Towards Tracking Semantic Change by Visual Analytics

Christian Rohrdantz[1] Annette Hautli[2] Thomas Mayer[2]
Miriam Butt[2] Daniel A. Keim[1] Frans Plank[2]

Department of Computer Science[1] Department of Linguistics[2]
University of Konstanz

June 21, 2011

Motivation

1. increasing amount of diachronic data electronically available
2. demand of historical linguists to process these corpora and see developments and patterns over time at-a-glance

Motivation

1. increasing amount of diachronic data electronically available
2. demand of historical linguists to process these corpora and see developments and patterns over time at-a-glance

Challenge

Tracking of overall developments of language and also allowing to delve into the details of the data.

Motivation

1. increasing amount of diachronic data electronically available
2. demand of historical linguists to process these corpora and see developments and patterns over time at-a-glance

Challenge

Tracking of overall developments of language and also allowing to delve into the details of the data.

Research question

Can we create tools that aid during the analysis of language change, can they test existing hypotheses of change and can they even generate new ones?

Research object

The object under investigation is semantic change (here: in English)

But what is semantic change?

- if a word changes its meaning over time, it has undergone semantic change.
- some types of semantic change:
  - *narrowing* (the meaning of a word becomes restricted), e.g. skyline
  - *widening* (the meaning of a word widens), e.g. horn
- semantic change in the last 20 years: words related to the computer and the internet

