

---

# Acquiring Applicable Common Sense Knowledge from the Web

---



Hansen A. Schwartz and Fernando Gomez  
School of Electrical Engineering and Computer Science  
University of Central Florida

# INTRODUCTION

---

General Goal: Acquire *common sense knowledge* from the Web that can be applied successfully to Natural Language Processing problems.

- Acquire words and phrases from the Web
  - use of *search phrases* to automatically search
  - use of a syntactic parser to increase accuracy
- Analyze relationship frequency data over WordNet
  - produces probabilities between concept and noun
- Apply successfully to word sense disambiguation

# INTRODUCTION

---

## Common Sense Knowledge (CSK)

Knowledge which...

- ... we use in everyday life without necessarily being aware of it.
- “...every person assumes his neighbors also possess.”  
(Panton et al. 2006)

## Examples

*keys are kept in one's pocket*

*keys are used to open a door*

# INTRODUCTION

---

## Motivation: Applications for NLP

1. *He put the batter in the refrigerator*

- lexical ambiguity

2. *She ate the apple in the refrigerator*

- syntactic ambiguity

- These problems can be solved via the knowledge:

*Food is commonly found in the refrigerator.*

# OVERVIEW

---

- Introduction
- • Background
- Noun Acquisition
- Concept Analysis
- Evaluation
- Conclusion

# BACKGROUND

---

## Related Work

- Acquisition of Lexical Relationships
  - VerbOcean (Chklovski and Pantel, 2004)
  - ConceptNet (Liu and Singh, 2004)
  - Manually built patterns (Hearst, 1992)
  - Noun-noun relationships used for SemEval-2007 Task 4 (Girju et al., 2007)

# BACKGROUND

---

## Related Work

- **CYC** (Lenat, 1995)
- **Use of the Web for word sense disambiguation**
  - Acquired *topic signatures* (Agirre et al., 2001)
  - Used directly in algorithm (Martinez et al., 2006; Schwartz and Gomez, 2008)

Motivation: This current work automatically creates a CSK database, where the type of knowledge is explicit.

# BACKGROUND

---

## Prepositions and Relationships

- Prepositions state a relationship between two entities:

a constituent of the sentence, and complement to the preposition

(Quirk et al., 1985)

<b>description</b>	<b>prepositions</b>
<i>on</i> surface or line	on, onto, atop, upon, on top of, down on
<i>in</i> area or volume	in, into, inside, within, inside of

# BACKGROUND

---

## Prepositions and Relationships

A relationship,  $e1\mathbf{R}e2$ , exists between entities  $e1$  and  $e2$  if one finds “ $e1$  is  $\mathbf{R}$   $e2$ .”

### Examples:

- *cup on table*
- *food in refrigerator*

# OVERVIEW

---

- Introduction
- Background
- • Noun Acquisition
- Concept Analysis
- Evaluation
- Conclusion

# NOUN ACQUISITION

---

## Creating Web Queries from Search Phrases

- Parameters of a *search phrase*:

- *nounA*
- *nounB*
- *prep*
- *verb* (defined as part of the phrase)

- Example search phrases:

place *nounA* *prep* *nounB*

*nounA* is located *prep* *nounB*

# NOUN ACQUISITION

---

## Web Search

- Algorithm

```
for each search_phrase
  for each prep
    for each det
      query = create_query(search_phrase, prep, det, nounB);
      samples = websearch(query);
```

