Distributional characterization of constructional meaning

Florent Perek
Albert-Ludwigs-Universität Freiburg / Université Lille III
florent.perek@gmail.com

Corpus Linguistics 2009 – Liverpool – 23th July 2009

Overview

- Goal
  - Design and test ways to derive the meaning of grammatical constructions from corpus data
- Outline
  - Theoretical background
  - First corpus study
    - Distributional comparison of constructions
    - Limits of this approach
  - Follow-up study based on LSA
  - Conclusion and prospects

Theoretical background

- Grammar in Cognitive Linguistics
  - Grammar = inventory of form-meaning pairs
  - No principled separation between syntax and lexicon
  - Syntactic patterns = form-meaning pairs
    - Argument Structure Constructions (Goldberg 1995, 2006)
    - Syntactic meaning
      - = schematized experience of an event type: transfer, movement, change of state, …
      - Most evident when verbs are used creatively e.g. *John sneezed the napkin off the table*
      - Predicts which verbs are allowed
    - Experimental evidence for constructions

Theoretical background

- Example: the ditransitive construction
  (from Goldberg 1995)

  e.g. *Mary gave her sister a penny.*
  Sam kicked Peter the ball.

  Syntax: Subject$_{\text{Agent}}$ V Object1$_{\text{Recipient}}$ Object2$_{\text{Theme}}$
  Semantics: Agent CAUSES Recipient TO RECEIVE Theme

Theoretical background

- The origin of constructional meaning
  - Current hypothesis: abstraction of verbal meaning
    - e.g. NP V NP NP (ditransitive)
      occurs with verbs of transfer: *give*, *send*, *hand*, …
      => the syntactic pattern is associated to a transfer meaning
    - Evidenced by the facilitating factor of a biased distribution (cf. Goldberg et al. 2004)
  - We investigated this idea with two corpus studies

Study 1

- Study 1
  - Hypothesis: distributional distance correlate with semantic distance
    - Supported by:
      - Goldberg's model: constructional meaning constrains the verbs occurring in the construction
      - Corpus studies
        - Stefanowitsch and Gries (2003), Gries et al. (2005)
        - The strongest collocates have the meaning closest to that of the construction
    - Constructions with different meaning should have different distributions, and conversely
    - Distributional differences should reflect semantic differences between constructions

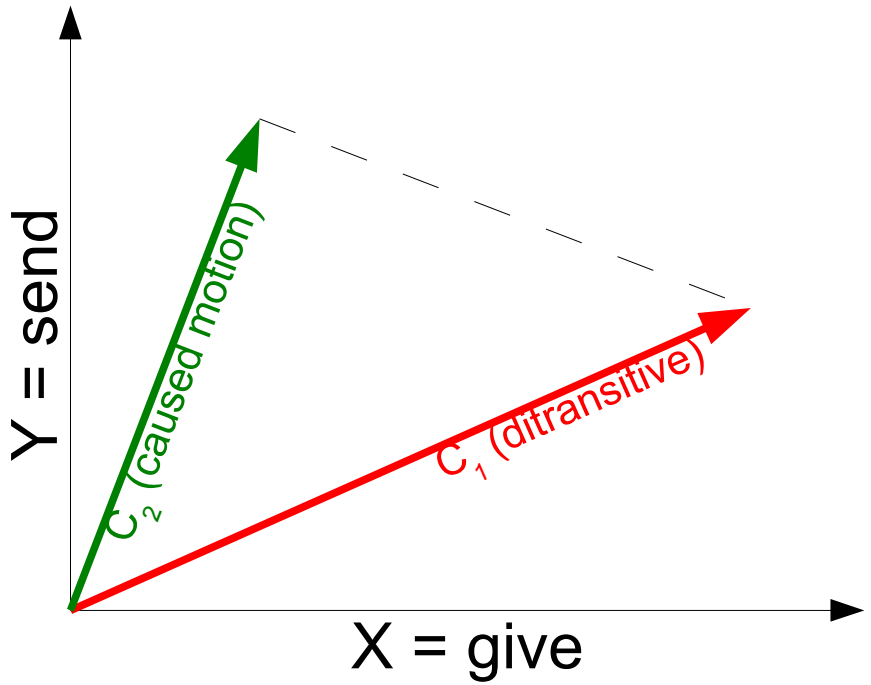
Study 1

- How to compare distributions?
  - Vector space approach to distributional similarity
    - Verbal distribution = vector (of frequencies)
    - The verbs define a multidimensional space
    - Distributional distance = distance between vectors