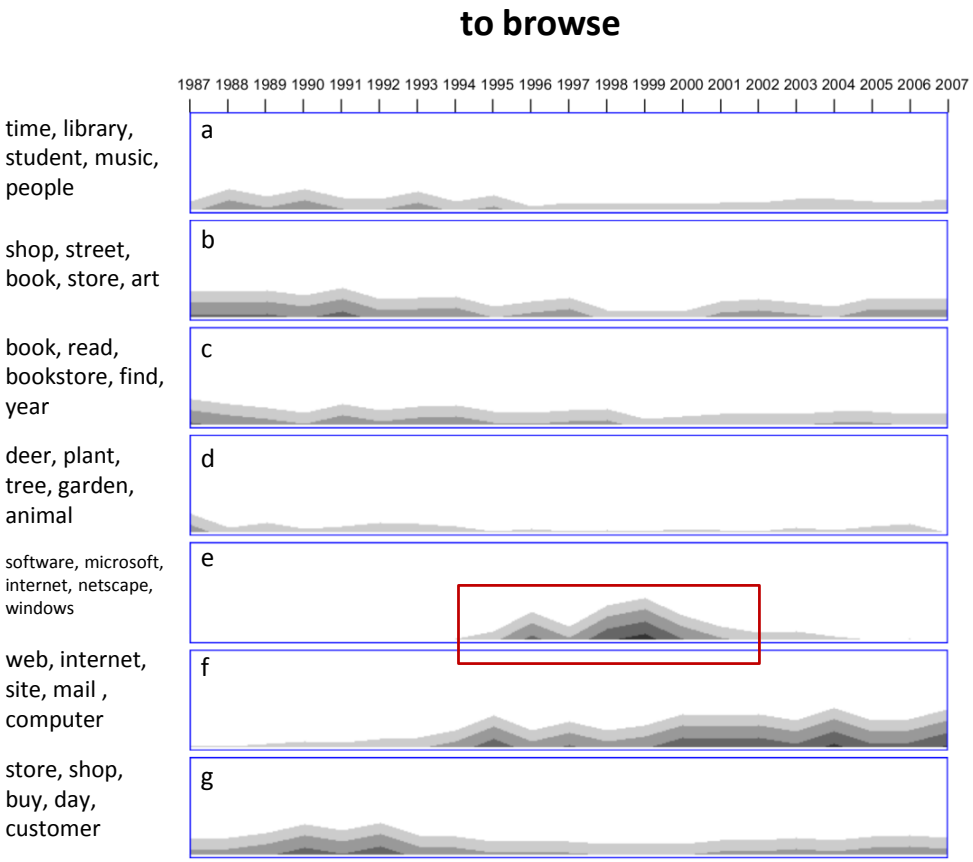
Methodology

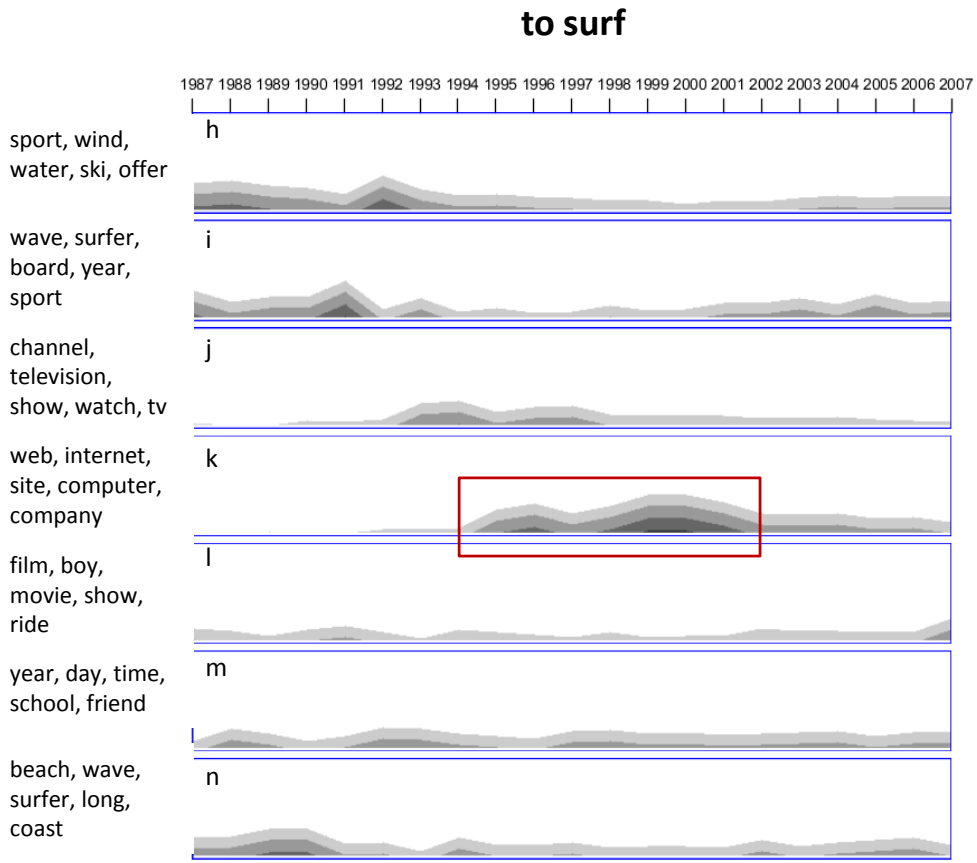
- search New York Times corpus
  - 1.8 million newspaper articles from 1987 to 2007
  - each article has a specific time stamp
- extract context of 25 words before and after the lexical item under investigation
- use statistics to model word senses on the basis of word contexts
  - Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
    - ★ not applied on documents but on contexts
  - we predefine the number of senses, each context is assigned to one sense
- add a visualization layer that graphically interprets the information from the statistical analysis and makes it accessible to historical linguists

