An In-Depth Look into the Co-Occurrence Distribution of Semantic Associates

Sabine Schulte im Walde
Natural Language Processing
University of Stuttgart

Alissa Melinger
School of Psychology
University of Dundee

International Conference on Linguistic Evidence 2008
Tübingen, January 31

Semantic Associates

- Semantic associates: words that are called to mind in response to a given stimulus

  cook → kitchen, bake, hot, soup, yummy, ...

- Cognitive science: investigate mechanisms underlying the semantic memory
  (representation and access of semantic information)

- Computational linguistics: empirical instantiations of semantic meaning and semantic relatedness

Distributional Hypothesis

- Semantic association is related to the textual co-occurrence of the stimulus-response pairs
- Cognitive Science: Miller (1969), Spence & Owens (1990); memory research, word recognition, semantic networks, ...
- Computational Linguistics:
  - » exploit connection between co-occurrence distributions and semantic relatedness in *automatic acquisition of semantic knowledge* from corpus data (Harris, 1968)
  - » use association norms as *test-bed* for distributional models of semantic relatedness

Distributional Hypothesis: Analyses

- What proportion of associate responses is observed in the context of their respective stimulus verbs?
- Replicate original experiment by Spence & Owens (1990)
- Break analysis down into various categories
- Descriptive approach, no inferential statistics

Overview

1. Data Collection
2. Co-Occurrence Method
3. Co-Occurrence Experiments
4. Conclusions

Data Collection

schneien \`to snow´

kalt	\`cold´
rodeln	\`sledge´
Schneemann	\`snowman´
weiß	\`white´
dämmern	\`dawn´

Experiment Data

- Stimuli: 330 German verbs
- Participants per verb: between 44 and 54
- Number of associations per target verb:
  range 0-16, average: 5.16
- Responses: 79,480 tokens for 39,254 types (all)
  15,788 tokens for 7,425 types (first only)

Data Preparation

association strength

klagen 'complain, moan, sue'			
Gericht	'court'	19	11
jammern	'moan'	18	6
weinen	'cry'	13	6
Anwalt	'lawyer'	11	1
Richter	'judge'	9	3
Klage	'complaint, lawsuit'	7	1
Leid	'suffering'	6	3
Trauer	'mourning'	6	1
Klagemauer	'Wailing Wall'	5	2
laut	'noisy'	5	0