Study 1

- Simple example with only 2 dimensions

Verb	Frequency in the ditransitive	Frequency in the caused motion
give	153	34
send	12	18



$$\vec{ditransitive} = \vec{C}_1 = \begin{bmatrix} 153 \\ 12 \end{bmatrix}$$

$$\vec{caused motion} = \vec{C}_2 = \begin{bmatrix} 34 \\ 18 \end{bmatrix}$$

$$similarity = \cos(\widehat{\vec{C}_1, \vec{C}_2})$$

$$\cos(\widehat{\vec{C}_1, \vec{C}_2}) \in [0,1]$$

1 => identity
the closer to 0, the more different

Study 1

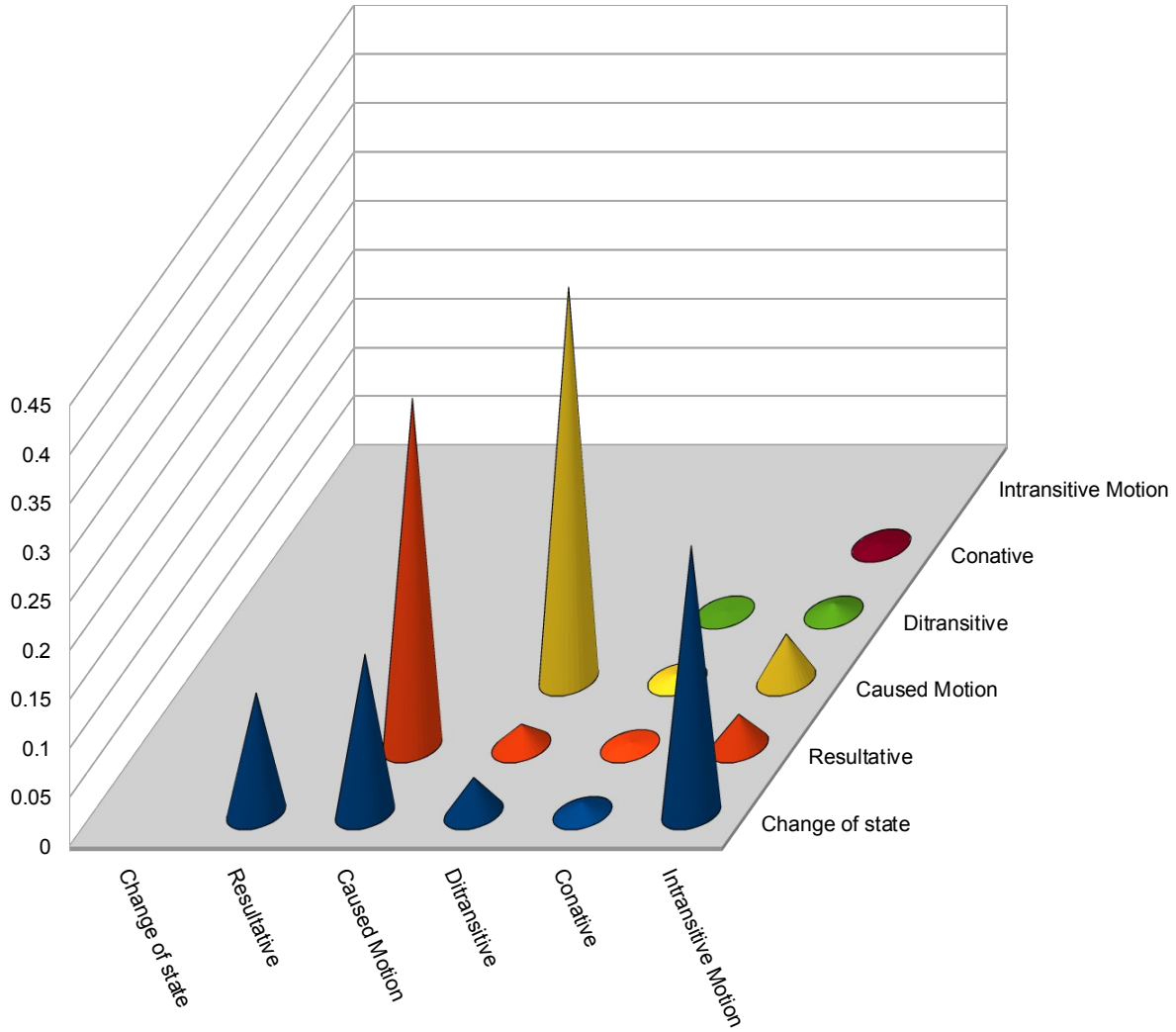
- Six constructions under study

Construction	Semantics	Syntax	Example	Freq.
Intransitive Motion	X_{Theme} MOVE Y_{Path}	Subject-Verb-Oblique$_{Path}$	The ball rolled down the hill	564
Change of state	X_{Theme} BECOME Y_{State}	Subject-Verb-Oblique$_{State}$	The pond froze solid	471
Caused motion	X_{Agent} CAUSE $Y_{Patient}$ TO MOVE Z_{Path}	Subject-Verb-Object-Oblique$_{Path}$	Bill broke the hell into the bowl	290
Resultative	X_{Agent} CAUSE $Y_{Patient}$ TO BECOME Z_{State}	Subject-Verb-Object-Oblique$_{State}$	Bill watered the tulip flat	175
Ditransitive	X_{Agent} CAUSE $Y_{Recipient}$ TO RECEIVE Z_{Theme}	Subject-Verb-Object1-Object2	Joe painted Sally a picture	307
Conative	X_{Agent} DIRECT ACTION AT Y_{Target}	Subject-Verb-Oblique$_{at}$	Bill kicked at the ball	178

- Manually identified in the ICE-GB (spoken)
- Extraction of the verbal distribution

