Application-Driven Definition Parsing

Ken Litkowski
CL Research
ken@clres.com
http://www.clres.com
http://www.clres.com/defpars/index.html

Purpose of Definition and MRD Analysis

- **To exploit fully the result of lexicographical expertise in constructing a dictionary, extracting information for NLP tasks from tagging to discourse analysis**
- **"Many computer scientists take the view that such details are too fine-grained to be of much interest for computing, or that the work has been done once (albeit on flawed sources) and is not worth doing again. It is therefore surprising to hear these same computer scientists complaining about the amount of lexical tuning needed to make a lexicon suitable for a particular application." [Hanks, *Two kinds of computational lexicography*, forthcoming]**

Outline

- **Historical perspective**
- **Using an overall model of the semantic structure of dictionaries as the context for definition parsing**
- **First tasks in definition parsing (assuming a marked-up MRD)**
  - **Coming to grips with the data (selecting what to use and preparing the data for use)**
  - **Preparing definitions for parsing (peculiarities)**
  - **What do you want to obtain from the parsing (importance of application)**
- **Main NLP-related tasks**
  - **Creation of LKB for use in parsing**
  - **Extraction of semantic hierarchies and other semantic relations**
  - **Analysis of definitions (consistency, conflation of entries)**
- **Applications**
  - **Mapping definitions between dictionaries**
  - **Word-sense disambiguation**
  - **Question-answering**

Historical MRD Analysis

- Quillian (late 60s)
- Olney, Webster's 7th (early 70s)
- Amsler (Webster's Pocket Dictionary) (late 70s)
- Evens & Smith (W7) (late 70s)
  - Markowitz, Ahlswede (early 80s)
- Calzolari (early 80s)
- IBM Group (W7, LDOCE) (mid and late 80s)
  - Chodorow, Klavans, Byrd, Boguraev
- Kilgarriff, Krovetz (late 80s, early 90s)
- Copestake (& Briscoe), Vossen, Meyer (at CMU) (early 90s)
- Wilks (at NMSU) (LDOCE)
  - Fass, Guo, Slator, Guthrie
- MS Group (90s)
  - Richardson, Dolan, Vanderwende
- Barriere

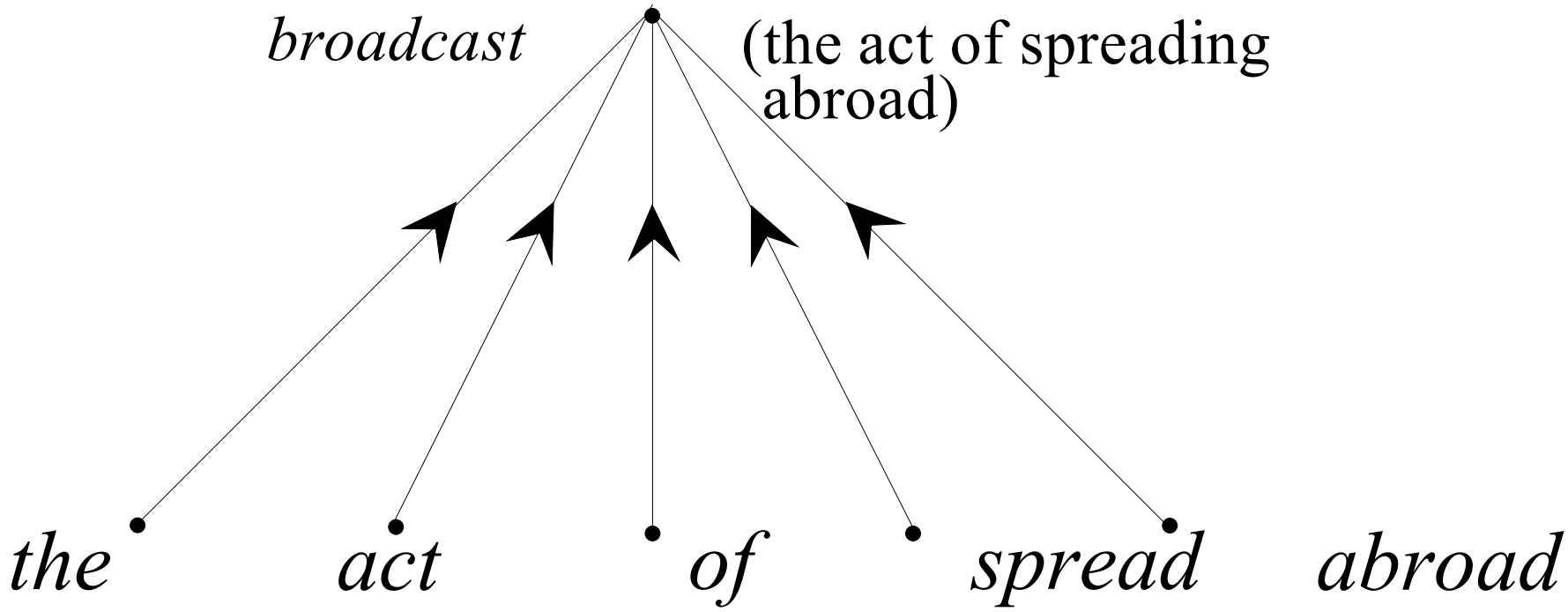
Historical Developments

- **Coping with the data (late 60s to late 70s)**
- **Extracting and building semantic hierarchies (some by hand) (late 70s, early 80s)**
- **More semantic relations, more information (e.g., lexical rules, HPSG structures, EuroWordNet) (90s)**
- **Back to the basics**