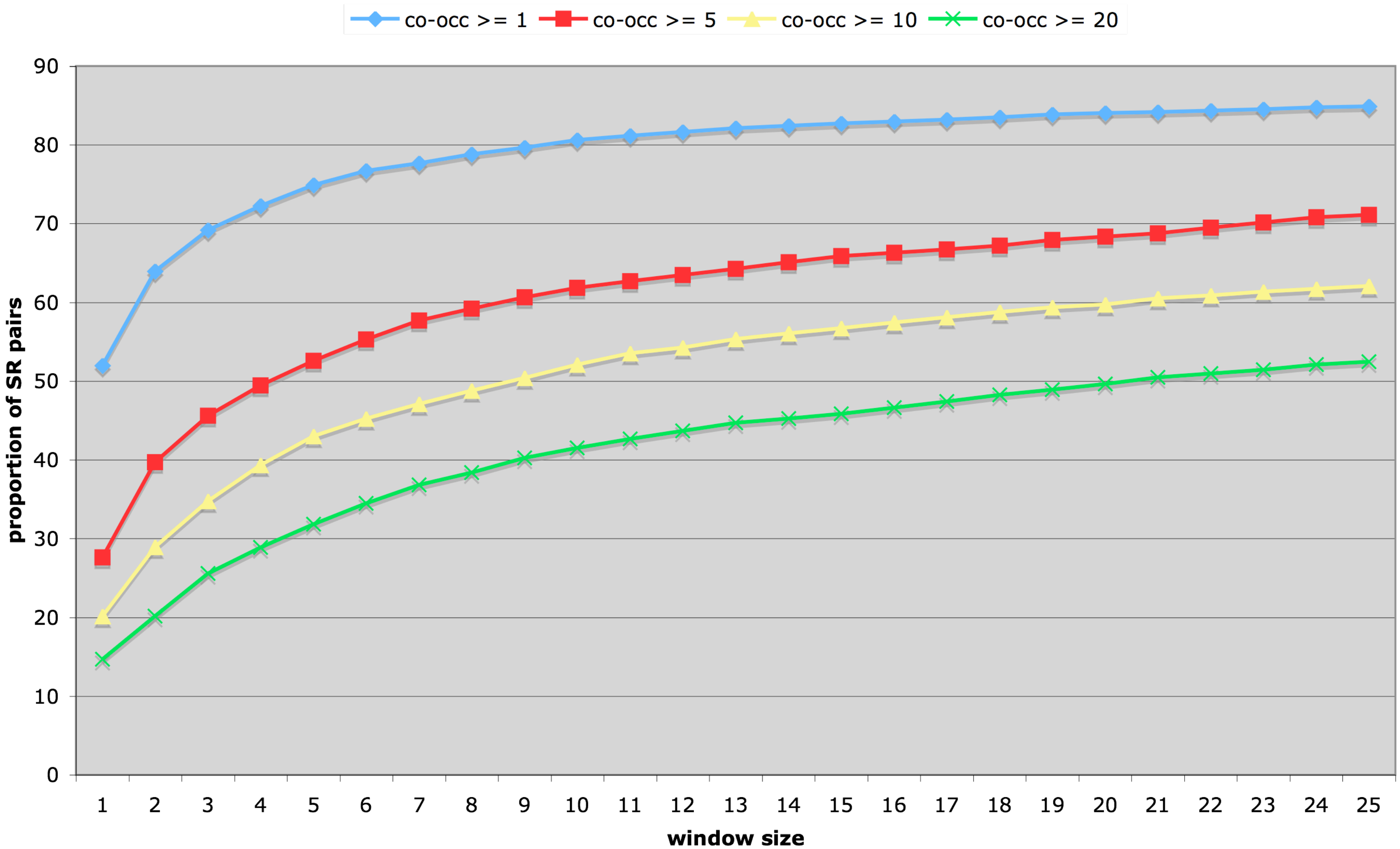
Co-Occurrence Method

Co-Occurrence Method

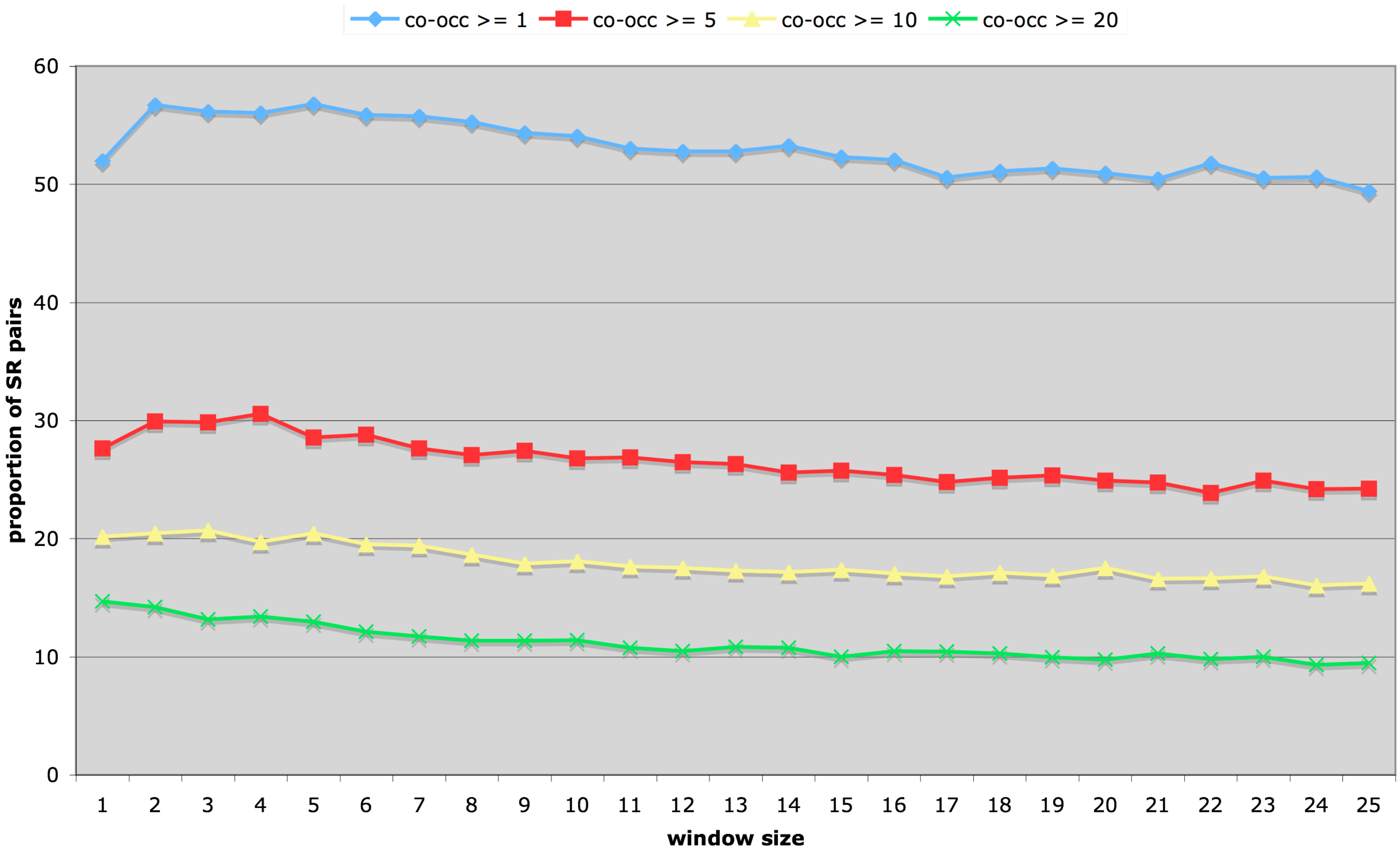
- What proportion of the 15,788 first response tokens is observed in the context of their respective target stimuli?
- Corpus of 200 million words of German newspaper text
- No punctuation, but function words
- Sliding context window with ±1 words to ±25 words
- Co-occurrence strength:
  How often did stimulus and response occur together?
- Cumulative view vs. non-cumulative view:
  total coverage vs. window-specific coverage