Our visualization approach

First approach

- aggregated view on the data

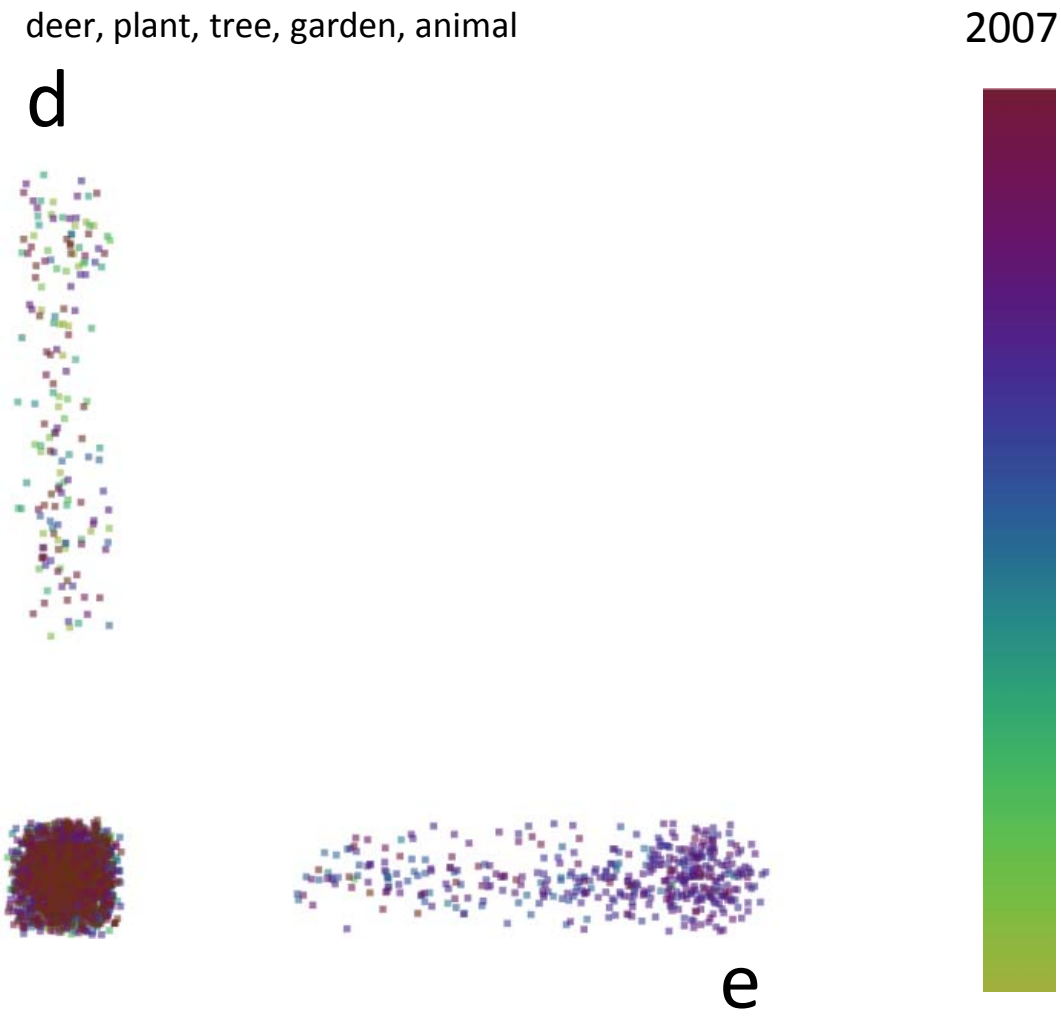




Our visualization approach

Second approach

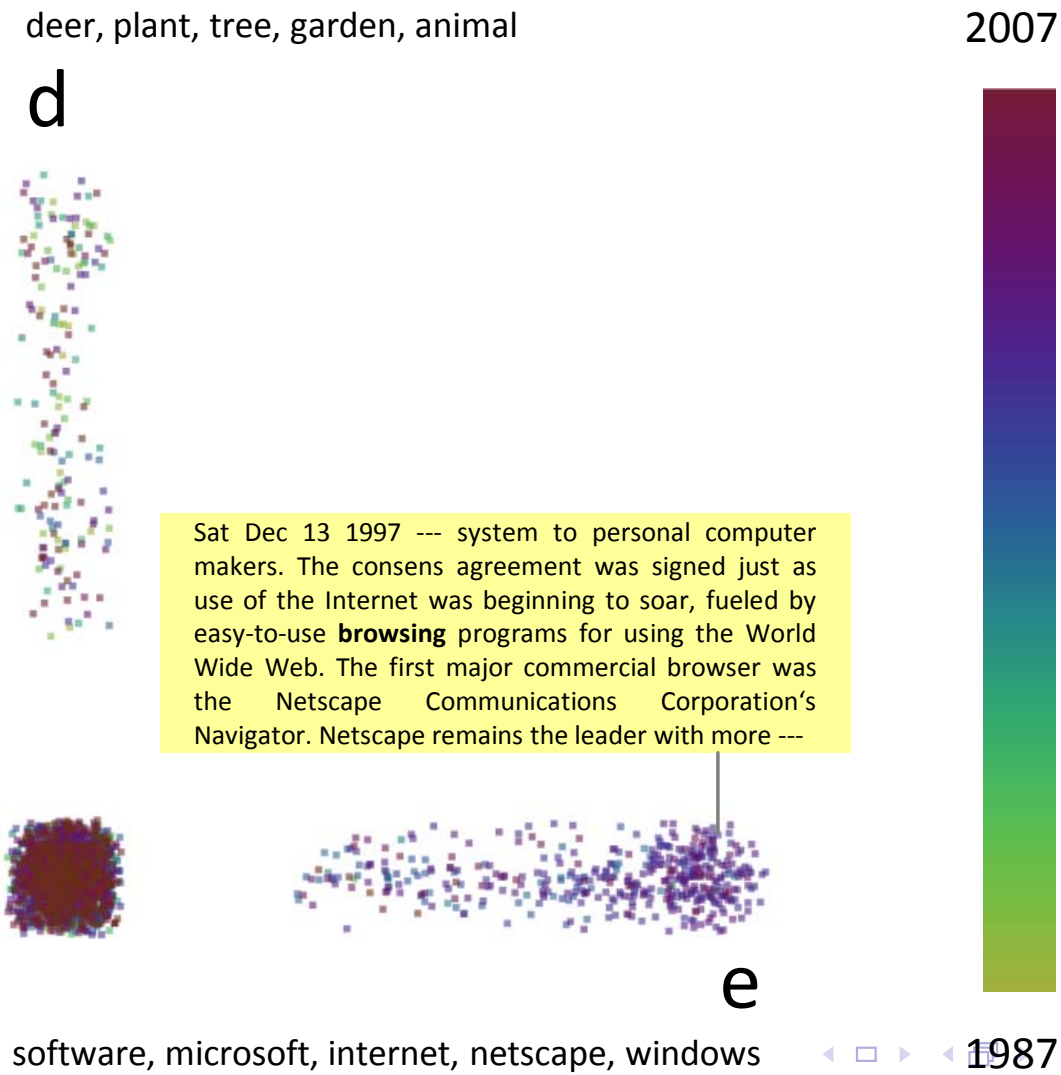
