



## Cognition Technology Resources Overview Semantic Map, System Architecture and Tools

Kathleen Dahlgren, Ph.D., Chief Technical Officer

Dan Albro, Ph.D., Chief Scientist

### I. Uses of Cognition's Semantic Natural Language Processing (NLP)

Traditionally, increased recall for Search technologies was expected to come at the expense of precision. Cognition introduces semantics to achieve simultaneously higher precision and recall.

Cognition's Semantic NLP™ technology operates at the level of the **word sense** rather than the word. As an example, instead of reasoning with the word "strike," Cognition's Semantic NLP reasons with a single meaning of that word within the context it was used. So, in the phrase "strike on the head," "strike" has one meaning, and in "worker's strike," it has another meaning. An individual word meaning, not the word itself, is stored in the Cognition's Semantic NLP taxonomy and mapped in the meaning thesaurus to the same meaning of other words, such as meanings of "hit" and "beat" that mean the same thing as "strike" in "strike on the head." The word meanings are determined, in part, by parsing, and in part by using the over 4.3 million sense contexts. Parsing technology can further increase performance by recovering argument structure, in other words, "who did what to whom".

Recall is improved in Cognition's semantics-based system because it is extended to include different words with the same meaning as the query terms. In the example above, retrievals for a query on "strike" meaning "hit" are expanded to include text with the words "hit" and "beat" if also used with the meaning "hit." A related source of improved recall is **taxonomic reasoning**. As an example, reference to the transportation sense of "vehicle" is expanded to all of the kinds of vehicles down the ontology, such as "car1," "SUV1," "Ford2," or "plane1," "jet1," "Boeing-747," etc. (Note: final digits are sense numbers). Since these relationships are based on word meanings, not the words themselves, precision is also enhanced. A query on the "hit" sense of "strike" will not retrieve to irrelevant text on the musical sense of "beat," nor will a query on the transportation sense of "vehicle" retrieve to irrelevant text on the carpentry sense of "plane."

Cognition's Semantic NLP employs additional technologies to improve recall and precision. By processing **derivational and inflectional morphologies**, all variant forms of a word (e.g., "initialize," re-initialize," "initialization," etc.) can be recognized as expressing the same concept. Millions of forms are mapped in this way. Also, acronyms are synonymous with their spell-out phrases, such as "SEC" and "Securities and Exchange Commission"; and synographs, or alternate spellings of the same word (e.g., "center" and "centre"), are similarly recognized.

Another source of precision enhancement is **phrasal reasoning**. Phrases such as “The Bill of Rights” are treated as atoms, so that there is no confusion with documents that happen to use the words “bill” and “right” in other contexts.

In summary, the Cognition’s Semantic NLP use of semantics enables NLP technology to **reason with meaning** rather than word patterns, and adds a valuable additional source of meaning to a parser-based technology.

## II. Cognition’s Semantic Map

Cognition Technologies’ lexical resources encode a wealth of morphological, syntactic and semantic information about the words of the English language and their relationships to each other. These resources were created and reviewed by lexicographers and linguists over a span of twenty-four years.