Co-Occurrence Experiments

Experiment 1: Basic Experiment *cumulative view*



Experiment 1: Basic Experiment *non-cumulative view*



Experiment 1: Basic Experiment

- Simplest analysis supports co-occurrence hypothesis:

  threshold of 1: 50% of SR pairs immediately adjacent,
85% total coverage;

  threshold of 5: 30% of SR pairs immediately adjacent,
70% total coverage;

  threshold of 20: 50% total coverage

- Non-cumulative view: more SR pairs in smaller than larger windows (decrease of 4-7%), but larger windows contribute as well

Exp 2:
Basic Experiment, corrected

Experiment 2: Basic, corrected

- Correct implicit assumption that two words co-occur in a corpus because they are semantically related.
- Establish a baseline:
  co-occurrence rate of unrelated words
- Artificial set of SR pairs: for each original SR pair type, response is replaced by another word, randomly chosen from corpus but matched for POS and corpus frequency;
  example: *abstürzen - Flugzeug* *(crash - airplane)* →
  abstürzen - Erkenntnis (crash - awareness),
  freq(Flugzeug) = 581, freq(Erkenntnis) = 582
- Correction by subtracting baseline from original values

Experiment 2: Basic, corrected

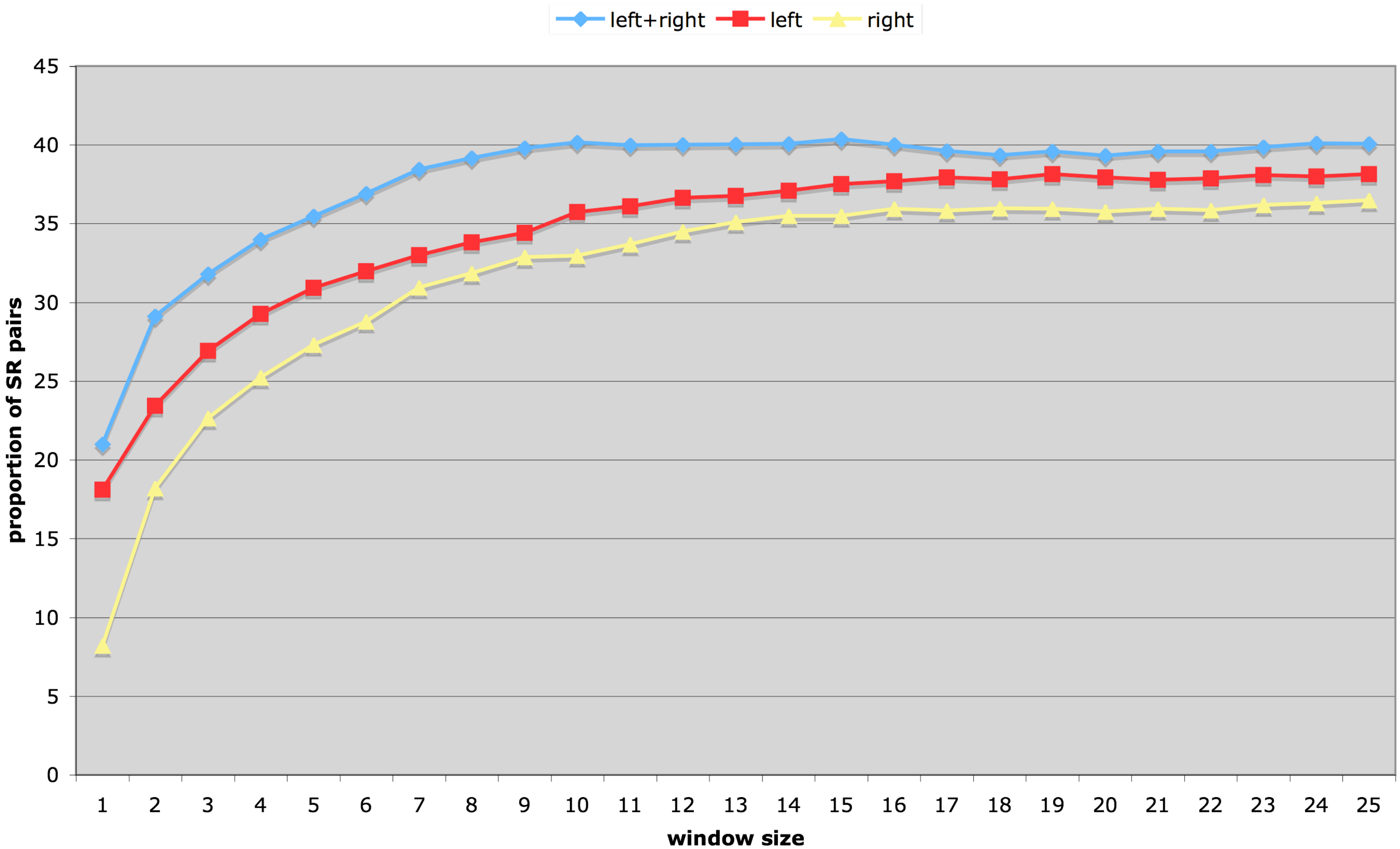
- Plot shapes of unrelated SR proportions are similar to basic experiment, but coverage is 12-44% lower.
- Relatively stable rates for unrelated SR proportions across all windows, with slight increase in large windows.
- Semantically related words co-occur in smaller windows relatively more often than semantically unrelated words.
- Taking baseline into account, still 34/20% (thresholds: 1/5) of SR pairs are immediately adjacent.
- Non-cumulative view: larger proportions in smaller window sizes.

