Evaluating Semantic Structure in Distributional Modelling

Thomas Wielfaert, Kris Heylen, Jakub Kozakoszczak,
Leonid Soshinskiy & Dirk Speelman

KU Leuven
Quantitative Lexicology and Variational Linguistics

Purpose of the talk

THEORETICAL

- Study the structure of lexical variation: mapping of meaning onto lexemes in different varieties.
- Analyse how this structure is apparent in usage data

METHODOLOGICAL

- Semantic Vector Spaces as a method for the quantitative, large-scale, corpus-based analysis of lexical semantics
- Interactive Visualisation of distributional models as an exploratory, visual analytic tool for lexicology
- Cluster quality measure for calibration of model parameters.

Overview

1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Cluster quality measures
5. Discussion and future work

Overview

1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Cluster quality measures
5. Discussion and future work

Linguistic Background

POLYSEMY: a word that is used to refer to different concepts. For instance the English noun *underground*[1]:

- *underground*$_1$: a secret group organized to overthrow a government or occupation force
- *underground*$_2$: an electric railway operating below the surface of the ground (usually in a city)

NEAR-SYNONYMY: multiple words that are used to refer to the same concept, e.g. SUBTERRANEAN TRANSPORT:

- *underground*$_2$: electric railway operating below the surface...
- *subway*$_1$: electric railway operating below the surface...
- *tube*$_5$: electric railway operating below the surface...

[1] Wordnet: http://wordnetweb.princeton.edu

Lectal variation

Geographical

- US: *underground*$_1$ *underground*$_2$ *subway*$_1$ *tube*$_5$
- UK: *underground*$_1$ *underground*$_2$ *subway*$_1$ *tube*$_5$

Register

In a London street conversation

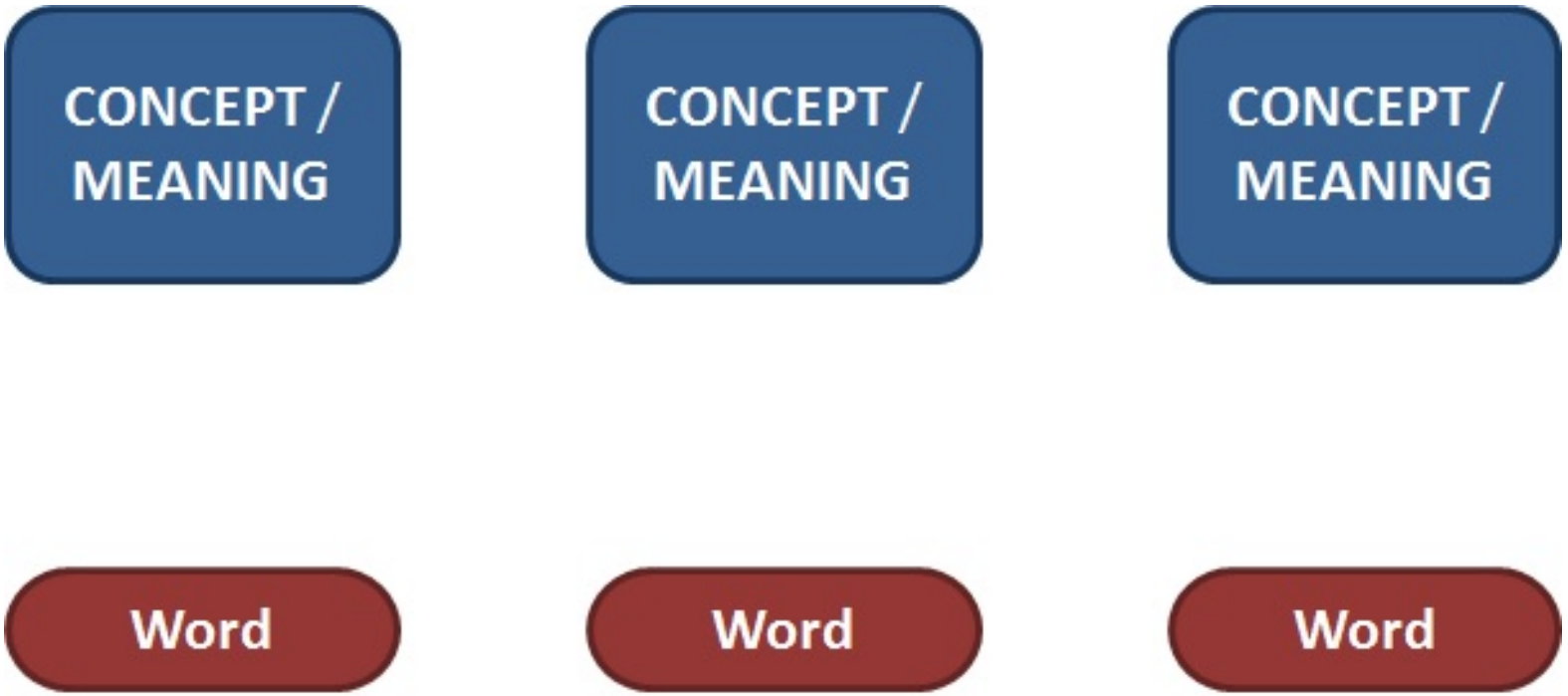
- about subter. transport, *tube* more likely than *underground*
- if *underground* used, meaning transport more likely than *rebel*

In a *Guardian* article

- about subter. transport, *underground* more likely than *tube*
- if *underground* used,meaning rebel more likely than *transport*

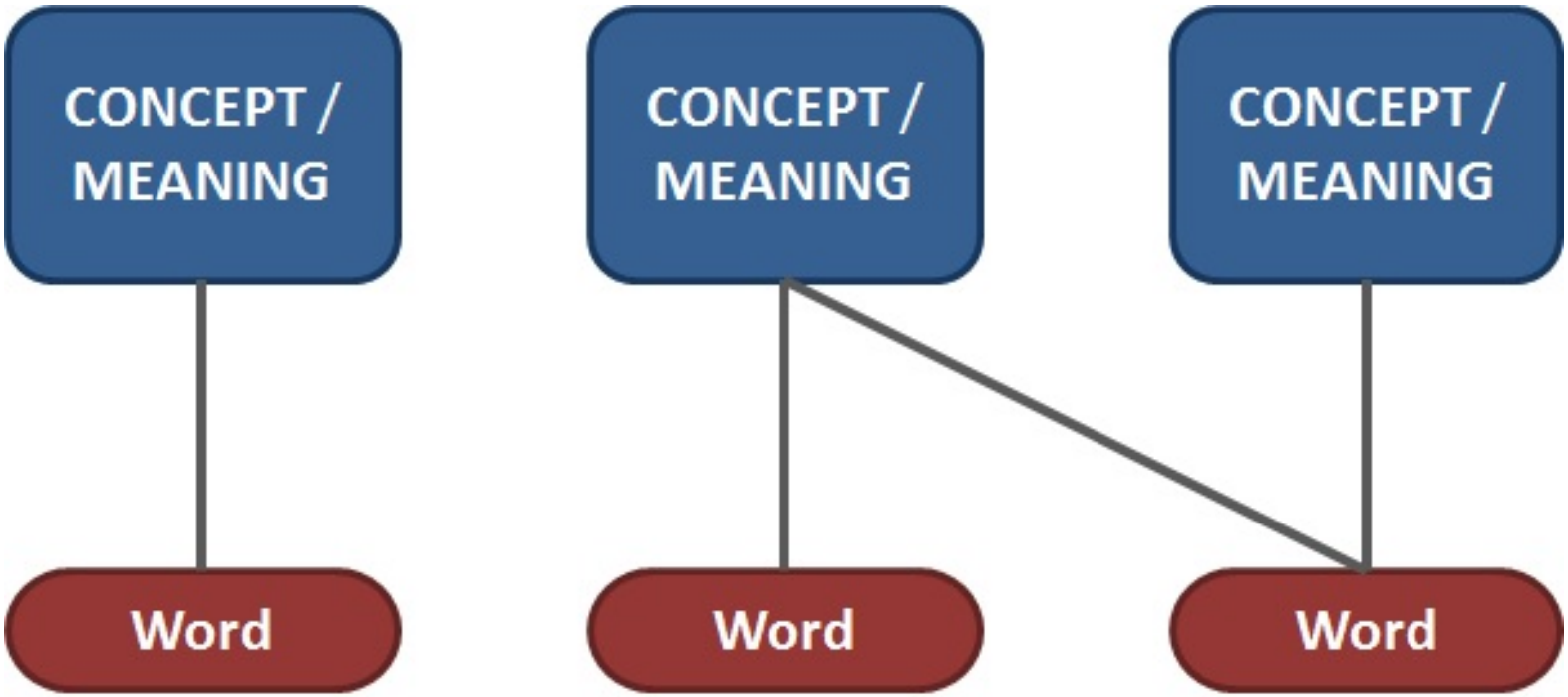
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):



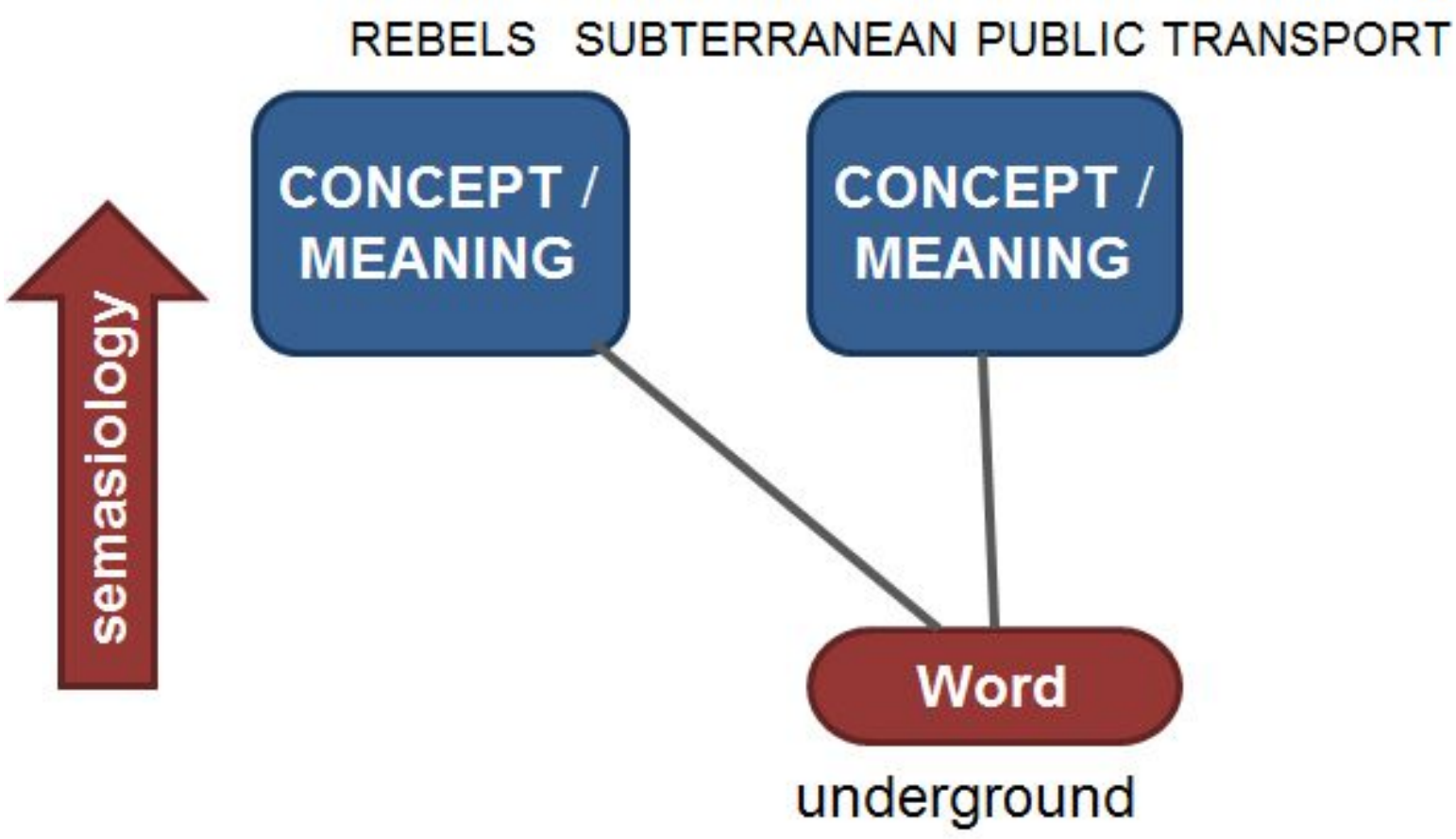
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):



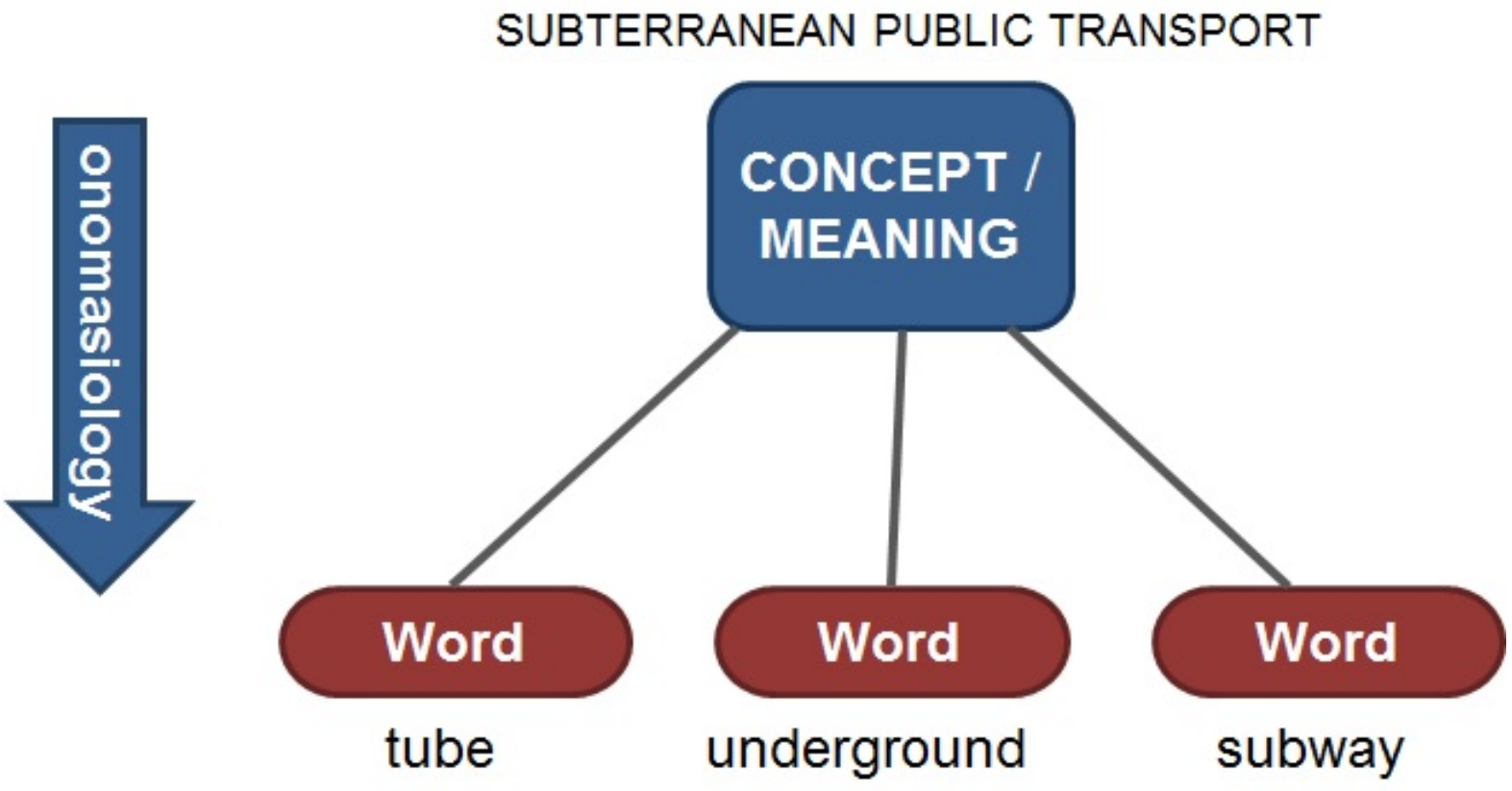
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):