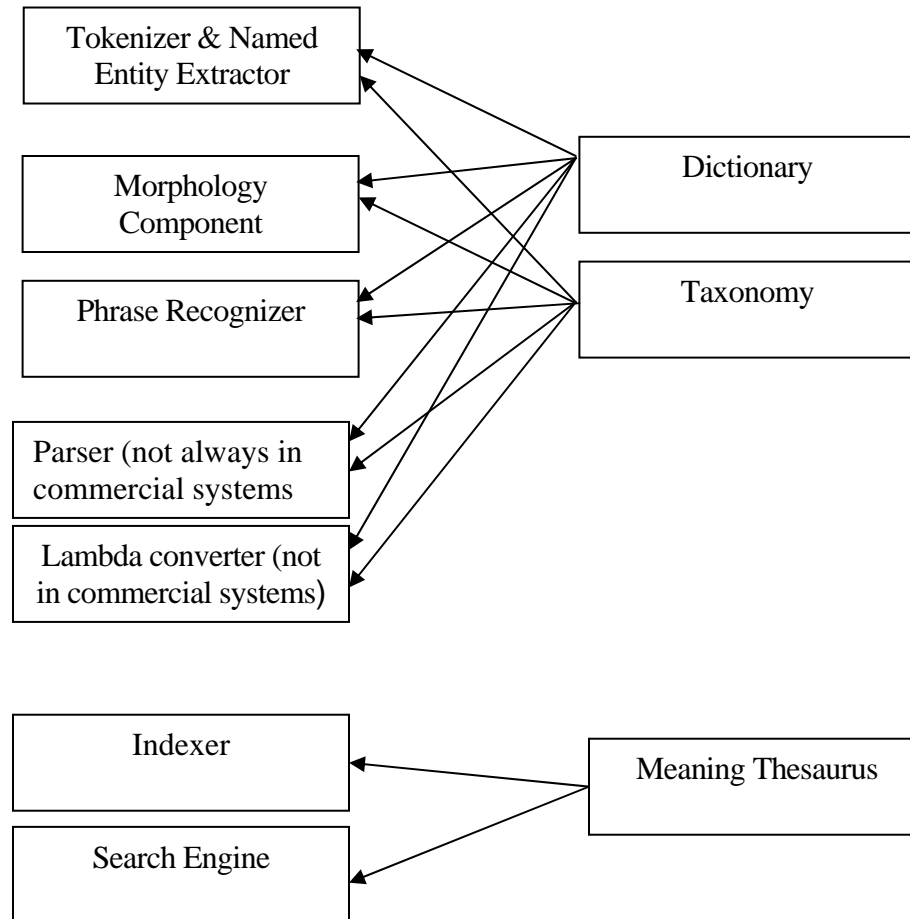
- **Word Stems** – [506,000]: Each word (including phrases and acronyms) in the Cognition lexicon is stored in a base (inflected) form.
- **Word Senses** – [536,000]: Key meanings of semantically ambiguous terms (such as, “strike” meaning “hit” versus “strike” meaning “labor dispute”) are stored as individual sub-entries within the word’s lexical entry. Importantly, each sub-entry may have distinct morphological, syntactic and semantic features; and distinct ontology/synonymy relationships in the semantic map. Different meanings are distinguished, but the same meaning may have various parts of speech (as in “love”, which can be both a noun and verb in the same meaning).
- **Taxonomy** – [7,500 nodes with 536,000 leaves]: All word meanings (senses) are placed in one or more positions in a semantic ontology. This allows Cognition’s Semantic NLP to reason from the general to the specific (e.g. knowing that one meaning of “tank” is a type of “container”) and plays a significant role in the technology’s syntactic and semantic features.
- **Meaning Thesaurus** – [75,000 groupings]: Word meanings with a fair degree of semantic equivalence are associated with each other (e.g. associating one sense of “car” with “automobile”). These relationships include synonymy but may go across syntactic categories (e.g. associating conceptually related nouns, adjectives and verbs, etc.). The parts of speech are marked, so that if it is desired to restrict a paraphrase relationship to a given part of speech, that can be done.
- **Sense Contexts** – [4,300,000 contexts for disambiguation of 17,000 ambiguous word stems]: Terms that may co-occur with ambiguous stems and aid in their disambiguation are stored for each sense of such stems.
- **Morphology Features** – [199 patterns]: Syntactic categories with regular and irregular inflectional and derivational morphology are encoded for each sense in the lexicon or

identified by the morphology processor. This allows Cognition's Semantic NLP to recognize tens of millions of word forms and associate them with their appropriate stems, as in "babies"- "baby", "re-run", "run", etc.