- individual plotting of the contexts of *to browse*



Our visualization approach

Second approach

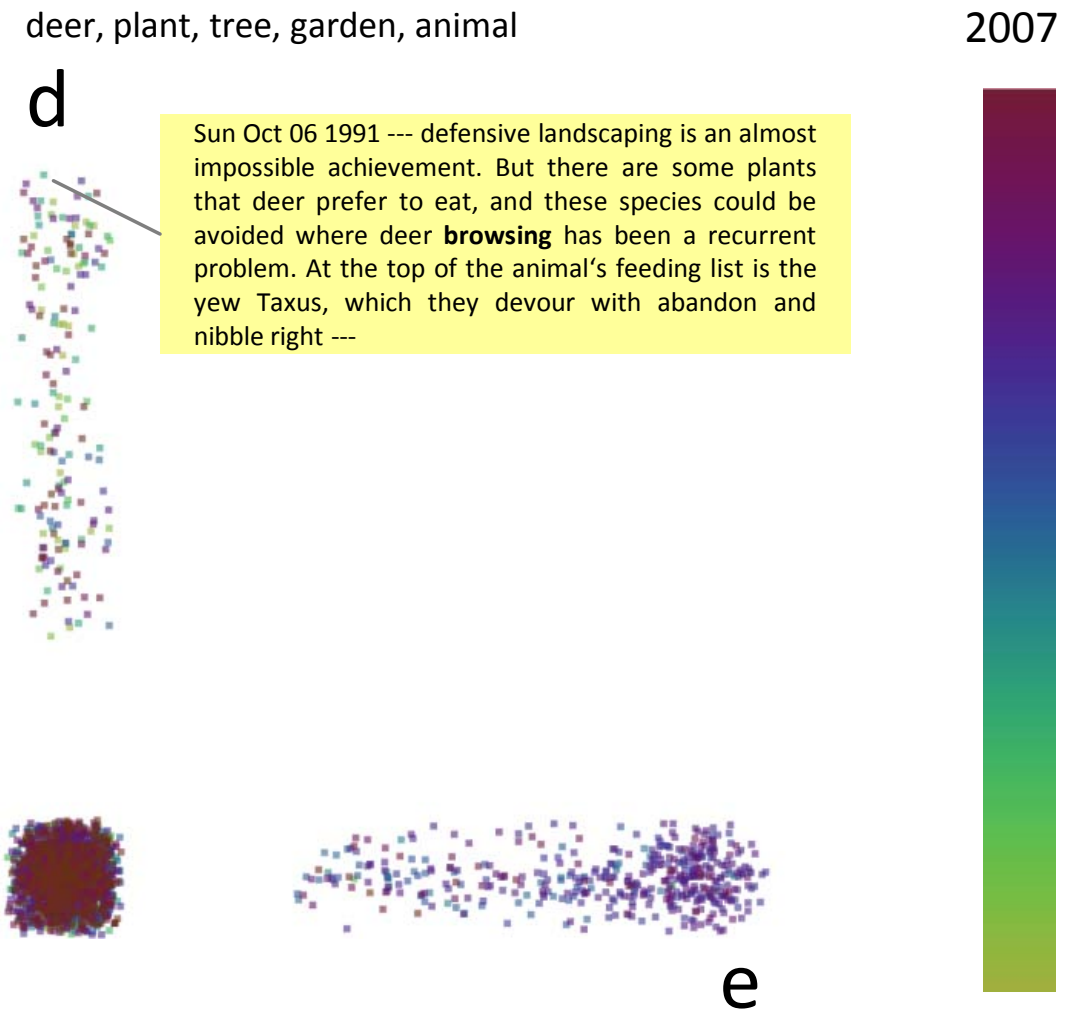
- individual plotting of the contexts of *to browse*