Study 1

- Comparison of all six constructions

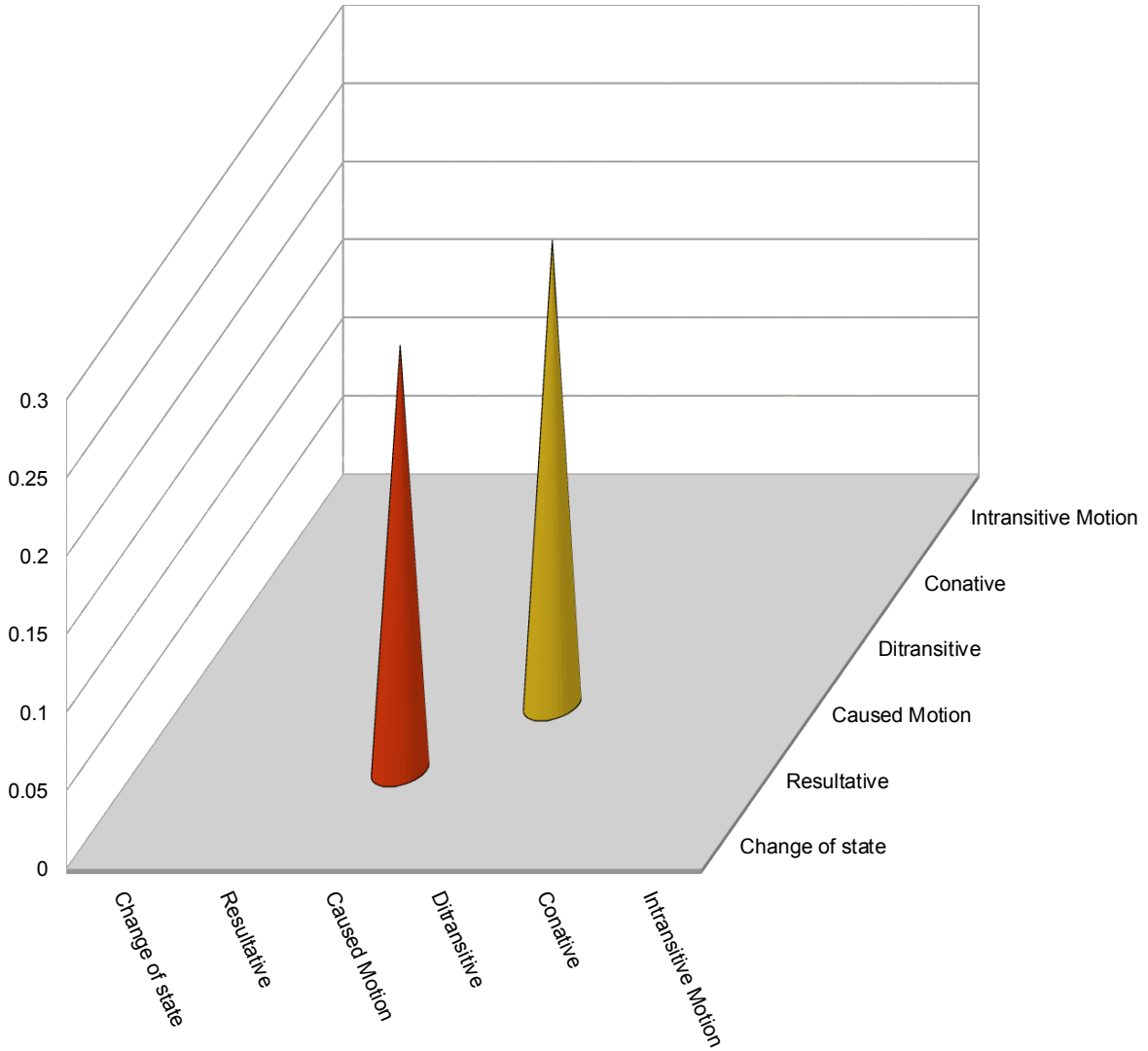


Study 1

- Possible improvement
  - Collexeme analysis (Stefanowitsch and Gries 2003)
    - Some verbs are more important collocates for some constructions than for others
    - Not rendered by raw frequencies
    - Collostruction strength instead of frequencies

Study 1

- With collexeme analysis



Interim conclusion

- Interim conclusion
  - Do reflect relevant meaning similarities
    - Change of state / change of location (cf. Goldberg & Jackendoff 2004)
    - Caused motion / transfer
  - But purely distributional
    - Does not take the meaning of verbs into account
    - Would the result be different if we do?
    - Two issues:
      - Representation of verbal meaning
      - Representation of constructional meaning derived from the former

Interim conclusion

- How to represent word meaning in a corpus?
  - Semantic annotations (e.g. WordNet ids)
  - Distributional characterization
- Latent Semantic Analysis
  - Used in data mining
  - Based on co-occurences of words in documents
  - Correlates with human judgements on semantic similarity, cf. Landauer, Foltz and Laham (1998)
  - Often used as an objective measure of semantic similarity, e.g. in Bencini & Goldberg (2000)

Interim conclusion

- Word meaning in LSA = a vector
  - Semantic distance = distance between vectors
  - Construction vectors can be derived from the vectors of the verbs in its distribution

$$\vec{C} = \sum (freq(V) \times \vec{V})$$

e.g. the ditransitive

Verb	Frequency
allow	2
ask	6
⋮	
write	1

$$\vec{ditransitive} = 2 \times \vec{allow} + 6 \times \vec{ask} + \ldots + 1 \times \vec{write}$$

- Sum of vectors = vector in the same space
- Same distance metrics than for words

Study 2

- Study 2
  - Corpus submitted to LSA
  - Two representations of constructional meaning:
    - Vectors calculated from word vectors + distribution
    - Vector of the construction symbol (=word) artificially inserted in the corpus
  - Questions:
    - To what extent do symbolic meaning and distributionally derived meaning correlate in the LSA space?
    - How semantic differences between constructions are reflected if verbal meaning is taken into account?