Digraph Model of Dictionaries

- **Main entries as nodes and "is-used-to-define" as arcs**
  - **Allows use of results from digraph theory**
    - **Examination of defining paths**
    - **Analysis of strong components ("vicious cycles")**
    - **Identification of "basis" set (primitives)**
  - **Litkowski (AJCL, 1978), Litkowski (ACL, 1980)**
- **Refinement of meaning of nodes and arcs**
  - **Nodes representing concepts**
  - **Arcs representing semantic relations (particularly ISA)**
- **Model prediction**
  - **"Concepts can be lexicalized and verbalized in more than one way" (realization in WordNet)**
  - **Can be used to guide analysis**

Simplest Digraph Model



Example of Vicious Cycle

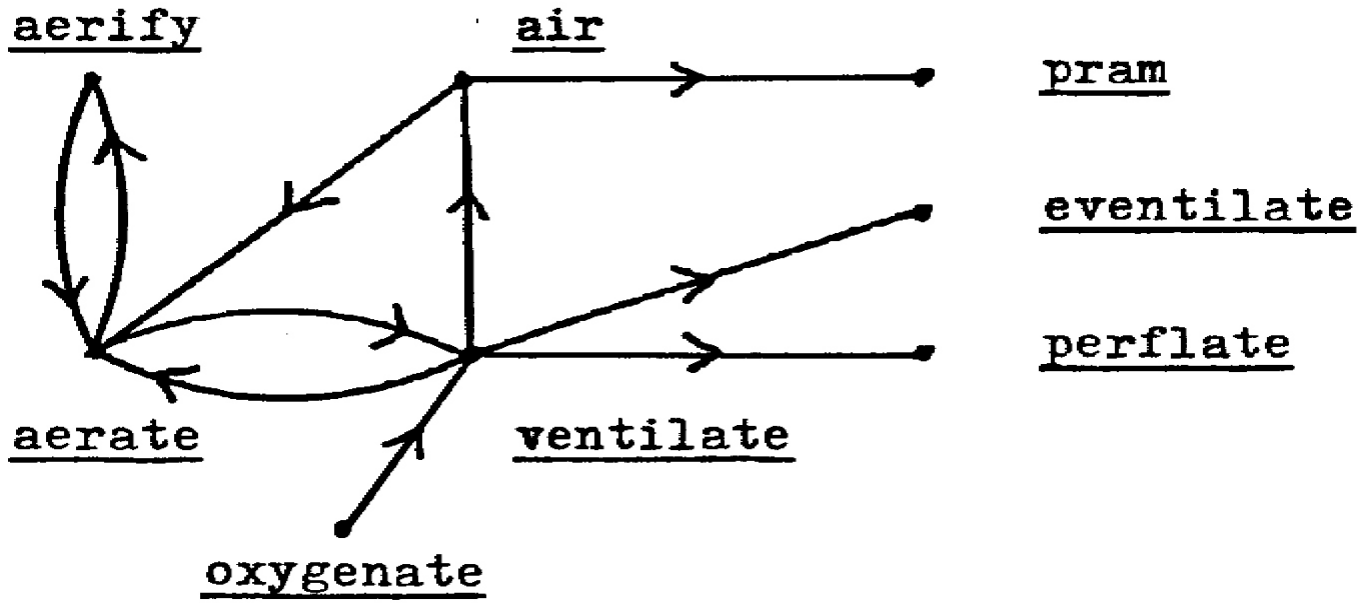
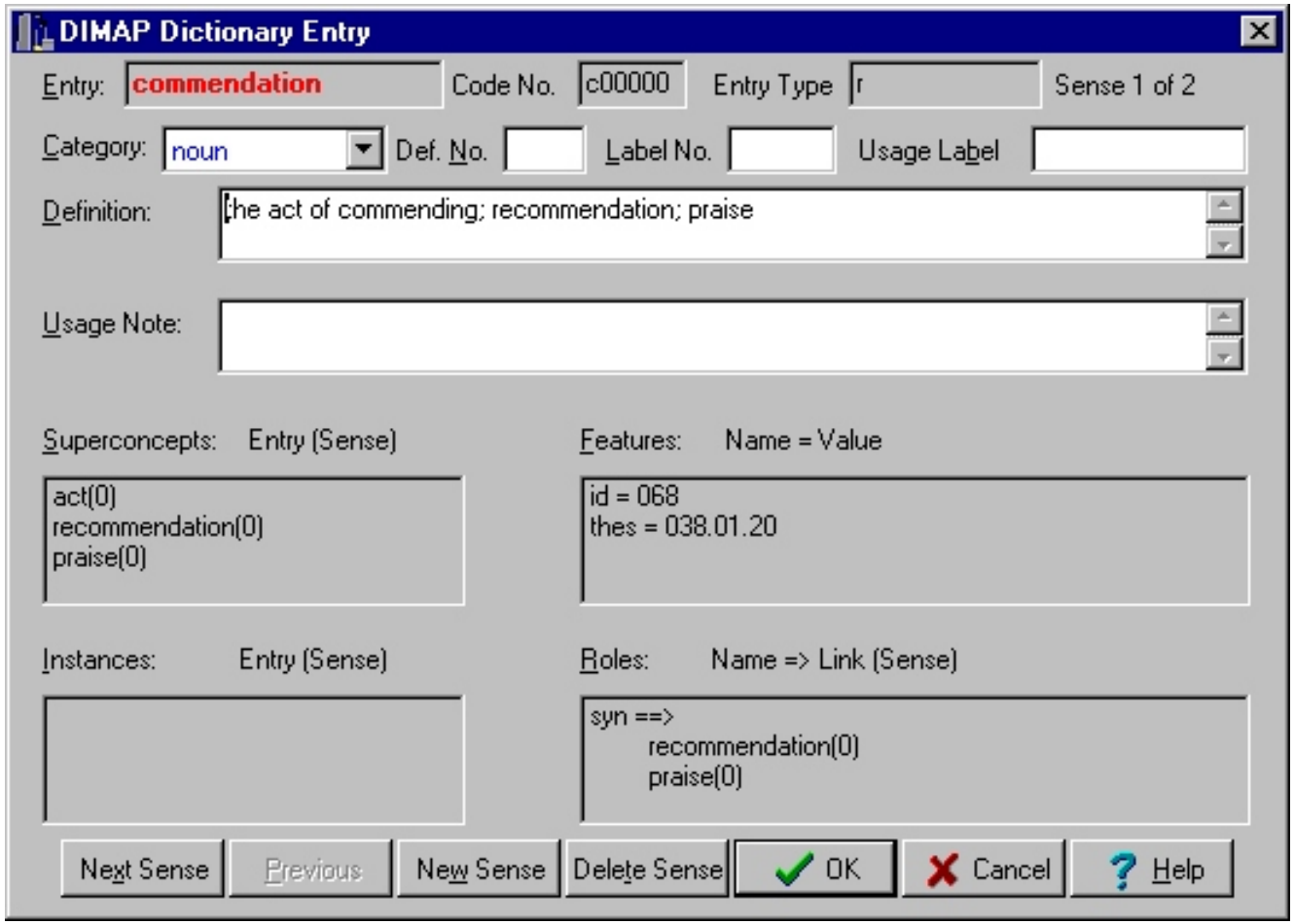


