# Towards a Meaning-Full Comparison of Lexical Resources

Ken Litkowski
CL Research
ken@clres.com
http://www.clres.com
http://www.clres.com/siglex99/index.html

# Meaning-Full

## Identification and analysis of semantic components and network through definition parsing

- Implementation of methods for modeling semantic structure of dictionaries (Litkowski, 1970s)
- Componential analysis techniques stemming from (Nida, 1970s)
- Steps toward implementation of methods for building a lexical knowledge base (Atkins, 1991)
- Methods for definition parsing and analysis are similar to those used by (Dolan, 1994) for definition clustering
- Componential techniques enable template and frame building (e.g., FrameNet)
- Components are similar to those built in Interlingua (e.g., Dorr)

# The SENSEVAL Problem

- Use of the unfamiliar Hector sense inventory
- Reliance of many competing systems on WordNet sense inventories
- Necessity of creating a map from WordNet to Hector senses
- Unknown (but judged negative) effect on performance for many systems

# The Computational Mapping Problem

- Use of SENSEVAL WordNet -> Hector mapping as a “gold standard”
- Implementation of mapping functionality inside dictionary maintenance software (DIMAP) to handle tests for syntactic, semantic, and collocational properties
- Use of Lesk-style word overlap methods (with and without stop lists) to provide a baseline against which to measure mapping success of other methods
- Definition parsing to identify semantic components; rich data structure to contain semantic network links, syntactic features, and collocational properties
- Use of “defining patterns” being developed in Dictionary Parsing Project for identification of semantic components
- Characterization of mapping between dictionaries of various types

# The Lexical Resources

The verbs of SENSEVAL: amaze, band, bet, bother, bury, calculate, consume, derive, float, hurdle, invade, promise, sack, sanction, scrap, seize, shake, slight

- Hector (5.7 senses per word, 18.4 words per sense)
- WordNet (WN) (3.7, 5.3)
- Webster's 3rd New International Dictionary (W3) (12.0, 9.9)
- American Heritage Dictionary (AHD) (6.2, 7.1)
- Oxford Advanced Learners Dictionary (OALD) (3.4, 8.7)
- Dorr's Lexical Knowledge Base (Dorr) (2.2)

# The WordNet - Hector Mapping

- 66 WordNet senses into 102 Hector senses
- 86 assignments made by lexicographer
- 9 WordNet senses given no assignment
- 40 WordNet senses given exactly one assignment
- 17 WordNet senses given 2 or 3 assignments
- WordNet senses contained 348 words
- Hector senses contained 1878 words

# Word Overlap Analysis

- Strict (no root-finding), with and without stop list (165 words consisting mainly of prepositions, pronouns, conjunctions, and common open-class words)
- Example: **bet**, WN 2 *(stake (money) on the outcome of an issue*) to Hector 4 (*(of a person) to risk (a sum of money or property) in this way*). Overlap on two words (*money, of*) (0.13 of its 15 words) without the stop list. With stop list, overlap of one (*money*, 0.07 of Hector). Lexicographer made three assignments (Hector 2, 3, and 4); our scoring as only 1 out of 3 correct
- 28 of 86 (32.6%) correct without stop list
- 31 of 86 (36.1%) correct with stop list, but only 23 of 86 (26.7%) when null assignments are removed
- 41 content words involved in mapping with stop list (1.8 words per assignment)
- 9 of 66 WordNet senses not assigned when using stop list

# Componential Analysis Technique

- Definition parsing to identify hypernyms (hyp), synonyms (syn), and other semantic relations (semrels)
- Semrels based on defining patterns (**manner**: in(dpat((~ rep01(det(0)) adj manner(0) sr(manner)))) to identify role (i.e., **manner**) and value (i.e., **adj**)
- Result of parsing is semantic network entries for each sense, with several relations **x R y** (with **R** equal to **hyp, syn, tsubj, tobj, instr**, **means**, **loc**, **purp**, **source**, **manner**, **has-constituents**, **has-members**, **is-part-of**, **locale**, and **goal**)
- Exclusion from viable matches of senses that conflict on syntactic or collocational properties
- Mapping based on matching **x**, **R**, **y**, with relaxation allowed on **x** and **y** to synset members and hypernym synsets (using WordNet), maximum of 2 levels
- Scoring of 5 points for **x** and **y** matches, 2 points for **R** matches