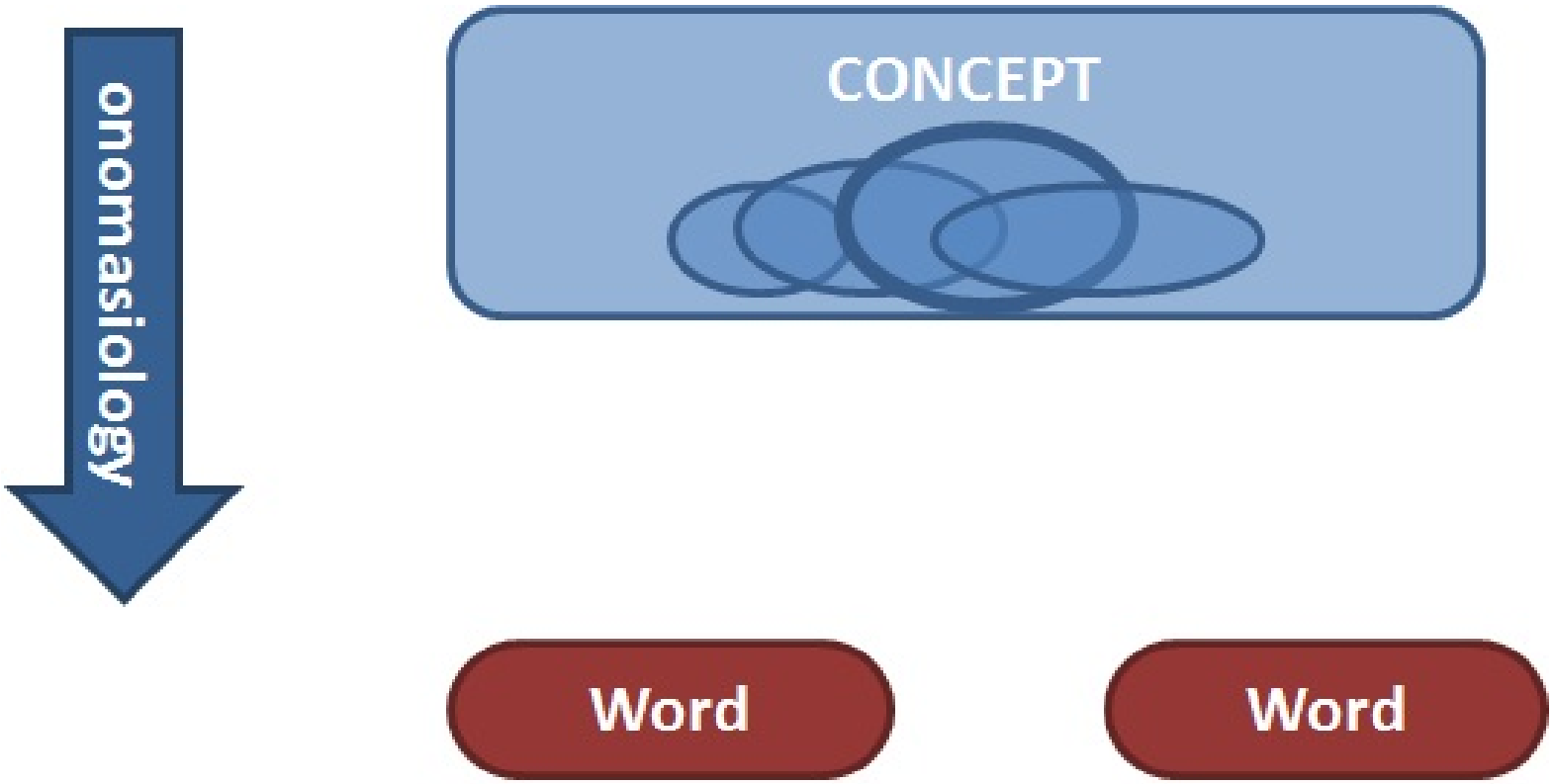
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):



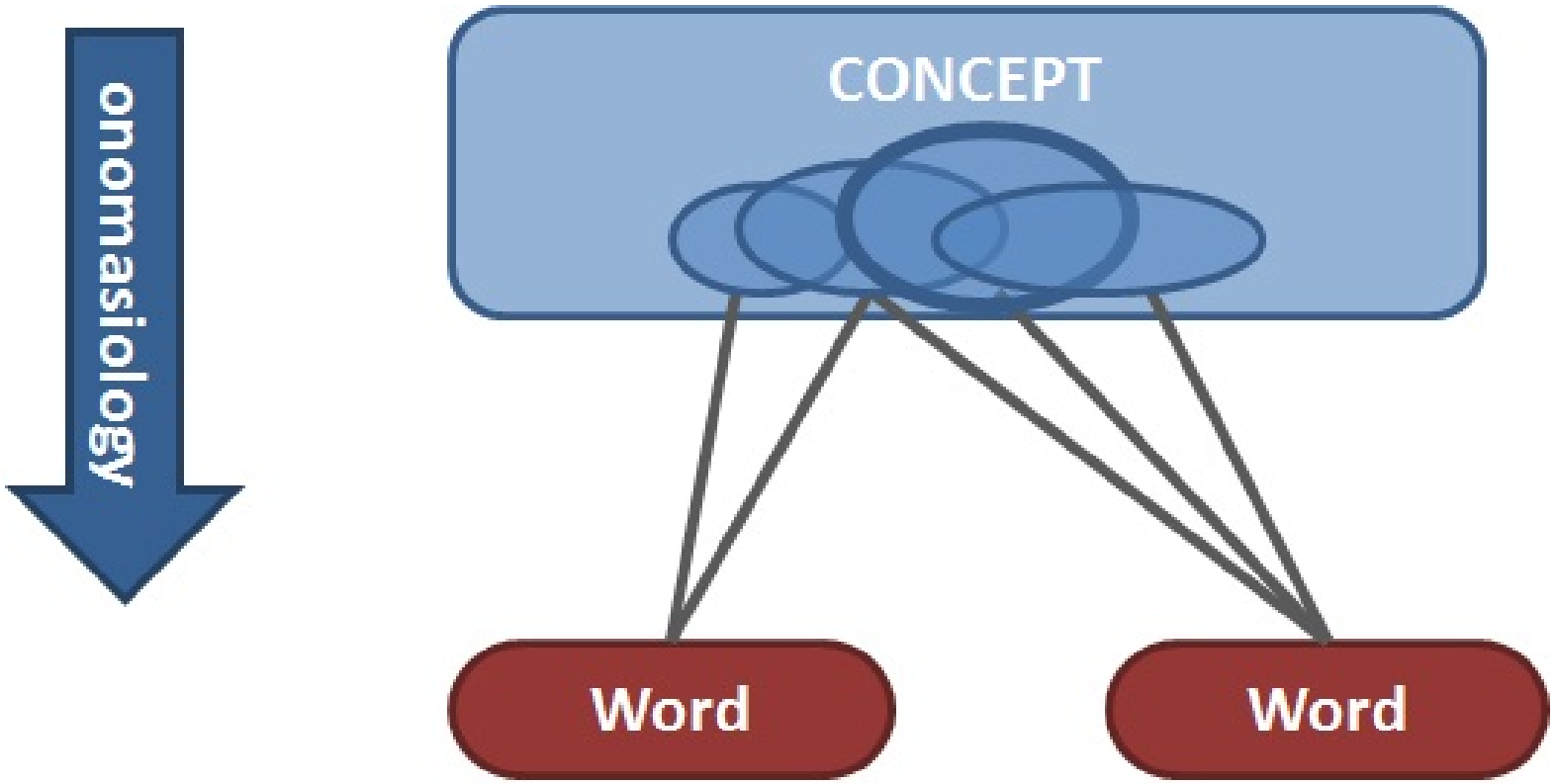
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):
PROTOTYPE STRUCTURE:



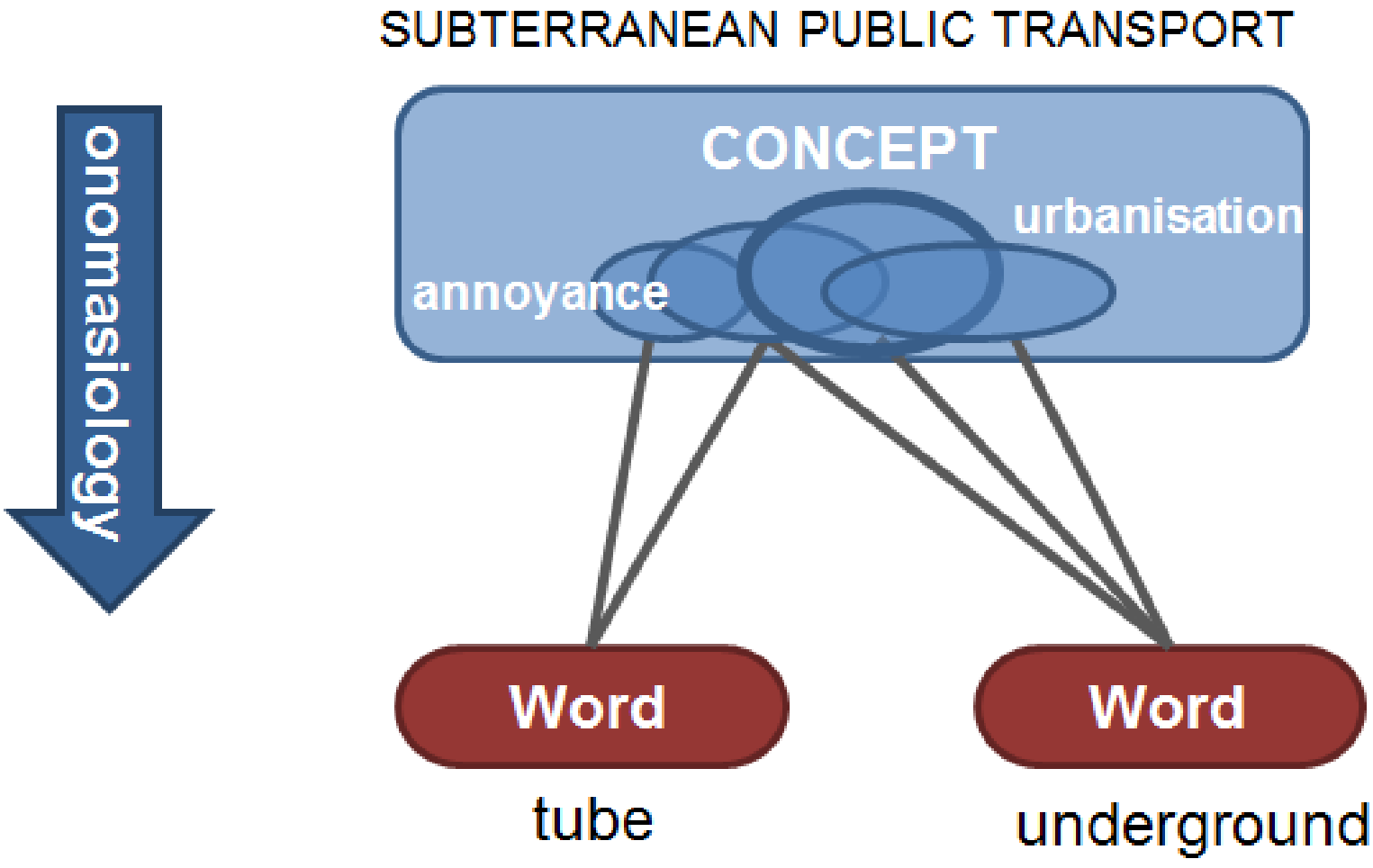
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):
PROTOTYPE STRUCTURE:



Linguistic Background

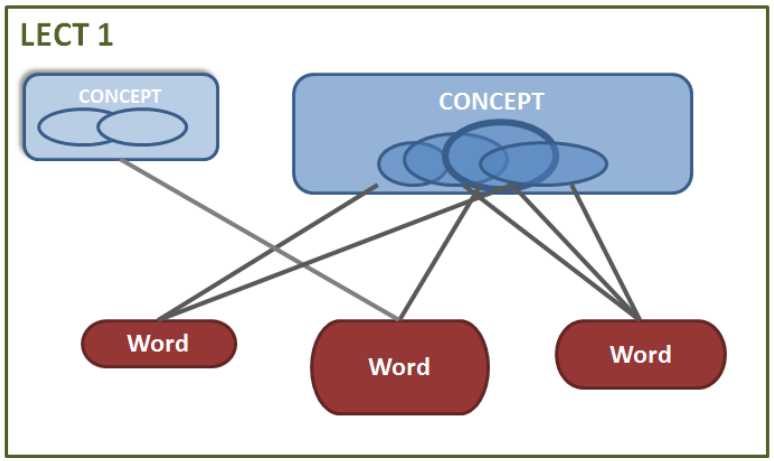
Structure of Lexical Variation (Geeraerts et al. 1994):
PROTOTYPE STRUCTURE:

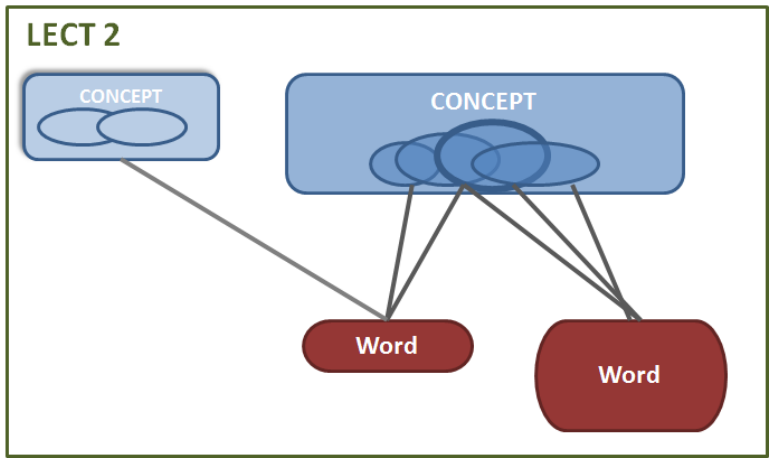


Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):

LECTAL VARIATION:

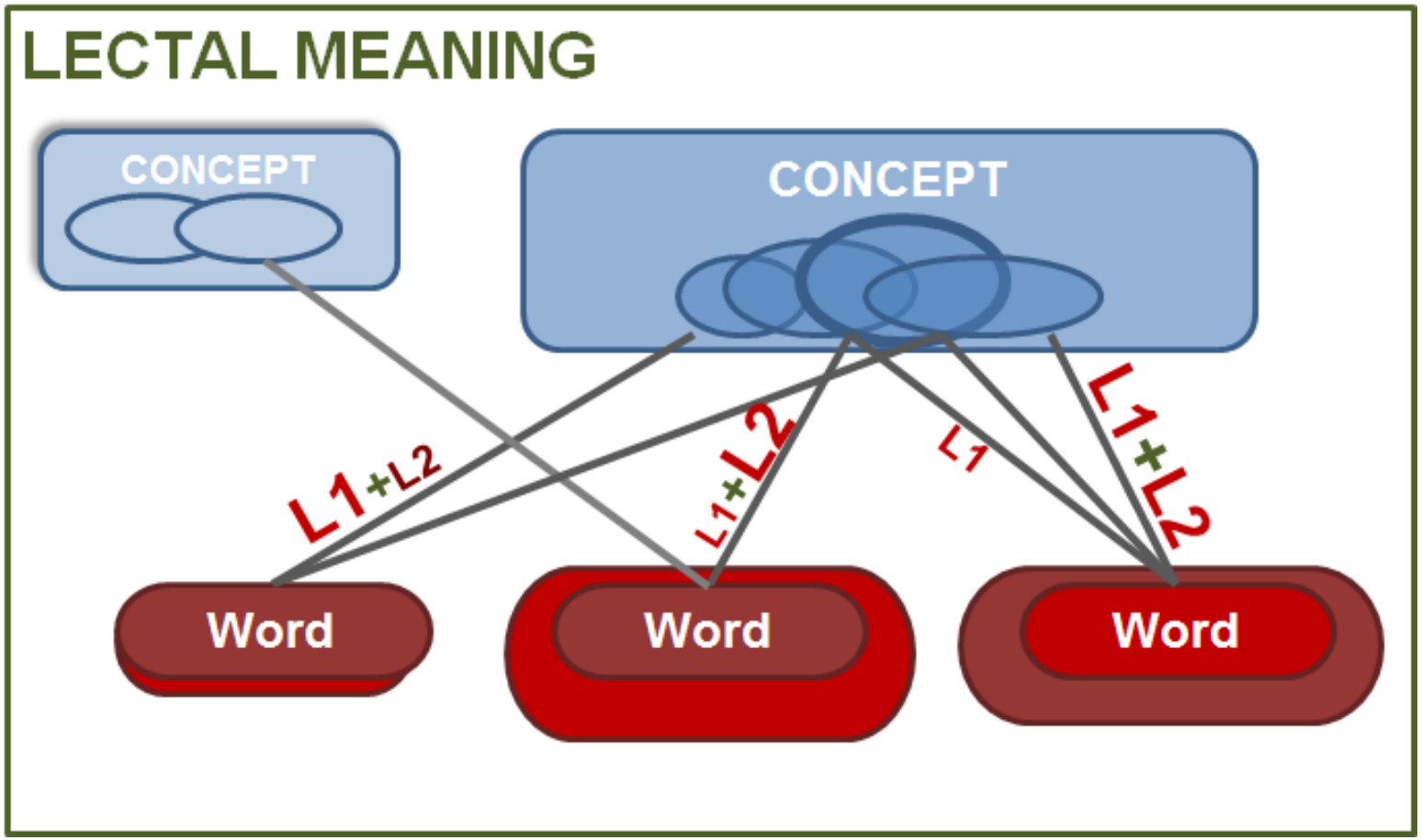




Linguistic Background

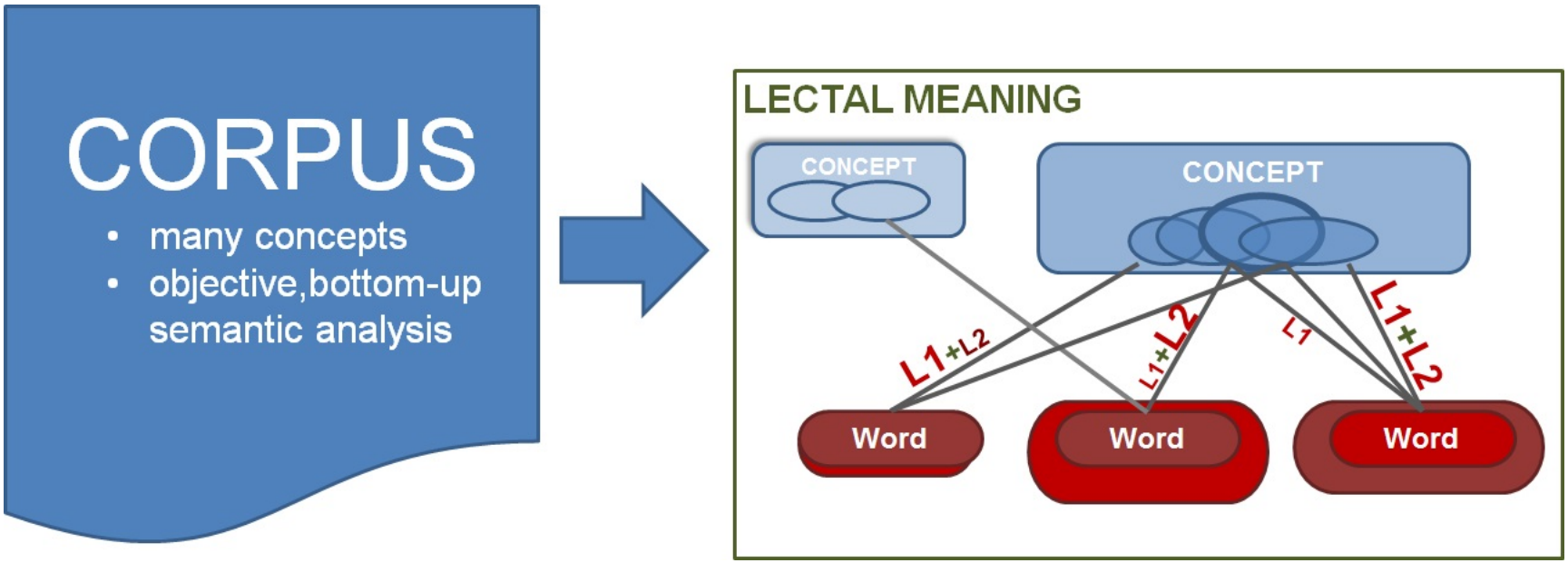
Structure of Lexical Variation (Geeraerts et al. 1994):

LECTAL MEANING:



Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):
USAGE-BASED STUDY:



Overview

1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Cluster quality measures
5. Discussion and future work

2. Semantic Vector Spaces

Linguistic origin: Distributional Hypothesis

- "You shall know a word by the company it keeps" (Firth)
- a word's meaning can be induced from its co-occurring words
- long tradition of collocation studies in corpus linguistics

Semantic Vector Spaces in Computational Linguistics

- standard technique in statistical NLP for the large-scale automatic modeling of (lexical) semantics
- aka Vector Spaces Models, Distributional Semantic Models, Word Spaces,... (cf Turney & Pantel 2010 for overview)
- generalised, large-scale collocation analysis
- mainly used for automatic thesaurus extraction:
  $\Rightarrow$ words occurring in same contexts have similar meaning

Type-level SVS

Collect co-occurrence frequencies for a large part of the vocabulary and put them in a matrix

	transport	*train*	*commute*	*ticket*	*scene*	*sugar*	*cream*	*now*
subway	120	424	388	82	12	11	3	189
underground	154	401	376	99	305	20	1	123
coffee	5	8	18	4	1	72	102	152

Type-level SVS

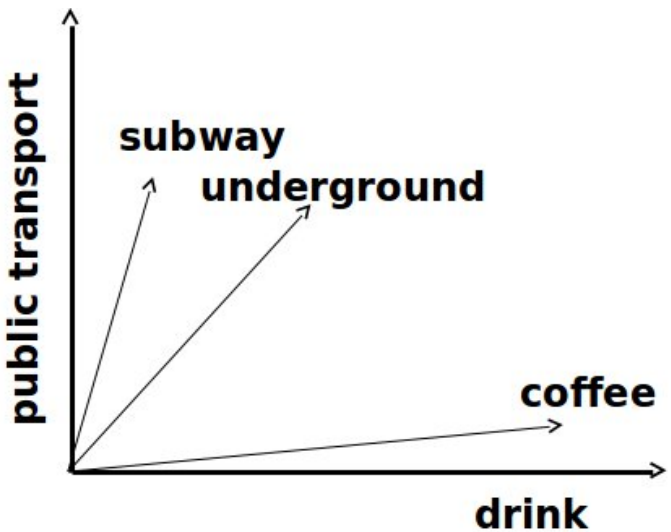
weight the raw frequencies by collocational strength (pmi)

	transport	*train*	*commute*	*ticket*	*scene*	*sugar*	*milk*	*now*
subway	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
underground	4.3	8.1	5.7	3.2	6.2	0.5	0.0	0.1
coffee	0.1	0.2	0.4	0.1	0.0	6.4	7.2	0.1

Type-level SVS

calculate word by word similarity matrix

	subway	underground	coffee
subway	1	.71	.08
underground	.71	1	.09
coffee	.08	.09	1



Token-level SVS

Make a vector for each occurrence of the variants

the teacher saw the dog chasing the cat

Token-level SVS

Make a vector for each occurrence of the variants

	3.2	4.3		0.8		7.1	
	5.1	2.2		3.7		0.1	
	0.2	3.5		2.3		0.3	
	3.1	1.9		2.9		4.1	
	4.7	0.2		1.3		3.1	
	2.2	3.1		4.1		3.8	
the	teacher	saw	the	dog	chasing	the	cat

Token-level SVS

Make a vector for each occurrence of the variants

					AVERAGE
3.2	4.3		0.8	7.1	3.9
5.1	2.2		3.7	0.2	2.8
0.2	3.5		2.3	0.3	1.6
3.1	1.9		2.9	4.1	3.0
4.7	0.2		1.4	3.1	2.3
2.2	3.1		4.1	3.8	3.3
teacher	saw	dog	chasing	cat	

Token-level SVS

Weighting

	3.2	4.3		0.8	7.1
	5.1	2.2		3.7	0.1
	0.2	3.5		2.3	0.3
	3.1	1.9		2.9	4.1
	4.7	0.2		1.3	3.1
	2.2	3.1		4.1	3.8
	teacher	saw	dog	chasing	cat
PMI weights	0.4	0.8		2.1	1.5

Context words are not equally informative for the meaning of dog.

Token-level SVS

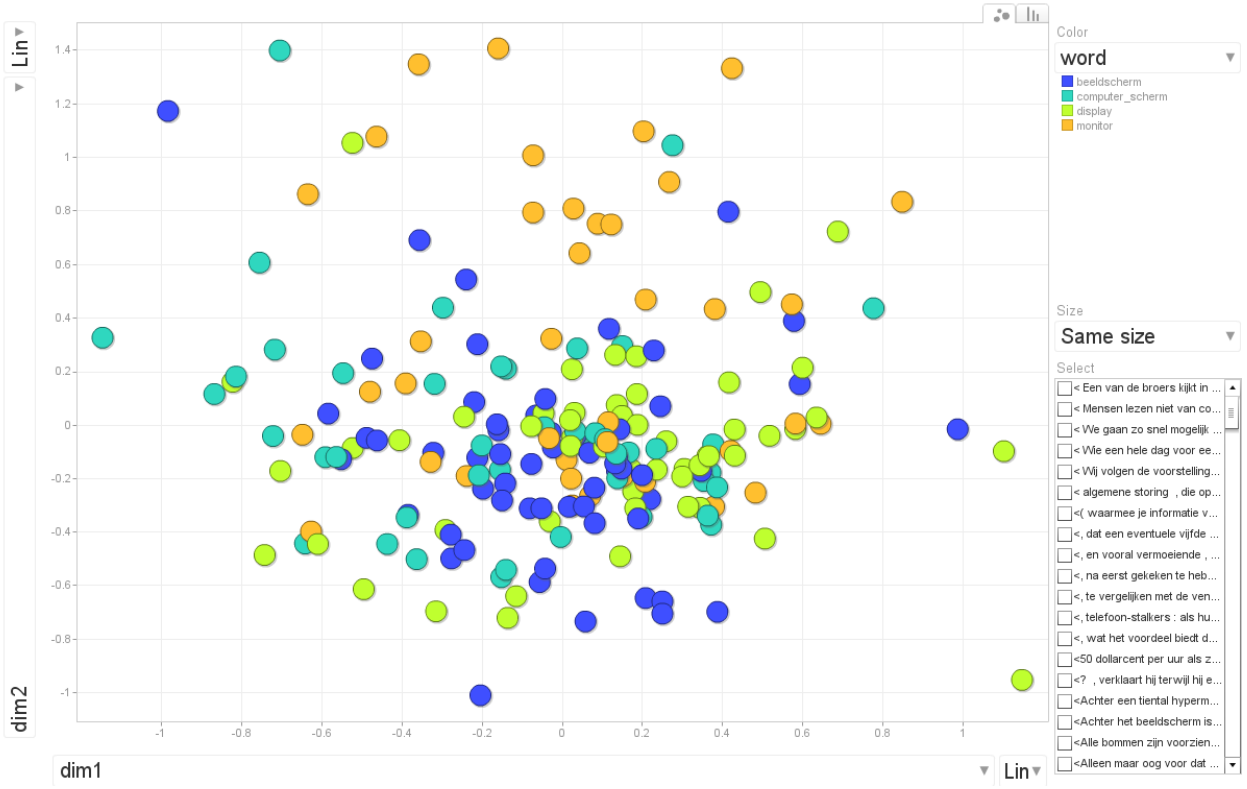
Weighted vectors

					WEIGHTED AVERAGE
3.2x0.4	4.3x0.8		0.8x2.1	7.1x1.5	4.3
5.1x0.4	2.2x0.8		3.7x2.1	0.2x1.5	3.0
0.2x0.4	3.5x0.8		2.3x2.1	0.3x1.5	2
3.1x0.4	1.9x0.8		2.9x2.1	4.1x1.5	3.8
4.7x0.4	0.2x0.8		1.4x2.1	3.1x1.5	2.4
2.2x0.4	3.1x0.8		4.1x2.1	3.8x1.5	4.4
teacher	saw	dog	chasing	cat	

Visual Analytics: Token clouds

Calculate similarity between all tokens

Version 1: use MDS and googlevis to plot interactively in 2D



Calibration problem

Semantic Vector Spaces, and especially token-level SVSs are parameter-rich.

Examples of parameters

- Bag-of-Words $\leftrightarrow$ Dependency Models
- Size of the context window for co-occurrences
- Size of the context window for weights
- Weighting scheme:
  Pointwise Mutual Information $\leftrightarrow$ Log-Likelihood Ratio
- Include $\leftrightarrow$ exclude highly-frequent (function words) words

Overview

3. Visual Analytics

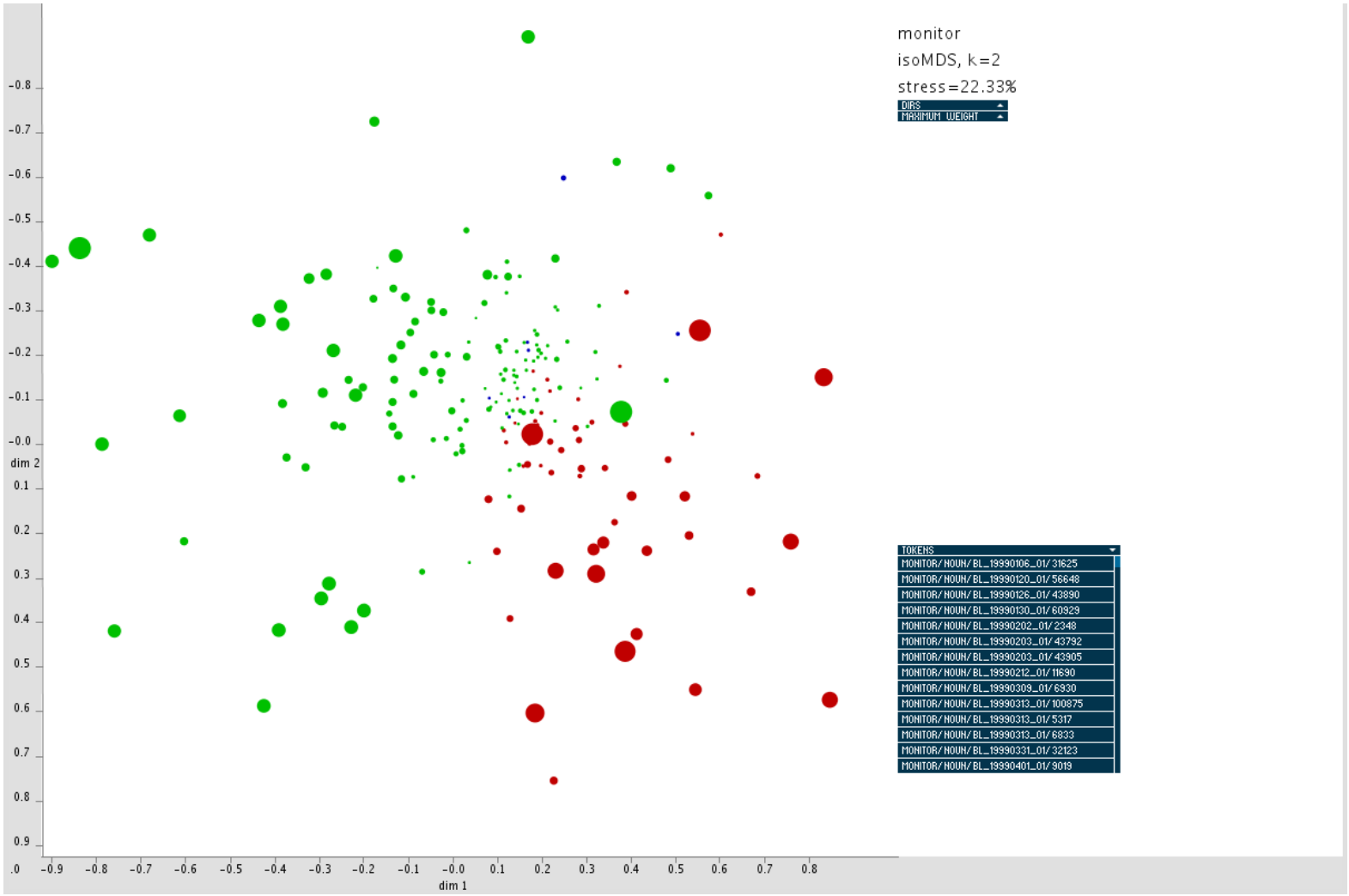
- Calibration could benefit from visual analytics of the different solutions.
- Using manually disambiguated data facilitates the visual evaluation as we can color–code the tokens for their different meanings.
- Misclassified tokens are quickly identified.
- We built our own, customisable tool to explore these token clouds.