Figure 7. Basic model, verb subgraph example subject to Rule 3.

Converting the MRD to Analysis Database

- **Unlikely to work directly from the MRD, so put it in a form for analysis, with view toward its use in applications (e.g., into CL Research's DIMAP)**
- **Conversion phase**
  - **Parts of speech, grammar codes (attributive, plural noun with singular verb, "a" or "an" based on pronunciation), typing (transitivity, subcategorization patterns)**
  - **Subject label, irregular inflections, variants, usage information (e.g., prepositions)**
  - **Definitions, (preliminary) sense hierarchy, encyclopedic information (e.g., for use in question-answering), illustrative uses**
  - **Runons (establishing derivational links)**
  - **"Kind" equations (translating idiomatic uses into regular expressions for recognition, e.g., in WSD)**
  - **...**

Dictionary Sense Entry



Preprocessing Definitions

Definitions are not in a form for immediate recognition by parser

- **Nouns → NPs**
- **Verbs → usually more than VPs (particularly transitive definitions, VPs + preps (?indicative of case roles)**
- **Adjectives → frequently adjectives or APs, but also frequently participial phrases**
- **Adverbs → adverbs, but mostly PPs**
- **Prepositions → PP + prep (or verb)**
- **Subordinating conjunctions → PP + complementizer**
- **Solution: put definitions into sentence frame in preprocessing phase**