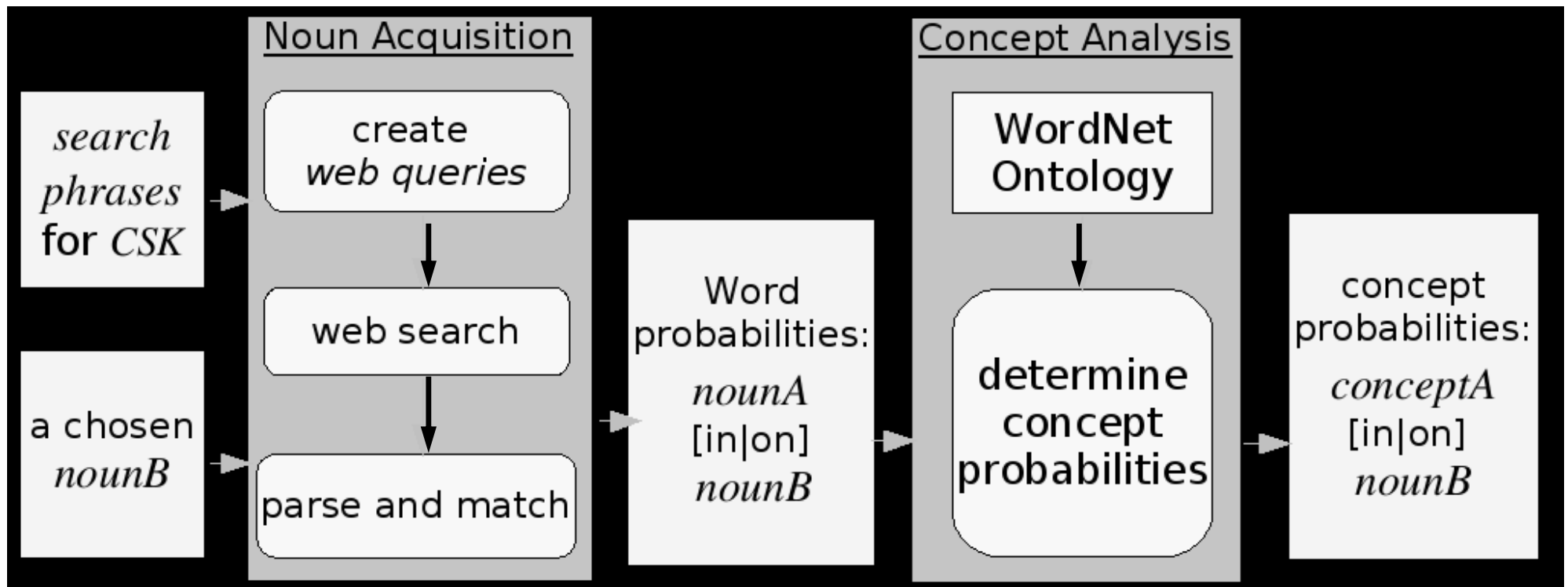
- Example

*search phrase:*            *place nounA prep nounB*  
*web query:*                *place \* in the refrigerator*

search using Google Search API (no longer supported), or Yahoo! Search Web Services ([developer.yahoo.com/search](http://developer.yahoo.com/search))

# NOUN ACQUISITION

Relationship during acquisition: *nounA* is [in | on] *nounB*



# NOUN ACQUISITION

---

## Parse and Match

- Match missing parameter using Charniak's Parser

*web query*: place \* in the refrigerator => place **something** in the refrigerator

(VP (VB place)  
(NP (**NN something**)  
(PP (IN in) (NP (DT the) (NN refrigerator))))))

**web query result**: He was told to place **the mixed batter** in the refrigerator

(S1 (S (NP (PRP He))  
(VP (AUX was) (VP (VBN told) (S (VP (TO to)  
(VP (VB place)  
(NP (**DT the**) (**JJ mixed**) (**NN batter**))  
(PP (IN in) (NP (DT the) (NN refrigerator))))))])

*nounA* = `batter'  
(head noun of matching phrase)

(Charniak, 2000)

# NOUN ACQUISITION

---

## Example Eliminations from parse:

...(CC and)  
(VP (VB place)  
  (PP (IN for) (NP (JJ several) (NNS hours)))  
  (PP (IN in) (NP (DT the) (NN refrigerator))))]

web query result:

...and place for several hours in  
the refrigerator

(VP (VB Place)  
  (NP (**NN something**))  
  (PP (IN on) (NP (DT the) (NN road))))]

web query:

place \* on the road

(S1 (S  
  (VP (VB Place)  
  (NP (DT the) (NN organization))  
  (PP (IN on) (NP (NP (DT the) (NN road))  
    (PP (TO to) (NP (NN recovery)))))) (. .)))

web query result:

Place the organization on the  
road to recovery.