3. Visual Analytics

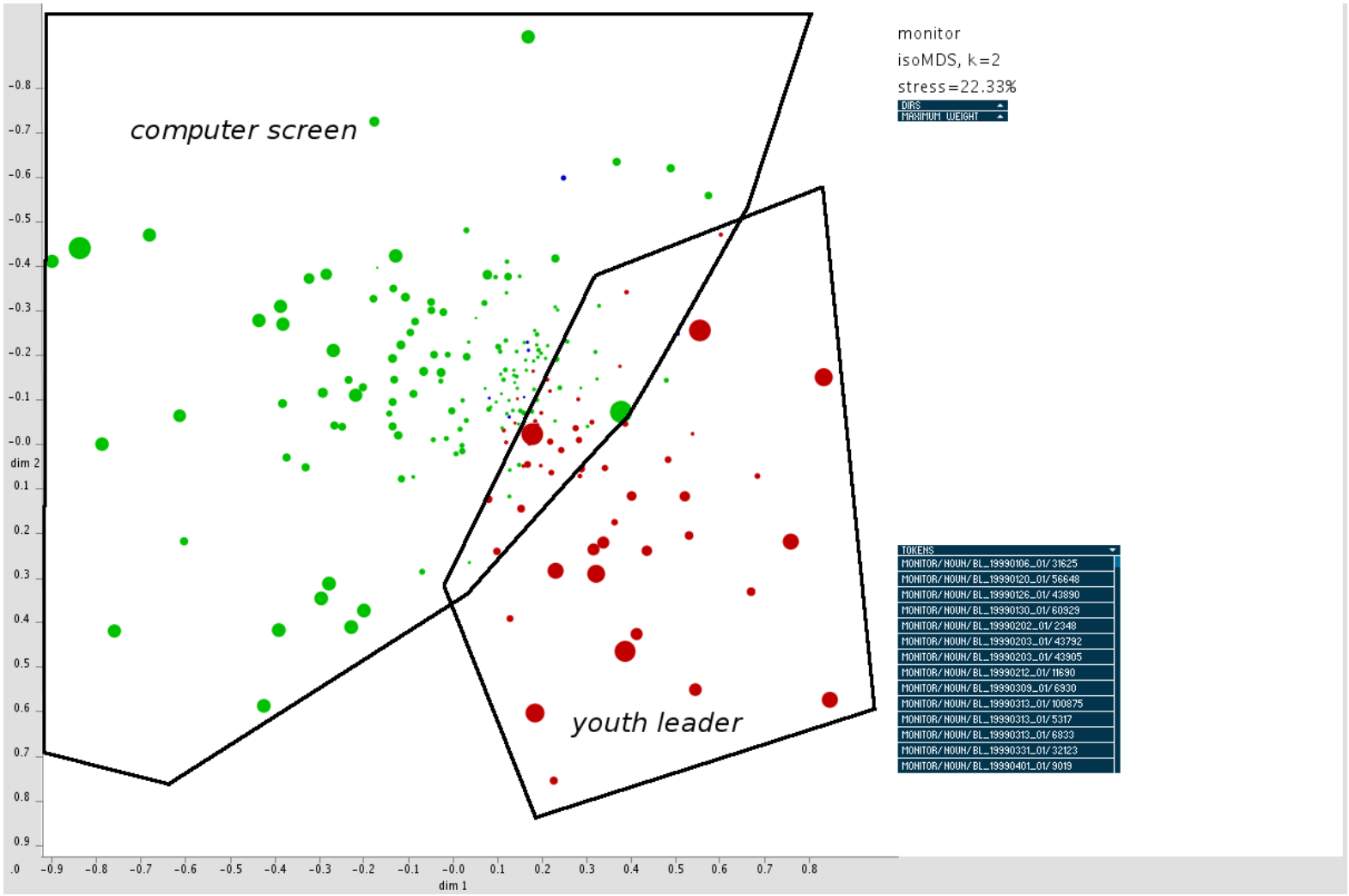
Dutch noun *monitor*

- Data from large Dutch newspaper corpora
  Leuven News Corpus (LeNC): 1.3 billion words
  Twente News Corpus (TwNC): 500 million words
- Manually disambiguted data for the concept of BEELDSCHERM (display)
- Semasiological analysis of a polysemous word with lectal variation (geographical and register):
  $monitor_1$: display, (computer) screen
  (Standard Dutch)
  $monitor_2$: supervisor of youth leasure activities
  'speelpleinmonitor'
  (Belgian Colloquial Dutch)