Study 2

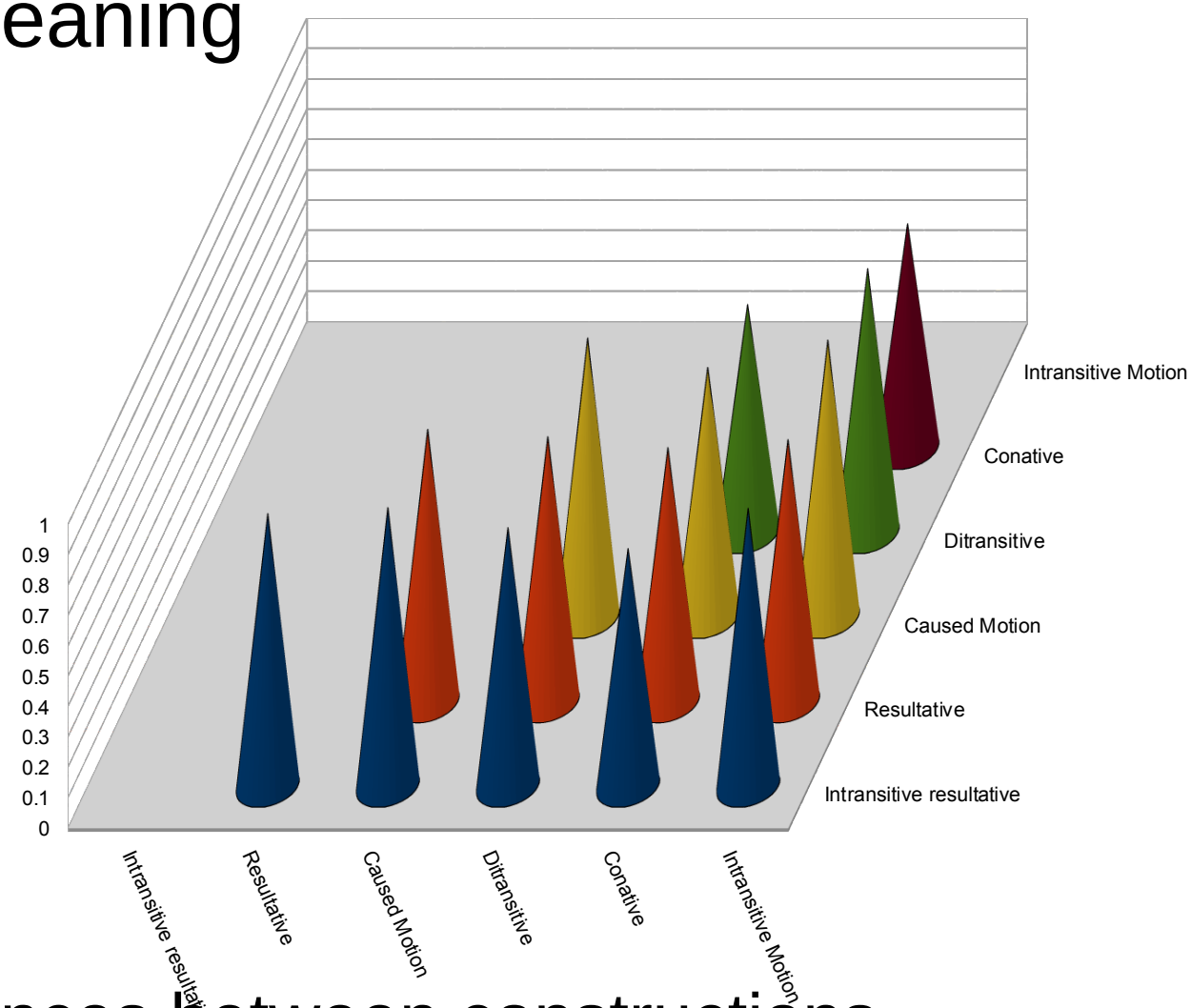
- Comparison of symbolically derived *vs.* distributionally derived constructional meaning
  - They are strongly similar

Construction	Similarity between construction symbol and distribution vector
Change of state	0.7981
Resultative	0.7391
Caused motion	0.7378
Ditransitive	0.7157
Conative	0.7649
Instransitive Motion	0.8322

- Seems to validate the view that constructional meaning originates with verbal meaning