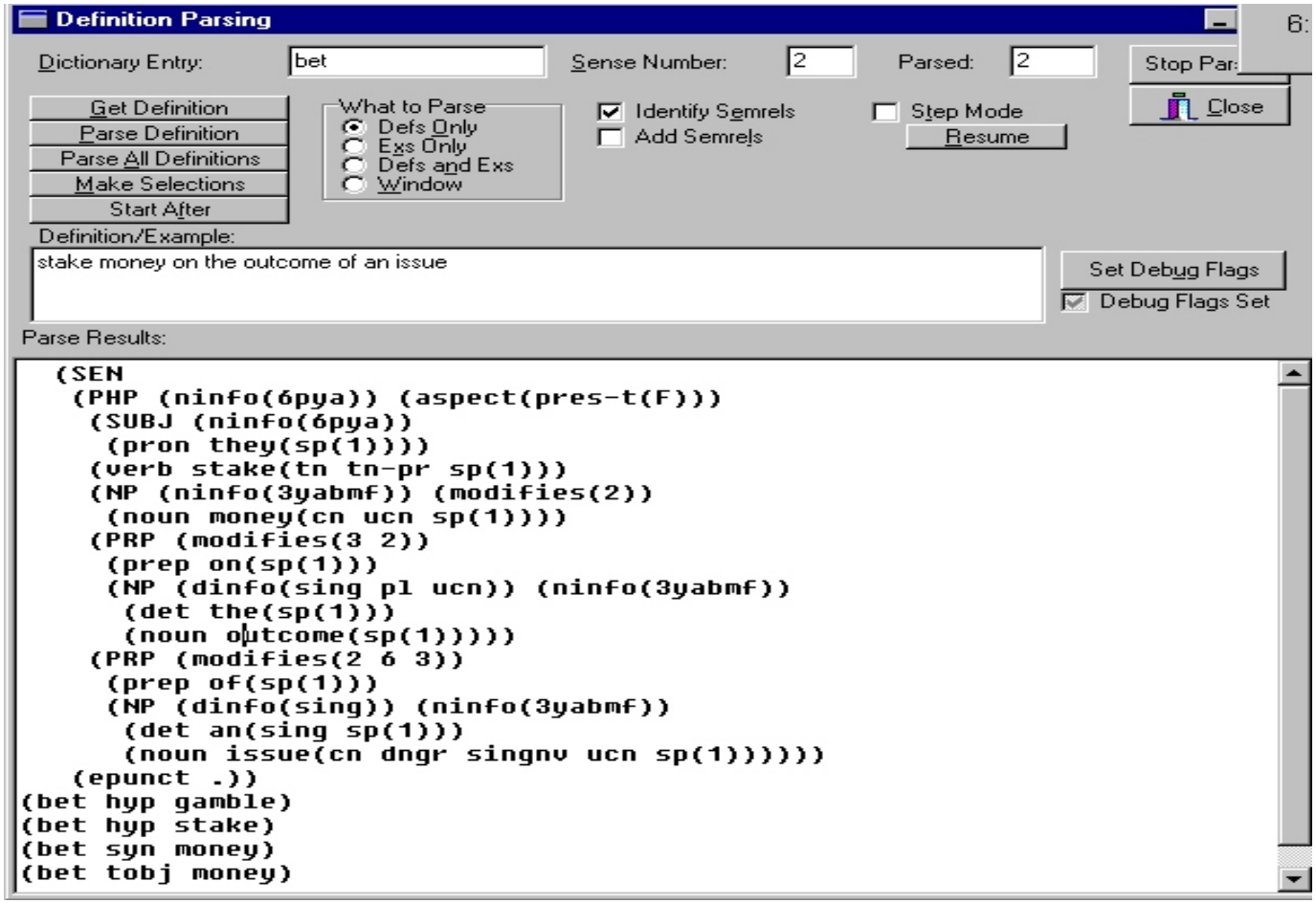
Definition Parsing Process

- **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**

Definition 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)**

Extraction of Semantic Relations



Identifying Semrels (MindNet)