Exp 3:
Window Direction

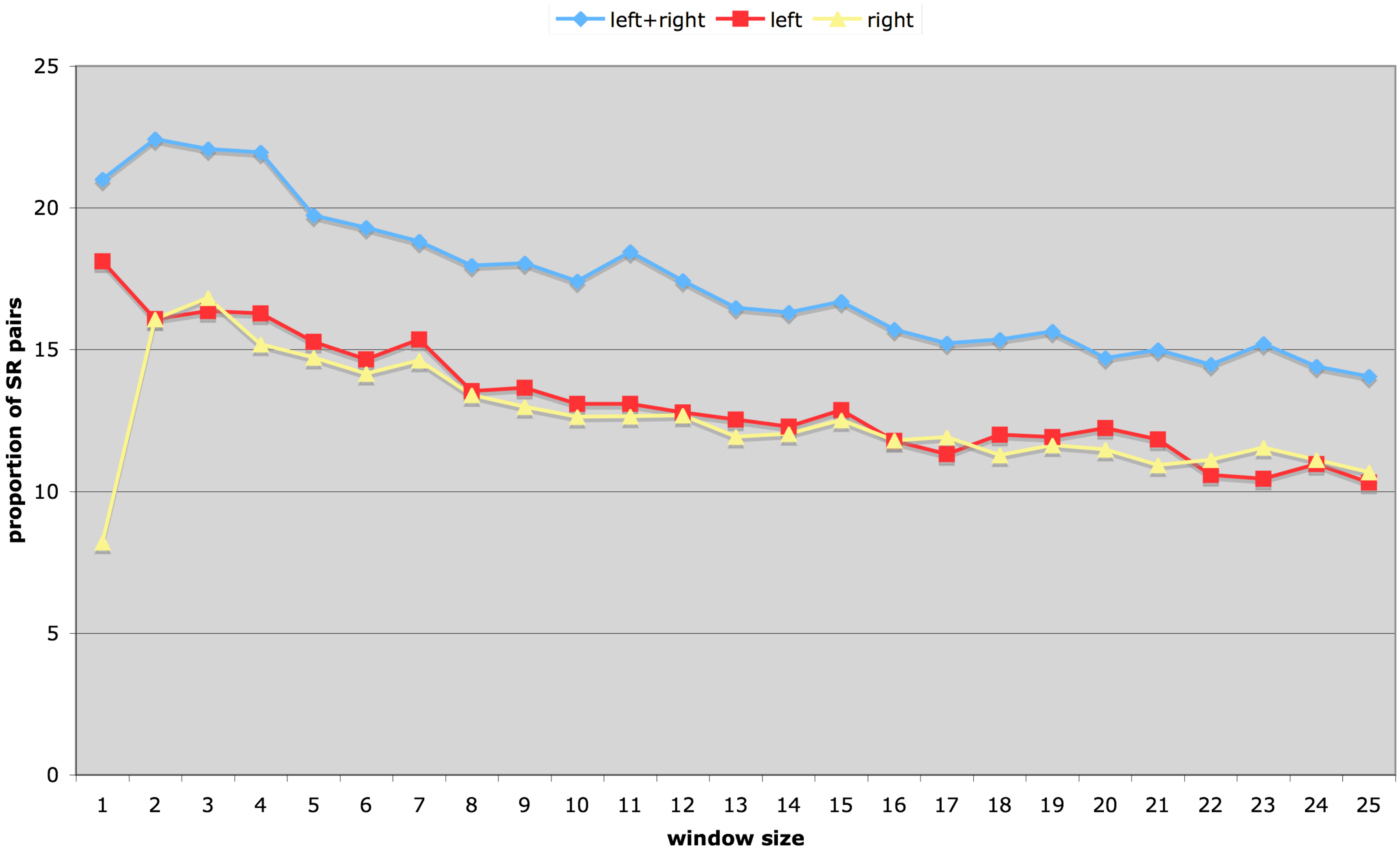
Experiment 3: Window Direction

- So far, context window conflates over responses preceding vs. following the target.
- Some views suggest that stimuli elicit continuations rather than preceding text, e.g. Plaut (1995).
- Church and Hanks (1990) included search direction into co-occurrence model, accounting for association pairs in fixed order (e.g., *bread and butter, sit on)*.
- Are certain window positions prominent for a particular type of SR relationship?
- Co-occurrence strength threshold of 5, corrected.

Experiment 3: Window Direction



Experiment 3: Window Direction



Experiment 3: Window Direction

- More responses precede than follow their targets.
- Difference emerges in window position 1:
  over-utilisation of position immediately preceding target,
  under-utilisation of position immediately following target.
- Pattern runs counter to hypothesis that targets trigger the production of continuations.
- Experiment should further distinguish parts-of-speech.

