Token Spaces: A Tool for Comparing Word Meaning.

Thomas Wielfaert & Kris Heylen
QLVL, KU Leuven

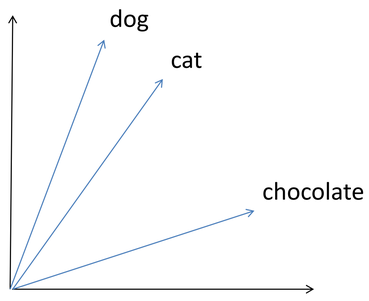
1. Semantic Vector Spaces: The Word-Context Matrix

Semantic Vector Spaces model word meaning in terms of frequency distributions of words over co-occurring context words. The idea is based on the Distributional Hypothesis in Linguistics: Words that appear in the same context tend to have a similar meanings (Harris 1954).

Type-level SVS (first order co-occurrences):

Example:
1) the **dog** was brought to the vet
2) the **cat** was brought to the vet
3) the **chocolate** lies in the cupboard

	lies	brought	vet	cupboard
dog	0	1	1	0
cat	0	1	1	0
chocolate	1	0	0	1

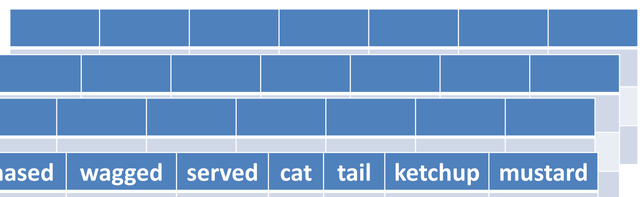


Optional weighting:
Not every word is as informative, for instance "vet" will be more informative than "was" in 1) and 2) for respectively *dog* and *cat*. We assign a weight to each context words based on how informative its appearance in the context of the target word is.

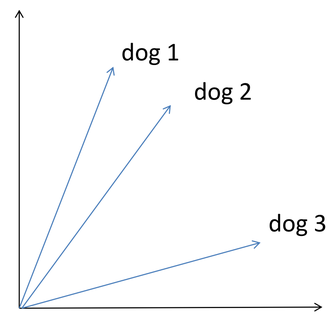
Literature: see Turney and Pantel (2010) for an overview of the types of Semantic Vector Spaces.

Token-level SVS (second order co-occurrences):

Example:
1) the **dog** chased the cat
2) the **dog** wagged his tail
3) the **hot dog** was served with ketchup and mustard



	chased	wagged	served	cat	tail	ketchup	mustard
dog 1	1	0	0	1	0	0	0
dog 2	0	1	0	0	1	0	0
dog 3	0	0	1	0	0	1	1



Disadvantage:
Large corpus results in a hyperdimensional space, uninterpretable for lexicographers and lexicologists.
Solution:
Statistical dimension reduction to 2 or 3 dimensions can be visualized in a scatterplot.

