Cet article a été publié dans les actes de la Conférence TALN-RECITAL 2013.

Interactive visualizations of Semantic Vector Spaces for lexicological analysis

Thomas Wielfaert Kris Heylen Dirk Speelman
QLVL, University of Leuven, Faculty of Arts, Blijde-Inkomststraat 21/3308, B-3000 Leuven, Belgium
{thomas.wielfaert, kris.heylen, dirk.speelman}@arts.kuleuven.be

Résumé

Visualisations interactives des espaces vectoriels sémantiques pour l'analyse lexicologique

Dans le domaine de la Linguistique Computationnelle, les modèles sémantiques distributionnels sont devenus les piliers pour modéliser la sémantique lexicale à grande échelle. La modélisation distributionnelle présente également un grand potentiel pour la recherche dans le domaine de la Linguistique proprement dite. Elle permet aux linguistes de baser leurs analyses sur des grandes quantités de données authentiques et d'élargir ainsi considérablement leur base empirique, ainsi que de détecter des motifs sémantiques potentiellement intéressants. Cependant, jusqu'à présent, il y a eu relativement peu d'applications, principalement en raison de la complexité technique et de l'absence d'une interface conviviale permettant aux linguistes d'explorer les résultats obtenus. Dans cet article, nous proposons une visualisation interactive d'une matrice de similarité distributionnelle basée sur le Positionnement Multidimensionnel. Nous présentons un prototype d'un outil de visualisation que nous avons construit en Processing [1]. Il permettra d'ouvrir de nouvelles voies pour l'analyse visuelle des modèles au niveau de l'occurrence (*token*) et nous l'appliquons à une petite étude de cas d'un mot polysémique néerlandais.

Abstract

Within Computational Linguistics, distributional models of semantics have become the mainstay of large-scale modelling of lexical semantics. Distributional modelling also holds a large potential for research in Linguistics proper : It allows linguists to base their analysis on large amounts of usage data, thus vastly extending their empirical basis, and makes it possible to detect potentially interesting semantic patterns. However, so far, there have been relatively few applications, mainly because of the technical complexity and the lack of a linguist-friendly interface to explore the output. In this paper, we propose an interactive visualization of a distributional similarity matrix based on Multi-Dimensional Scaling. We present our prototype for a visualization tool built in Processing [1] which opens up new possibilities for the visual analysis of token-based models and apply it to a small case study of a Dutch polysemous word.

Mots-clés : visualisation, sémantique distributionnelle, espaces vectoriels sémantiques, niveau d'occurence, positionnement multidimensionnel, sémantique lexicale, lexicographie, néerlandais.

Keywords: visualization, distributional semantics, semantic vector spaces, token-level, multidimensional scaling, lexical semantics, lexicography, Dutch.

[1] http ://www.processing.org

1 Introduction

Distributional approaches to lexical semantics model a word's meaning in terms of the contexts it appears in. In Computational Linguistics this idea has been implemented in the form of high-dimensional co-occurrence matrices that can be manipulated with linear algebra. Under the heading of Vector Space Models (VSMs), Semantic Vector Spaces (SVSs) or Word Spaces, distributional models have since the middle of the 1990s, become the mainstay in statistical NLP for the large-scale analysis of lexical semantics (see Turney and Pantel (2010) for an overview). They have been applied to different tasks like Thesaurus Extraction, Word Sense Disambiguation or Lexical Entailment, and are traditionally evaluated in terms of precision, recall and F-score against a gold standard. On the other hand, distributional approaches have been used in Linguistics proper and Applied Linguistics to study lexical semantics for over half a century in the form of collocational analysis. However, these studies analyse co-occurrence frequencies on a much smaller scale than the computational Vector Space Models and tend to look at individual words in their most typical usage contexts. The identification and interpretation of more general patterns is typically left to the linguists themselves. The possibilities of Vector Space Models for large-scale semantic modelling and automatic pattern finding in Big Data have not been widely adopted in theoretical and applied linguistic subdisciplines. However, Vector Space implementations could greatly contribute to the analysis of lexical semantics in lexicology (the linguistic study of word meaning), provided that they are made accessible to lexicologists through an interactive visualization of their output that also allows easy access to the original input data and that enables the linguists to gauge effect of different parameter settings in the models. In this paper, we report on work in progress in the development of such a visualization and the features and functionalities that we added in response to the needs experienced in a practical case study. As such, the work in this paper is situated on the cross roads of Distributional Semantic Modelling in Computational Linguistics, Visual Analytics and (Applied) Linguistics. Although the creation of a tool for linguistic analysis is our first priority, such a tool could also allow computational linguists to tap into the expert knowledge of lexicologists and lexicographers to perform an in-depth, qualitative evaluation of their models that can complement the traditional precision and recall scores obtained on standard tasks.

Our visualization of Semantic Vector Spaces was guided by two traditional subtasks in lexicology : (1) identifying the different senses of a word in a large set of attestations of that word, and (2) charting the variation in word meaning and word choice between different varieties of the same language. Consequently, our vector spaces are situated in the Word Sense Induction paradigm and have vector representations for individual occurrences (tokens) of a word. In order to study lexical *variation*, word tokens are represented by vectors from multiple corpora that are representative of different varieties of the same language. Differences in word choice for the same concept are analysed by simultaneously modelling tokens from a set of near-synonymous words rather than from one word type at the time. This way we recreate the typical input for a lexicological analysis, viz. a long list of concordances of a set of near-synonyms, in which the lexicologist has to identify structure in terms of word senses and regional patterns. The idea is now that a token-level semantic vector space can help the lexicologist to discern the structure and patterns quicker and more easily through statistical modelling of word distributions. Importantly, the aim is NOT to provide a ready-made analysis of the semantic structure, but rather a starting point for analysis by a human expert. That expert can then interact with and improve upon the discovered semantic structure by inspecting the tokens' concordances. It is therefore important that the lexicologist does not experience the semantic vector space as a black box, but rather can see the effect of different parameter settings and play with them. Such more direct assessments of the semantics captured by distributional models is also desirable from a Computational Linguistic point of view : Baroni and Lenci (2011) point out that "To gain a real insight into the abilities of DSMs (*Distributional Semantic Models*, A/N) to address lexical semantics, existing benchmarks must be complemented with a more intrinsically oriented approach, to perform direct tests on the specific aspects of lexical knowledge captured by the models".