- **Syntax Features: [3,246 patterns]:** Syntax features spell out the syntactic sub-categorization frames for words, 3,215,335 morphological and syntax features are encoded in the lexicon.
- **Selectional Restrictions** – [45,812 encodings]: Ontological restrictions on arguments (e.g. that "vehicles" are the typical objects of one sense of the verb "drive") are stored in the sense entries for words that take arguments.
- **Acronyms** – [19,122]: Acronyms are stored with their spell-outs. Each acronym sense may have many spell-outs. Different spell-outs for ambiguous acronyms are encoded in separate senses.
- **Phrases** – [191,000]: Multi-word expressions are stored with their own lexical features and semantic relationships. To the extent that they are compositional, the particular senses of the individual words in each phrase may be indicated. (Note: the Cognition's Semantic NLP "reader" module recognizes and regularizes additional phrases, such as names, dates, phone numbers, etc., that may not be stored in the lexicon, yielding an indefinite number of phrases that can be recognized.)
- **Synographs** – [17,000]: Common and/or dialect-dependent alternate spellings are stored for word stems.
- **Domains** – [17 domains, 581,110 encodings]: Individual word senses are marked as preferred for particular topic domains, such as "legal" or "military", etc. The number of domains is unrestricted so that more may be added as desired. Domain markings indicate that a given sense should be preferred over others for documents in the domain.
- **Naive Semantic Features** – [Approximately 50 feature types, 540,684 encodings]: A variety of commonsense knowledge, such as "cats have tails", "hands have five fingers", "the function of a chair is sitting", "the consequence of buying X is owning X", etc., may be stored with individual word senses.

### **III. Cognition's Semantic NLP Architecture**

The primary components of the Cognition's Semantic NLP system are outlined below. The sentence parser is commercially available, but not active on the current Cognition Websites. A final component, the lambda expression generator, can be demonstrated, but is not yet commercially available.



- **Tokenizer & Named Entity Extractor:** The Tokenizer breaks the text or query into document sections, sentences, and words. The Tokenizer is also responsible for converting character codings, and other such tasks. The Named Entity Extractor recognizes and regularizes a variety of common patterns, such as human name strings, dates, citations, telephone numbers, etc.
- **Morphology Component:** The Morphology Component is more in-depth than a typical stemmer. In addition to removing prefixes and suffixes, it determines the derivation of word forms and computes derived syntactic and semantic features. It enables Cognition’s Semantic NLP to recognize many millions of word forms. The Morphology Component handles regular and irregular inflectional and derivational morphology. A few examples are provided below:
  - Nouns with irregular plural morphology, such as “mouse-mice,” “tooth-teeth”
  - Nouns with regular changes in the plural, such as “baby-babies”
  - Regular verb inflections, such as “raze – razed – razing” and “ship – shipped – shipping”
  - Verbs with irregular past tense forms, such as “catch – caught – catching”

- Regular derived forms:

inter-	communicate	intercommunicate
-tion	communicate	communication
inter- + -tion	communicate	intercommunication
-ize	actual	actualize
re-	marry	remarry

- **Phrase Recognizer:** The Phrase Recognizer combines words into phrases for more accurate interpretation.
- Compounds - the Phrase Recognizer interprets lexical phrases, such as “movie set”, “heart attack”, “net revenue”, etc.
- Idioms - the Phrase Recognizer pieces together idioms, including morphological variants of them (which are linked to other word meanings in the meaning thesaurus).
- **Word Meaning Interpreter & Parser:** The Word Meaning Interpreter primarily uses word collocation information to determine the meanings of words within context, but is also informed by the preceding components and other considerations, such as topic domain. The Parser (application determined via parameter settings) uses syntactic structure and sub-categorization and selectional restrictions to aid in disambiguation.
- **Lambda Expression Generator:** This system module, not yet commercially available, applies formal semantic rules to the output of the syntactic parser to generate lambda expressions for each processed phrase. The output of this module is an expression that indicates argument structure for the sentence, that is, what entities are being talked about, and what they did to each other.

#### IV. Lexicography Tools

- **Lexicon Tool**

Entries in the dictionary include single words, multiple-word phrases and ontological nodes. Each entry has one or more senses (meanings) with associated morphological, syntactic and semantic features.

The Lexicon Tool is the primary means for editing the Cognition’s Semantic NLP dictionary. The tool runs as a server and guarantees that no word is edited by more than one lexicographer at a time. It also makes changes to words available to all lexicographers immediately after they have been saved into the database. The tool allows users to enter, edit or delete words and the primary features listed below; and includes a few additional features, such as the ability to view sections of the ontology.

