

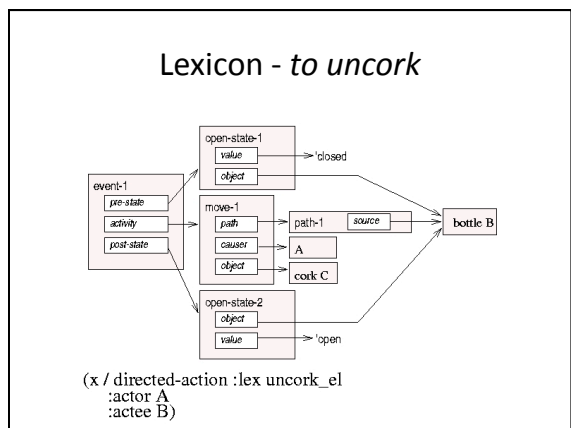
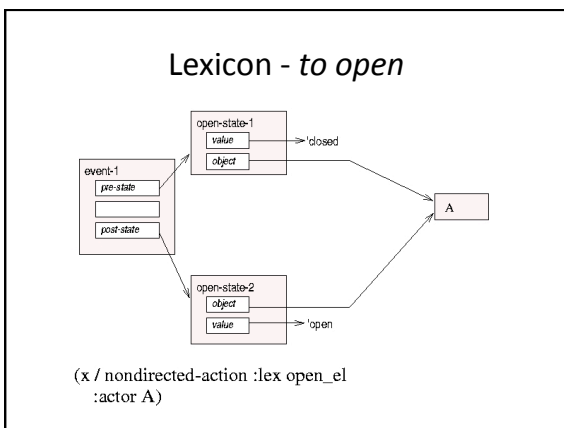
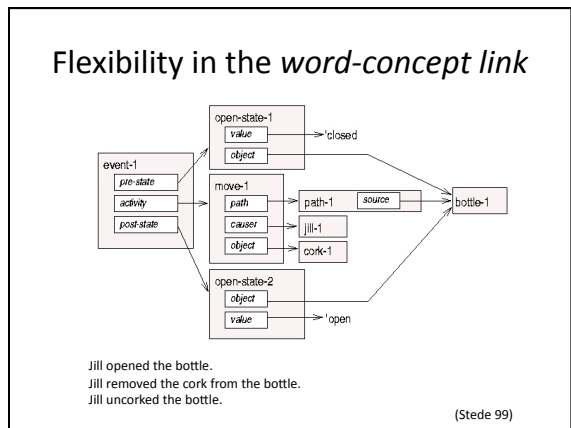
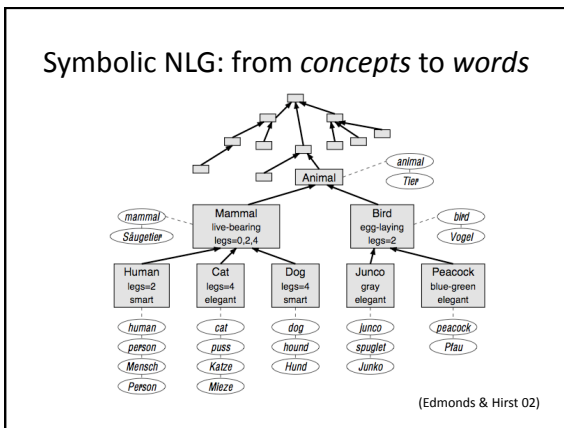
## Lexical similarity and distributional similarity

*A guided tour from old-school language generation to contemporary paraphrasing*

Manfred Stede  
Applied Computational Linguistics  
Center for Cognitive Science  
University of Potsdam/Germany

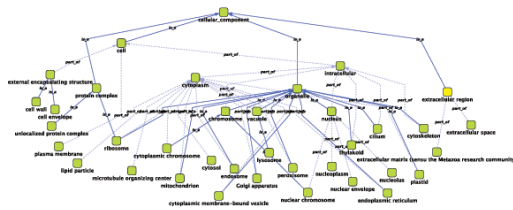
## Natural language generation (NLG)

- **Cognitive** inspiration:  
models of human language production
- **Linguistic** inspiration:  
generation models that implement our knowledge of grammar and lexicon
- **Engineering** inspiration:  
implementations that produce language output for some application (dialog systems, etc.)