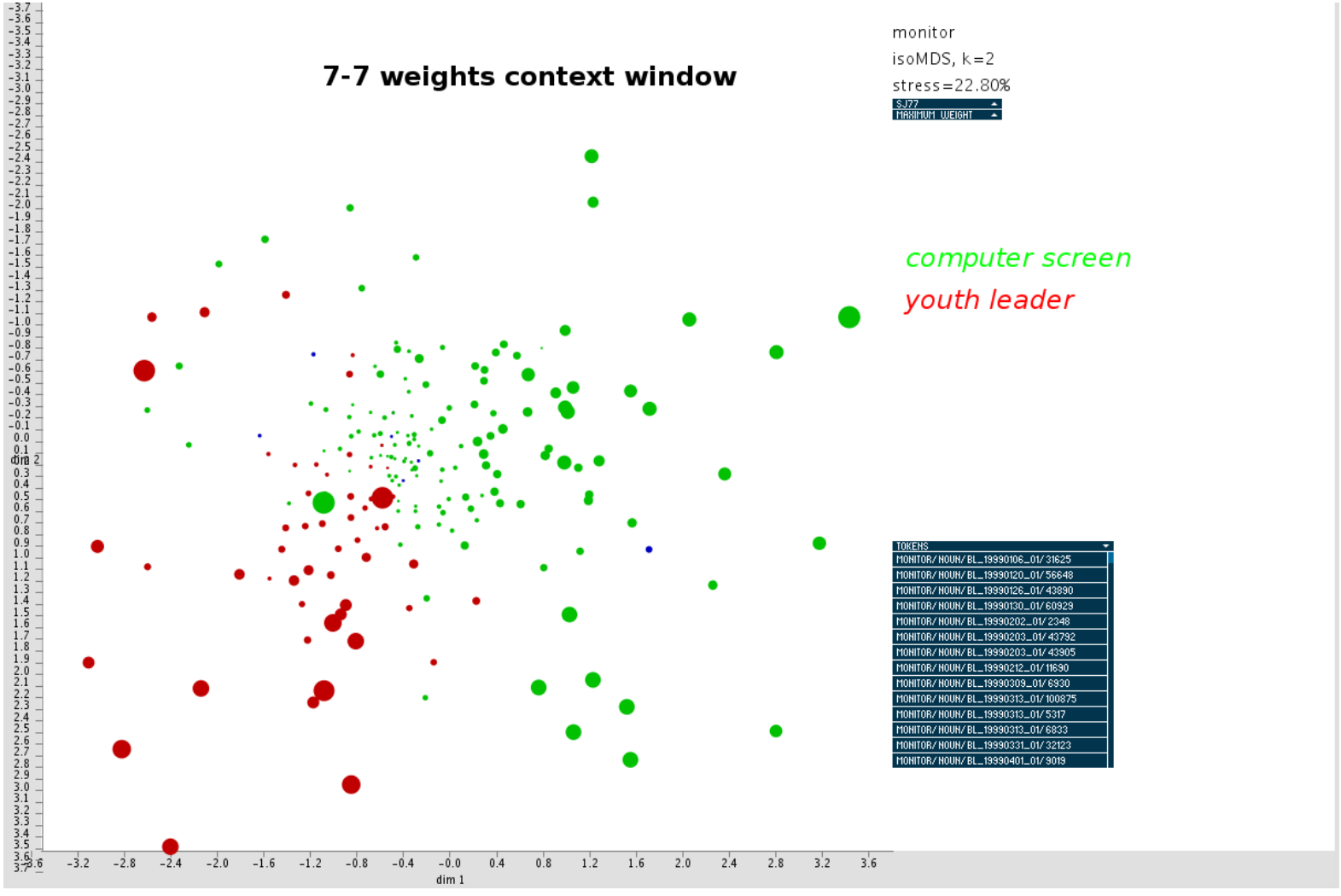
3. Visual Analytics



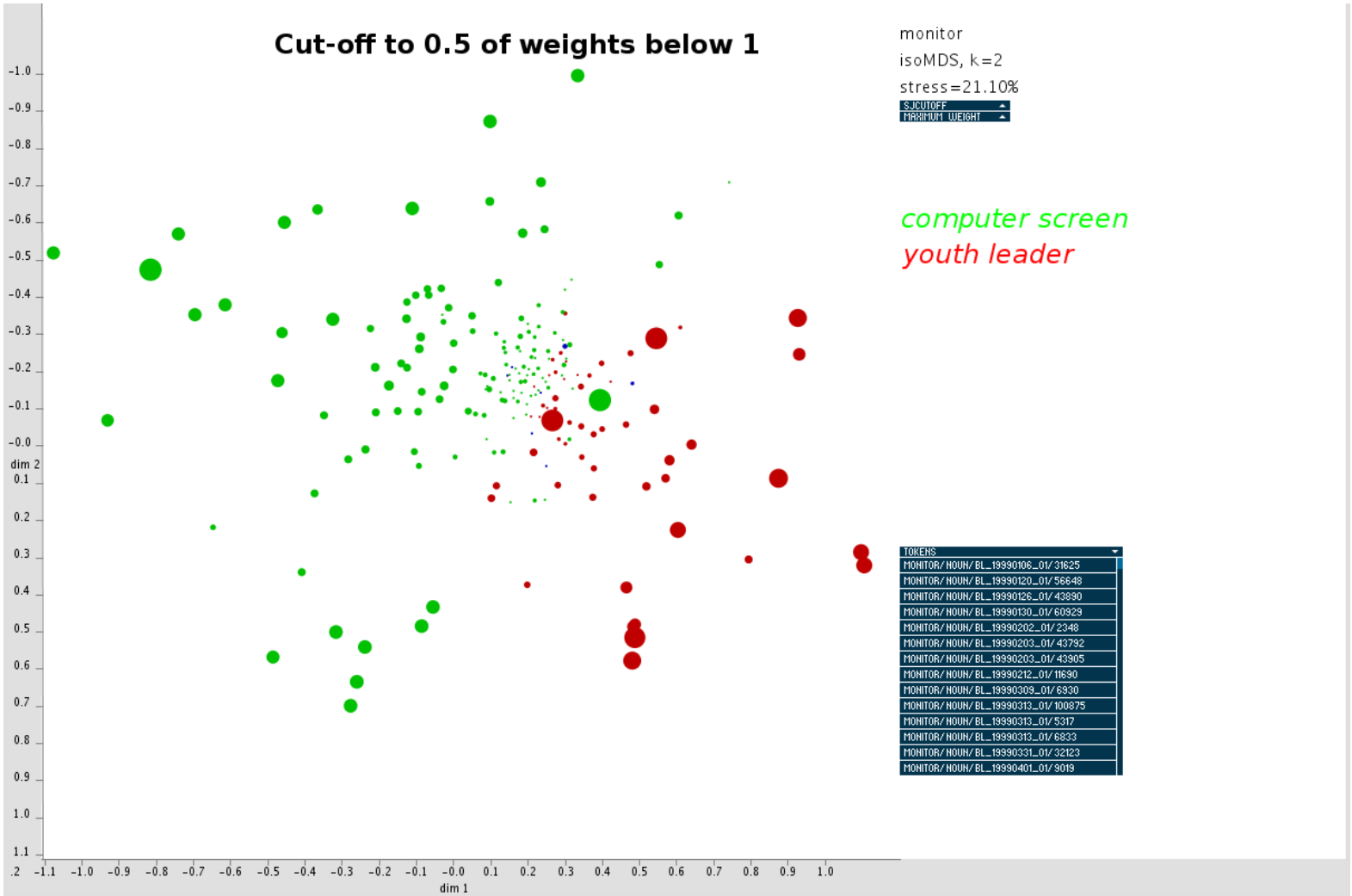
3. Visual Analytics



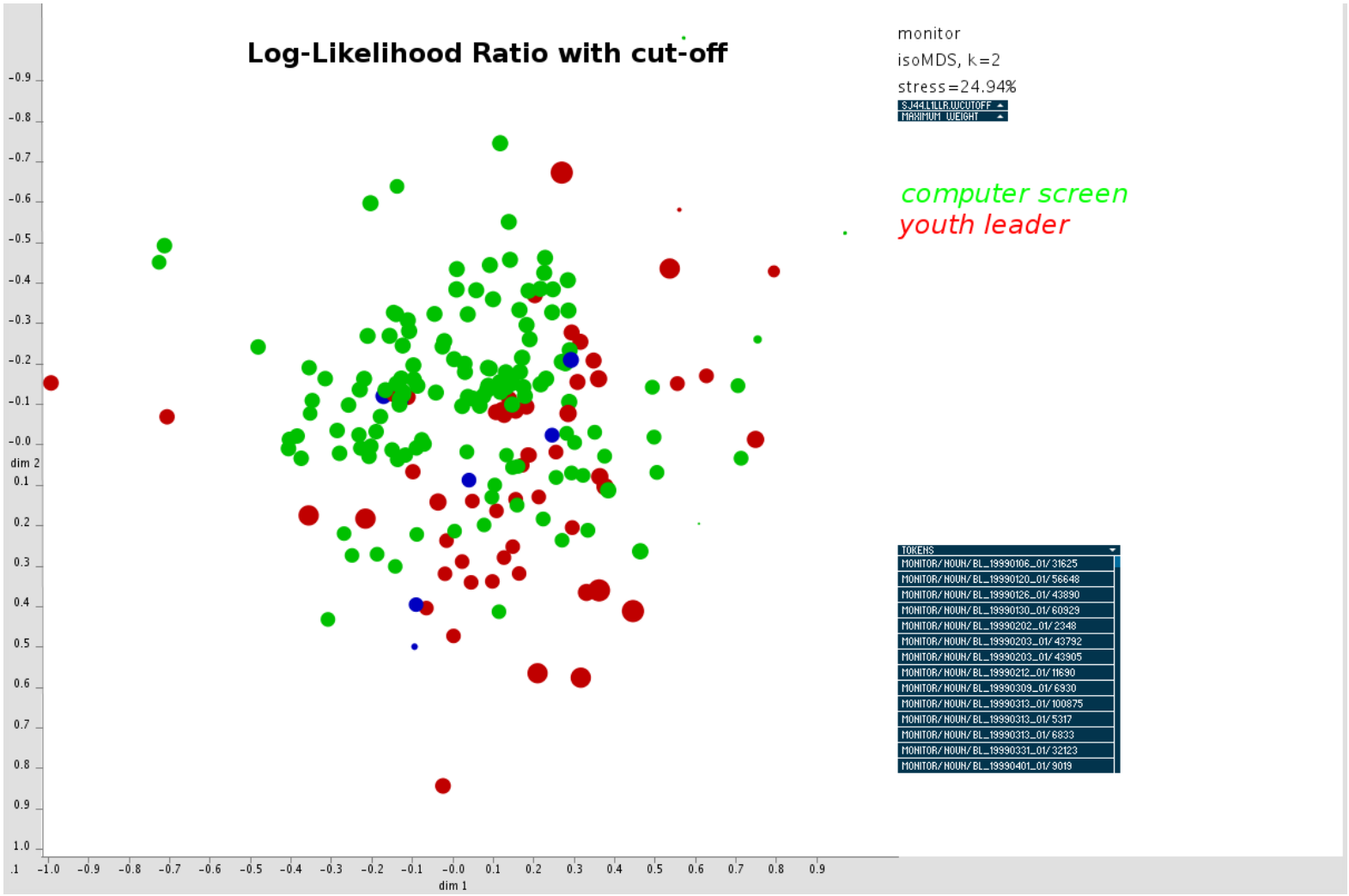
3. Visual Analytics



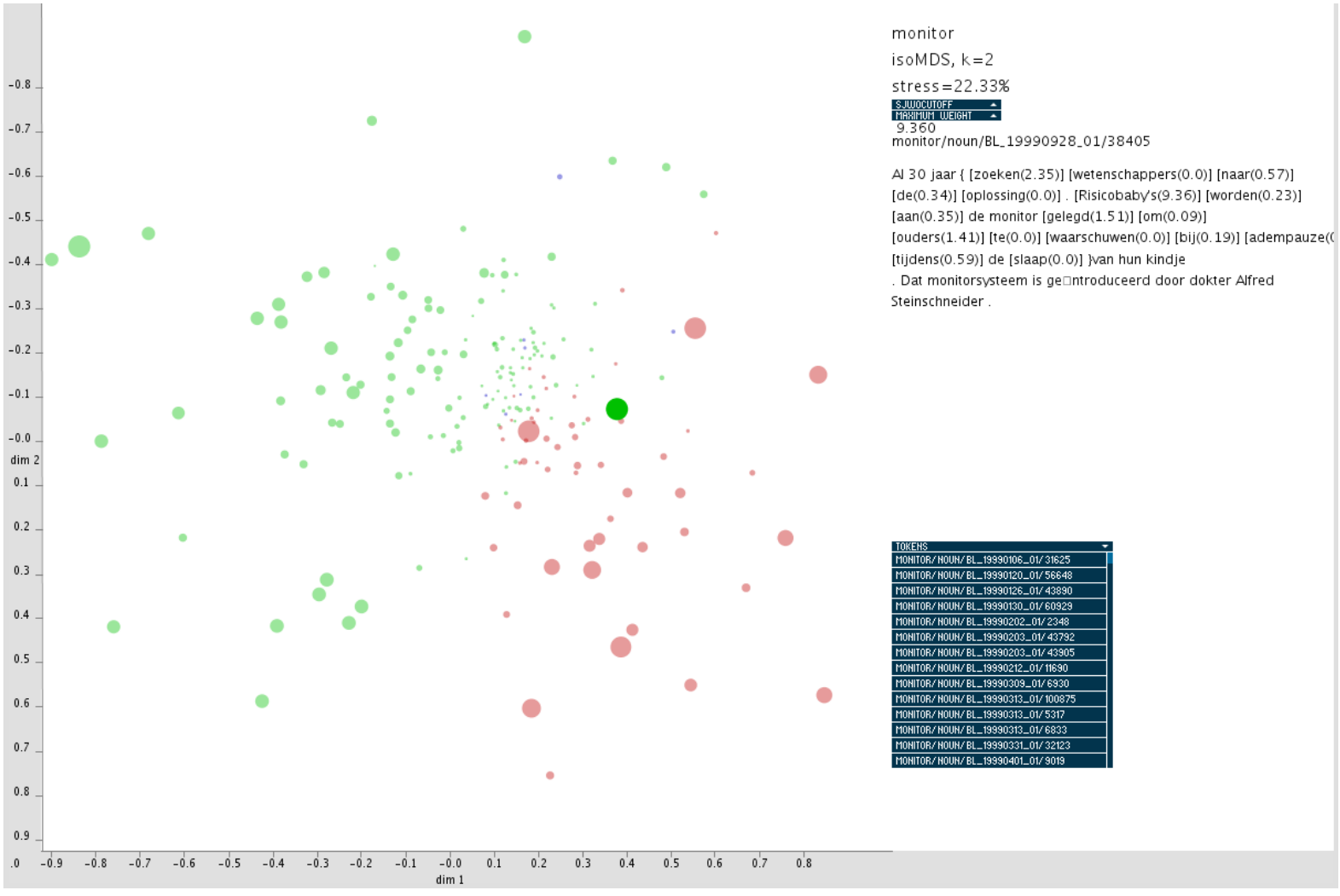
3. Visual Analytics



3. Visual Analytics



3. Visual Analytics



3. Visual Analytics

SemEval WSI data

- SemEval: competition for semantic analysis systems.
- Data from 2010 Word Sense Induction task.
- Test data disambiguated for 50 nouns.

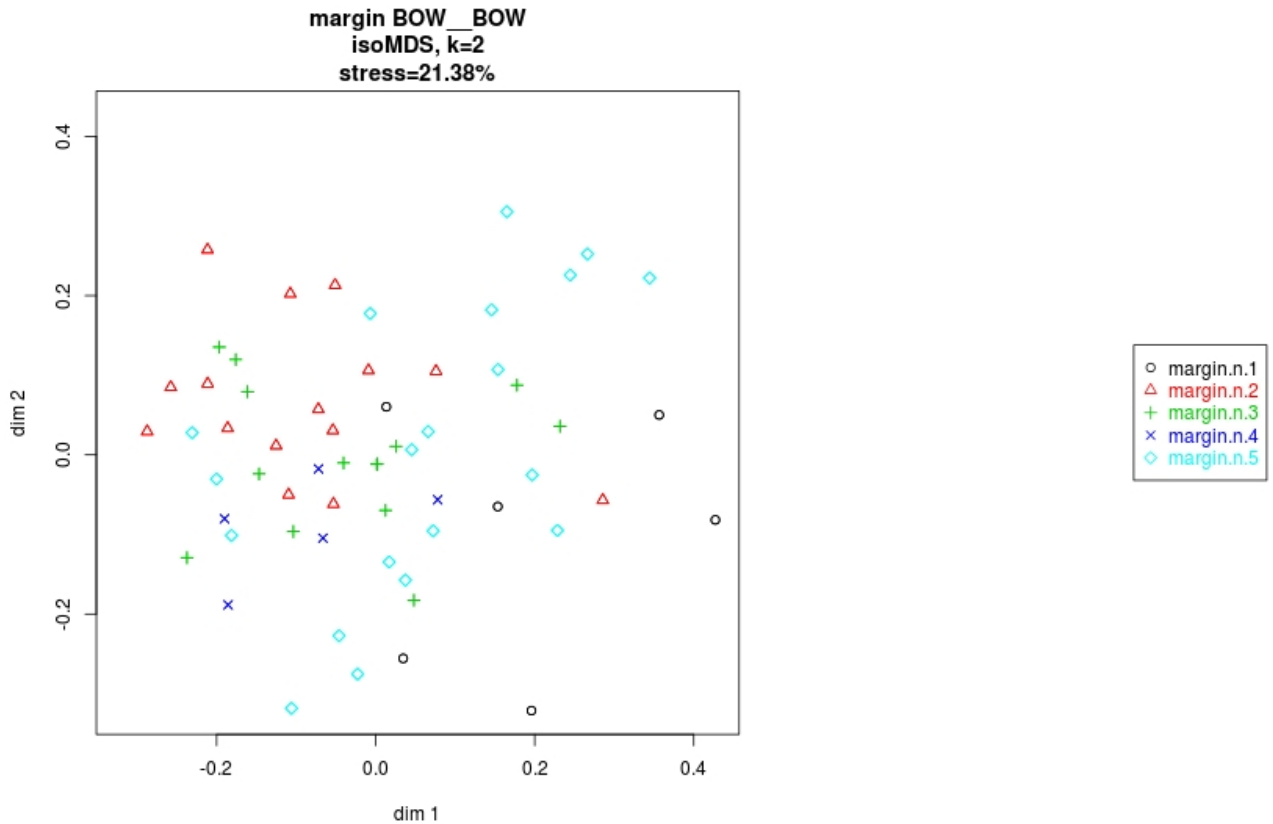
3. Visual Analytics

English noun *margin*

- *margin*$_1$ the boundary line or the area immediately inside the boundary
- *margin*$_2$ an amount beyond the minimum necessary
- *margin*$_3$ the amount of collateral a customer deposits with a broker when borrowing from the broker to buy securities
- *margin*$_4$ (finance) the net sales minus the cost of goods and services sold
- *margin*$_5$ a permissible difference; allowing some freedom to move within limits