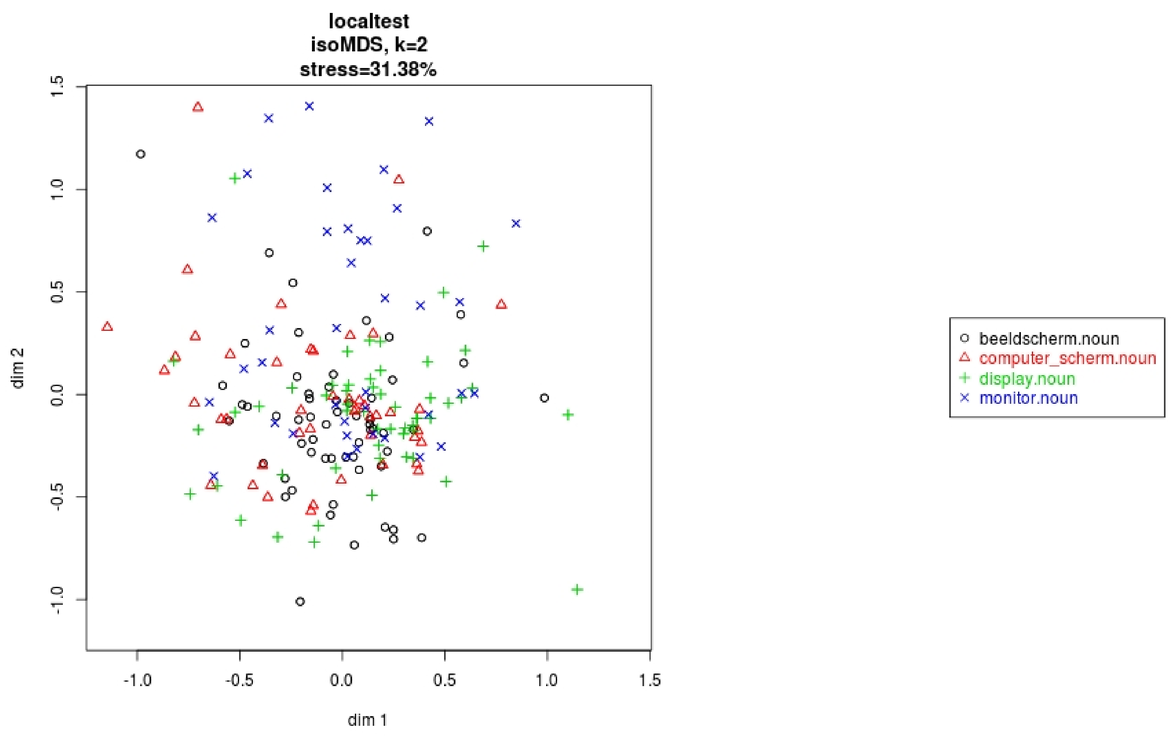
2. Data collection & parameters

Large corpora:
- TwNC : Dutch newspaper material, 500M words, 1999-2005
- LeNC: Belgian newspaper material, 1.3G words, 1999-2005
- Automatically lemmatized , POS tagged and parsed with Alpino (van Noord, 2006)

Test dataset:
800 tokens for the concept of COMPUTER SCREEN, consisting of 4 word types:
beeldscherm, *computerscherm*, *monitor* and *display*
Manually disambiguated by Daems (2012): "gold standard".

Parameter settings:
- Type vector window: 4 words on each side of the target word
- Token vector window: 10 words on each side of the target word
- Number of context words considered: 6328
- Vector dimensionality: 5430 context features
- Weighting scheme: Positive Pointwise Mutual Information (PPMI)
- Similarity measure: cosine

3. Token Space Visualization for all COMPUTER SCREEN word types



R scatterplot for the four word types after applying the statistical dimension reduction technique (isoMDS). Note that the stress is still quite high: 31%.