(Charniak, 2000)

## *nounA* in bowl

<b>nounA</b>	<b>freq.</b>	<b>nounA</b>	<b>freq.</b>	<b>nounA</b>	<b>freq.</b>
food	219	flour	183	dough	171
ingredient	132	water	119	mixture	83
sugar	82	hand	60	potato	58
rice	55	butter	53	eggs	53
noodle	47	chicken	39	meat	38
piece	37	apple	35	milk	34
yogurt	34	slice	33	hands	32
something	32	spoon	31	tomato	30
couscous	28	yolk	28	ball	27
bean	24	cube	24	egg_white	24
fish	24	soup	23	vegetable	23
cereal	22	coin	22	bulgur	21
mushroom	21	soul	21	raisin	20
shrimp	20	stone	20	chocolate	19
egg	19	salt	18	spinach	18
bread	17	foot	17	fruit	17
head	17	money	17	pasta	17
cornstarch	16	cucumber	16	egg_yolk	16
onion	16	pebble	16	strawberry	16

## *nounA* in pocket

<b>nounA</b>	<b>freq.</b>	<b>nounA</b>	<b>freq.</b>	<b>nounA</b>	<b>freq.</b>
money	687	hand	295	cash	187
firework	78	something	55	dollar	53
ball	45	hands	41	key	37
coin	30	pedometer	27	card	26
battery	24	item	21	penny	20
phone	20	music	17	buck	16
implant	16	profits	15	wallet	15
camera	14	device	14	pen	14
anything	12	finger	11	insert	11
things	11	transmitter	11	dime	10
stick	9	baseball	8	book	8
cards	8	case	8	change	8
drop	8	glasses	8	information	8
magazine	8	profit	8	receiver	8
tooth	8	unit	8	bag	7
box	7	cent	7	deck	7
glove	7	gun	7	letter	7

## *nounA* on table

<b>nounA</b>	<b>freq.</b>	<b>nounA</b>	<b>freq.</b>	<b>nounA</b>	<b>freq.</b>
book	34	hands	32	shoes	29
elbow	27	flat	27	card	26
plate	23	cards	22	food	22
object	22	cup	18	glass	17
head	17	item	16	pencil	16
foot	15	lamp	15	candle	14
chair	14	dish	14	hand	14
beaker	13	cloth	13	box	12
bread	12	jar	12	key	12
napkin	12	piece	12	report	12
bet	11	bottle	11	bowl	11
briefcase	11	celtic	11	chip	11
money	11	paper	11	penny	11
something	11	ranger	10	unit	10
bible	9	forearm	9	gun	9
newspaper	9	pen	9	bag	8
ball	8	board	8	camera	8
candlestick	8	case	8	coin	8
face	8	information	8	microphone	8

# NOUN ACQUISITION

---

Frequency becomes Probability

$$p_w(nA, \mathbf{R}, nB)$$

This is the probability of  $nA$  being returned to a query for the relationship,  $\mathbf{R}$ , with  $nB$ .

# CONCEPT ANALYSIS

---

Relationships between concept and word

*conceptA* is [in | on] *nounB*

Concept as synset in WordNet (Miller et al., 1993)

(*batter-1, hitter-1, slugger-1, batsman-1*)  
“(baseball) a ballplayer who is batting”

---

probabilities for  
noun senses and  
synsets

$$p_{ns}(nAs, \mathbf{R}, nB) = \frac{p_w(\text{lemma}(nAs), \mathbf{R}, nB)}{\text{senses}(\text{lemma}(nAs))}$$

$$p_{syn}(syns, \mathbf{R}, nB) = \sum_{nAs \in syns} p_{ns}(nAs, \mathbf{R}, nB)$$

# CONCEPT ANALYSIS

---

## Incorporating the ontology

$$P_c(cA, \mathbf{R}, nB) = p_{syn}(syns(cA), \mathbf{R}, nB) + \sum_{h \in hypos(cA)} P_c(h, \mathbf{R}, nB)$$

Function recurs based on the idea that a concept subsumes the probability all of its hyponyms.