- **Plain English definition:** This is what is displayed to the user in the Search interface if the user chooses to view the Search concepts.
  - **Ontological attachment(s):** Every sense is attached to one or more nodes in the ontology. For example, the first sense of “dog” is attached to “pet\_node” and “canine\_mammal”.
  - **Syntactic features:** There are currently approximately 3,447 unique syntax and morphology features. Each sense has two or more features associated with it. The syntactic features include main category features (noun, verb, etc.), morphological features for classifying the different forms of a word (e.g. “wind,” “winding,” “wound”), and sub-categorization features (such as, intransitive for verbs). For example, the first sense of “dog” has the category feature “noun”, a morphological feature indicating its plural form, and no sub-categorization features.
  - **Semantic features:** Each sense may have semantic features, such as “domain” features (used to prefer a sense in a particular domain) and selectional restrictions (for use with a parser). Selection features help guide word sense disambiguation. For example, the verb “charge” in the meaning “indict” requires a sentient object and a crime as the oblique object, whereas “charge” in the meaning “electrify” requires an electrical device as object and a form of energy as oblique object. Also, the sense may have naïve semantic or commonsense knowledge features, such as haspart(1,tail) for “cat.”
- **Meaning Thesaurus Tool**

The Meaning Thesaurus Tool enables the lexicographer to look up words in the lexicon, view the meanings, select a meaning and view all of the concept groups that the meaning is a member of. The lexicographer may add to the concept groups, delete from them and create new ones. This can be done in parallel with more than one concept group at a time.

The Tool runs as a server and guarantees that no concept group is edited by more than one lexicographer at a time. It also makes changes available to all lexicographers immediately after they have been saved into the database.

## V. Customer Tools

A number of tools have been developed to enable customers to index their documents, customize their specific jargon such as product names and search their documents. Cognition’s Semantic NLP employs client-server communication for optimal ease of use and efficiency on large document bases. A stand-alone, non-client-server version of Cognition’s Semantic NLP is also available. At the core of the system is the Cognition’s Semantic NLP Server, which can be configured for indexing, searching, or both.

- **Indexing Tool**

The Cognition's Semantic NLP Indexer GUI is the primary interface used to create and index a Cognition's Semantic NLP Project. A Cognition's Semantic NLP Project is simply a list of documents (in any of the formats Cognition's Semantic NLP handles), together with a set of parameters to be used when they are indexed and searched. It is straightforward and easy to use, though as with any software, good results will depend on training (or reference to the user's guide) and practice.

- **Automatic Dictionary expansion**

Cognition's Semantic NLP dictionary can be automatically expanded to include large numbers of customer vocabulary words. Any file of terms and words used in the customer's business along with the categories of the terms can be merged automatically with Cognition's Semantic NLP dictionary. Recently, Cognition's Semantic NLP significantly expanded its vocabulary in medicine and molecular biology by adding over 130,000 words using semi-automated techniques.

- **Customer Dictionary expansion**

The customer may add words in ontological classes if desired. The customer may force Cognition's Semantic NLP into the desired meaning of a word, and the customer may force Cognition's Semantic NLP to consider a word a last name, a first name, or not a name, as desired.

## **VI. Product Features of Cognition's Semantic NLP**

- **Relevance ranking.** Cognition's Semantic NLP returns a list of retrievals containing documents in which the query concepts were found together in a sentence. Those with exact word matches to the query terms come first. The next group is documents in which there were exact matches to some query terms in the body of the document, but other query terms only matched conceptually. Which terms match exactly is indicated. The last group is documents in which there were conceptual matches to all of the query terms. Within each group, documents are listed according to the number of sentences in which all query terms were found.
- **Spelling correction.** In the Search interface, unrecognized words are noted and the user is given a list of alternative spellings to select from. The user may also leave the word as is. If the user desires, Cognition's Semantic NLP will not search on words that are unrecognized.
- **Specific retrievals highlighted.** When the user clicks on the "Highlighted Text" link corresponding to a retrieved document, Cognition's Semantic NLP highlights the relevant section. Additional relevant retrievals within a document are indicated by a pointing-hand

figure at the end of a highlighted section. If there is no pointing-hand figure at the end of a highlighted section, it tells the user that there are no additional relevant results. In other words, it behaves like a clipping service, which is most useful with larger documents.