Study 2

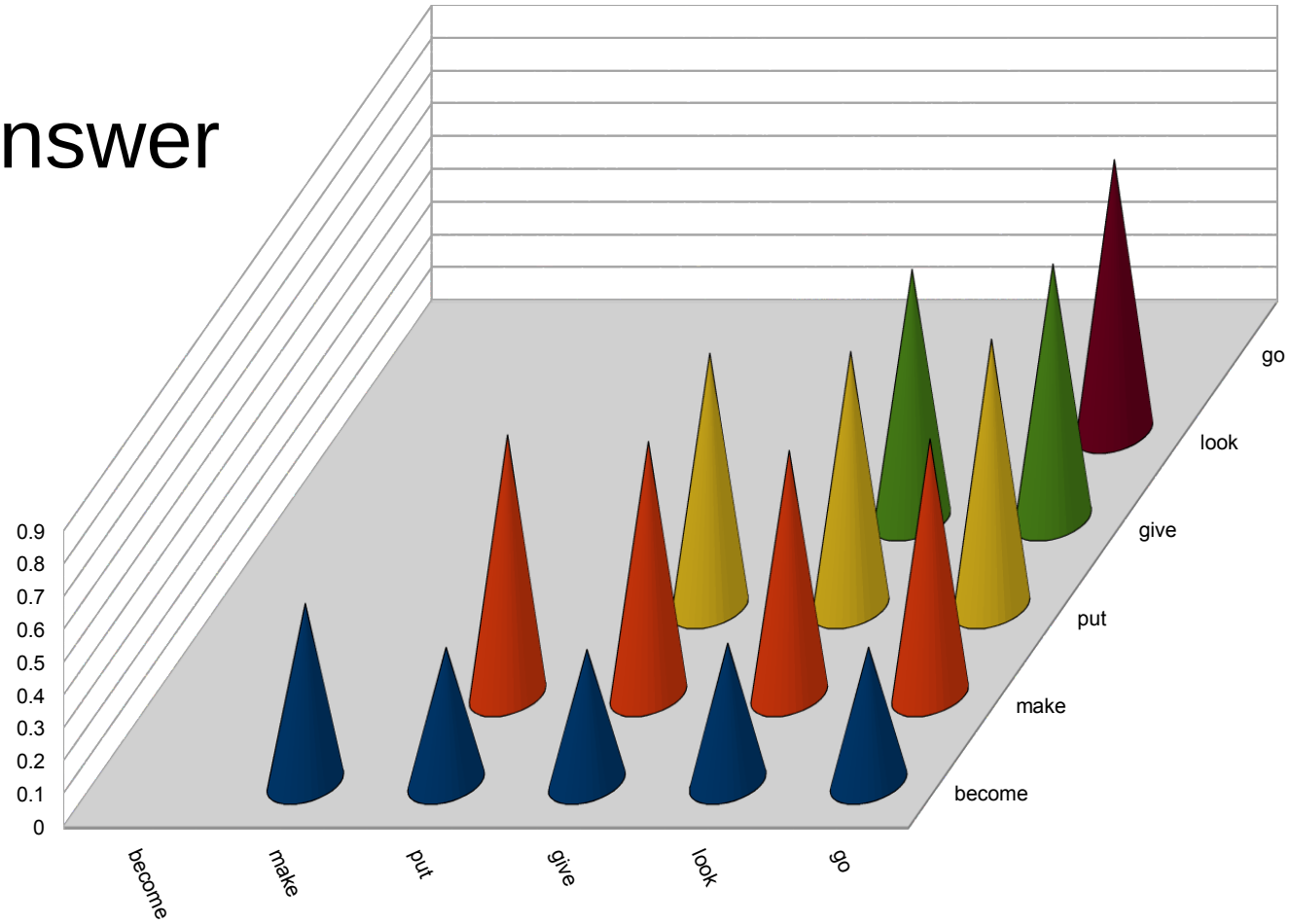
- Comparison of distributionally derived constructional meaning