### Backbone: Hierarchical Knowledge Base



### The Hyperonym Problem

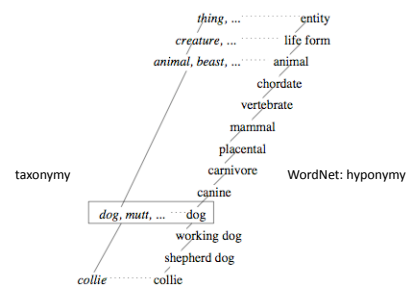
- When usage conditions for *poodle* are fulfilled, those for *dog*, *animal*, *thing* are fulfilled as well
- Rule: Always prefer the most specific term, unless
  - want to avoid using the same term over and over
  - want to convey connotations, e.g.:  
*collie* – {*dog* | *mutt*} – {*animal* | *beast*} – *thing*
  - want to avoid basic-level effects (Rosch 78, Reiter 91)  
*Quick! There's a {tiger shark – shark} in the water!*

### The Hyperonym Problem

- |   |  |
|---|--|
| <ul style="list-style-type: none"> <li>• <b>Hyponymy</b></li> <li>• „is a“</li> <li>• spaniel – dog</li> <li>• rose – flower</li> <li>• mango – fruit</li> <li>• kitten – cat</li> <li>• queen – monarch</li> <li>• waiter – man</li> </ul> | <ul style="list-style-type: none"> <li>• <b>Taxonomy</b></li> <li>• „is a kind/type of“</li> <li>• spaniel – dog</li> <li>• rose – flower</li> <li>• mango – fruit</li> <li>• ? kitten – cat</li> <li>• ? queen – monarch</li> <li>• ? waiter – man</li> <li>• usually &lt;=5 levels, often fewer</li> </ul> |
|---|--|

(Cruse 77, 86)

### Two hierarchies



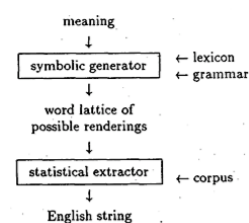
(Stede 00)

### Summary: Lexical choice =

- computing the **best match** on the grounds of a motivated feature set for word meaning and context

### II. Statistical NLG

- The pioneer: Nitrogen (Langkilde & Knight 1998)



## Nitrogen

- Backbone: *Sensus* ontology & mapping to WordNet synsets
- Ranking of alternatives by means of unigram and bigram statistics

## Nitrogen: Example

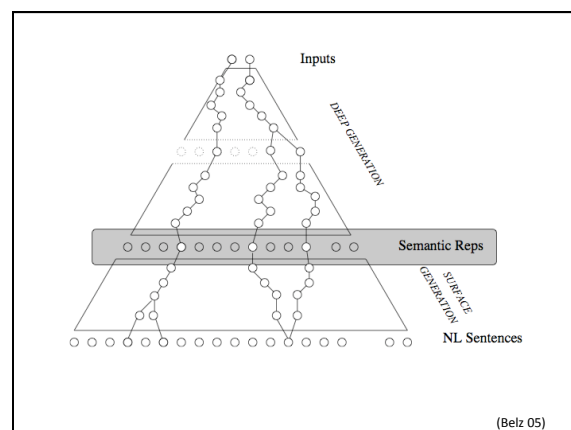
(A / |workable|  
 :DOMAIN (A2 / |sell<cozen|  
 :AGENT I  
 :PATIENT (T / |trust,reliance|  
 :GPI THEY))  
 :POLARITY NEGATIVE)

I cannot betray their trust  
 I will not be able to betray their trust  
 I am not able to betray their trust  
 I are not able to betray their trust  
 I is not able to betray their their trust  
 I cannot betray the trust of them  
 I cannot betray trust of them  
 I cannot betray a trust of them  
 I cannot betray trusts of them  
 I will not be able to betray trust of them