- **Specific words highlighted.** When the user clicks on the “Highlighted Text” link corresponding to a retrieved document, Cognition’s Semantic NLP also color-codes the specific words which matched the query within the relevant highlight section. As the user hovers the cursor over a matched word, the corresponding query term is indicated. Sometimes a given word may correspond to more than one query term. In this case the word is highlighted according to the first query term matched, but the hover-over text indicates all matched terms.
- **Linguistic Boolean Search.** Cognition’s Semantic NLP searches can be formed using fully-recursive Boolean expressions with AND, OR, WITH, AND NOT, and NOT WITH operators. The expressions connected with the Boolean operators are interpreted for meaning. See the Help function at <http://medline.cognition.com> or <http://wikipedia.cognition.com> for complete instructions and examples of use for Linguistic Booleans.
- **Fuzzy Search.** Cognition’s Semantic NLP searches can be formed using wildcards and fuzzy operators (*e.g.*, “/Liebowicz/n” matches names that sound like “Liebowicz”, and “<dr\*g>i” matches words that start with “dr” and end with “g”, regardless of case) such that proper names and other words can be matched approximately. See the Help function at <http://medline.cognition.com> or <http://wikipedia.cognition.com> for complete instructions and examples of use for fuzzy search.
- **Formats.** Cognition’s Semantic NLP indexes documents in any of the formats supported by Oracle’s Stellant™ product. HTML, XML, OCR'd text and plain ASCII text are indexed natively. Documents in Microsoft Word™, PowerPoint™, RTF, or WordPerfect™ are converted to HTML before being indexed. Some engineering may be required for XML. Documents in PDF are converted to plain text before being indexed. The user may view retrievals in documents converted to HTML or plain text with the specific retrieved sections highlighted, but may also choose to view the original file without highlighting.
- **Customer tags and meta-tags.** Cognition’s Semantic NLP can search in customer-specified tags, if desired, with some engineering assistance from Cognition Technologies.
- **Indexing local directories.** The Cognition’s Semantic NLP Indexer interface permits the user to select individual files or whole directories for indexing.
- **Spider.** Cognition’s Semantic NLP Indexer GUI includes an interface for creating a list of Web files to index. The user enters a URL from which to start, a desired depth to crawl, and parameters to include or exclude particular URLs.

- **Authentication.** The Spider (referenced above) can be directed to use passwords or cookies to enter sites that require authentication, so that the user can index these sites.
- **Languages.** Cognition's Semantic NLP searches in any language handled by Unicode.
- **Search and retrieval pages.** Customizable Search and retrieval ASP pages are provided for the user. The user can make the Search and retrieval look any way desired.
- **Partial updating.** The user may select any number of files to re-index, rather than having to re-index an entire document base when individual documents are added or changed.
- **Automatic updating.** A console-level indexing command is provided so that system administrators can automatically update new files or changed files on a regular basis, initiated by the computer clock.
- **Load balancing.** The Cognition's Semantic NLP indexing interface automatically distributes document indexing across as many servers as the administrator selects.
- **Brokering.** Administrative tools enable the system administrator to manually control indexing and searching load, and membership access to document bases. The tools send queries to servers in response to load and allocate databases to specified servers. Customer-specific criteria that may involve user parameters, such as subscription membership, are supported.
- **Categorization.** The interface queries users for document categories and saves whole retrieval lists or individual files into the categories. New categories can be created on the fly. Subsequent searches can be restricted to categories.
- **User defined ontology.** Users can add ontological classes by editing a standard file. In this way users can define search into classes unknown to Cognition's Semantic NLP, such as company widget names or phrases. For example, an electronics company could add a category "video-recorder" with specific video-recorder names as category members. Using this new category, the user could use the search term "video-recorder" and retrieve to specific video recorder names mentioned in indexed documents.
- **User control of names.** Users can force preference for name or non-name interpretations of words, such as Bush and Stone, which can either be names or common words.