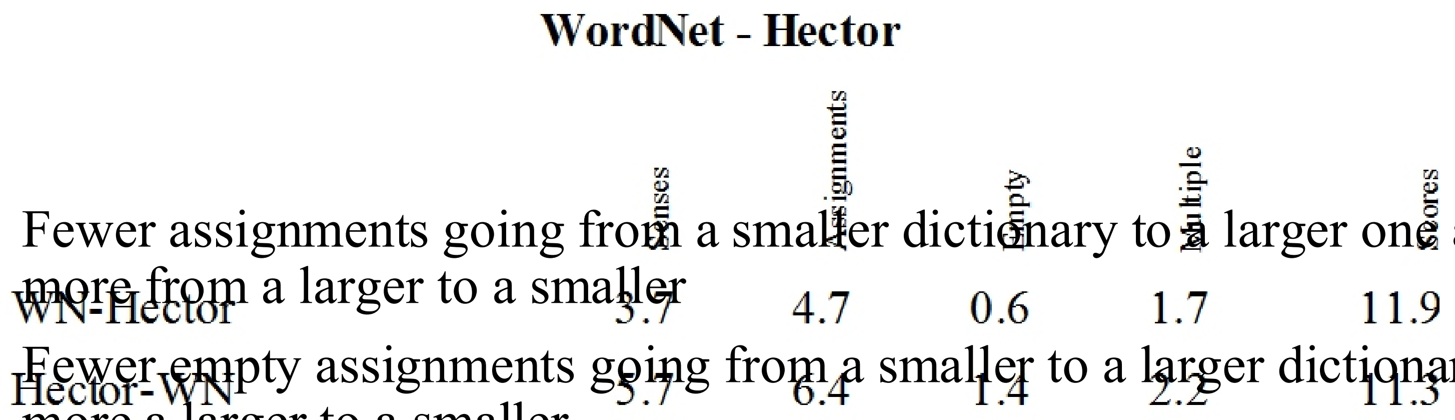
# Componential Analysis Results

- 35 of 86 (40.7%), compared to 23 of 86 in word overlap analysis when null assignments removed
- 4 “errors” arose from making assignments where lexicographer had made none, suggesting some basis for mapping
- 228 hits responsible for scores in the selected assignments (compared to 41 hits in word-overlap analysis when stop list was used)
- Results are based on use of still impoverished identification of semrels (0.86 per sense in Dictionary Parsing Project, compared to 3.26 per sense achieved by MindNet)

# Dictionary Mappings

- Number of senses, number of assignments in target dictionary, number of senses for which no assignment could be made, number of multiple assignments, and score of the assignments
- WordNet <-> Hector
- W3 <-> OALD
- W3 <-> AHD

# WordNet - Hector Mappings

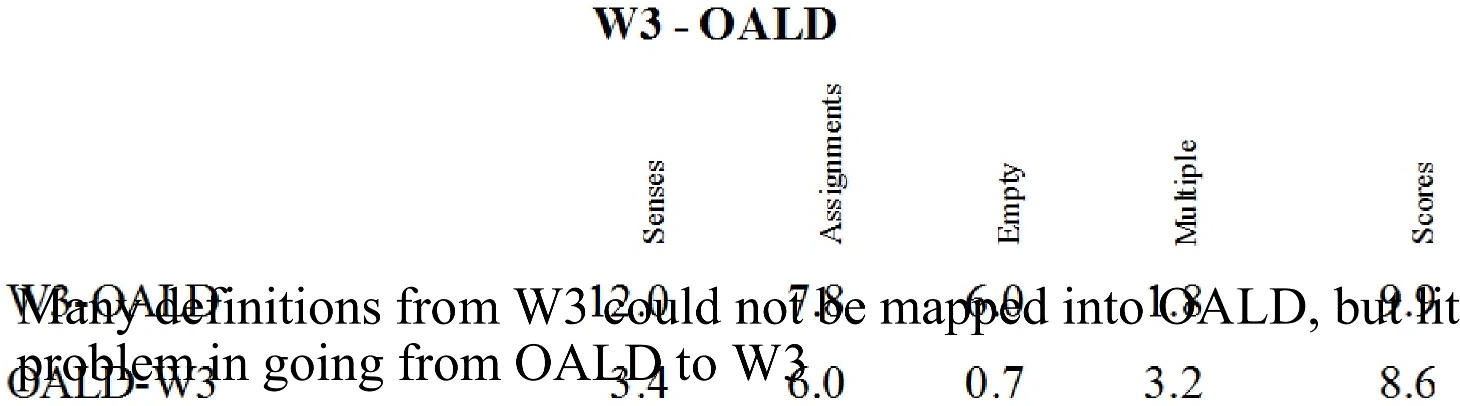


**WordNet - Hector**

| | Senses | Assignments | Empty | Multiple | Scores |
|---|---|---|---|---|---|
| WN-Hector | 3.7 | 4.7 | 0.6 | 1.7 | 11.9 |
| Hector-WN | 5.7 | 6.4 | 1.4 | 2.2 | 11.3 |

- Fewer assignments going from a smaller dictionary to a larger one and more from a larger to a smaller
- Fewer empty assignments going from a smaller to a larger dictionary and more a larger to a smaller
- More multiple assignments going from a larger to a smaller dictionary