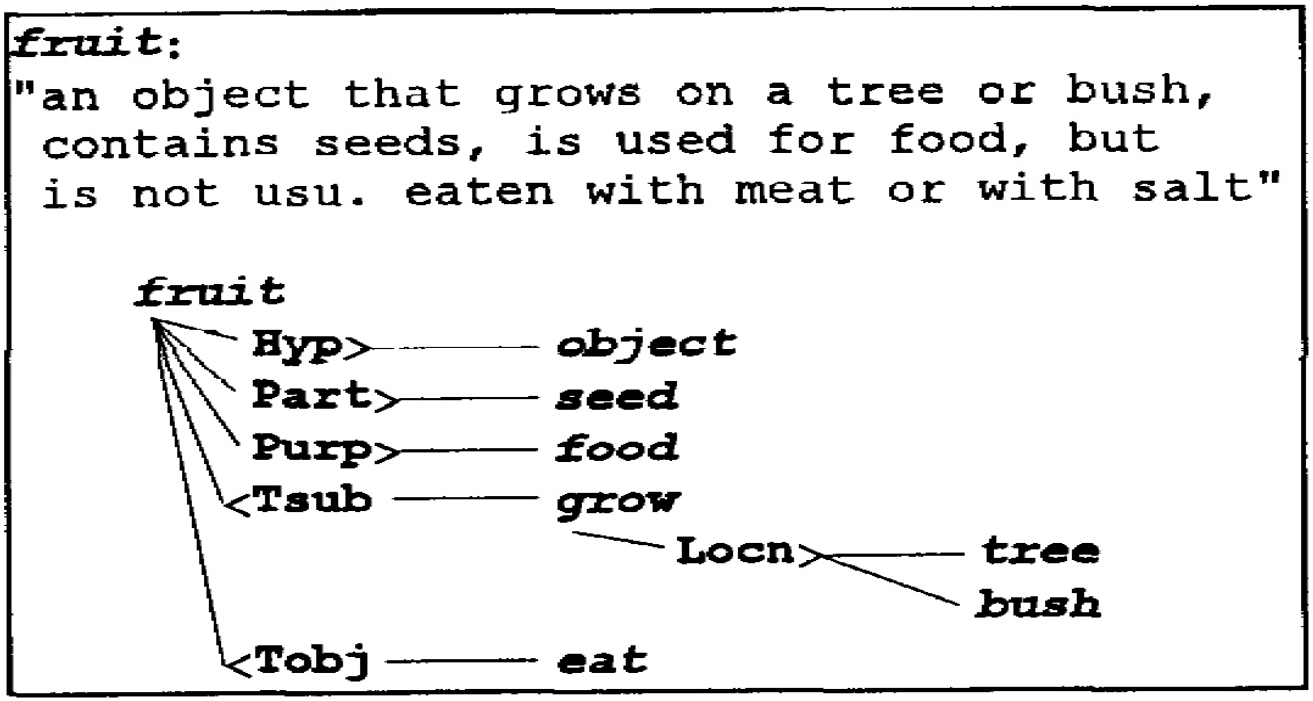


Figure 2.5. Semantic relation structure for the definition of **fruit**

Identifying Inverted Semrels

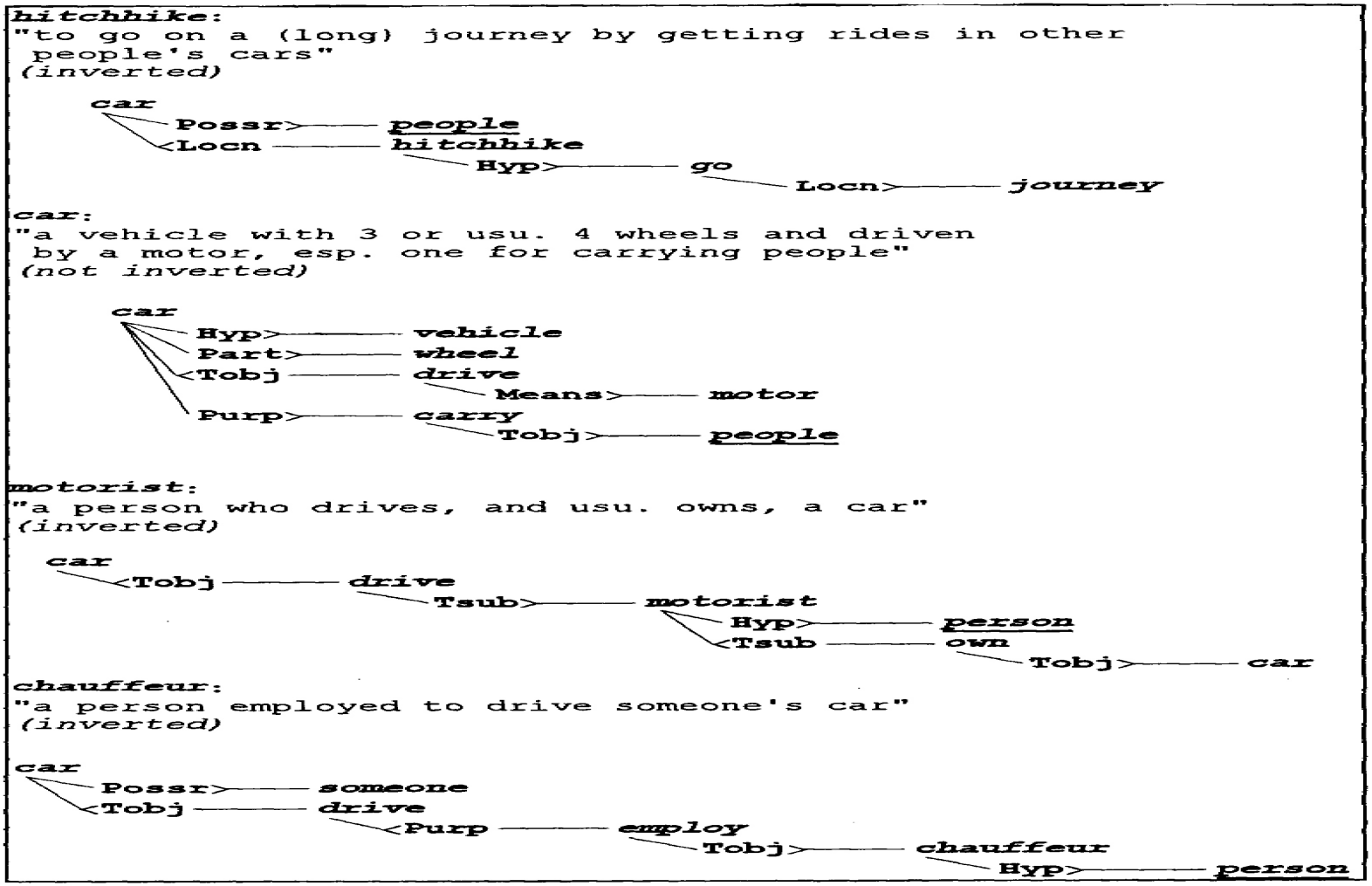


Figure 3.1. Some inverted semrel structures associated with **car** *that also contain* **people** *or* **person**