For this interactive and intrinsic assessment we first construct a number of token vector spaces where we vary parameters like weighting and context size. Next, we make two dimensional visualizations of the semantic structure that is captured by the high-dimensional vector spaces and we present these visualization in an interactive chart where the lexicologist can inspect the concordances and track the effect of the different parameters. The two dimensional visualization relies on a Multidimensional scaling analysis of the token-by-token similarity matrix that is calculated from the underlying token vectors. For the interactive functionality of the charts, we have tried out different packages including the Python Image Library and Google Motion Charts. As a response to their short-comings in desired functionality, we developed our own tool in the Java-based programming language Processing. The usability and functionality of the tool was tested in a case study where we look at a set of near-synonyms in Belgian and Netherlandic Dutch that refer to the concept COMPUTER SCREEN.

The rest of this paper is structured as follows. In the next section we discuss related approaches in distributional semantics and visual analytics. Section 3 introduces our corpus and our implementations of a token-level semantic vector space. In Section 4 we discuss the different visualizations we experimented with. In Section 5 we present our interactive visualization tool and discuss how its functionality was adapted in response to needs in the case study. Finally, Section 6 wraps up with conclusions and prospects for future research.

2 Related work

A number of approaches have applied distributional semantic models to the linguistic study of lexical variation. Diachronic variation, or meaning change, has been modelled on a type and a token-level. Gulordava and Baroni (2011), Cavallin (2012), Cook and Hirst (2011) try to detect new meanings of word types in general. Cook and Stevenson (2010) look specifically at the change in semantic orientation of word types. On a token-level, Sagi et al. (2009) aims to model semantic generalisation and specialisation processes. Rohrdantz et al. (2011) and Lau et al. (2012) try to identify tokens with new meaning using topic models. Regional variation in word semantics is modelled by Peirsman et al. (2010).

The visualization of distributional semantic models has most notably been undertaken by Rohrdantz et al. (2011) They explicitly aim to combine distributional semantics and visual analytics to track changes in word meaning. More specifically, they train a topic model on text snippets in which words occur that have undergone recent semantic change like *mouse* or *surf*. They then plot for each topic how the number of occurrences changed over time. This visualizes how topics that represent newer meanings, have more high-scoring tokens in more recent years.

Another approach to visualize high-dimensional semantic space was presented by Kievit-Kylar and Jones (2012). In their "Word-2-Word" visualization tool, words are represented as nodes and word similarities are represented by directed edges between these nodes. The software implements 22 common similarity-based metrics, both corpus- and WordNet-based, and the visualizations allow direct user input. They present various case studies varying from an artificial language, over the CHILDES corpus to a visual analysis of Bush and Obama State of the Union Adresses.

The most direct precursor of our approach is Heylen et al. (2012) who propose a token-level, interactive visualization of the regional variation in the semantics of Dutch near-synonyms. They present an interactive scatter plot using Google Chart Tools, but point out that these charts are not designed to display larger chunks of texts that are needed for informative concordances. It is this approach whose interactive visualization we try to optimize in this paper.

3 Corpus and Token-level SVS

The corpus for our study consists of Dutch newspaper materials from 1999 to 2005. For Netherlandic Dutch, we used the 500M words Twente Nieuws Corpus (Ordelman, 2002) [2], and for Belgian Dutch, the Leuven Nieuws Corpus (aka Mediargus corpus, 1.3 million words [3]). The corpora were automatically lemmatized, part-of-speech tagged and syntactically parsed with the Alpino parser (van Noord, 2006).

Token-level SVSs have been developed for modelling word meaning in context for different tasks (see Dinu et al. (2012) for an overview). The token-level SVS we use here is a fairly basic implementation and constitutes an adaptation of the early approach proposed by Schütze (1998). He models the semantics of a token as the frequency distribution over its so-called second order co-occurrences. These second-order co-occurrences are the type-level context features of the (first-order) context words co-occurring with the token. This way, a token's meaning is still modelled by the "context" it occurs in, but this context is now modelled itself by combining the type vectors of the words in the context. This higher-order modelling is necessary to avoid data-sparseness : any token only occurs with a handful of other words and a first-order co-occurrence vector would thus be too sparse to do any meaningful vector comparison. Note that this approach first needs to construct a type-level SVS for the first-order context words that can then be used to create a second-order token-vector.

In our study, we therefore first constructed a type-level SVS for the 573,127 words in our corpus with a frequency higher than 2. Since the focus of this study is visualization rather than finding optimal SVS parameter settings, we initially chose a number of settings that gave good results in our previous studies Peirsman et al. (2008); Heylen et al. (2008); Peirsman et al. (2010). For the context features of this SVS, we used two bag-of-words approaches, one with a window of 4 to the left and right around the targets and one with a window of 7 words on both sides of the context word. Furthermore, we experimented with two different weighting schemes : Positive Pointwise Mutual Information (PPMI), where negative PMI values are set to zero on the one hand and an adapted version of Log Likelihood Ratio (LLR) on the other. The context feature set was restricted to the 5430 words, that were among the 7000 most frequent words in the

[2] Publication years 1999 up to 2002 of *Algemeen Dagblad*, *NRC*, *Parool*, *Trouw* and *Volkskrant*

[3] Publication years 1999 up to 2005 of *De Morgen*, *De Tijd*, *De Standaard*, *Het Laatste Nieuws*, *Het Nieuwsblad* and *Het Belang van Limburg*

corpus, (minus a stoplist of 34 high-frequent function words) AND that occurred at least 50 times in both the Netherlandic and Belgian part of the corpus. The latter was done to make sure that Netherlandic and Belgian type vectors were not dissimilar just because of topical bias from proper names, place names or words relating to local events.