Our visualization approach

Second approach

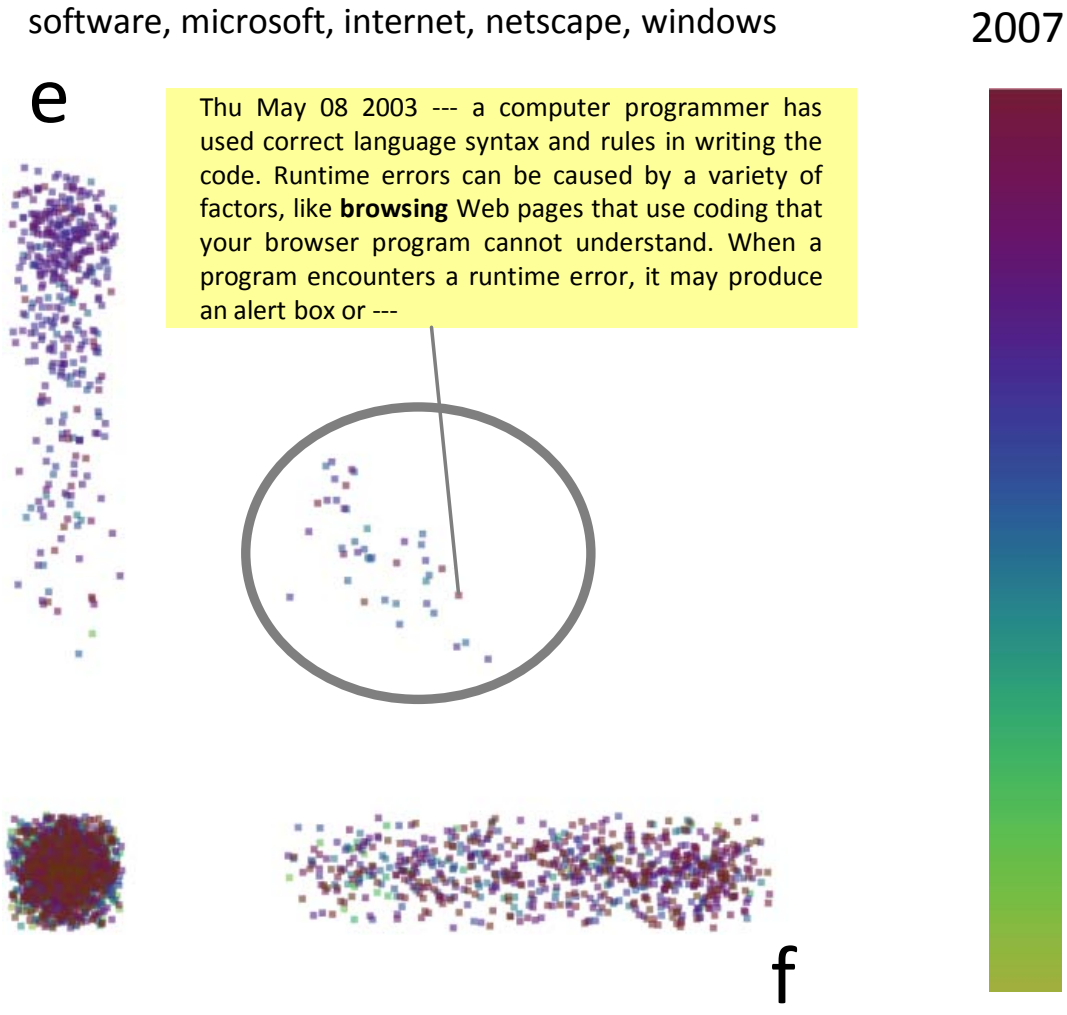
- individual plotting of the contexts of *to browse*