Building Frames from Definitions

Basic Frames for Definitions 1 and 2 of "change"

Definition 1: become different in one or more respects without becoming something else

```
[BECOME DIFFERENT (FROM-STATE .NE. TO-STATE)
   ((SUBJ) (PAT v AGT)
           (* D * (PAT v AGT)
                  (ESSENTIAL ATTRIBUTES ...)
                  (ACCIDENTAL ATTRIBUTES ...)
                      ("RESPECT" (TIME1) (FROM-STATE ...)
                                 (TIME2) (TO-STATE ...))))]
```

Definition 2: become something materially different from before

```
[BECOME DIFFERENT (FROM-STATE .NE. TO-STATE)
   ((SUBJ) (PAT v AGT) (TIME1) (FROM-STATE)
           (* D * (PAT v AGT) ...))
   (("RESULT") (TIME2) (TO-STATE)
           (* D * ("RESULT") ...))]
```

Syntagmatic Variation

Intransitive Uses of "change"

Use	Definition
follow (vt 4b)	change in constant relation to
gel (vi)	change into a gel
gelatinize (vi)	change into a jelly
graduate (vi 2)	change gradually
hold (vi 1b(1))	not change
melt (vi 1a)	change from a solid to a liquid state usu. by the action of heat
push (vi 5b)	change in quantity or extent
quarter (vi 4)	change from one quarter to another - used of the moon
range (vi 6)	change within limits
reform (vi)	change for the better
resinify (vi 1)	change into a resin
rote (vi)	change by rotation
run (vi 11b)	change to a liquid state
run into (vt 1a)	change into
solate (vi)	change to a sol
specialize (vi 3)	change adaptively
transfer (vi 2)	change from one vehicle or transportation line to another