Example: (*money-3*) is-a (*currency-1*), so

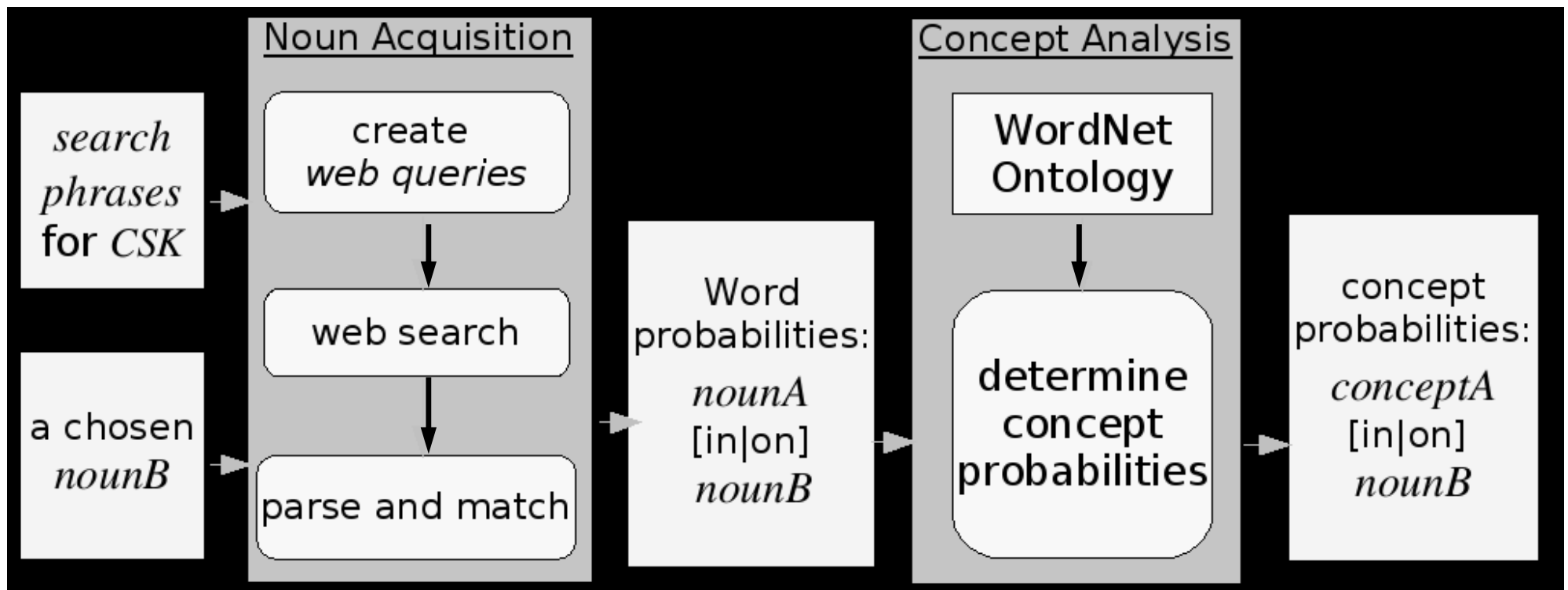
$P_c(\text{currency-1}, R, \text{'pocket'})$  subsumes

$P_c(\text{money-3}, R, \text{'pocket'})$

# CONCEPT ANALYSIS

Relationship probabilities after analysis:

*conceptA* is [in | on] *nounB*



# OVERVIEW

---

- Introduction
- Background
- Noun Acquisition
- Concept Analysis
- • Evaluation
- Conclusion

# EVALUATION

---

## Disambiguation System

- *CSK* not sufficient by itself
  - Only one type of relationship
  - Intended to be used to improve WSD
- Integrated into **GWSD** (Sinha and Mihalcea, 2007)
  - High all-words results
  - Compatible with WordNet
  - Results from 4 graph metrics, easy to integrate our knowledge as a 5<sup>th</sup> metric.

# EVALUATION

---

## Disambiguation System

- We use the  $P_c(\text{concept}, \mathbf{R}, nB)$  values, where *concept* corresponds to a sense of the target word.
  - $nB$  matches “*prep det nB*” within the sentence
  - suggests all senses with  $P_c$  value greater than  $0.75 \max P_c$  over all senses
- Voting combines the 4 GWSD predictions with CSK suggestions.
  - Ties were broken with lowest sense number among those tied.

# EVALUATION

---

## Experimental Corpus

- Needed annotated corpus with instances of nouns as prepositional complements.
- Annotated sentences from Wikipedia matching “*prep det lemma*”
  - *Lemma* is one of the 30 *nounBs* for which we acquired relationships.
  - 342 sentences with one target noun annotated per sentence. [in, on]
  - Assigned all appropriate WordNet senses: 26.3% instances were given multiple senses due to fine-grained nature of WordNet. (Ide and Wilks, 2006)
  - Only polysemous nouns.