Our visualization approach

Second approach

- individual plotting of the contexts of *to browse*



Evaluation

- generally difficult (if not impossible) to fully evaluate statistical approaches to meaning change
- one attempt: compare the findings from the visualization with information from dictionaries from different time periods
  - Longman Dictionary from 1987 (LONG)
  - WordNet from 1998 (WN)
  - Collins dictionary from 2007 (COLL)

Evaluation

	to browse		to surf		messenger		bookmark	
	# of word senses		*# of word senses*		*# of word senses*		*# of word senses*	
	DIC	VIS	DIC	VIS	DIC	VIS	DIC	VIS
1987 (LONG)	2	3	1	1	1	2	1	1
1998 (WN)	5	4	3	3	1	3	1	2
2007 (COLL)	3	4	3	2	1	4	2	2

Table: Evaluation of visualized senses against dictionary senses

- in general, the number of our senses corresponds to the information coming from the dictionary
- in the case of "messenger" the visualization proves to be even more detailed

Evaluation

	messenger	
	# of word senses	
1987	LONG: a person who brings a message	VIS: bike messenger messenger (genetics)
1997	WN: a person who carries a message	VIS: bike messenger messenger (genetics) religious messenger
2007	COLL: a person who brings a message	VIS: bike messenger messenger (genetics) religious messenger instant messenger

Table: Sense development of *messenger* from 1987 to 2007

Future work

- test the approach on a broader range of terms, texts and languages
- overcome some issues of historical corpora
  - e.g. deal with scriptural variances in diachronic and synchronic data
- provide more ways for interactive visualizations
- enable for parameter tuning
- collapse overlapping senses

Conclusion

- novel and promising interactive visualization approach that
  - facilitates investigations into language change using new technology
  - can verify existing hypotheses about change

Conclusion

- novel and promising interactive visualization approach that
  - facilitates investigations into language change using new technology
  - can verify existing hypotheses about change

Research question

Can we create tools that aid during the analysis of language change, can they test existing hypothesis and even generate new ones?

Conclusion

- novel and promising interactive visualization approach that
  - facilitates investigations into language change using new technology
  - can verify existing hypotheses about change

Research question

Can we create tools that aid during the analysis of language change, can they test existing hypothesis and even generate new ones?

Yes!

Challenge

How can we improve the existing models to make the system more fine-tuned and flexible to other input parameters?

Thank you!

Latent Dirichlet Allocation (LDA)

- topic model developed by Blei, Ng and Jordan (2002)
- instead of classifying documents as topics, we classify contexts as belonging to senses
- each context is assumed to be a mixture of senses (similar to probabilistic latent semantic analysis)
- predefined number of senses (usually topics)
- contexts have probabilities for belonging to certain senses
  - senses are described by key words (as we saw earlier)
  - other contexts with similar keywords are classified as belonging to the same sense with a certain probability