In a second step, we took a random sample of 100 Netherlandic and a 100 Belgian newspaper issues from the corpus and extracted all occurrences of 476 nouns in 218 semi-automatically constructed Dutch synsets by Ruette et al. (2012). For each occurrence, we built a token-vector by averaging over the type-vectors of the words in a window of 10 words to the left and right of the token. We experimented with two averaging functions. In a first version, we followed Schütze (1998) and just summed the type vectors of a token's context words, normalizing by the number of context words for that token :

$$\vec{o_i^w} = \frac{\sum_{j \in C_i^w}^{n} \vec{c_j}}{n}$$

where $\vec{o_i^w}$ is the token vector for the i^{th} occurrence of noun w and C_i^w is the set of n type vectors $\vec{c_j}$ for the n context words in the window around that i^{th} occurrence of noun w. However, this summation means that each first-order context word has an equal weight in determining the token vector. Yet, not all first-order context words are equally informative for the meaning of a token. In a sentence like "While walking to work, the teacher saw a dog barking and chasing a cat", $bark$ and cat are much more indicative of the meaning of dog than say $teacher$ or $work$. In a second, weighted version, we therefore increased the contribution of these informative context words by using the first-order context words' PMI or LLR values with the noun in the synset. PMI and LLR can be regarded as measures for informativeness and target-noun/context-word weights were available already from our large type-level SVS. The PMI or LLR of a noun w and a context word c_j can now be seen as a weight $pmi_{c_j}^w$ or $llr_{c_j}^w$. In constructing the token vector $\vec{o_i^w}$ for the ith occurrence of noun w with PMI as weight, we now multiply the type vector $\vec{c_j}$ of each context word with the PMI weight $pmi_{c_j}^w$, and then normalize by the sum of the pmi-weights :

$$\vec{o_i^w} = \frac{\sum_{j \in C_i^w}^{n} pmi_{c_j}^w * \vec{c_j}}{\sum_j^n pmi_{c_j}^w}$$

The same can be done with LLR as weight. The token vectors of all nouns from the same synset were then combined in a token by second-order-context-feature matrix. Note that this matrix has the same dimensionality as the underlying type-level SVS (5430). By calculating the cosine between all pairs of token-vectors in the matrix, we get the final token-by-token similarity matrix for each of the 218 synsets[4].

3.1 A polysemous case

Heylen et al. (2012) discusses the interesting case of the concept COMPUTER SCREEN in Dutch which is represented by three near-synonyms : *computerscherm*, *beeldscherm* and *monitor*. The

[4] string operations on corpus text files were done with Python 2.7. All matrix calculations were done in Matlab R2009a for Linux

latter appeared to have interesting properties : the first meaning of *monitor* is '(computer) screen' or 'display' in both Netherlandic and Belgian Dutch. In Belgian Dutch, however, *monitor* is polysemous and can also refer to a *supervisor of youth leisure activities*, henceforth *youth leader*[5].

To demonstrate the sensitivity of Semantic Vector Spaces models we limit our case study to the *monitor* tokens. In a token-level SVS for these tokens, one can expect two distinct token clouds : one with all the *screen* tokens and a second one with the *youth leader* tokens. The latter should only contain tokens from Belgian Dutch newspapers and we expect more instances from 'popular' newspapers than from 'quality' newspapers because the *youth leader* meaning is explicitly marked as Belgian Dutch in dictionaries and is thus not regarded as Standard Dutch.

To verify whether the spatial position of a token corresponds to the actual meaning, we needed manually disambiguation of the tokens.[6] Using their context, the tokens were manually assigned to three categories : 'display', 'youth leader' and 'other'. It is explicitly not the aim to apply time consuming manual disambiguation on a larger scale. We solely use this manual approach to analyse the effects of the different parameters on the result.

4 Towards an interactive visualization

The token-by-token similarity matrix is a high-dimensional matrix, which is not very informative by itself. To visualize it in 2 or 3 dimensional representation, we need to apply a dimension reduction technique. One way to visualize a similarity matrix is to apply Multidimensional Scaling (MDS) (Cox and Cox, 2000). MDS reduces the high-dimensional space to a 2 or 3 dimensional representation while respecting the original distances in high-dimensional space in the best possible way. We apply Kruskal's non-metric Multidimensional Scaling in R using the ISOMDS function from the MASS package. The degree of faithfulness to the original distance for the MDS solution is called stress. The stress values in this case study are roughly between 20 and 25% which is quite high, but still acceptable.

Our first attempt to visualize the token clouds was an R scatter plot (see Figure 1)[7]. This type of visualization allows to quickly eyeball the different types in the token cloud, but the level of interaction is rather low which make it fairly difficult to analyse a specific token[8]. In the next step, we integrated more of meta-data, such as country and newspaper and assign a colour code for the former categories so that in addition to word type, each country and newspaper gets its own colour. Heylen et al. (2012) uses Motion Charts from the Google Chart Tool to colour-code the data points in an interactive chart using the R-package GOOGLEVIS as an interface between R and the Google Motion Charts. These interactive charts for 218 Dutch near-synonyms (Ruette et al., 2012) can be explored online as well[9]. The chart for COMPUTER SCREEN[10] in Figure 2 shows

[5] A recent example from the Belgian 'popular' newspaper *Het Nieuwsblad* found in a report on a court case following an accident during a youth leisure activity : "Volgens het [H]of van [B]eroep was er onvoldoende toezicht door de monitoren." (According to the Court of Appeal there was insufficient supervision by the youth leaders.) `http://www.nieuwsblad.be/article/detail.aspx?articleid=DMF20130403_00528125`

[6] Jocelyne Daems deserves our eternal gratitude for performing this drudgery task as part of her M.A. project.