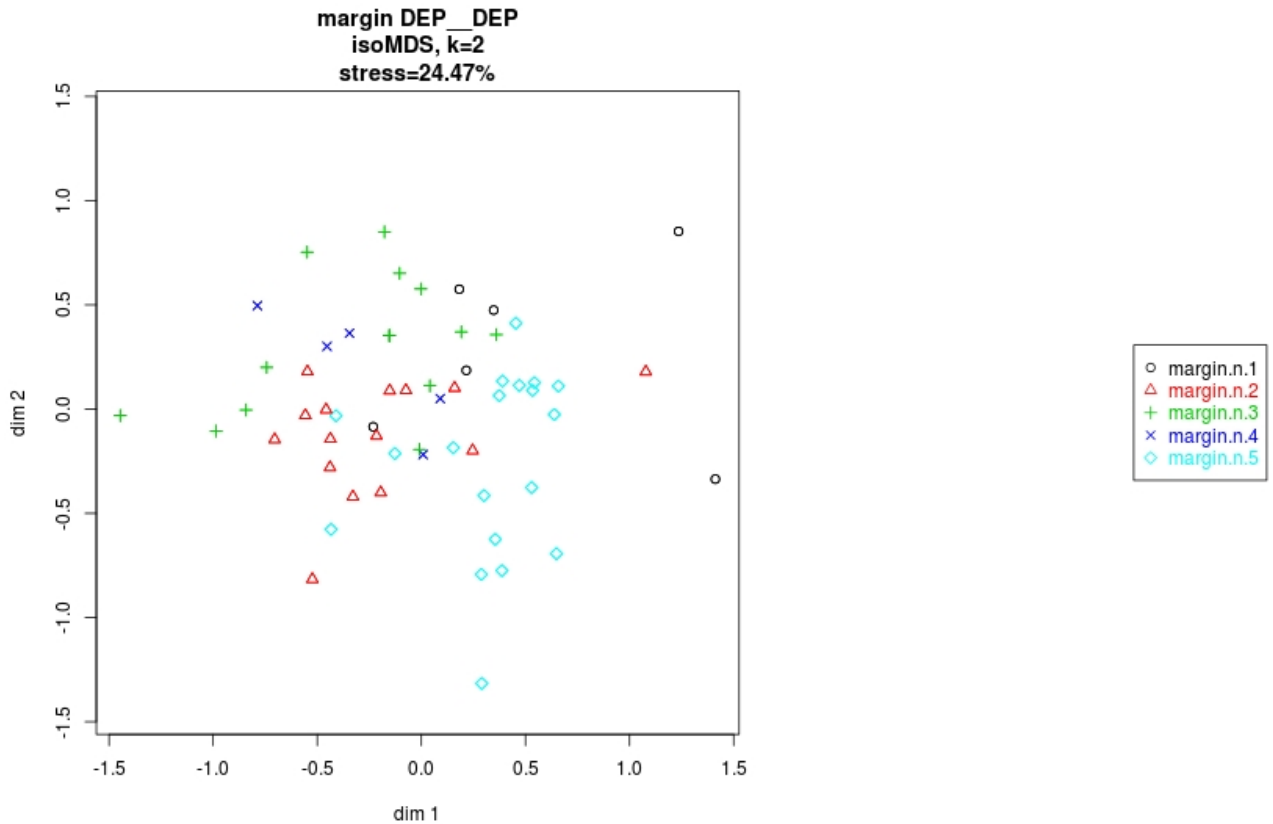
3. Visual Analytics

Bag-of-Words model



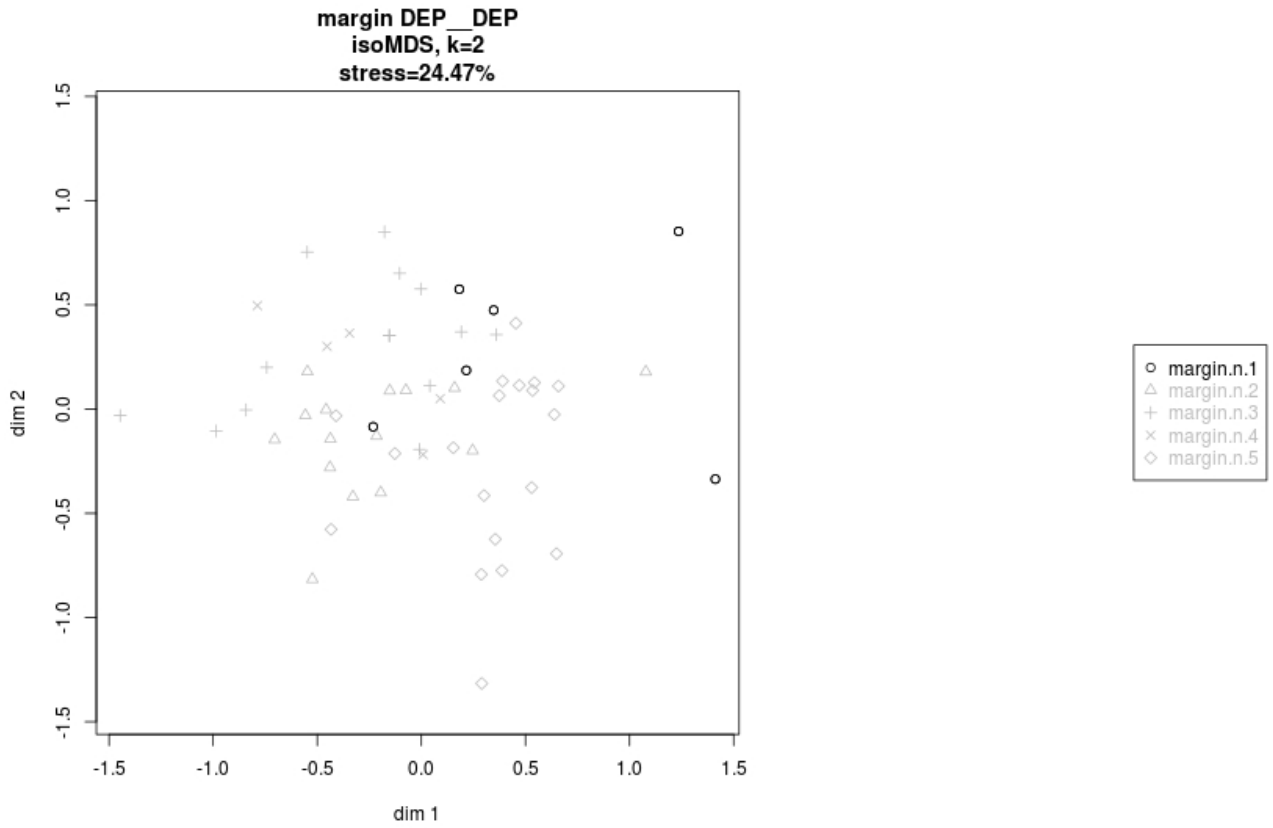
3. Visual Analytics

Dependency model



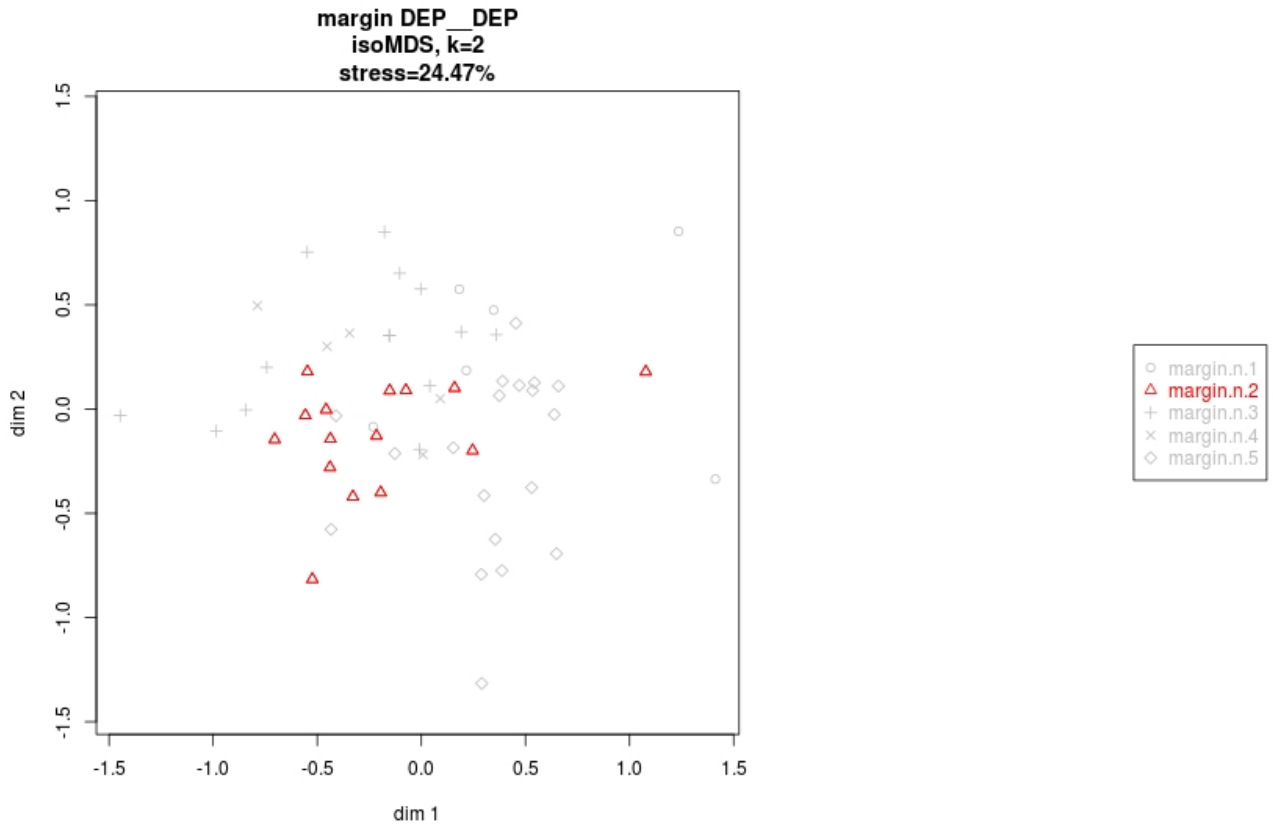
3. Visual Analytics

Dependency model: sense 1



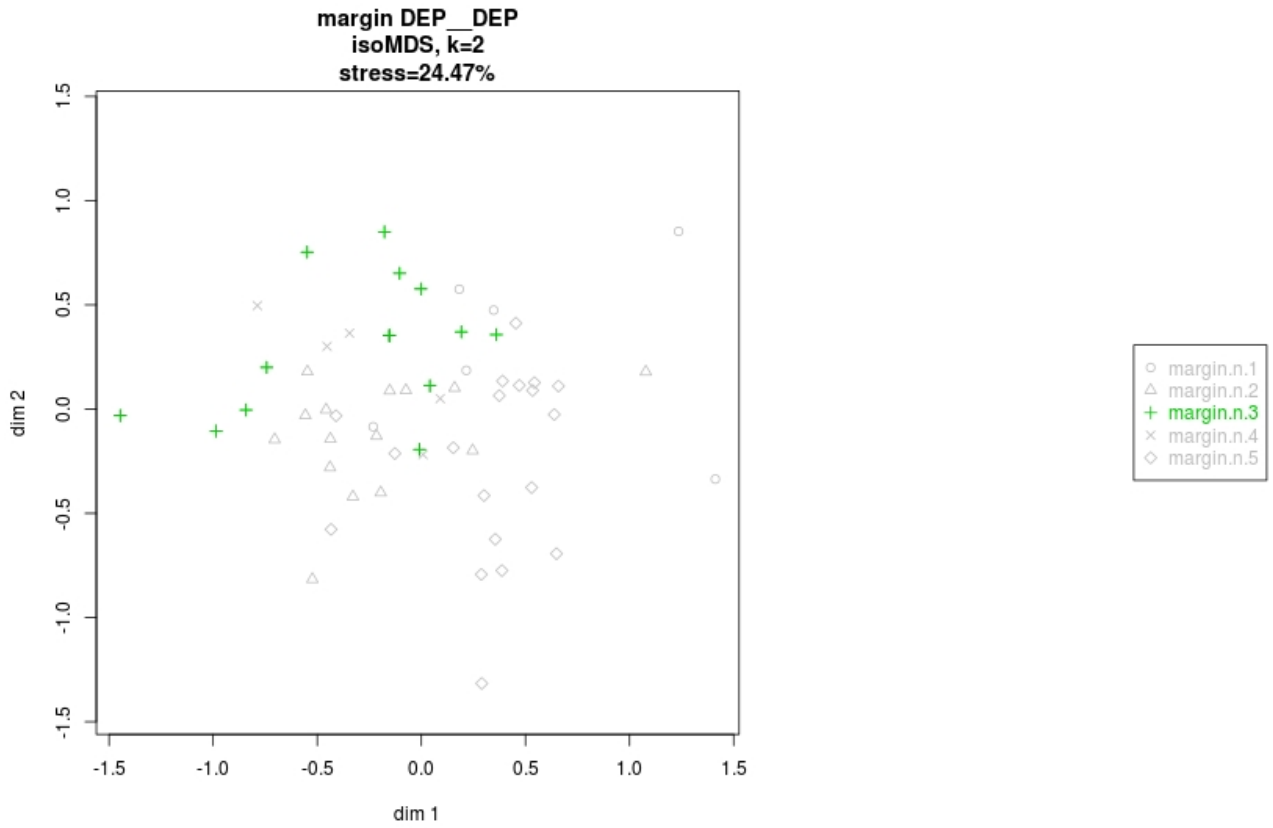
3. Visual Analytics

Dependency model: sense 2



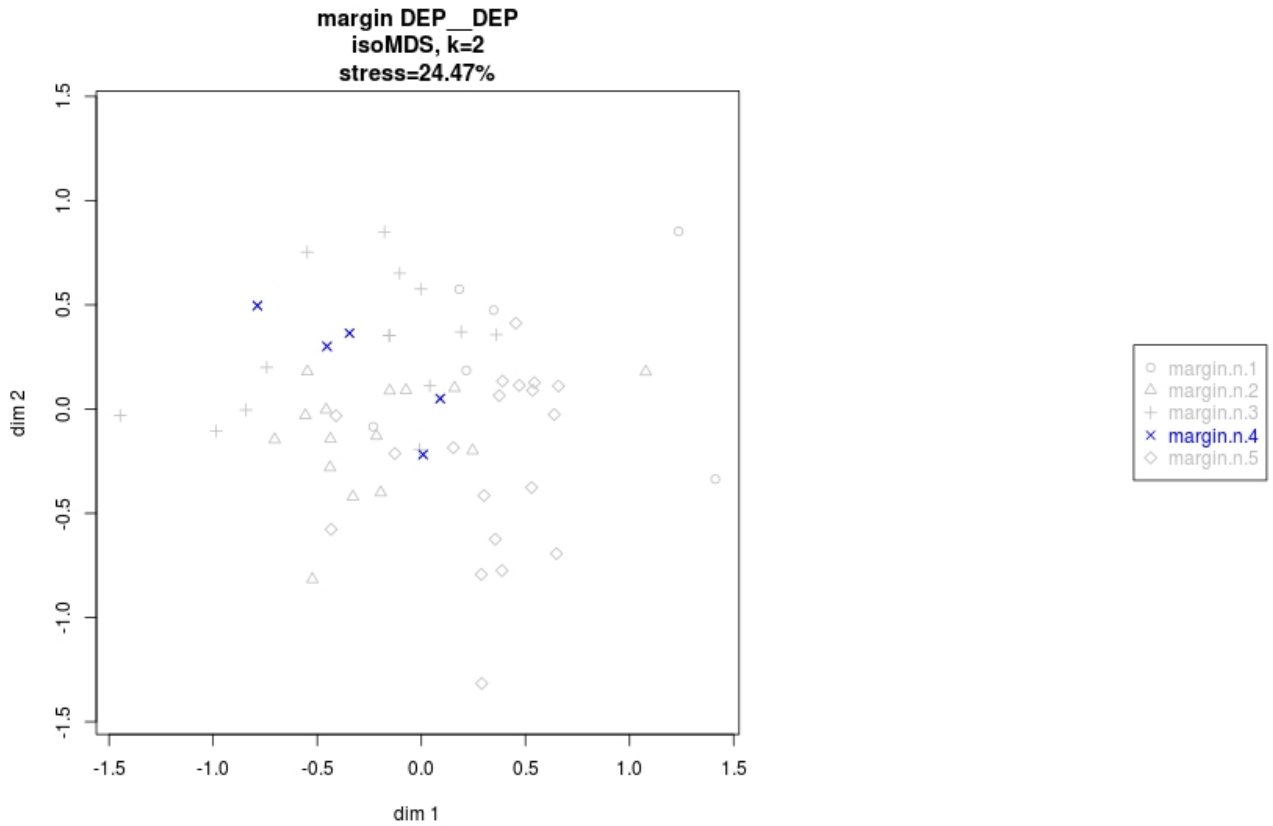
3. Visual Analytics

Dependency model: sense 3



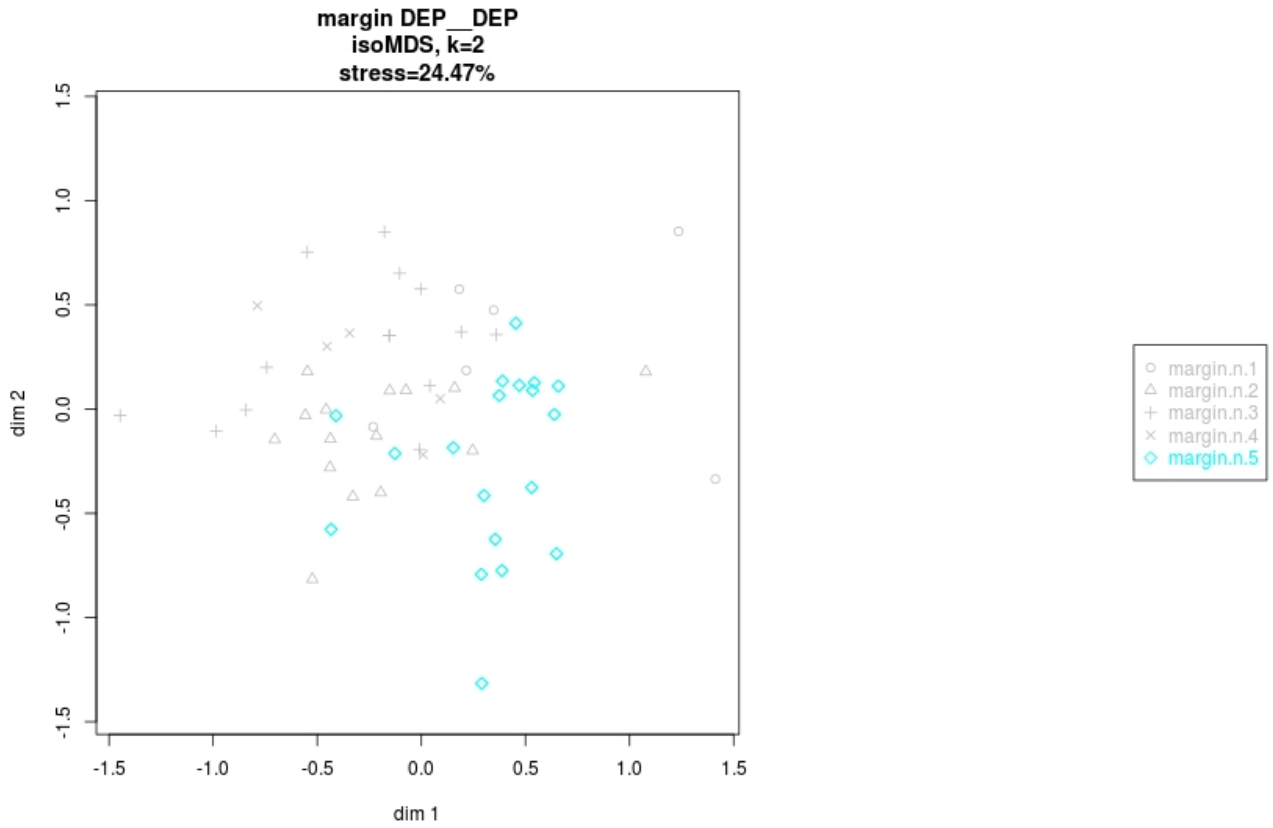
3. Visual Analytics

Dependency model: sense 4



3. Visual Analytics

Dependency model: sense 5



Overview

4. Cluster quality measures

Aggregate cluster quality

- First proposed by McClain and Rao (1975) to evaluate clustering in marketing research.
- Speelman and Geeraerts (2009) proposed a similar measure for dialectometry.

$$\text{clusterqual: } \frac{S_W/N_W}{S_B/N_B}$$

S_W: within distances
N_W: number of distances between pairs
S_B: between distances
N_B: number of distances between pairs

4. Cluster quality measures

Nearest neighbours quality

- Instead of aggregating over n instances, we look at the k-nearest neighbours for each token.

$$\text{kNN quality: } \frac{\sum_{i=1}^{n} \frac{S_i}{k}}{n}$$

S: number of tokens belonging to the same cluster as i
n: total number of tokens

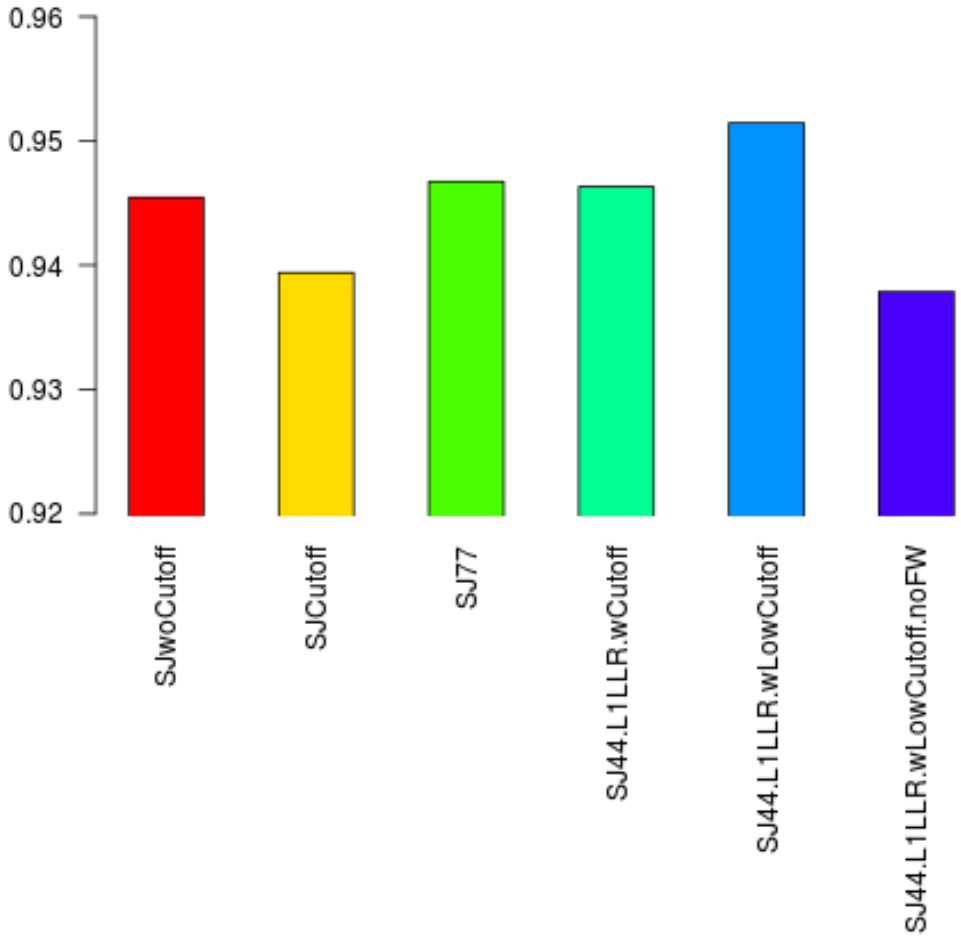
4. Cluster quality measures

Comparing the measures

- clusterqual: low value indicates better quality (divide small S_W by large S_B)
- kNN-qual: 'purity' percentage; high values are better