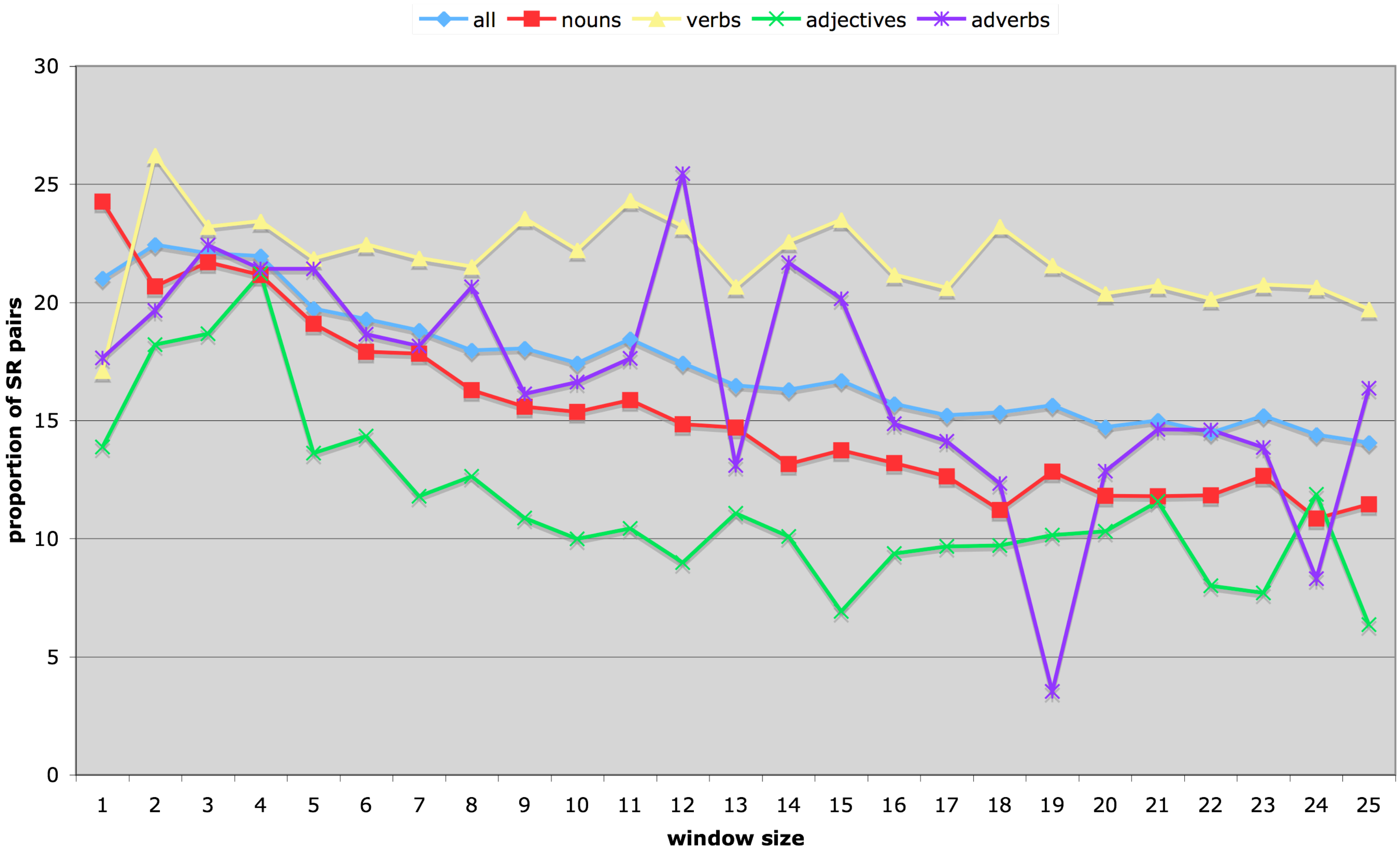
Exp 4:
Response
Part-of-Speech

Experiment 4: Response Part-of-Speech

- Are SR pairs more likely to co-occur in the corpus when the response comes from a particular part-of-speech?
- Co-occurrence strength of parts-of-speech
- Co-occurrence distribution of parts-of-speech, e.g. nouns in argument positions