Peculiarities of Definitions

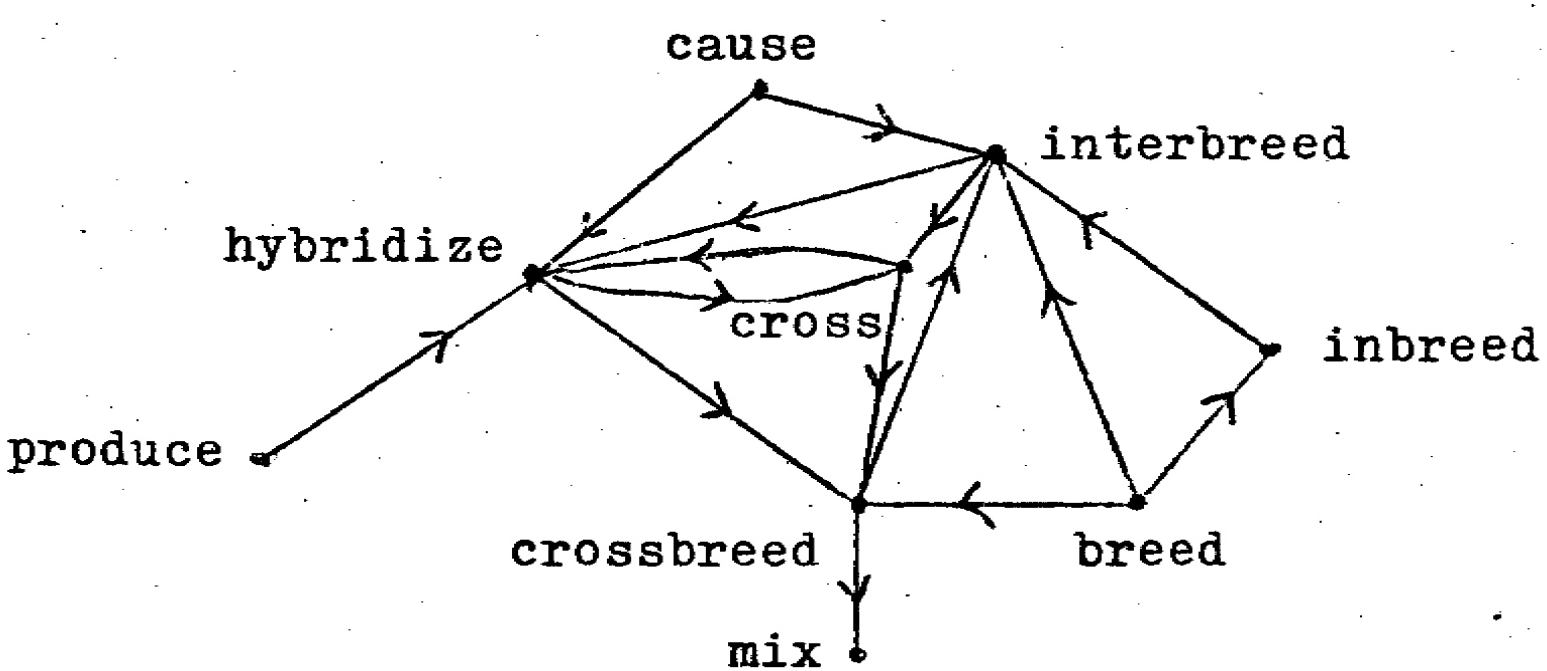
Will affect semantic interpretation

- **Empty heads (”a kind of”, “a part of”, etc.): special processing to get at appropriate hypernym or meronym**
- **“etc” problem: dairy (”a shop or company that sells milk, butter, etc.”)**
  - **Cruse’s covert categories (no fit in traditional hierarchies such as WordNet)**
- **Deictic references: dam (2) (”a body of water confined by such a barrier”); damask (2) (”table linen of this material”)**
- **“something”: dare (”to have the necessary courage or boldness for something”)**
- **“particular” and “specified”: date (”a particular day”)**
- **“usually” and “typically”: decanter (”a vessel, usually an ornamental bottle, from which wine, water, etc., is served at table”)**

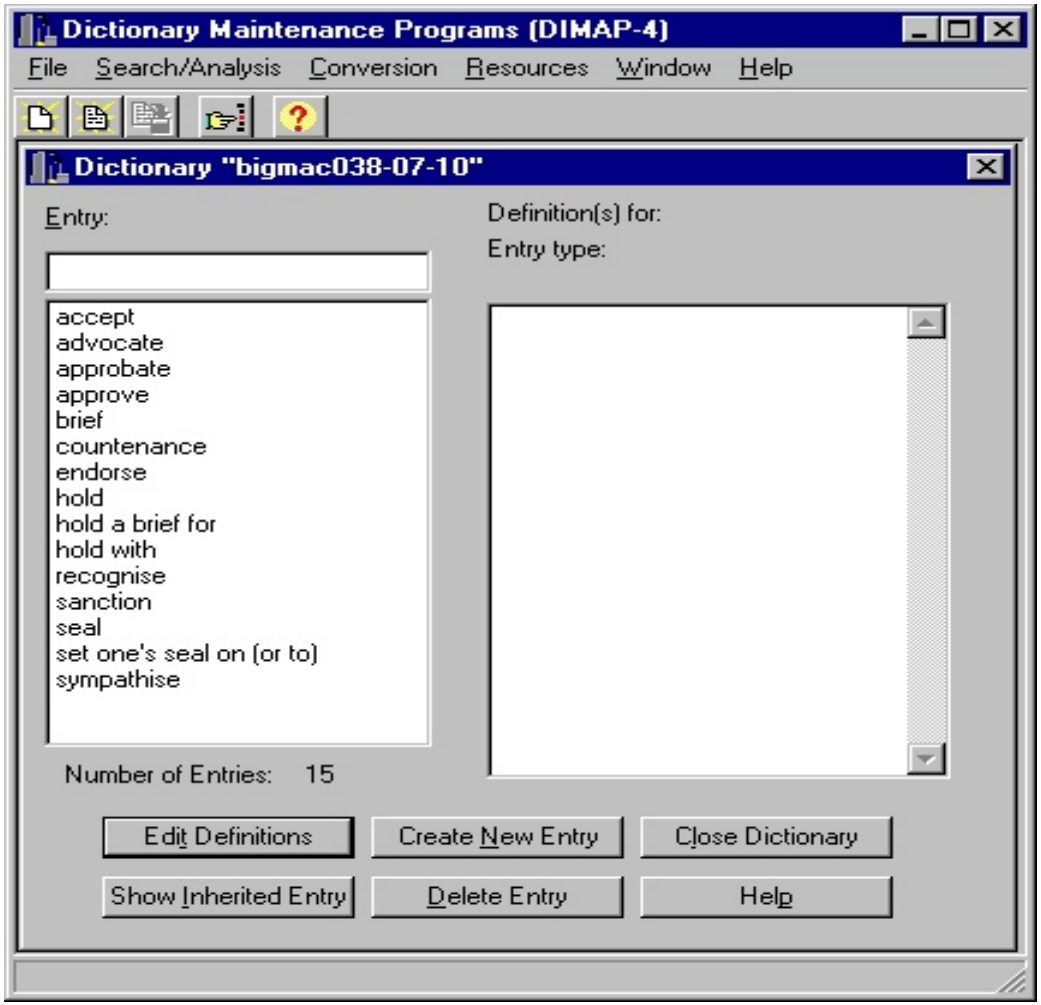
Application: Definition Mapping

- **Among main dictionary and derived dictionaries**
  - **Link senses of derived dictionaries to main dictionary (90 percent success using word overlap methods along with some type restrictions)**
  - **Identify core senses of main dictionary based on senses within smallest dictionary**
- **Within an entry**
  - **Establish sense hierarchy and how components of meaning are added in sense extensions**
- **Across entries and senses having same hypernyms**
  - **Automatic creation of thesaurus**
  - **Identify syntagmatic variation**
- **Analyze thesaurus entries**
  - **Establish consistency**
  - **Identify variations in lexicalizations and verbalizations**