[7] This is the token cloud for the concept of COMPUTER SCREEN with 3 near-synonymous types : *beeldscherm*, *computerscherm* and *monitor*

[8] A more interactive version of the R plot can be consulted online : `https://perswww.kuleuven.be/~u0038536/committees/data/beeldscherm/`

[9] `https://perswww.kuleuven.be/~u0038536/googleVis`

[10] `https://perswww.kuleuven.be/~u0038536/googleVis/wSOCC/beeldscherm.html`

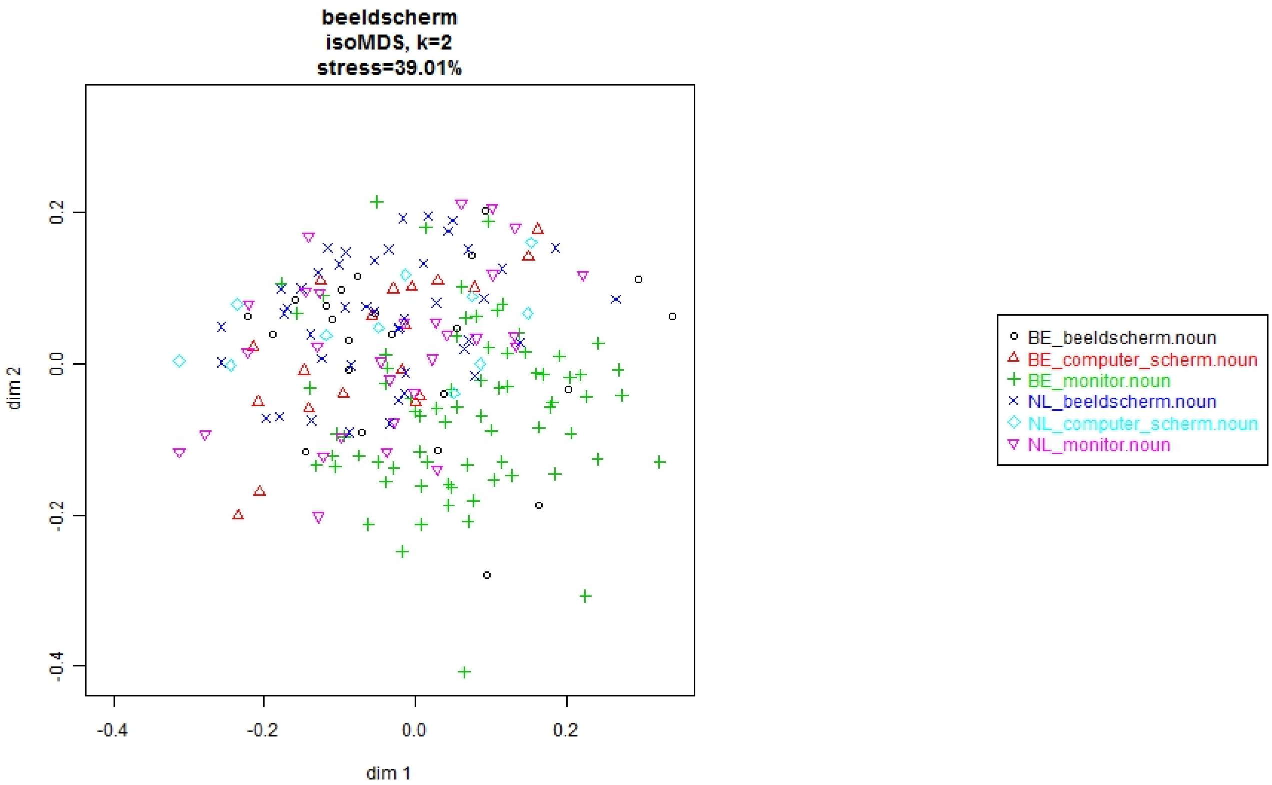


FIGURE 1 – Visualization of COMPUTER SCREEN

that the Belgian *monitor* tokens form a separate cloud on the left-hand side[11] of the plot.