  - No sharp differences between constructions

Study 2

- A possible answer

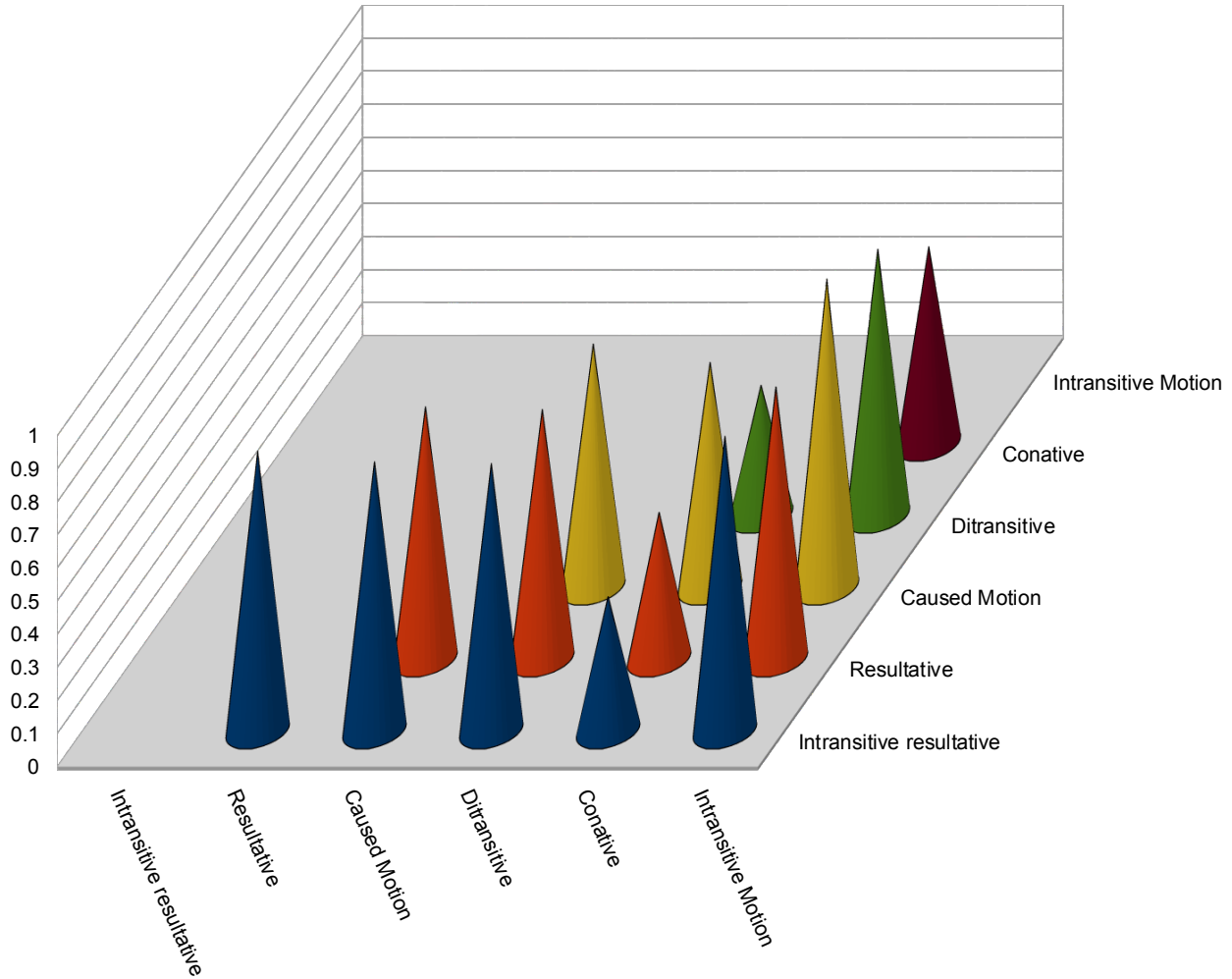


  - The most frequent words are not distinctive enough
  - They appear in two wide a range of contexts
  - LSA does not capture sharp semantic differences

Study 2

- Solution: use “stopwords”
  - Frequent verbs are simply ignored in the analysis
  - Semantic differences between constructions are captured by the less frequent verbs

Study 2



- A slight improvement
  - Conative evidently different from other constructions
  - But still no clear differences

Conclusion

- Conclusion
  - Distribution captures semantic differences between constructions
  - Less so clear with an account of verbal meaning
    - Seems to capture the meaning of constructions
    - But not semantic differences between constructions
- Prospects
  - Bigger corpus to derive word vectors from
  - Use dictionary-based semantic distance
    - more reliable, especially for highly frequent verbs

References

Bencini, G. and Goldberg, A. (2000). The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language* 43(4):640–51.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language.* Oxford University Press, Oxford.

Goldberg, A. E., D. Casenhiser, and N. Sethuraman (2004). Learning argument structure generalizations. *Cognitive Linguistics* 15(3):289–316.

Goldberg, A. E. and Jackendoff, R. (2004). The English resultative as a family of constructions. *Language* 80(3):532–568.

Gries, S. T., Hampe, B., and Schönefeld, D. (2005). Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4):635–76.

Landauer, T. K., P. W. Foltz and D. Laham (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes* 25(2&3):259–284.

Perek, F. (2008). Towards a constructional approach to automatic argument structure acquisition: the case of oblique phrases. Master's thesis, Université Charles de Gaulle Lille III, Villeneuve d'Ascq, France.

Stefanowitsch, A. and S. T. Gries (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2:209-43.