# W3 - OALD

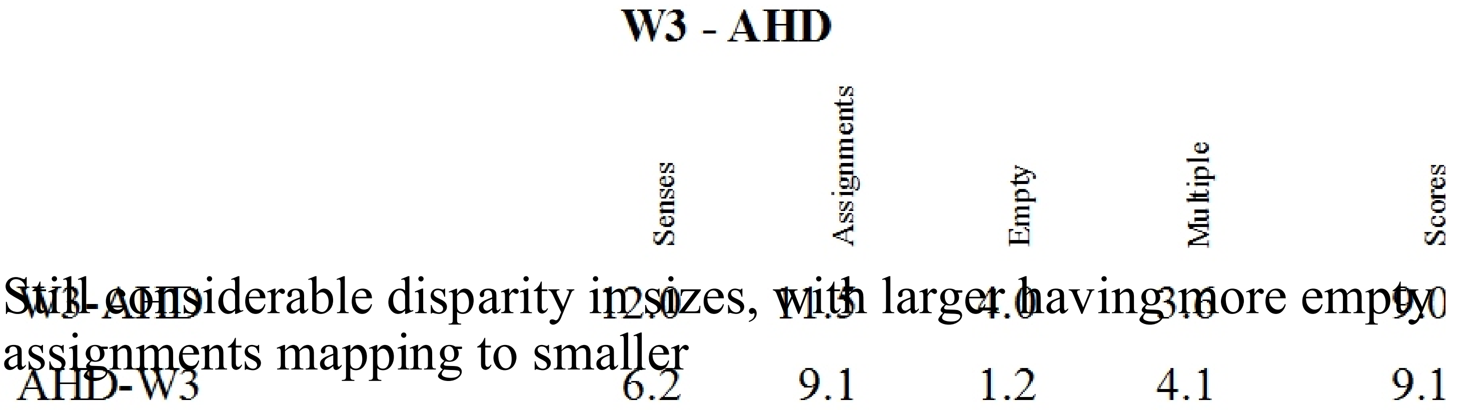


W3 - OALD

| | Senses | Assignments | Empty | Multiple | Scores |
|---|---|---|---|---|---|
| W3-OALD | 12.0 | 7.8 | 6.0 | 1.8 | 9.9 |
| OALD-W3 | 3.4 | 6.0 | 0.7 | 3.2 | 8.6 |

- Many definitions from W3 could not be mapped into OALD, but little problem in going from OALD to W3
- Many multiple assignments going from OALD to W3, indicating a lack of specificity in OALD

# W3 - AHD



W3 - AHD

| | Senses | Assignments | Empty | Multiple | Scores |
|---|---|---|---|---|---|
| W3-AHD | 12.0 | 11.5 | 4.0 | 3.6 | 9.0 |
| AHD-W3 | 6.2 | 9.1 | 1.2 | 4.1 | 9.1 |

- Still considerable disparity in sizes, with larger having more empty assignments mapping to smaller
- Lower scores than for WORDNET-Hector indicates lesser recognition of defining patterns

# Dorr's Lexical Knowledge Base

- Contains thematic grids which characterize the thematic roles of obligatory and optional semantic components, frequently identifying accompanying prepositions (encoded as transitivity type and roles in DIMAP, e.g., **instr** component)
- Some mappings from WordNet to Dorr for *float* and *shake* (for which there were multiple senses), illustrating mapping capability for lexical resources of different types
- Many semantic (theta) roles not yet recognizable in DIMAP defining patterns
- "verbs that incorporate thematic elements in their meaning would not allow that element to appear in the complement structure." (Olsen et al. 1998)
- Suggests identification of semantic components that are lexicalized and which are transmitted through to the thematic grid
- Example: **shake**, "to bring to a specified condition by or as if by repeated quick jerky movements," transmits "goal" to the thematic grid (2 senses in Dorr)

# Discussion and Conclusions

- Componential analysis method works, bringing back prepositions (removed by stop list) in identifying semrels
- Success due in part to consideration of senses as part of a network rather than just in isolation
- Considerable room for improvement as semrel defining patterns are elaborated
- Method allows for componential analysis of differences between definitions (lumpers vs. splitters)
- No need for "gold standard" (any intuitive mapping can be developed and analyzed)

# Future Work

- Definition comparison functionality is embedded with Senseval parsing functionality, allowing parsing of target words in corpus samples (i.e., lexicographer's workstation)
- Allows analysis of structure of a single word's senses and analysis of a synonym's defintions (see also Dolan, 1994)
- Defining patterns relevant not only to definitions but also to free text, allowing identification of thematic roles and "definitional" relations between sentence constituents
- Ability to map categories, concepts, or definitions between dictionaries, ontologies, and terminology databases 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”)
- Richer set of semrels (and resultant semantic network) enables richer lexical chaining and analysis of lexical cohesion