

The Dictionary Parsing Project: Steps Toward a Lexicographer's Workstation

Ken Litkowski

CL Research

ken@clres.com

<http://www.clres.com>

<http://www.clres.com/dppdemo/index.html>

Dictionary Parsing Project

Purpose: to create publicly available semantic networks and ontologies based on parsing dictionary definitions

<http://www.clres.com/dpp.html>

Participants

- **CL Research (Ken Litkowski)**
 - Use of CL Research's DIMAP (DIctionary MAintenance Programs) to maintain dictionaries, parse definitions, and analyze parse results, inventories of semantic relations
- **USC Information Sciences Institute (Eduard Hovy, Bruce Jakeway)**
 - Conversion of raw files, development of Perl scripts, ontology building, chunking data
- **Micra, Inc. (Pat Cassidy)**
 - Preparation of raw data from Webster's Revised Unabridged Dictionary (1913), updating files, importing WordNet data for more current words
- **Franklin Electronic Publishers, Inc. (Ned Irons)**
 - Development of parser, porting to different platforms
- **Many advisors from computational lexicology and lexicography**
 - Experience from previous MRD research, identification of defining patterns for semantic relations

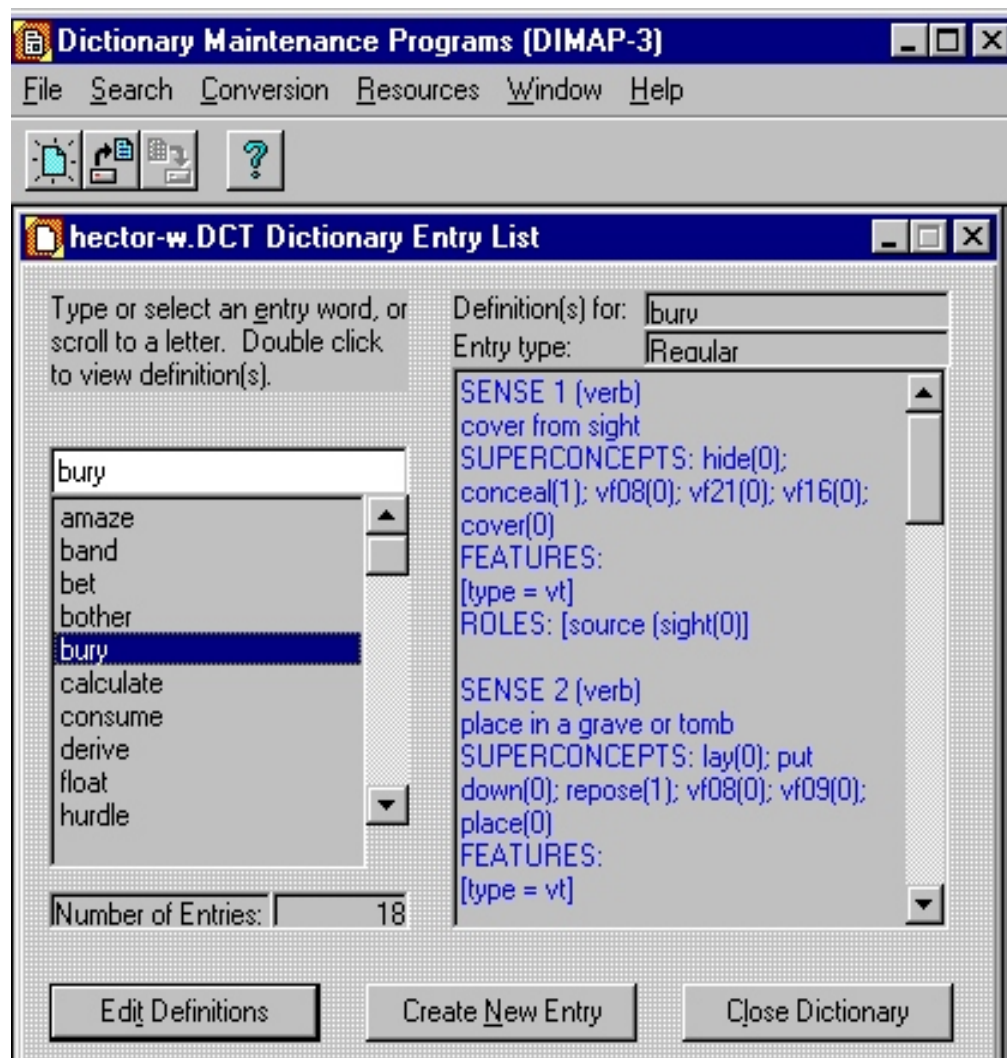
Franklin Sentence Parsing Program

- **ATN-style grammar with 350 productions:**
 - start state, condition for transition, end state (with action functions to produce parse tree nodes and annotations)
 - adds parsing goals dynamically based on subcategorization patterns for lexical entries
 - parsing dictionary based on Oxford Advanced Learner's Dictionary (4th ed.)
- **Parsing dictionary is easily extensible (currently exploring acquisition of lexical properties on the fly)**
- **Parses 400 definitions or 100 sentences per minute on 266 Mhz Pentium II with 64 MB RAM**
- **C source code available upon request (possible GNU GPL)**
 - compiles under Visual C++ 6.0, Borland C++ Builder3, Linux, BSD Unix, Sun4
 - 120 pages of documentation
- **Used by CL Research in Senseval (All-words category)**
 - 68% precision, 67% recall at coarse-grained level (highest)
 - Best overall improvement over baseline among all systems

CL Research DIMAP (Dictionary Maintenance Programs)

- **NLP dictionary creation and maintenance**
 - Usual Windows file functionality, with additional dictionary merging, subdictionary creation, uploading other dictionaries, converting DIMAP dictionaries to LISP, Prolog, or user-defined formats
 - Automatic creation of subdictionaries from integrated MRD or WordNet or identification of open compounds or capitalized phrases
- **Flexible underlying data structure for entries**
 - Multiple senses, each with usual paper dictionary data (definitions, usage notes, etc) and special fields for hypernyms, hyponyms, AV feature structures, semantic roles, semantic interpretation rules
