standard so the researcher does not have to keep track of different standards and can focus on using the relevant information.

### **output mathematically-derived values from input values**

To take a simple example, say that a buildings inspector wants to find information on all the buildings over a certain number of square feet from a collection of documents containing information about buildings.

**output absolute values from relative references** Say that a knowledge worker in the military intelligence domain wants to find all of the information about a particular person in a geographical region at a particular time in absolute values such as latitude and longitude, but all of the information in the documents is unstructured language about

## conclusion

**conventional search technology** This technology can not meet the three challenges that language presents:

- It fails the multiple forms challenge, in which one concept takes different forms: conventional search technology has no knowledge of how many different ways a single concept can be expressed. Because it fails this challenge, this technology fails to return all of the relevant information.
- It fails the multiple meanings challenge, in which one word or group of words has several different meanings: conventional search technology can only recognize word forms divorced from their meaning. Because it fails this challenge, this technology returns information that is not relevant, which wastes time and the costs associated with it because the information must be waded through
- It fails the no forms challenge, in which information exists in unstructured form from which new information must be derived and presented in a standard format: conventional search technology can not begin to derive information that is not obvious in the text, much less manipulate it and present it in a standard format.

**thematix** Microlanguage's next-generation intelligent language processing technology, Thematix, is specifically designed to meet and overcome these challenges:

- It overcomes the many forms challenge by using knowledge about all the ways that a single concept can be expressed and finding all of those expressions in the text, without confusing
- It overcomes the many meanings challenge by reading words and groups of words in context to determine their correct meaning so that only relevant information is returned.
- It overcomes the no forms challenge by going several steps beyond any conventional technology and deriving new information from unstructured text that does not directly indicate that information and presenting it in a standard format.