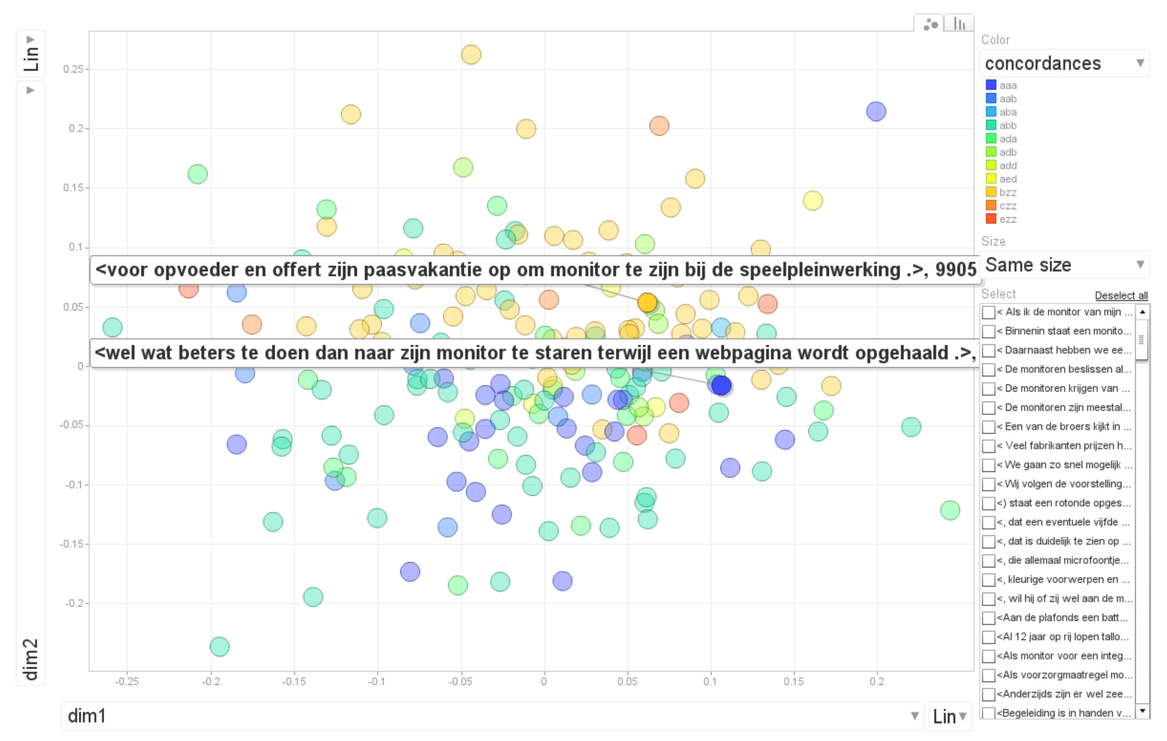
4. The polysemy of *monitor*

In Dutch, the word *monitor* can refer to two different concepts:
1) Monitor or computer screen, similar to English, Standard Dutch in BE and NL (concordance code starts with 'a')

Example:
Sneller internet willen we allemaal: [iedereen(0.0)] [heeft(0.0)] [wel(0.35)] [wat(0.0)] [beters(0.17)] [te(0.0)] [doen(0.0)] [dan(0.46)] [naar(0.57)] [zijn(0.0)] monitor te [staren(0.0)] [terwijl(0.13)] [een(0.58)] [webpagina(0.0)] [wordt(0.23)] [opgehaald(0.0)].
Faster internet we want all: everyone has something better to do than staring at his monitor while a web page is being retrieved.
Note the low weights for the informative context words *staren* (staring) and *webpagina* (web page). As a result, the token is positioned close to the "youth leader" tokens, despite being a true COMPUTER SCREEN.

2) A supervisor of (youth) leisure activities, slightly informal Dutch in BE (concordance code 'bzz')

Example:
Bart [Moens(0.0)] [studeert(0.0)] [voor(0.48)] [opvoeder(4.13)] [en(0.59)] [offert(0.0)] [zijn(0.0)] [paasvakantie(3.77)] [op(0.25)] [om(0.09)] monitor [te(0.0)] [zijn(0.02)] [bij(0.19)] [de(0.34)] [speelpleinwerking(6.61)].
Bart Moens studies education and sacrifices his Easter holiday to be leader at the playground activities.



5. References

Daems, J. (2012). Sociolectometric studies on a pluricentric language: theory and state of the art. Complemented by personal research on convergence and divergence in the Dutch lexicon. Unpublished report.
Harris, Z. (1954). Distributional structure. *Word*, 10(23): 146-162.
Heylen, K., Speelman, D. and Geeraerts, D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 16-26.
Turney, P.D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1), 141-188.
van Noord, G. (2006). At Last Parsing Is Now Operational. *Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles (TALN06).* Leuven, Belgium: Presses universitaires de Louvain, 20-42.

Thomas Wielfaert, Kris Heylen thomas.wielfaert@arts.kuleuven.be kris.heylen@arts.kuleuven.be
University of Leuven/Faculty of Arts/Quantitative Lexicology and Variational Linguistics