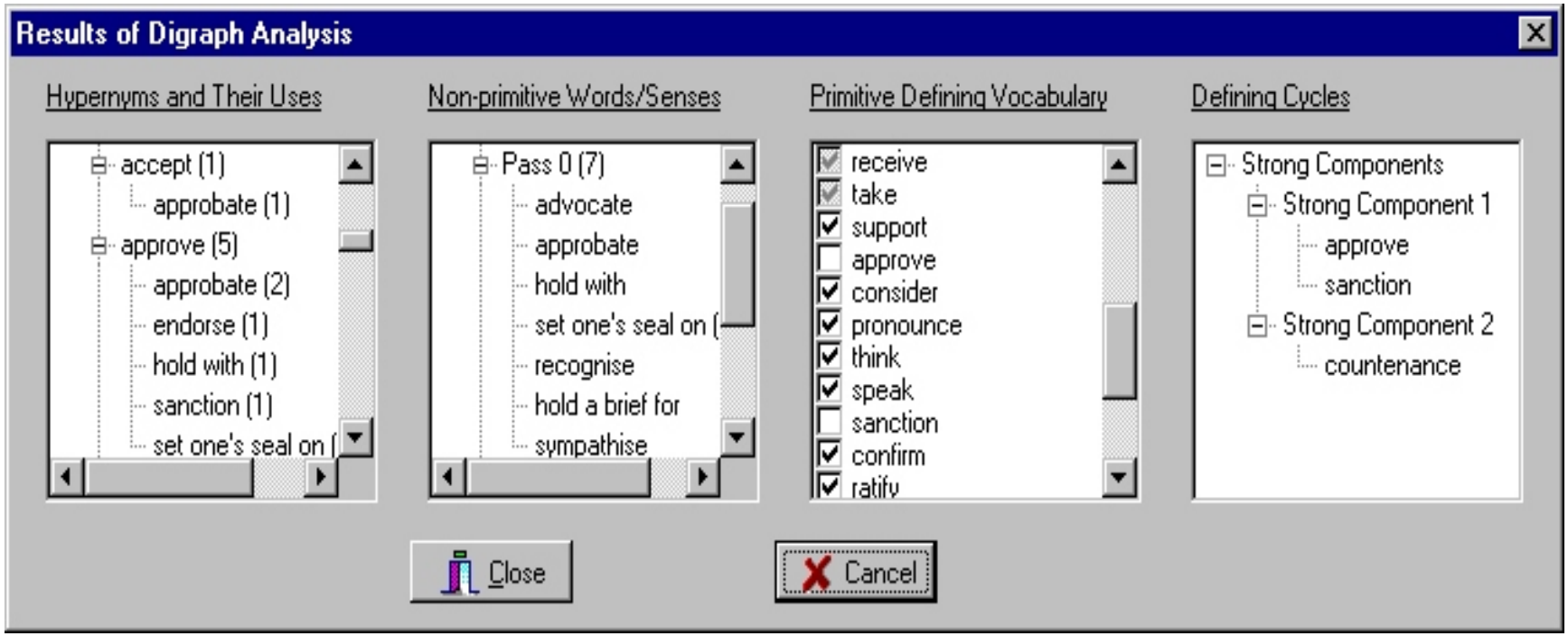
Identification of Semi-Primitives



Thesaurus Analysis (1)



Thesaurus Analysis (2)



Thesaurus Analysis (3)

- Assumes linkage between thesaurus and definitions or selection of a set of words (synset)
- Parse definitions to build ISA relations or identify them by hand
- Eliminate non-primitives (words that are not themselves used hypernymically), iteratively
- Identify strong components (words mutually reachable by hypernym paths)
- Identify hypernyms not in set and used to define set (semi-primitives)
- Examine definitional patterns of semi-primitive definitions (paradigmatic variation)
- E.g., one sense of 'nature' (with synonyms of 'wilderness', 'the great outdoors', 'the wild', a sense of 'waste') → "<{large? AREA}|STATE> of <uncultivated|desolate|remote> LAND"
- Extension to all senses with same hypernym

Application: Testing Dictionary Entries in 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**
- **Pushes toward identifying and using collocational cues**

Application: Question-Answering

- **TREC QA Track: identify short-phrase (≤50 bytes) answer**
- **Basic approach of using semantic relation triples [discourse entity (nouns), semantic relation (prepositions), attachment point (verbs, nouns)] and matching question structure to document sentence representation**
- **Using WSD techniques, augmented by**
  - **Using synonyms and hypernyms on discourse entity**
  - **Using semrel structures for matching on semantic relations and attachment points, particularly verbs**
- **Pushes toward methods for computing on words and relations**

Final Words

- **Fascinating and challenging look into the semantic structure of dictionaries and their definitions**
- **Continuing exploitation of more and more information present in dictionaries**
- **Finding out what is most needed and what will work in NLP applications**