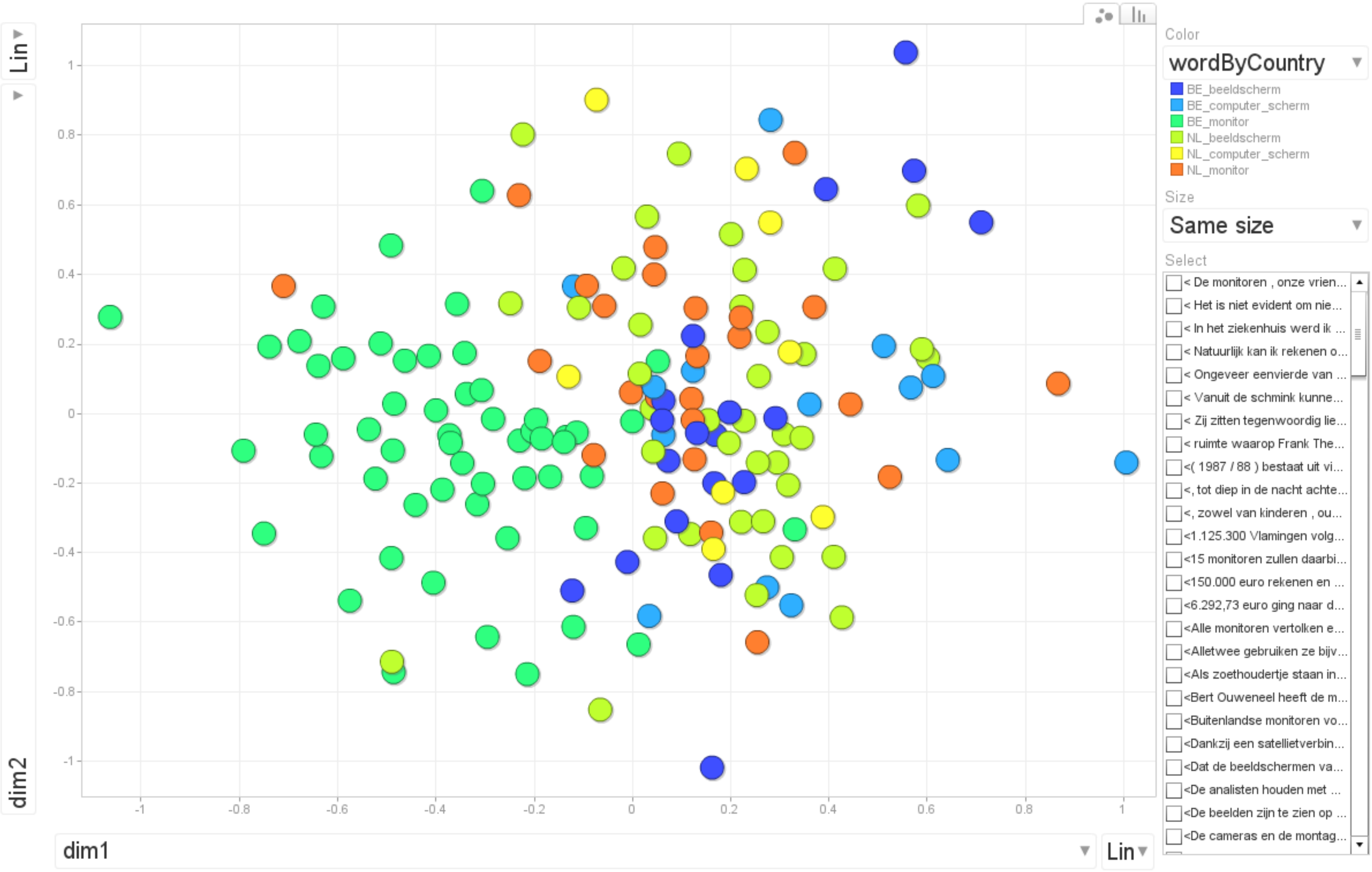
5 Case study : *monitor* tokens

As we base our tests on the manually disambiguated *monitor* tokens from both corpora, we use this example as a 'touchstone' to try different parameter settings. The visualizations become more informative by displaying the underlying properties of a token in the data points of the scatter plot. Despite the adaptability of the Google Charts it is not possible to display additional (textual) information next to the plot that could enhance the interpretation of the clouds. Therefore, we propose an early version of our own visualization tool, built in Processing[12].

For the evaluation of our manually disambiguated *monitor* tokens, we colour-code the tokens based on the manually assigned senses rather by country, newspaper or word type. The red dots are the 'youth leader' tokens, the green ones represent the 'display' tokens and we use blue for other categories. For now, we can ignore the blue, other tokens because of their small number

[11] The isoMDS algorithm can freely rotate the solution, so one should be careful with describing the plot in terms of left and right.

[12] Processing is a Java-based programming language for quick visual applications It allows to build GUIs with a fraction of the code Java Swing would require. Processing has a Javascript mode which relies and HTML5 instead of Java. See `http://www.processing.org` and `http://www.processingjs.org` for more information.



Color

wordByCountry ▼

- 🟦 BE_beeldscherm
- 🟦 BE_computer_scherm
- 🟩 BE_monitor
- 🟩 NL_beeldscherm
- 🟨 NL_computer_scherm
- 🟧 NL_monitor

Size

Same size ▼

Select

- □ < De monitoren , onze vrien...
- □ < Het is niet evident om nie...
- □ < In het ziekenhuis werd ik ...
- □ < Natuurlijk kan ik rekenen o...
- □ < Ongeveer eenvierde van ...
- □ < Vanuit de schmink kunne...
- □ < Zij zitten tegenwoordig lie...
- □ < ruimte waarop Frank The...
- □ <(1987 / 88) bestaat uit vi...
- □ <, tot diep in de nacht achte...
- □ <, zowel van kinderen , ou...
- □ <1.125.300 Vlamingen volg...
- □ <15 monitoren zullen daarbi...
- □ <150.000 euro rekenen en ...
- □ <6.292,73 euro ging naar d...
- □ <Alle monitoren vertolken e...
- □ <Alletwee gebruiken ze bijv...
- □ <Als zoethoudertje staan in...
- □ <Bert Ouweneel heeft de m...
- □ <Buitenlandse monitoren vo...
- □ <Dankzij een satellietverbin...
- □ <Dat de beeldschermen va...
- □ <De analisten houden met ...
- □ <De beelden zijn te zien op ...
- □ <De cameras en de montag...

FIGURE 2 – Google Motion Chart for COMPUTER SCREEN

and diverse nature. Interested readers can explore the online version [13] of the plot themselves.