- Preprocessing step: automatic assignment of POS, relying on an empirical dictionary (Schulte im Walde, 2003)

V	N	ADJ	ADV
34%	56%	7%	1%

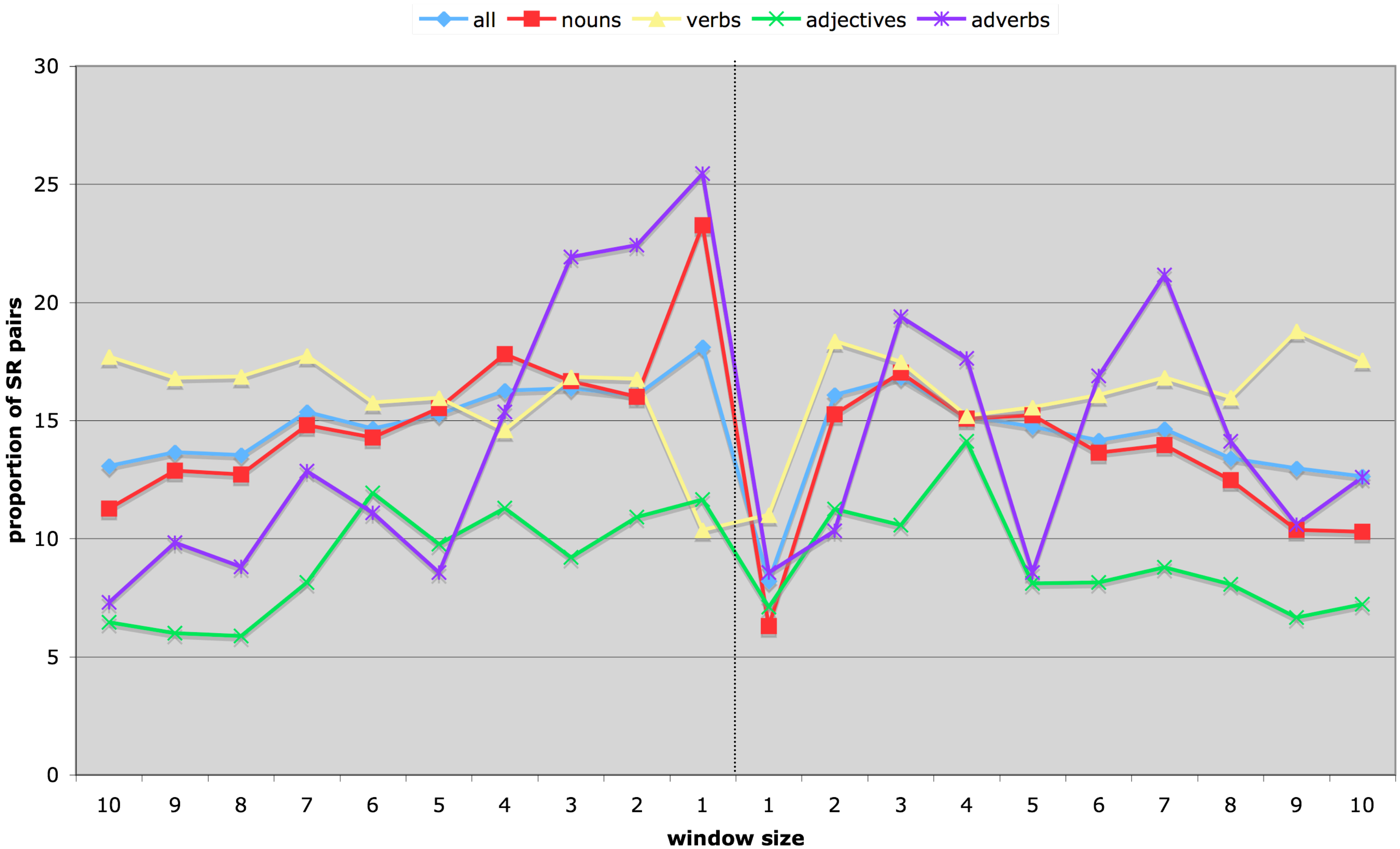
Experiment 4a: Response Part-of-Speech



Experiment 4a: Response Part-of-Speech

- Nouns peak at ±1 words (adjacency)
- Verbs peak in ±2 words
- Adjectives peak at ±4 words
- Adverbs have several ups and downs
- Differences in POS distributions also in later windows