- **Definition analysis functionality**
 - Parsing definitions, comparing definitions across dictionaries, examining inheritance hierarchies, regular expression searches on all fields, identifying semantic primitives using graph-theoretic model
- **Integrated text parsing**

DIMAP Dictionary



DIMAP Dictionary Entry

DIMAP Dictionary Entry [X]

Entry: Code No. Entry Type Sense 1 of 6

Category: Def. No. Label No. Usage Label

Definition:

Usage Note:

Superconcepts: Entry (Sense)

Features: Name = Value

Instances: Entry (Sense)

Roles: Name => Link (Sense)

Semantic Interpretation Rule:

Left Hand Side (Pattern)

Right Hand Side (Logical Form)

Next Sense Previous Sense New Sense Delete OK Cancel Help

Definition Parsing

The screenshot shows a software window titled "Definition Parsing". At the top, there are input fields for "Dictionary Entry:", "Sense Number:", and "Parsed:" (containing the number 0). To the right of these fields are buttons for "Stop Parsing" and "Close". Below the input fields is a vertical stack of buttons: "Get Definition", "Parse Definition", "Parse All Definitions", "Make Selections", and "Start After". To the right of these buttons is a group box labeled "What to Parse" containing four radio buttons: "Defs Only" (selected), "Exs Only", "Defs and Exs", and "Window". Further right are checkboxes for "Identify Semrels" and "Add Semrels", a "Step Mode" checkbox, and a "Resume" button. Below the "What to Parse" group box is a large text area labeled "Definition/Example:". To the right of this text area are buttons for "Set Debug Flags" and a checked checkbox labeled "Debug Flags Set". At the bottom of the window is a large area labeled "Parse Results:" with a vertical scrollbar on the right side.

Definition Parsing Process

- Definitions placed into sentence frames appropriate to the part of speech, with special consideration given to selectional restrictions (usually parenthesized structures), usage notes ("used with *up*"), and specialized wording ("typically", "usually", "to a specified condition")
- Parse results (annotated parse tree) analyzed to identify extract hypernyms, synonyms, and other semantic relations (semrels)
- Use of defining patterns (e.g., manner: in(dpat((~ rep01(det(0)) adj manner(0) sr(manner)))) to identify semrels (hyernym, synonym, instrument, means, location, purpose, source, manner, has-constituents, has-members, is-part-of, locale, and goal)
- Identified semrels are placed in dictionary being parsed, where they are then available for subsequent analysis built into DIMAP functionality
- 400 definitions per minute on a 266 MHz Pentium II with 64 MB RAM