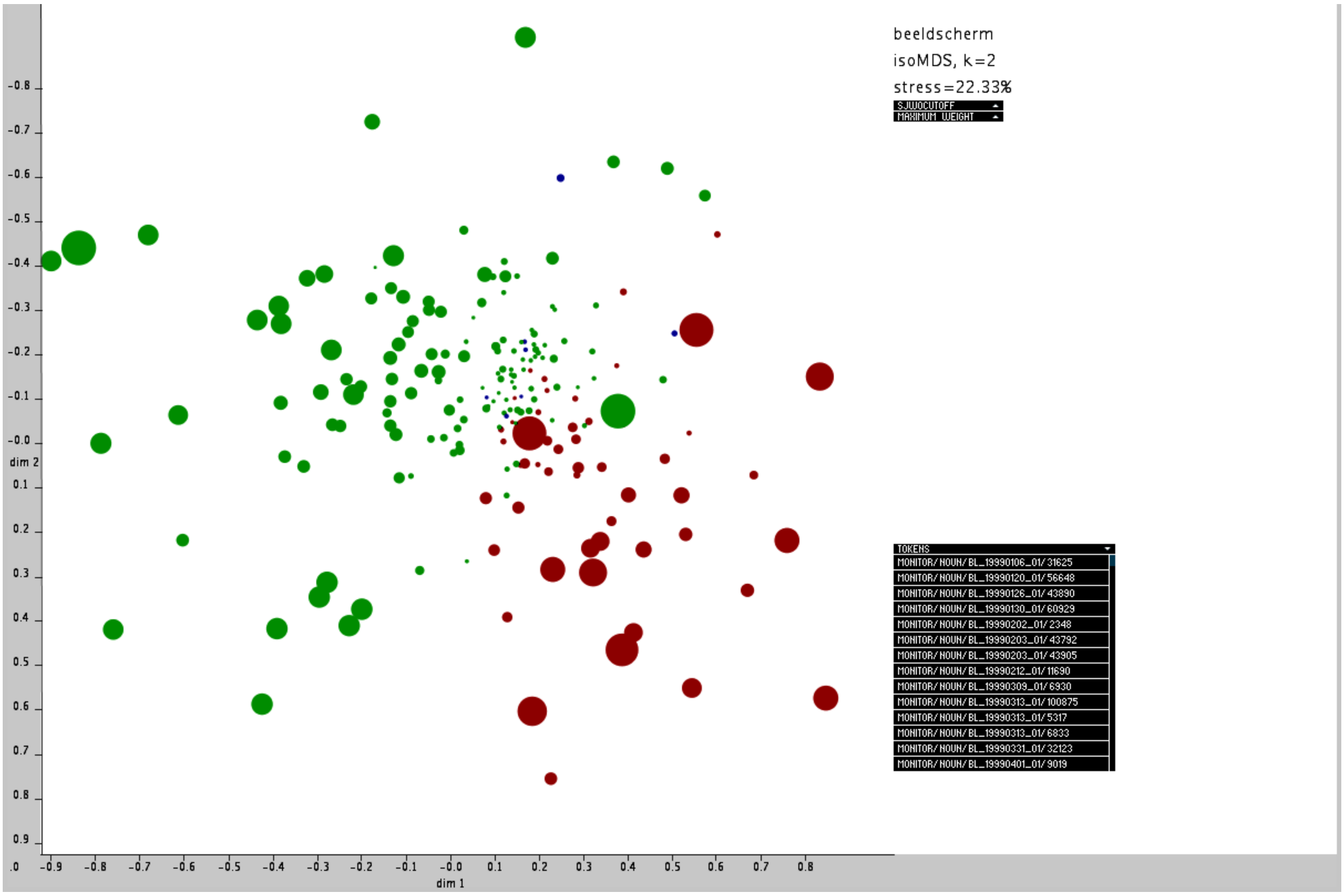

beeldscherm
isoMDS, k=2
stress=22.33%
SJWOCUTOFF
MAXIMUM WEIGHT
TOKENS
MONITOR/ NOUN/ BL_19990106_01/ 31625
MONITOR/ NOUN/ BL_19990120_01/ 56648
MONITOR/ NOUN/ BL_19990126_01/ 43890
MONITOR/ NOUN/ BL_19990130_01/ 60929
MONITOR/ NOUN/ BL_19990202_01/ 2348
MONITOR/ NOUN/ BL_19990203_01/ 43792
MONITOR/ NOUN/ BL_19990203_01/ 43905
MONITOR/ NOUN/ BL_19990212_01/ 11690
MONITOR/ NOUN/ BL_19990309_01/ 6930
MONITOR/ NOUN/ BL_19990313_01/ 100875
MONITOR/ NOUN/ BL_19990313_01/ 5317
MONITOR/ NOUN/ BL_19990313_01/ 6833
MONITOR/ NOUN/ BL_19990331_01/ 32123
MONITOR/ NOUN/ BL_19990401_01/ 9019
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
-0.0
dim 2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
.0
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
-0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
dim 1

FIGURE 3 – Processing implementation of the *monitor* tokens

In Figure 3 we can see three drop down menus, right of the plot. With the first one, we can choose the dataset we want to visualize. The selection shown in Figure 3 is ‘SJwoCutoff’, where ‘SJ’ refers to the name of the sample and ‘woCutoff’ (without cut-off) means that we used the original weights rather than replacing values below the cut-off with a (low) standard value. For the other parameters we used our standard settings : 10-10 first-order co-occurrences window, 4-4 second-order co-occurrences window and Positive Pointwise Mutual Information for the weighting. The second drop down menu can be used to change the size of the dot depending on the weights assigned to the words within the 10-10 context window for each token. We have implemented three options : the maximum weight, the sum of weights and the absolute number of non-zero weights. Exploring these parameters allows the user to get a better insight in the impact of the weighting on the token’s position.