Experiment 4b: Response Part-of-Speech



Experiment 4: Response Part-of-Speech

- **Noun** responses often occur directly before target verbs, and seldom directly but nevertheless close after.
  Co-occurrence rates of nouns decrease in both directions.
  → NPs directly preceding/following verbs
- Distribution of **verb** responses peaks at -2 and +2 words.
  Verbs have strong co-occurrence rates across windows.
  → conjunction/subcategorisation in either order
- **Adjectives** peak at +4 words, decrease in larger windows
  → position within NPs following verbs
- **Adverbs** peak at -1 words, but occur across windows.
  → high frequency, modify many verbs, flexibility in position

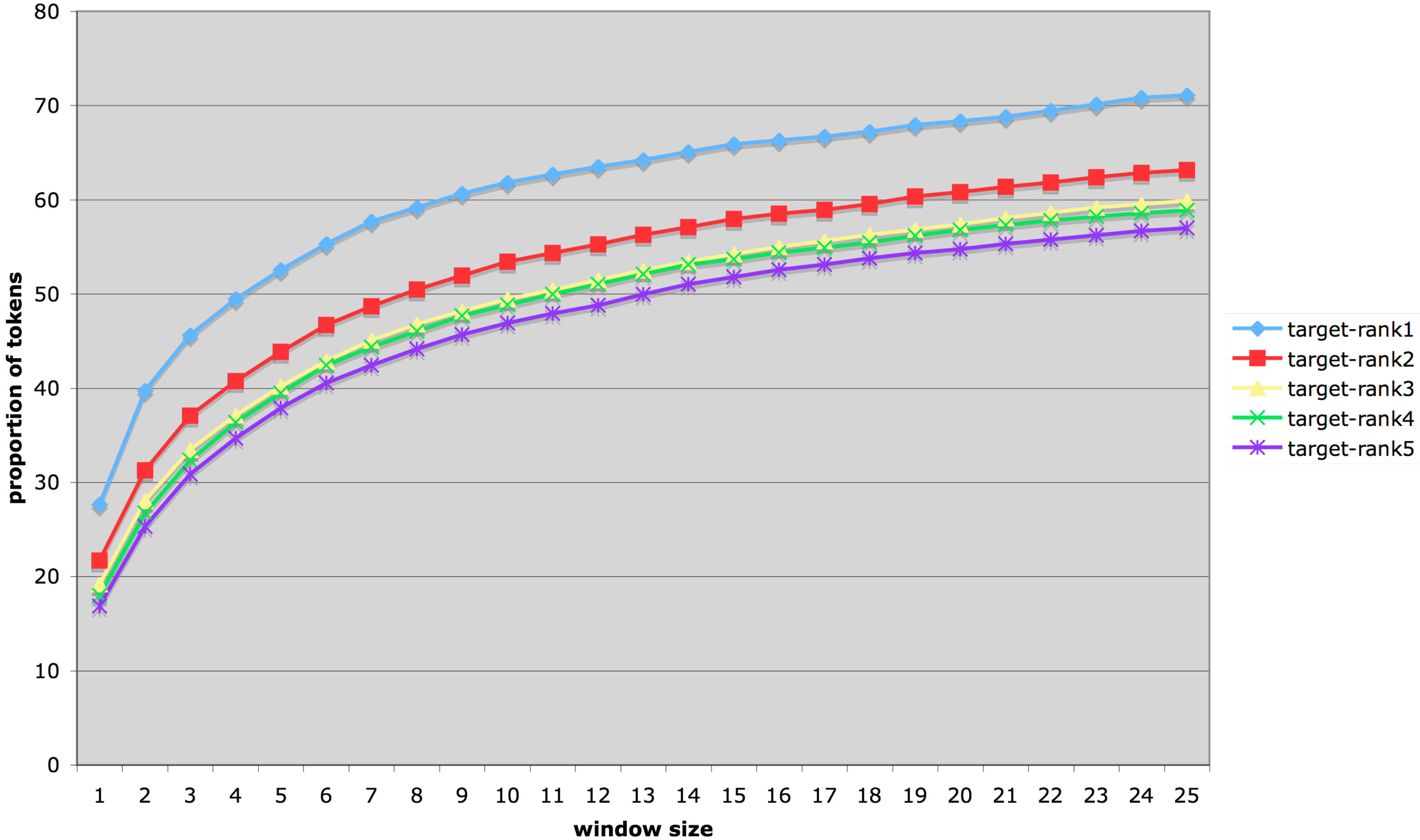
Exp 5:
Association Chains

Experiment 5: Association Chains

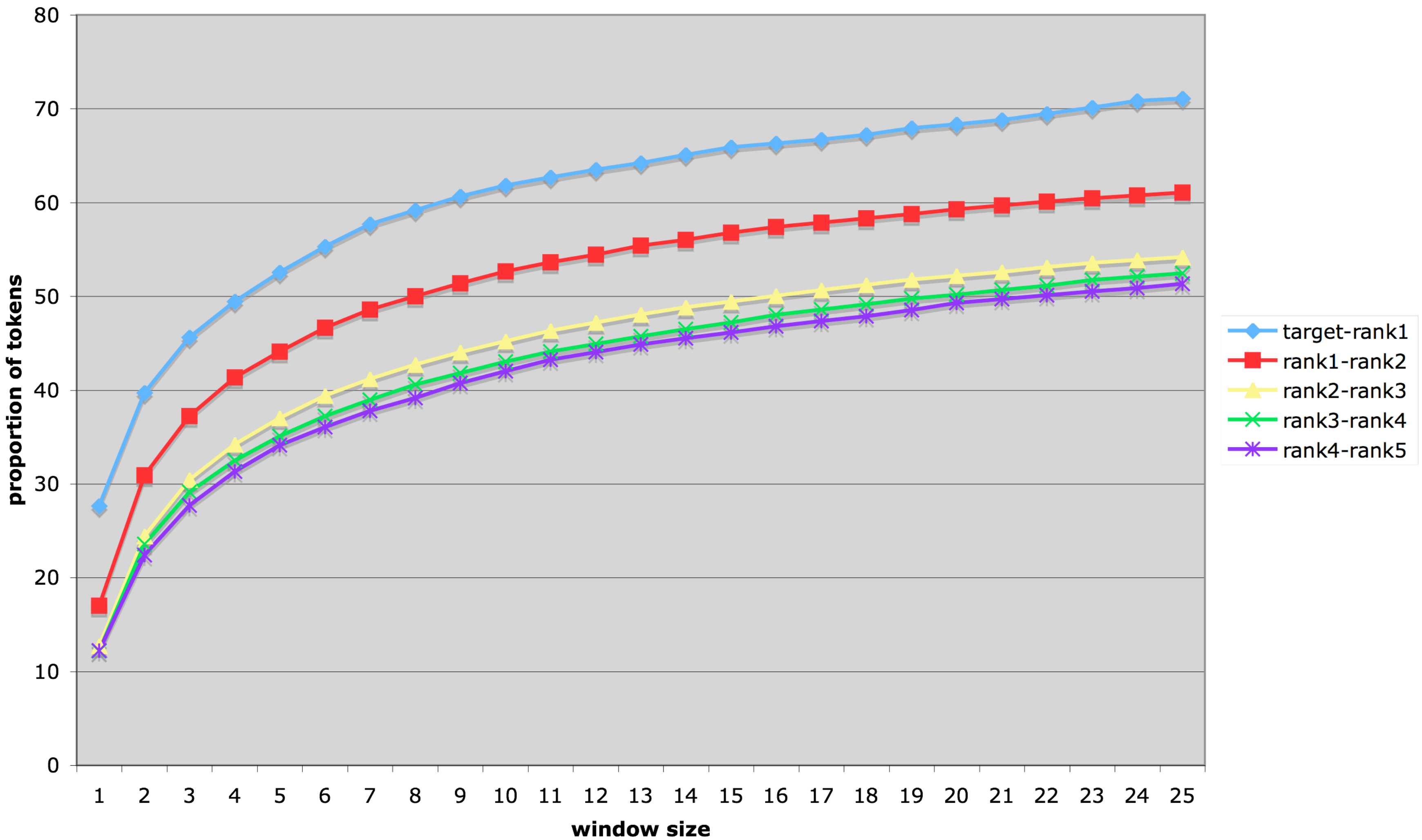
- Single vs. multiple responses to stimuli
- Association chain effect: n^{th} response is associated to the $(n-1)^{th}$ response rather than the stimulus;

  example: *storm* → *lightning*, *Zeus*, ...
- To what extent are *n+1* responses linked to the n^{th} responses rather than to the target, as indexed by co-occurrence rates?
- Use first five responses instead of first only

Experiment 5: Association Chains



Experiment 5: Association Chains



Experiment 5: Association Chains

- First response exhibits stronger co-occurrence patterns with target than any of the later responses.
- Difference mostly due to small windows.
- Similar patterns (and values) for *rankX-rankY* and for *target-rankY.*
- Later responses are related, via co-occurrence, to their n-1 responses, but they are still as related to the target.
- Thus, multiple responses could provide a richer picture of target semantics than single responses only, by indexing additional meaning components.

Conclusions

- Basic experiment + correction
- Functional relationships between stimuli and responses
- Association chain effects

- Cognitive Science: more complete picture of the co-occurrence distributions of semantic associates
- Computational Linguistics: combining part-of-speech distinctions of word-word pairs with positional information (window distances, syntactic functions) might improve automatic acquisition of semantic word-word relations