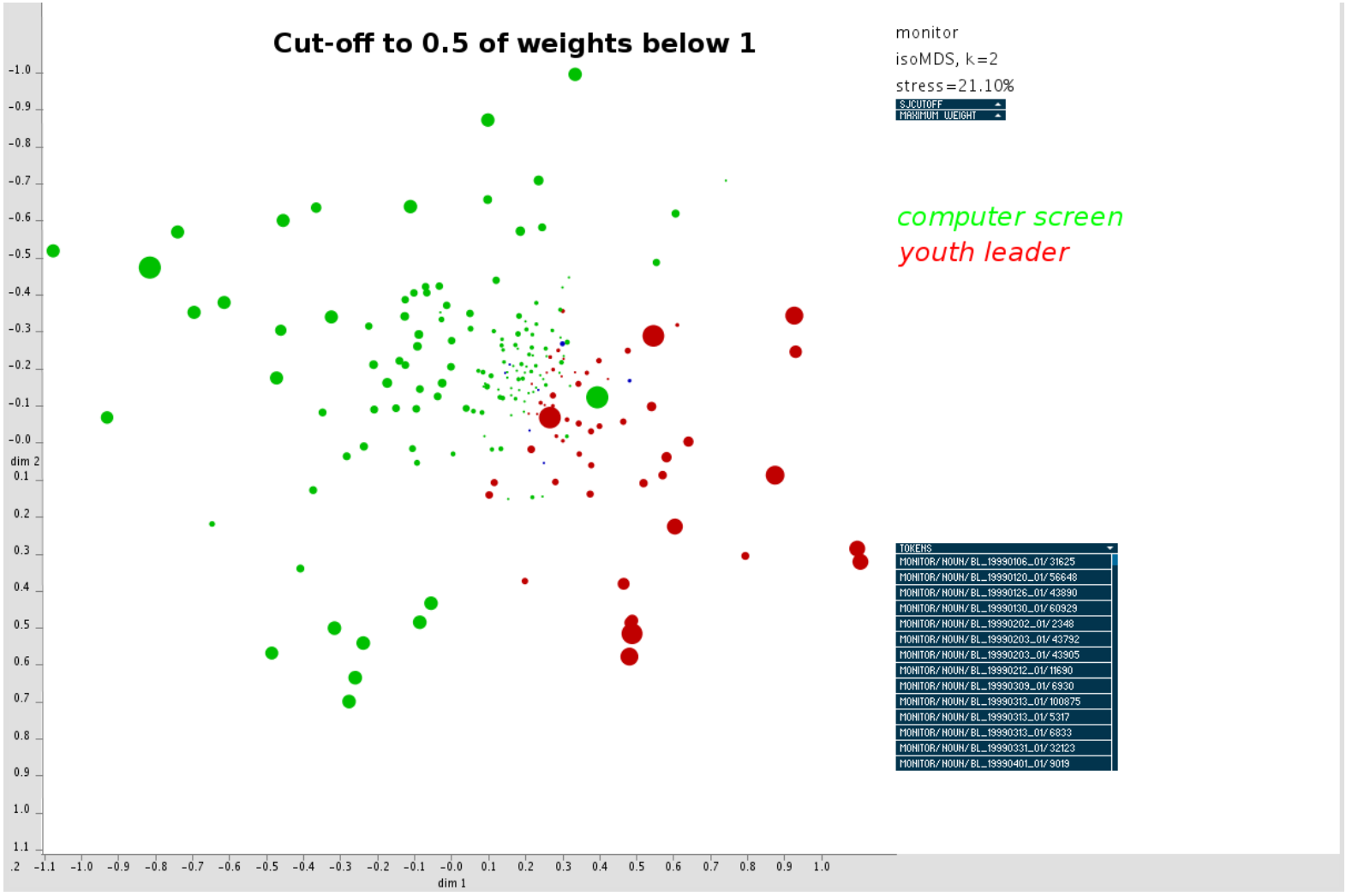
4. Cluster quality measures

clusterqual: *monitor* tokens



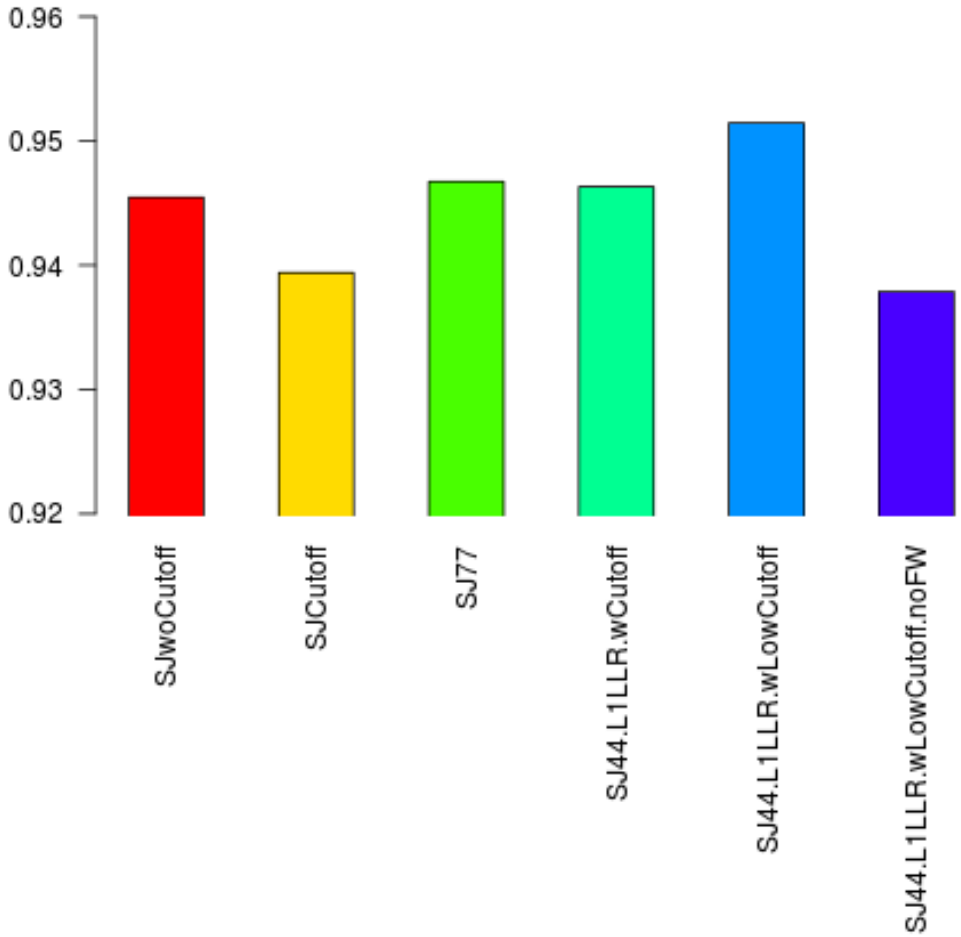
4. Cluster quality measures

Parameter settings with good clusterqual



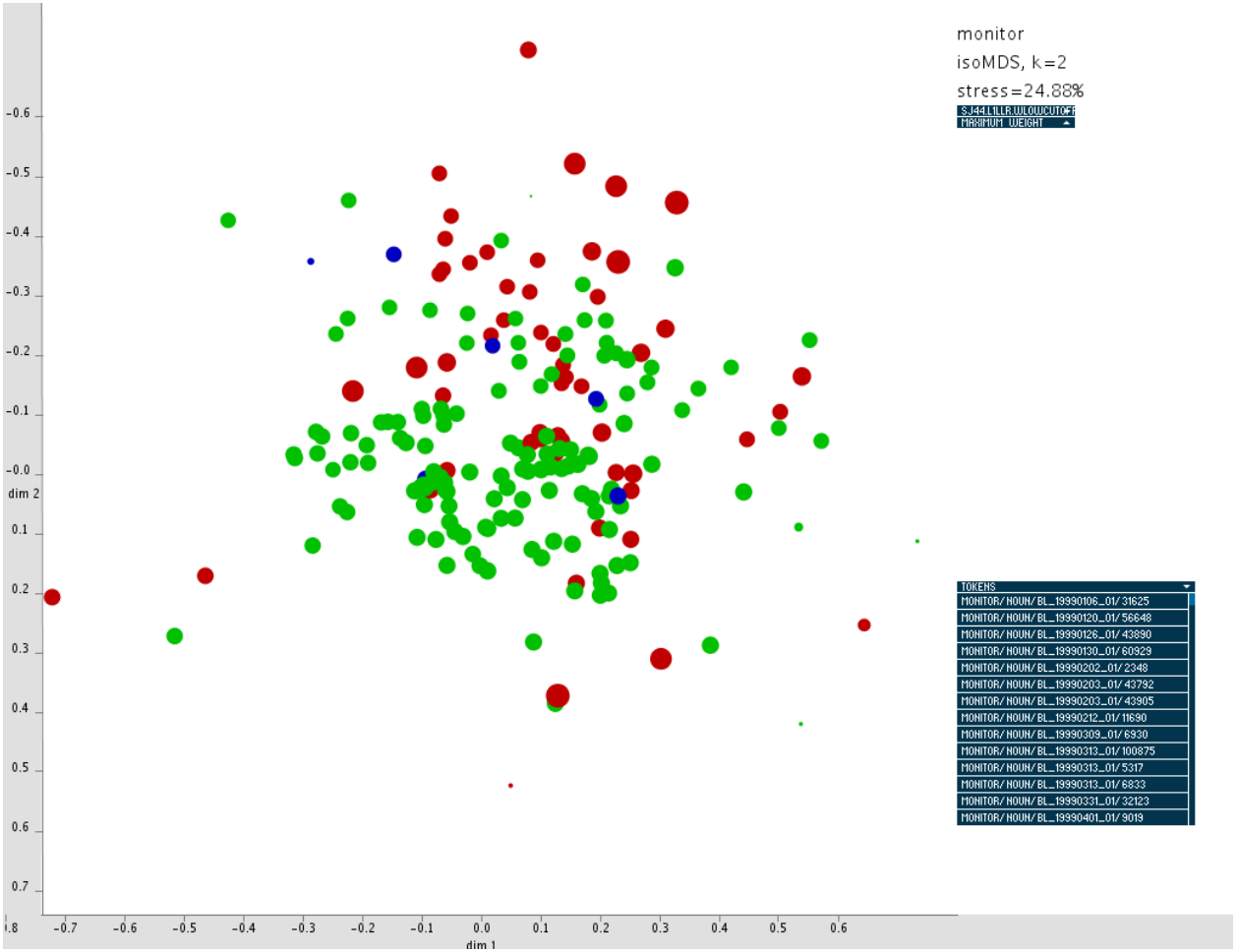
4. Cluster quality measures

clusterqual: *monitor* tokens



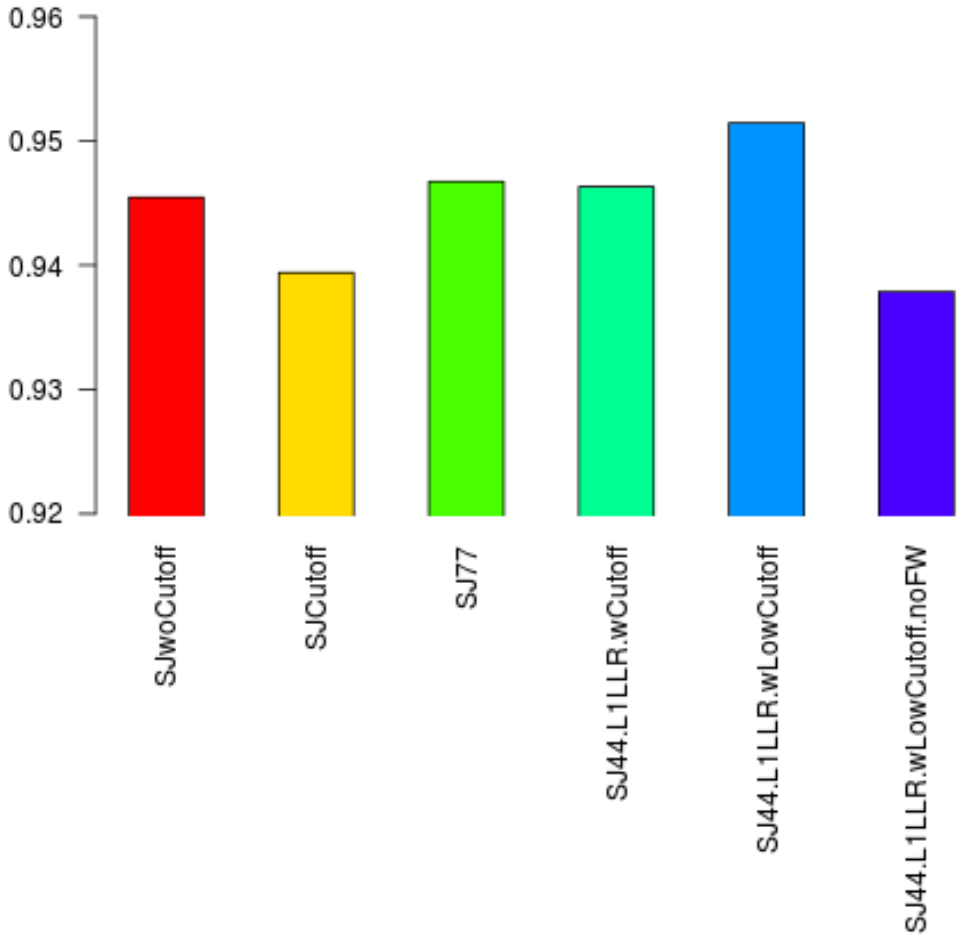
4. Cluster quality measures

Parameter settings with bad clusterqual

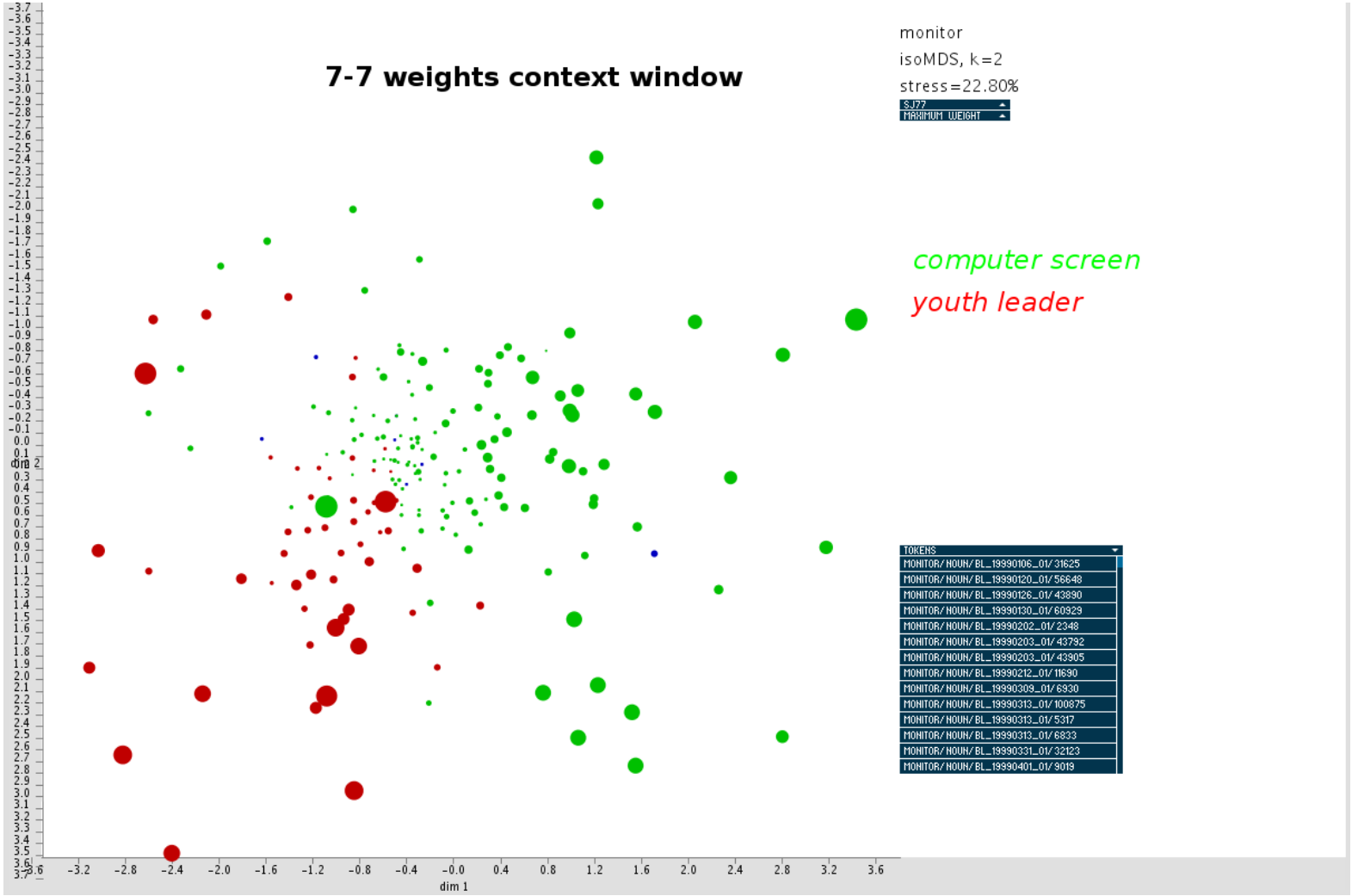


4. Cluster quality measures

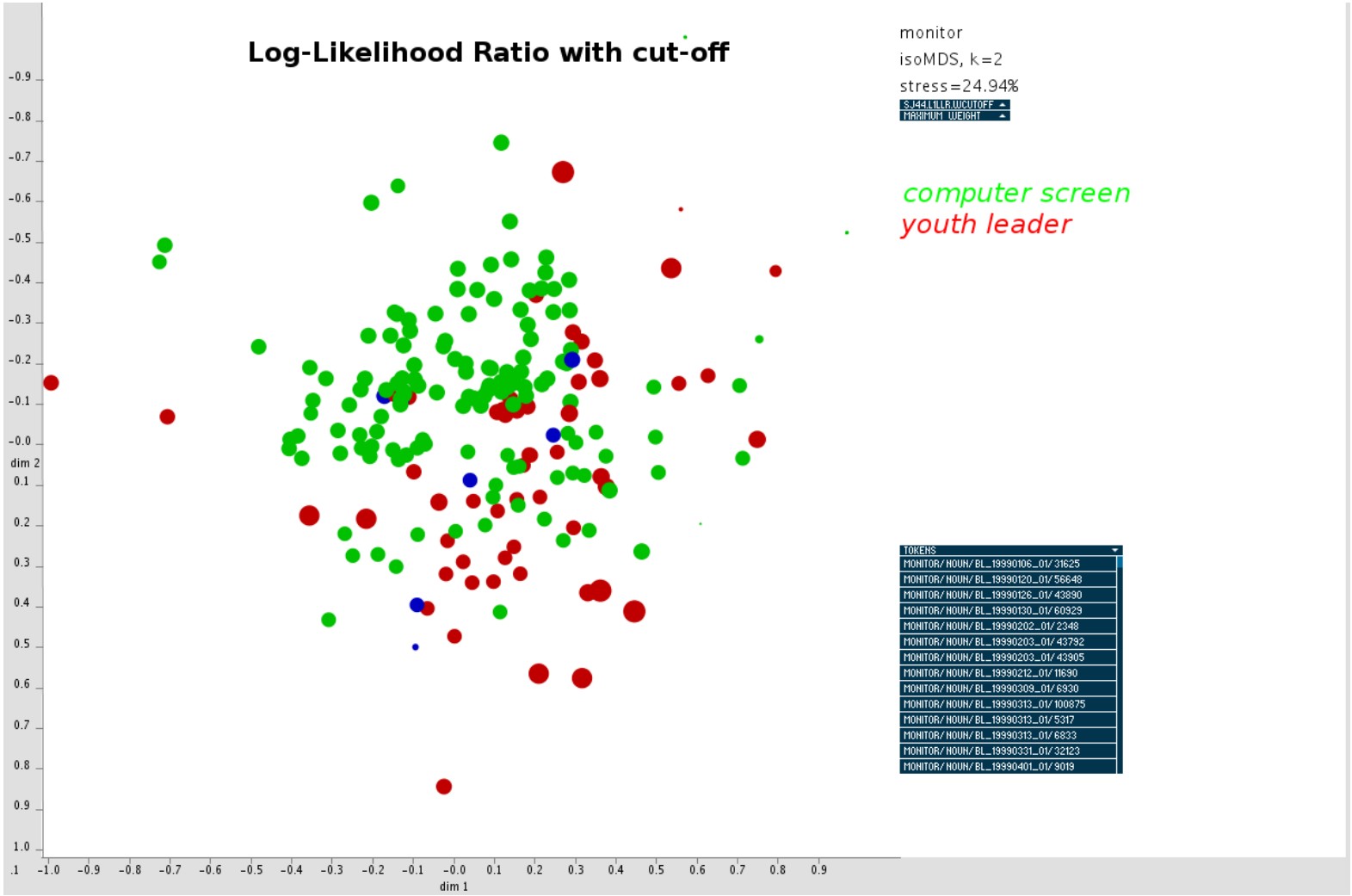
clusterqual: 'monitor' tokens



4. Cluster quality measures



4. Cluster quality measures



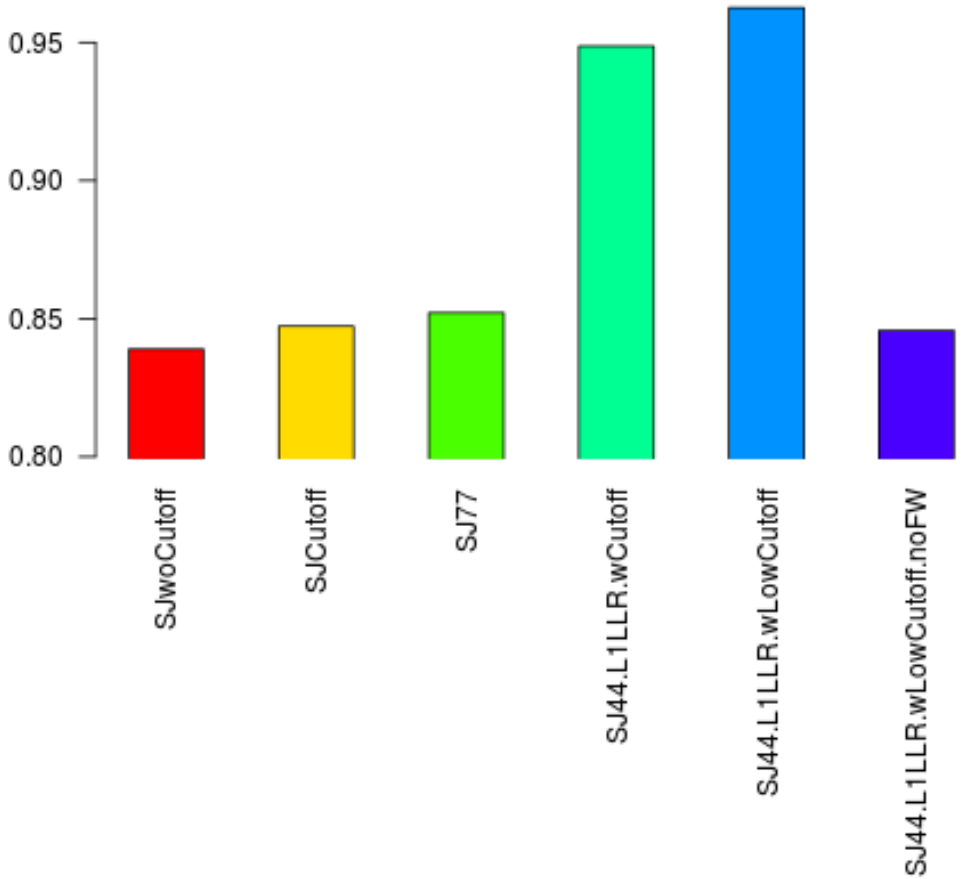
4. Cluster quality measures

Issue

- Visual intuition is not reflected in the clusterqual results.
- Are we actually looking at the same data?
  High-dimensional space $\leftrightarrow$ 2D co-ordinates.

4. Cluster quality measures

clusterqual after MDS: *monitor* tokens



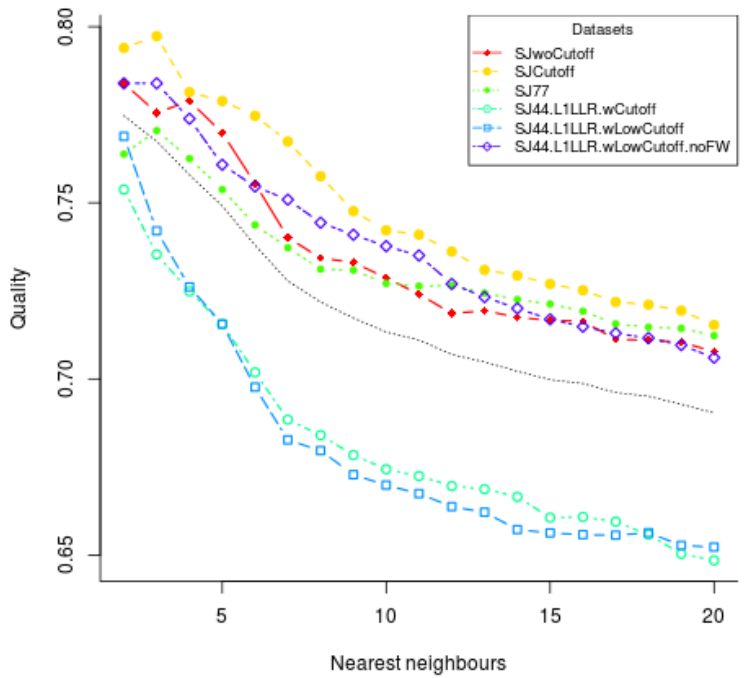
4. Cluster quality measures

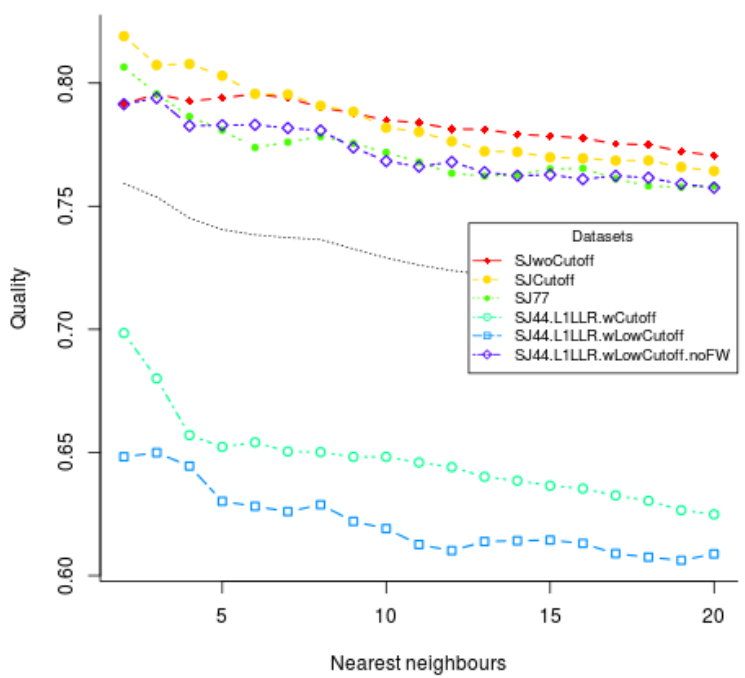
clusterqual after MDS

- The visual and mathematical analysis for the two parameter settings now correspond.
- However, the differences between the other solutions are completely flattened out.

4. Cluster quality measures

kNN-qual: *monitor* tokens





4. Cluster quality measures

Comparison of the rankings

clusterqual	kNN-qual
SJ44.L1LLR.wLowCutoff.noFW	SJCutoff
SJCutoff	SJ44.L1LLR.wLowCutoff.noFW
SJwoCutoff	SJwoCutoff
SJ44.L1LLR.wCutoff	SJ77
SJ77	SJ44.L1LLR.wCutoff
SJ44.L1LLR.wLowCutoff	SJ44.L1LLR.wLowCutoff

4. Cluster quality measures

clusterqual for SemEval WSI nouns

word	bow-bow	dep-dep
access.n	0.464	0.456
accounting.n	0.811	0.827
address.n	0.779	0.708
body.n	0.502	0.495
camp.n	0.702	0.697