Unigram frequencies:  
 reliance – 567    reliances - 0  
 trust - 6100    trusts - 1083

## From n-grams to syntax trees

- Bangalore/Rambow 00: *FERGUS*
- Input: dependency tree
- Tree chooser uses stochastic tree model to associate TAG trees with nodes (supertagging)
- Meaning = WordNet synsets; avoid sense disambiguation by using *supersynsets*
- Lexical choice = choosing the most appropriate supersynset member
- Choose lexeme of daughter node by maximizing  $p(l_d | l_m, s_m)$
- (Choose lexeme of root node by picking the most frequent one)



## n-gram based near-synonym choice

- mistake, error, fault
- job, task, chore
- duty, responsibility, obligation
- difficult, hard
- material, stuff
- put up, provide, offer
- decide, settle, resolve, adjudicate

- Edmonds 97
  - training: 1 year of WSJ
  - co-occurrence counts
- Inkpen 07
  - training: Waterloo terabyte web corpus
  - PMI (Turney 01)

Set	No. of cases	Accuracy			
		Base-line	Edmonds method	Stat. method (Docs)	Stat. method (Words)
1.	6,630	41.7%	47.9%	<b>61.0%</b>	59.1%
2.	1,052	30.9%	48.9%	<b>66.4%</b>	61.5%
3.	5,506	70.2%	68.9%	69.7%	<b>73.3%</b>
4.	3,115	38.0%	45.3%	64.1%	<b>66.0%</b>
5.	1,715	59.5%	64.6%	68.6%	<b>72.2%</b>
6.	11,504	36.7%	48.6%	52.0%	<b>52.7%</b>
7.	1,594	37.0%	65.9%	74.5%	<b>76.9%</b>
<b>AVG</b>	31,116	44.8%	55.7%	65.1%	<b>66.0%</b>

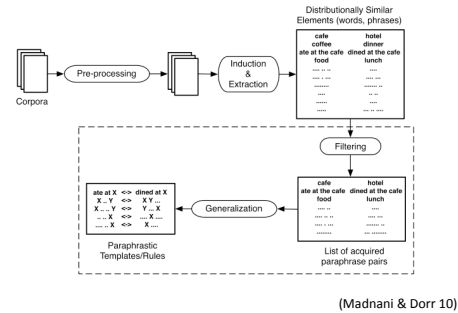
## Summary: Lexical choice =

- computing the best match on the grounds of a motivated feature set for word meaning and context
- choosing from a given synset the **most typical** word in context, as computed from a corpus; context = neighbourhood in surface string or dependency tree

### III. Paraphrasing

- „A paraphrase is an alternative surface form in the same language, expressing the same semantic content as the original form.“ (Madnani & Dorr 10)
- Goal: Automatic acquisition of paraphrases, or of the ability to produce paraphrases, from corpora
- Also known as an instance of *text-to-text generation*
- What for?
  - query expansion in information retrieval
  - question answering (different expressions in Q and A)
  - compute textual entailment
  - text similarity (e.g., plagiarism detection)
  - detect redundancy (multi-document summarization)
- Levels
  - lexical
  - phrasal
  - sentential

### Distributional similarity



### Distributional similarity

- *You shall know a word by the company it keeps.* (Firth 1957)
- The vision (Lin 98, after Nida 75)
  - A bottle of **tezgüino** is on the table.
  - Everyone likes **tezgüino**.
  - **Tezgüino** makes you drunk.
  - We make **tezgüino** out of corn.
- => *Tezgüino* is similar to *beer, wine, wodka, ...*

### Distributional similarity

- What corpus to use?
  - „comparable text“ (e.g., Barzilay et al 01, Pang et al 03, Shimohata & Sumita 05)
  - one corpus as large as possible (e.g., Lin 98, Lin & Pantel 01)
- What is *context*?
  - syntactic structures: constituent trees or dependency triples (e.g., Lin 98, Gasperin et al. 01, Glickman & Dagan 03, Pang et al. 03)
  - surface word windows (e.g., Rapp 04, Kolb 08, Grigonyte et al 10)