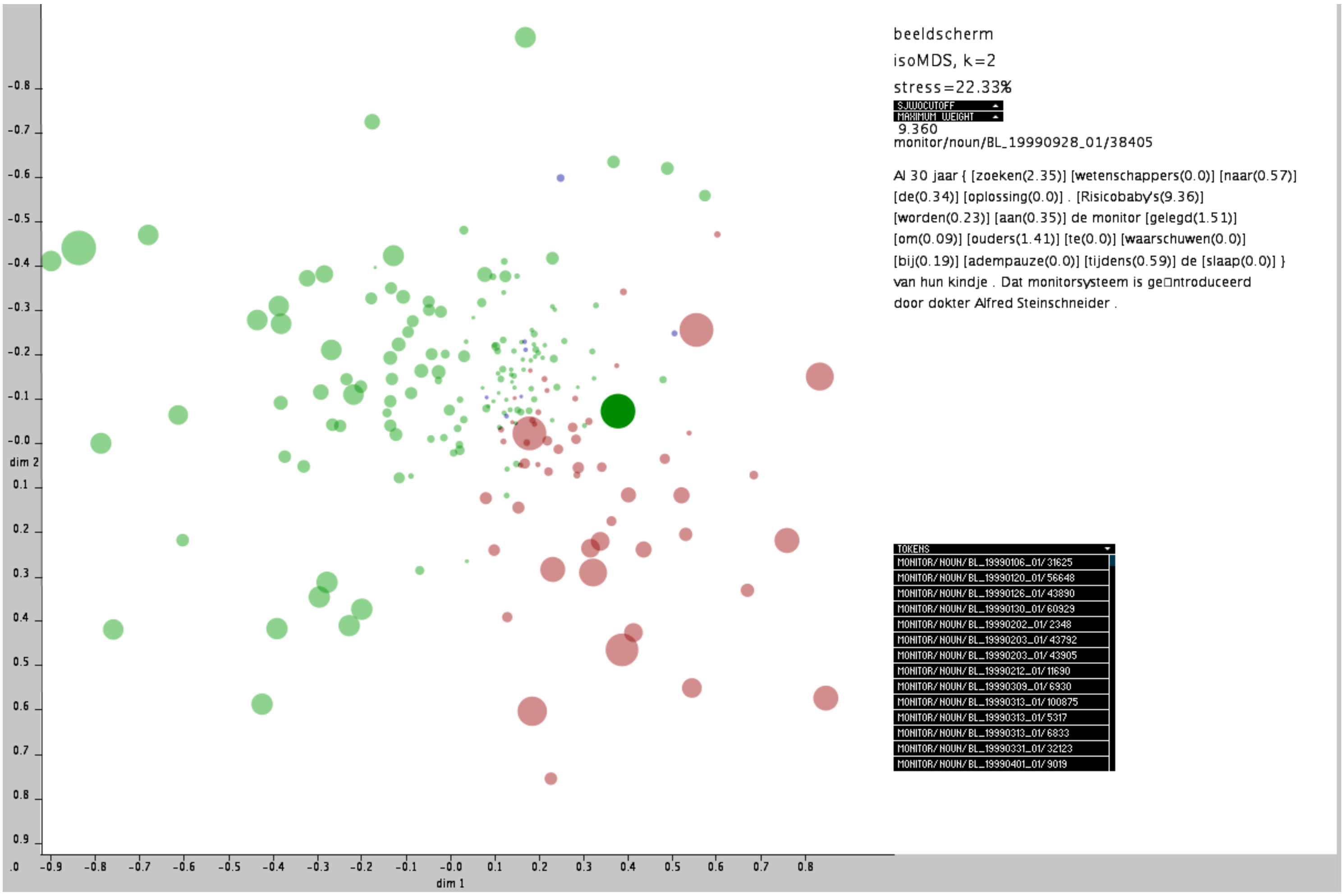
5.1 Misclassification analysis

Analysing a token’s position is particularly useful in the case of ‘misplaced’ tokens. In Figure 3 we can spot a very large green dot between the red ones in right half of the token space. This

[13] `https://perswww.kuleuven.be/~u0083608/tcViz/`

means that a ‘display’ token is closer to the ‘youth leader’ than to the ‘display’ sense despite a high-weighted context word. For improving the algorithm, it is indispensable to get an insight in the factors that pushed this token out of the ‘display’ token cloud where it belongs. We can analyse the token properties by clicking on the dot (see Figure 4) and examine the broader context of this token. The curly brackets in the text delimit the scope of the context window while the square brackets define the context words accompanied by their respective weight.

For the selected token in Figure 4, the context word *risicobaby* [14] (risk baby) has been assigned a weight of 9.36. As the size of the dot corresponds to the maximum weight over all the context words, *risicobaby* is the major factor that puts the token in the wrong cloud. Other context words with a weight above 1 should also be considered, namely *zoeken* (search), *leggen* (lay) and *ouder* (parent). Focussing on the context words weighted above 1 is somewhat arbitrary, but the function words, which are not considered very informative in bag-of-word models, tend to have a weight below 1 with PMI weighting.



beeldscherm

isoMDS, k=2

stress=22.33%

MAXIMUM WEIGHT 9.360

monitor/noun/BL_19990928_01/38405

Al 30 jaar { [zoeken(2.35)] [wetenschappers(0.0)] [naar(0.57)] [de(0.34)] [oplossing(0.0)] . [Risicobaby's(9.36)] [worden(0.23)] [aan(0.35)] de monitor [gelegd(1.51)] [om(0.09)] [ouders(1.41)] [te(0.0)] [waarschuwen(0.0)] [bij(0.19)] [adempauze(0.0)] [tijdens(0.59)] de [slaap(0.0)] } van hun kindje . Dat monitorsysteem is ge□ntroduceerd door dokter Alfred Steinschneider .

TOKENS

MONITOR/ NOUN/ BL_19990106_01/ 31625
MONITOR/ NOUN/ BL_19990120_01/ 56648
MONITOR/ NOUN/ BL_19990126_01/ 43890
MONITOR/ NOUN/ BL_19990130_01/ 60929
MONITOR/ NOUN/ BL_19990202_01/ 2348
MONITOR/ NOUN/ BL_19990203_01/ 43792
MONITOR/ NOUN/ BL_19990203_01/ 43905
MONITOR/ NOUN/ BL_19990212_01/ 11690
MONITOR/ NOUN/ BL_19990309_01/ 6930
MONITOR/ NOUN/ BL_19990313_01/ 100875
MONITOR/ NOUN/ BL_19990313_01/ 5317
MONITOR/ NOUN/ BL_19990313_01/ 6833
MONITOR/ NOUN/ BL_19990331_01/ 32123
MONITOR/ NOUN/ BL_19990401_01/ 9019

FIGURE 4 – Token cloud visualization with misclassified *risicobaby* token selected

The visualization with variable dot size allowed us to quickly eyeball misclassified tokens with misleading context words such as *risicobaby*. The word *risicobaby* has a very high weight (9.36) compared to the second most informative context word, which is *zoeken* (2.35). With a total number of 12 occurrences in our Dutch news corpora, *risicobaby* is a low-frequent word. The high weight of *risicobaby* illustrates two more general aspects. First, we use token-level SVSs because

[14] *risicobaby’s* is plural in the original text, but the corpora have been stemmed.

they are, in contrast with type-level SVSs, sensitive to polysemy of the co-occurrences and are capable of modelling subtle meaning differences. Nevertheless, our model is still sensitive to the properties of an individual type vector, i.e. *risicobaby*. Apparently, the context words of *risicobaby* overlap more with the typical context words a 'youth leader' than with those of a 'display'. Second, the large weight of *risicobaby* is an artefact of the PMI weighting ; Pointwise Mutual Information is sensitive to low-frequency events (Manning and Schütze, 1999). Low-frequent word types are thus more probable to get a high weight than high-frequent ones. In general, this property is not problematic because low-frequent words can be very informative co-occurrences. In this specific case though it associates the token with the wrong meaning and there are no other informative context words that could compensate.