Definition Parsing

Dictionary Entry: Sense Number: Parsed:

☒ Defs Only
 ☐ Exs Only
 ☐ Defs and Exs
 ☐ Window

☒ Identify Semrels
 ☐ Add Semrels
 ☐ Step Mode

Definition/Example:

☒ Debug Flags Set

Parse Results:

```

(SEN
  (PHP (ninfo(6pya)) (aspect(pres-t(F)))
    (SUBJ (ninfo(6pya))
      (pron they(sp(1))))
    (verb stake(tn tn-pr sp(1)))
    (NP (ninfo(3yabmf)) (modifies(2))
      (noun money(cn ucn sp(1))))
    (PRP (modifies(3 2))
      (prep on(sp(1)))
      (NP (dinfo(sing pl ucn)) (ninfo(3yabmf))
        (det the(sp(1)))
        (noun outcome(sp(1)))))
    (PRP (modifies(2 6 3))
      (prep of(sp(1)))
      (NP (dinfo(sing)) (ninfo(3yabmf))
        (det an(sing sp(1)))
        (noun issue(cn dngr singnv ucn sp(1)))))
    (epunct .))
  (bet hyp gamble)
  (bet hyp stake)
  (bet syn money)
  (bet tobj money)

```

Examination of Parsing Results

- **Examination of parsing results to make corrections to parsing system**
 - Identifying parser problems
 - Identifying words unknown to the parser
- **Listing identified semrels**
- **Identifying senses where no semrels were found**
- **Performing consistency analysis against WordNet (e.g., do hypernyms found from parsing match WordNet hypernyms)**
- **Definition comparison (mapping between two dictionaries), using word overlap or componential analysis method (see Litkowski, SIGLEX99)**
- **Analysis of dictionary digraph to identify primitives (based on ISA links)**

Lexicographer Functionality

Definition Parsing

Dictionary Entry: Sense Number: Parsed:

What to Parse
☒ Defs Only
☐ Exs Only
☐ Defs and Exs
☐ Window

☐ Identify Semrels ☐ Step Mode
☐ Add Semrels

Definition/Example:

☒ Debug Flags Set

Parse Results:

Debugging Specifications

☐ Live parses after each word
☐ Live parses after word:
☐ CheckBox3

Printing Preferences (Base Dictionary + Extension)

☐ Parse Trees (*.par)
☐ Unable to parse (*.nop)
☐ Bad Parses (*.bad)
☐ STUB Parses (*.stb)
☐ Parses w/ Unknown Words (*.unk)
☐ Semantic Relations (*.sem)
☐ No Semrels Found (*.nos)
☐ WordNet Differences (*.wna)

Definition Comparison

The screenshot shows a Windows-style dialog box titled "Map Definitions Between Two Dictionaries". It contains several input fields and checkboxes for configuring a word mapping process. The "Map Dictionary:" field is set to "hect-dor" and the "Into Dictionary:" field is set to "hector-w". Under the "Map:" section, the "Selected Word" checkbox is checked and the word "shake" is entered in the adjacent text box. In the "Using:" section, the "Componential Analysis Method" checkbox is checked, while "Word Overlap Method" and "With Stop List" are unchecked. There is also an unchecked "With User Reference:" checkbox with an empty text field next to it. At the bottom right, there are three buttons: "Begin Mapping", "Cancel", and "Close". The "Close" button features a small icon of a window with a red 'X'. A large, empty rectangular area at the bottom of the dialog is labeled "Results:" on the left side.

Map Definitions Between Two Dictionaries

Map Dictionary:

Into Dictionary:

Map:

☐ **All Words**

☒ **Selected Word**

Using:

☐ **Word Overlap Method** ☐ **With Stop List**

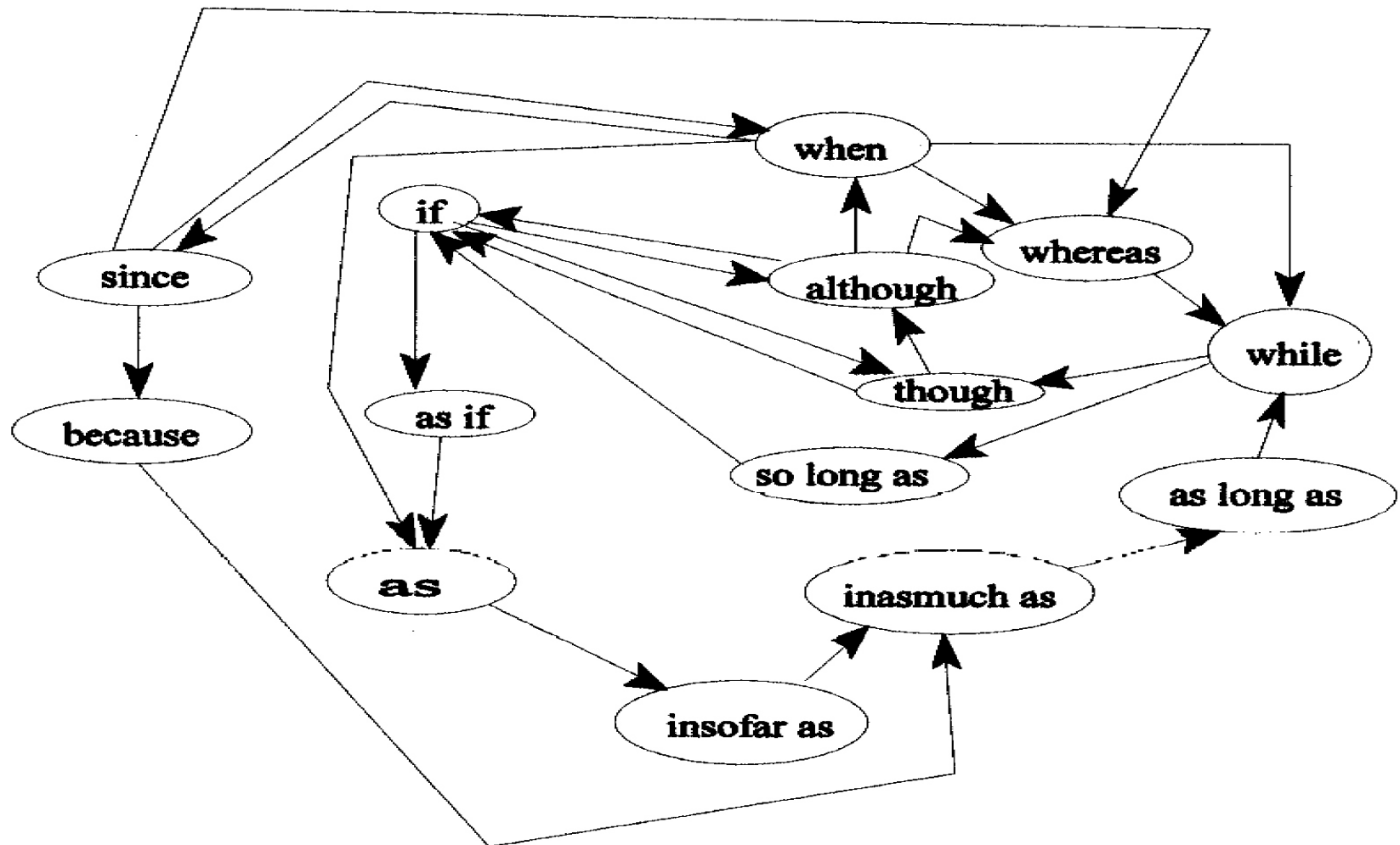
☒ **Componential Analysis Method**

☐ **With User Reference:**

Results:

Digraph of Primitive Subordinating Conjunctions

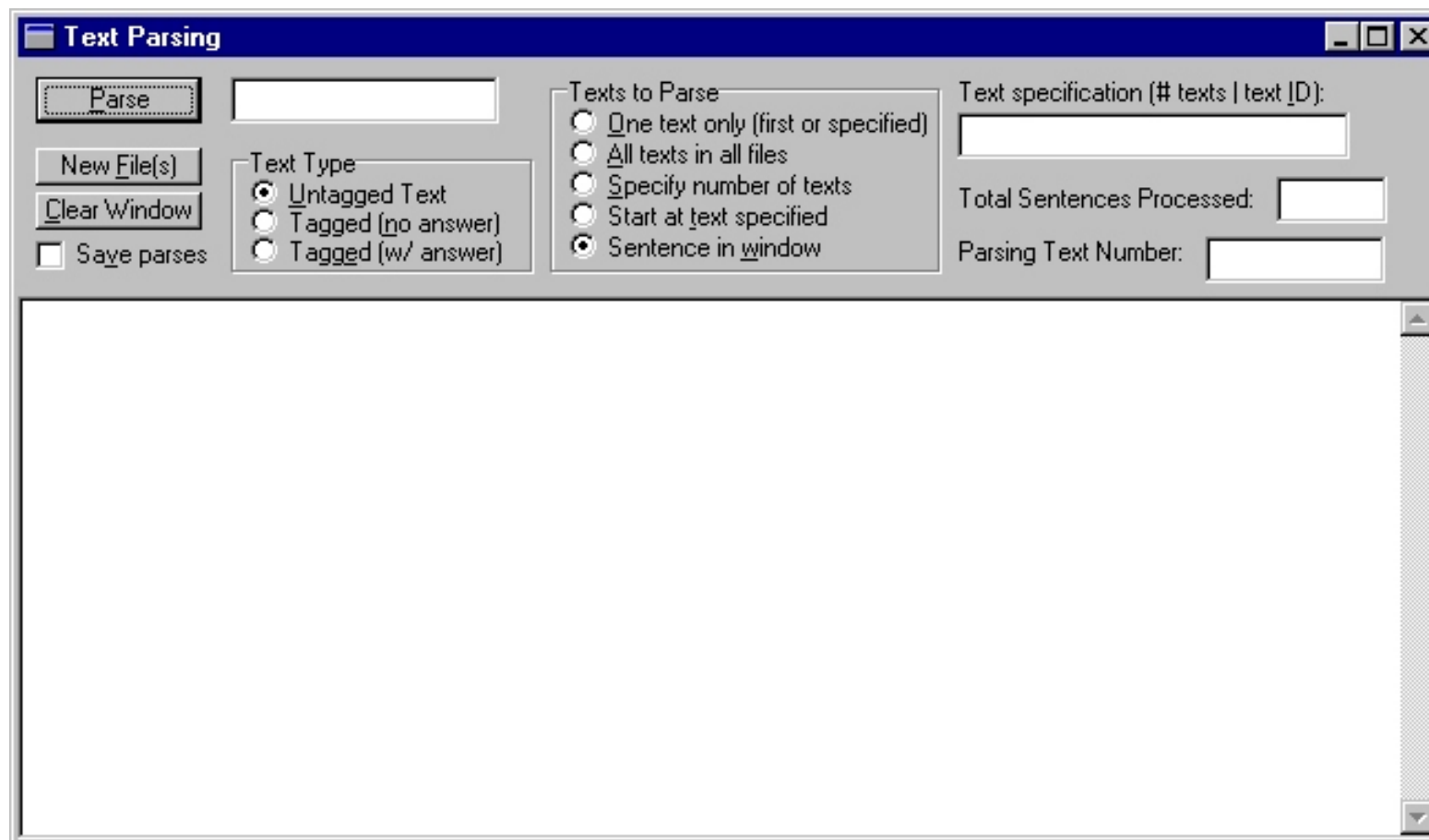
Figure 2: Dominant Subordinating Conjunction Strong Component



Testing Dictionary Entries (Word- Sense Disambiguation)

- **Given a corpus sample, how well does characterization of current sense set (including identified semrels) allow WSD**
- **Use SENSEVAL model (with target words tagged or untagged)**
- **Examine individual sentences or entire corpora**
- **Accompanying scoring program to determine effect of improvements**
- **90 sentences per minute on a 266 MHz Pentium II with 64 MB RAM**

Testing Corpus Instances (Senseval)



DPP Status

- **Parsing definitions to build semantic network (like thesaurus or MindNet) automatically (0.86 semrels per sense, compared to 3.26 for MindNet)**
- **Identifying backbone of hierarchy with genus terms and part-of relations**
- **Filling in details of network with many types of semantic relations, conceptually oriented, including purpose, means, manner, source, destination, locale and location**
- **Ability to map categories, concepts, or definitions between dictionaries and ontologies based on parsing their descriptions**
 - (“if it quacks like a duck, moves like a duck, has the parts of a duck, chances are that it’s a duck”)
- **Inventories of semantic relations (UMLS, WordNet, EuroWordNet, Micra, Wordsmyth, Webster’s 3rd prepositions)**