### Lin 98: Computing word similarity via distributions

- 64 million word corpus
- 56.6 million dependency triples (8.7 million unique) (have **subj** I), (dog **obj-of** have), ...
- Similarity between two objects = amount of information contained in the commonality between the objects divided by the amount of information in the individual descriptions of the objects

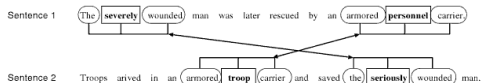
$$\frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}$$

### Lin 98: „Corpora are better than reference handbooks“

- The method yields for *adversary*:
  - enemy, foe, ally, antagonist, opponent, rival, detractor, neighbor, supporter, competitor, partner, trading partner, accuser, terrorist, critic, Republican, advocate, skeptic, challenger
- Webster's Collegiate Thesaurus lists for *adversary*:
  - Synonyms = *opponent, antagonist, anti, con, match, opposer, oppugnant*
  - Related words = *assaulter, attacker*
  - Contrasted words = *backer, supporter, upholder.*
  - Antonyms = *ally*
- => Webster's misses synonyms such as *enemy, foe, rival, competitor, challenger*

## Shimohato & Sumita 05

- Monolingual comparable text
- Three criteria: identical context (1-word l&r), prevention of outside appearance, POS identity
- 22% of found word pairs are also in WordNet synset (with both words in WordNet individually)
- Non-synonyms: express = say, present = report, decrease = drop



## Example: DISCO (Kolb 08)

- [www.linguatools.de/disco](http://www.linguatools.de/disco)
  - Word similarity calculation for six languages, based on large corpora
  - Uses Lin's (98) measure, with dependency relation replaced by word distance
  - (For English, this yields essentially the same results as dependency triples do)
- aircraft (5835)
  - plane (0.1501)
  - planes (0.1408)
  - bombers (0.1046)
  - fighter (0.1007)
  - helicopter (0.0911)
  - jets (0.0898)
  - helicopters (0.0895)
  - aeroplane (0.0884)
  - pilots (0.0848)
  - pilot (0.0808)
  - jet (0.0784)
  - airliner (0.0782)

## Color terms in DISCO

- green (9382)  
blue - yellow - red - brown - white - pink - grey - black - purple - dark - orange - pale - thick - coloured - silk - bright - scarlet - cream - gold - crimson
- red (11167)  
white - blue - black - yellow - pink - green - brown - grey - purple - scarlet - orange - dark - bright - coloured - pale - crimson - gold - silk - silver - thick
- mauve (164)  
purple - maroon - pink - crimson - lilac - violet - turquoise - blue - beige - scarlet - yellow - peach - fawn - lime-green - green - orange - rose-pink - reddish - amber - khaki

## But, be cautious!

- „If the pair of words *teacher* and *instructor* is considered to be more semantically equivalent than, say, the pair *teacher* and *musician*, then the distributions of the first pair will also be more alike than that of the latter pair.“  
(Madnani & Dorr 10, after Firth 54)
- Second-order word similarity in DISCO<sub>BNC</sub>
  - *teacher* (8039) / *instructor* (504): 0.26
  - *teacher* (8039) / *musician* (575): 0.31

## In fact, be very cautious!

- old (45.002)  
ancient (0.0767)  
young (0.0629)  
little (0.0604)  
older (0.0579)  
big (0.0567)  
medieval (0.0533)  
oldest (0.0525)  
olds (0.0516)  
Old (0.0511)  
elderly (0.0505)  
traditional (0.0490)
- The hyperonym problem is back, and mixed with an antonym / co-hyponym / relatedness problem

## Synonyms and Antonyms

- Traditional hypothesis:  
  
Antonyms often appear in the same sentence – which distinguishes them from synonyms  
(Charles & Miller 89, Justeson & Katz 91, Fellbaum 95)

## Synonyms and Antonyms

- But (near-)synonyms often occur in the same sentence, too
  - amplification  
*they always **asserted** and **affirmed** that they...*
  - avoiding repetition  
*the **car** started moving and hit two other **vehicles***
  - definition  
*the **hedgehog**, often also called **woodchuck**, can be found...*