We could remediate the frequency effect of PMI in two ways : use a larger context window for the weighting, for instance 7-7 instead of 4-4 and hope that a bigger window increases the number of informative context words and thus lowers the relative weight of *risicobaby*. We explored this option with the 'SJ77' dataset and we found that the weight of *risicobaby* indeed lowered (8.8), but not enough to 'push' it out of the 'youth leader' cloud. Another possible approach is to use a different association measure such as Log Likelihood Ratio instead of PMI. Again we can vary the window size or try different cut-off values. We invite the reader to compare these different parameter settings on our website.

6 Discussion

With our new tool to visualize similarity matrices we continued the work Heylen et al. (2012). Developing our own visualization tool allows us to imitate the visual appeal of the Google Chart Tools while remediating its shortcomings for text analytics. Our main goal remains to open up the black box of Semantic Vector Spaces, not only to computational linguists, but also to lexicographers and lexicologists with an interest in empirical quantitative data analysis. Of course, this is still a work in progress and there are some possible improvements.

First and foremost, it is clear that there is almost an infinite number of parameters that can be used in order to get more coherent token clouds. However, we need a more formal and efficient measure to evaluate the different solutions. We already tried a fairly simple and naive algorithm called CLUSTER QUALITY we borrowed from Speelman and Geeraerts (2009). The basic idea is that for every token, we sum the distances to all the other tokens that belong to the same manually disambiguated cluster and divide it by the distances to the tokens that belong to a different cluster. When we then average over all clusters, the result is the cluster quality value. Unfortunately, we found that this algorithm is so sensitive to outliers that the cluster quality of the real similarity matrix can not reliably be distinguished from a distance matrix created with random numbers. A possible solution is taking only the k-nearest neighbours into account. The value of k is something that needs to be determined in future research. This approach is of course only useful if we have manually disambiguated data. Obtaining these is very labour intensive and contrary to our goal to use unsupervised learning for lexical semantics. For the analysis of the larger dataset, we might have to experiment with traditional cluster analysis algorithms. The number of clusters should then ideally be decided by the user and entered as a parameter into the visualization tool.

Secondly, we need to further explore the possibilities to make the SVS's black box mechanisms more accessible. We can now visualise the weights of the first-order co-occurrences and explore some simple calculations based on these weights, but there are more underlying parameters that could be unveiled. We could for instance make the properties of the second-order co-occurrence vectors explicit.

Thirdly, creating 3D plots is still on our wish list. Not only is it visually more appealing, it could also reduce the stress of the isoMDS solution. For our *monitor* tokens, the stress values are moderate (roughly ranging between 0.2 and 0.25), but we restricted the plots to one word type in order to explore the parameter settings.

Finally, the plots do not allow real user input, mainly because the Semantic Vector Spaces are calculated in advance. There are technical and practical limits for using real time calculations that are beyond the scope of this paper. However, we could for instance let the user move wrongly positioned tokens and recalculate the MDS co-ordinates based on these corrections. The ISOMDS algorithm is fast and requires little resources which makes a real time calculations feasible. This kind of feature could potentially benefit other fields which make use of high-dimensional similarity matrices.

References

Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cavallin, K. (2012). Automatic extraction of potential examples of semantic change using lexical sets. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 370–377.

Cook, P. and Hirst, G. (2011). Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language Information and Computation (PACLIC 25)*, pages 265–274, Singapore.

Cook, P. and Stevenson, S. (2010). Automatically identifying changes in the semantic orientation of words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 28–34, Valletta, Malta.

Cox, T. F. and Cox, M. (2000). *Multidimensional Scaling, Second Edition*. Chapman and Hall/CRC, 2 edition.

Dinu, G., Thater, S., and Laue, S. (2012). A comparison of models of word meaning in context. In *Proceedings of the 2012 Conference of the the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 611–615.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, pages 67–71.

Heylen, K., Peirsman, Y., Geeraerts, D., and Speelman, D. (2008). Modelling Word Similarity. An Evaluation of Automatic Synonymy Extraction Algorithms. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2008)*, pages 3243–3249, Marrakech, Morocco. ELRA.

Heylen, K., Speelman, D., and Geeraerts, D. (2012). Looking at word meaning. an interactive visualization of semantic vector spaces for dutch synsets. In *Proceedings of the EACL 2012 Joint*

Workshop of LINGVIS & UNCLH, pages 16–24, Avignon, France. Association for Computational Linguistics.

Kievit-Kylar, B. and Jones, M. (2012). Visualizing multiple word similarity measures. *Behavior research methods*.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.*, pages 591–601. Association for Computational Linguistics.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Ordelman, R. (2002). Spoken document retrieval for historical video archives - dutch speech recognition in the echo project. Technical report, University of Twente, Parlevink Group.

Peirsman, Y., Geeraerts, D., and Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4) :469–491.

Peirsman, Y., Heylen, K., and Geeraerts, D. (2008). Size matters : tight and loose context definitions in English word space models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 34–41, Hamburg. ESSLLI.

Rohrdantz, C., Hautli, A., Mayer, T., Butt, M., Plank, F., and Keim, D. A. (2011). Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (Short Papers)*, pages 305–310, Portland, Oregon, USA. Association for Computational Linguistics.

Ruette, T., Geeraerts, D., Peirsman, Y., and Speelman, D. (2012). Aggregating dialectology and typology : linguistic variation in text and speech, within and across languages. In Szmrecsanyi, B. and Wälchli, B., editors, *Linguistic variation in text and speech, within and across languages*. Mouton de Gruyter, Berlin.

Sagi, E., Kaufmann, S., and Clark, B. (2009). Semantic density analysis : Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS : GEometical Models of Natural Language Semantics*, pages 104–111, Athens, Greece.

Schütze, H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24(1) :97–123.

Speelman, D. and Geeraerts, D. (2009). The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing*, 2 :221–242.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *CoRR*, abs/1003.1141.

van Noord, G. (2006). At Last Parsing Is Now Operational. In *Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles (TALN06)*, pages 20–42, Leuven, Belgium. Presses universitaires de Louvain.