campaign.n	0.69	0.698
cell.n	1.392	1.352
challenge.n	1.382	1.411
chip.n	0.975	0.975
class.n	0.483	0.471
...	...	

4. Cluster quality measures

word	bow-bow	dep-dep
function.n	0.553	0.499
gap.n	0.94	0.92
gas.n	1.477	1.442
guarantee.n	0.434	0.433
house.n	1.918	1.933
idea.n	0.481	0.487
innovation.n	1.222	1.203
legislation.n	0.974	0.982
margin.n	0.299	0.292
mark.n	1.776	1.742

Overview

5. Discussion and future work

To reduce or not to reduce?

- Does Multidimensional Scaling actually preserve the underlying semantics of a solution or should we use original distances?
- Singular Value Decomposition (SVD) has been shown to be a disputable choice for word similarity studies (Gamallo & Bordag 2011).
- Correlation between MDS stress and clusterqual discrepancy.

5. Discussion and future work

SemEval unbalanced samples

- Comparison over different types results in unreliable clusterqual results.
- Our cluster evaluation measures appear to be sensitive to unbalanced samples.

5. Discussion and future work

Other cluster quality indices

- Whole rang of other indices implemented in R *clusterCrit* package.

Summary of the talk

THEORETICAL

- Study the structure of lexical variation: mapping of meaning onto lexemes in different varieties.
- Analyse how this structure is apparent in usage data

METHODOLOGICAL

- Semantic Vector Spaces as a method for the quantitative, large-scale, corpus-based analysis of lexical semantics
- Interactive Visualisation of distributional models as an exploratory, visual analytic tool for lexicology
- Cluster quality measure for calibration of model parameters.

For more information:

http://wwwling.arts.kuleuven.be/qlvl

thomas.wielfaert@arts.kuleuven.be

kris.heylen@arts.kuleuven.be

dirk.speelman@arts.kuleuven.be