(cf. French &amp; Labiouse 02)

## Experiments on syn / ant distribution (Peter Kolb)

- Corpus
  - German newspaper and fictional text
  - 202 million tokens, 11 million sentences
  - lemmatisation and pos with TreeTagger (Schmid 02)
- Lexical resource
  - GermaNet Version 4.1
  - 2086 pairs of N, V, A antonyms (with >20 occurrences in the corpus)
  - 14439 pairs of N, V, A synonyms (>20)

## Distributional similarity (DISCO) of antonyms and synonyms

- Antonyms
  - 2086 pairs
  - 43.5% completely dissimilar (0.0%)
  - the others are very similar; overall average is 0.0132 (e.g., *terraced house / house*)
- Synonyms
  - 14439 pairs
  - 40% completely dissimilar (0.0%)
  - overall average 0.0191

## Classification experiment - German

- Corpus: newspaper and fiction, 400 million tokens
- For syn, ant, hyp, cohyp, hol, 1000 word pairs for each relation, randomly selected from GermaNet 4.1 (antonymy defined between synsets)
- Features
  - distributional similarity
  - collocation strength relative to position (11 word window)
  - co-occurrence in documents
  - morphology (prefixes, partial string match)
- WEKA ClassificationViaRegression (best)
  - 5-way classification: 39.29% accuracy
  - 2-way classification (syn+hyp / ant+cohyp+hol): 62.9%

## Classification experiment - English

- Corpus: BNC, 116 million tokens
- For syn, ant, hyp, cohyp, hol, 800 word pairs for each relation, randomly selected from WordNet 3 (antonymy defined between words)
- Same features as in German, minus morphology
- WEKA ClassificationViaRegression (best)
  - 5-way classification: 42.4% accuracy
  - ant / syn: 71.3%
  - ant / hol: 79.4%
  - ant / kohyp: 83.6%
  - ant / hyp: 79.4%

## Distinguishing syn / ant: Lin et al 03

- **Method 1:** detect incompatibility with AltaVista query patterns: „from X to Y“, „either X or Y“
- **Method 2:** using a bilingual dictionary, intersect the set of w's distributionally similar words with the words that have the same translation as w
- P=93.3%, R=39.2%
- use a threshold to decide synonymy
- P=86.4%, R=95.0%

(Evaluation with randomly selected 80 pairs of syn/ant from Webster's Collegiate Thesaurus)

### Distinguishing syn / ant: Turney 08

- Reduce computation of synonymy, antonymy, association to the common problem of finding *analogy patterns*: A:B :: C:D
- For the syn/ant distinction, similar approach as Lin et al 03, but without handcrafted patterns
- Accuracy of generated patterns: 75%
- Baseline: guessing majority class – 65.4%

### Summary: Lexical choice =

- computing the best match on the grounds of a motivated feature set for word meaning and context
- choosing from a given synset the most typical word in context, as computed from a corpus; context = neighbourhood in surface string or dependency tree
- choosing from the set of words that have **similar distribution** in a corpus

### Conclusion

- In language generation, studying near-synonymy means lexical choice
- **Symbolic NLG**
  - dimensions for describing word meaning: denotation vs. connotation etc
  - how do we characterize a *context*
  - word–concept boundary and linking
  - ...
- **Statistical NLG**
  - „WordNet plus n-gram frequencies“
- **Paraphrasing**
  - distributional similarity => „lexical fields“
  - yields robustness (i.e., recall) for several practical applications
  - no solution for precision (i.e., relation differentiation) in sight
- => original choice problems were largely abandoned (not solved)

### Remember: be cautious!

- |                           |                            |
|---------------------------|----------------------------|
| • DISCOBNC                | • DISCOwikipedia           |
| • <b>synonymous</b> (368) | • <b>synonymous</b> (1586) |
| axiomatic                 | obsolete                   |
| irrelevant                | interchangeable            |
| insolvent                 | pejorative                 |
| untenable                 | useless                    |
| inoperative               | meaningless                |
| obsolete                  | identical                  |