# EVALUATION

---

## Experimental Corpus and Baseline

	<b>insts</b>	<b>agree</b>	<b>F1<sub>h</sub></b>	<b>F1<sub>rnd</sub></b>	<b>F1<sub>MFS</sub></b>
<b><i>on</i></b>	131	79.9	84.7	28.2	71.0
<b><i>in</i></b>	211	80.8	91.9	27.2	67.8
<b>both</b>	342	80.5	89.2	27.6	69.0

**insts**: number of annotated instances

percentages:

**agree**: inter-annotator agree %

**F1** values (precision = recall):

h: human annotation, rnd: random baseline,

MFS: most frequent sense baseline.

$$\frac{\sum_{i \in C} (|S_i^h \cap S_i^a|)}{\sum_{i \in C} (|S_i^h \cup S_i^a|)} \div 3/2$$

# EVALUATION

---

## Results

	without <i>CSK</i>		with <i>CSK</i>	
	<b>F1<sub>all</sub></b>	<b>F1<sub>indeg</sub></b>	<b>F1<sub>all</sub></b>	<b>F1<sub>indeg</sub></b>
<i>on</i>	62.6	63.4	64.9	67.2
<i>in</i>	68.7	69.7	71.6	72.5
<b>both</b>	66.4	<b>67.3</b>	69.0	<b>70.5</b>
<b>ties</b>	37	0	66	72

**F1** values with and without acquired *CSK*:

**all**: using all 4 graph metrics

**indeg**: using only the indegree metric

# EVALUATION

---

## More Results

- 54.7% of instances received at least 1 suggestion from CSK
- 24.5% of instances received multiple suggestions from CSK

Other options when using indegree metric predictions with CSK:

- F1 value when using MFS backoff for ties: 70.2
- Precision when not predicting ties: 71.9% (on 270 instances)

# CONCLUSION

---

Effective method of discovering relationships

- unique requirement to match syntactic parse of web query

Produced Relationship between concept and word

- used WordNet to produce *conceptA***R***nounB* probability.

Successfully incorporated into WSD system

- 4.5% error reduction for top results

# CONCLUSION

---

## Future Work

- Exhaustively acquire *CSK* for all nouns
- Acquire other forms of *CSK*
  - => Test on standard corpora
- Study and improve the effectiveness of the parse
- Improvements to concept analysis
- Improvements to application via alternative voting schemes

# REFERENCES

---

- Eneko Agirre, Olatz Ansa, and David Martinez. 2001. Enriching wordnet concepts with topic signatures. In *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburg, USA.
- Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 132–139, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain.
- Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of SemEval-2007*, pages 13–18, Prague, Czech Republic, June. Association for Computational Linguistics.
- Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539–545.
- Nancy Ide and Yorick Wilks, 2006. *Word Sense Disambiguation: Algorithms And Applications*, chapter 3: Making Sense About Sense. Springer.
- Douglas B. Lenat. 1995. CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
- H. Liu and P Singh. 2004. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22:211–226.
- David Martinez, Eneko Agirre, and Xinglong Wang. 2006. Word relatives in context for word sense disambiguation. In *Proceedings of the 2006 Australasian Language Technology Workshop*, pages 42–50.
- George Miller, R. Beckwith, Christiane Fellbaum, D. Gross, and K. Miller. 1993. Five papers on wordnet. Technical report, Princeton University.
- Kathy Panton, Cynthia Matuszek, Douglas Lenat, Dave Schneider, Michael Witbrock, Nick Siegel, and Blake Shepard. 2006. Common sense reasoning : From cyc to intelligent assistant. In Y. Cai and J. Abascal, editors, *Ambient Intelligence in Everyday Life*, pages 1–31.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.
- Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.
- Hansen A. Schwartz and Fernando Gomez. 2008. Acquiring knowledge from the web to be used as selectors for noun sense disambiguation. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 105–112, Manchester, England, August.
- Ravi Sinha and Rada Mihalcea. 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the IEEE International Conference on Semantic Computing*. Irvine, CA, September.

**Acknowledgement:** This research was supported by the NASA Engineering and Safety Center under Grant/Cooperative Agreement NNX08AJ98A.



# Thank You!

---

[www.eecs.ucf.edu/~hschwartz/CSK/](http://www.eecs.ucf.edu/~hschwartz/CSK/)

- Frequency Data
- Experimental Corpus