Available online at www.sciencedirect.com

ScienceDirect

Lingua 157 (2015) 153–172

Lingua

www.elsevier.com/locate/lingua

Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis

K. Heylen *, T. Wielfaert, D. Speelman, D. Geeraerts

QLVL – KU Leuven, University of Leuven, Belgium

Received 4 February 2014; received in revised form 2 December 2014; accepted 3 December 2014
Available online 3 February 2015

Abstract

This paper demonstrates how token-level Word Space Models (a distributional semantic technique that was originally developed in statistical natural language processing) can be developed into a heuristic tool to support lexicological and lexicographical analyses of large amounts of corpus data. The paper provides a non-technical introduction to the statistical methods and illustrates with a case study analysis of the Dutch polysemous noun 'monitor' how token-level Word Space Models in combination with visualisation techniques allow human analysts to identify semantic patterns in an unstructured set of attestations. Additionally, we show how the interactive features of the visualisation make it possible to explore the effect of different contextual factors on the distributional model.
© 2014 Elsevier B.V. All rights reserved.

Keywords: Distributional models; Lexical semantics; Statistical analysis; Visual analytics

1. Introduction

Compared to other linguistic disciplines, corpus-based analyses have a strong and long tradition in lexical semantics. Ever since the rise of philology and the emergence of large-scale dictionary projects in the 19th century, lexical semanticians have relied on contextual clues in attested language use to infer and organise the different senses and uses of a word. And while in the 1950s syntax research turned away from usage data, the ideas of John Rupert Firth (1957), Zelig Harris (1954) and Warren Weaver (1955) led to approaches that saw real language data as the natural empirical basis for semantic descriptions. (For a more extended history of recent corpus-linguistic approaches to lexical semantics, see Geeraerts, 2010: 165–178.) Initially, collecting and analysing corpus data was mainly manual labour, but with the advent of computers and ever larger electronic corpora, lexicologists and lexicographers now have enormous amounts of naturally occurring usage data available to base their descriptive work on. To analyse this wealth of data, scholars of lexical semantics widely use statistical analysis tools. More specifically, statistical methods have been introduced to facilitate two distinct steps in the analysis. On the one hand, statistical methods are used for identifying contextual clues in the corpus data that are indicative of a given lexeme's meaning. These include co-occurring words (collocations) and syntactic patterns (colligations). On the other hand, statistical approaches are employed for classifying the occurrences of a lexeme into distinct usages and senses based on these contextual clues.

* Corresponding author. Tel.: +32 16329998; fax: +32 16324713.
E-mail address: kris.heylen@kuleuven.be (K. Heylen).

http://dx.doi.org/10.1016/j.lingua.2014.12.001
0024-3841/© 2014 Elsevier B.V. All rights reserved.

Table 1
Methods of sense identification in lexical semantics.

	Identifying contextual clues	Classifying occurrences
Philology	Manual	Manual
Collocation analysis	Statistical	Manual
Behavioural profiles	Manual	Statistical
Word Space Models	Statistical	Statistical

The first approach (i.e. the introduction of statistical methods for the identification of contextual clues) has been mainly associated with the British tradition in corpus linguistics. Pioneered by John Sinclair (1991), this approach described lexical meaning as a function of the typical words (collocations) and syntactic patterns (colligations) that a word co-occurs with. Church and Hanks (1989) introduced statistical measures (*t*-score), to identify salient and informative collocations and colligations based on frequency distributions in text. These measures were subsequently refined (see Evert, 2004 and Wiechmann, 2008 for an overview), and they are now widely used in various linguistic subdisciplines.

The second approach (i.e. statistical clustering of usages) has an outspoken presence in recent developments in Cognitive Semantics. Specifically, the so-called Behavioural Profile approach has introduced multivariate statistical techniques to classify occurrences of a word automatically into distinctive senses and usages, based on corpus evidence.[1] Gries (2006) uses hierarchical cluster analysis to group occurrences of the verb *to run* into different senses based on contextual features like transitive use or co-occurring spatial prepositions. Glynn (2010) applies Correspondence Analysis to visualise how occurrences of the verb *to bother* are grouped into distinct usages based on syntactic behaviour and semantic characteristics like affect.

Interestingly, these statistical methods have been used independently of each other in these different traditions. Collocation-based analyses have statistically automated the identification of contextual clues but leave the classification of occurrences and typical contexts to manual analysis. Almost as a mirror image, behavioural profile analyses have statistically automated the classification of a lexeme's occurrences and typical contexts into senses, but predominantly use datasets with manually coded contextual features as input. Table 1 classifies different approaches in lexical semantics by whether they identify contextual clues and senses manually or through statistical analysis. Whereas classical philological studies did both steps manually, collocation studies and behavioural profile analyses have each by and large automated one of the two steps but not both.

In this paper, we will introduce Word Space Models (a.k.a. Semantic Vector Space Models) as a logical extension of the statistical state-of-art in support of lexical semantic analysis: a technique that essentially combines collocational measures and multivariate methods in a systematic way to explore lexical semantic structure in large corpora, it completes the pattern that emerges from the introduction of statistical tools in corpus-based lexical semantics, as summarised in Table 1.

The introduction of statistical methods to both steps in the data analysis process is, we think, not only a logical extension, but also a necessary one and this, for two reasons. First, additional support of statistical pattern finding techniques is the only way for lexicologists and lexicographers to cope with the data deluge that they face as they pursue their traditional descriptive work. Given the available wealth of corpus data, it is simply unfeasible to hand-code, classify or describe thousands upon thousands of concordances of a lexeme. Statistical data analysis can help to take a representative sample of different usages for further scrutiny. Secondly, so-called Big Data also suggests an extension of the traditional focus of lexicological and lexicographical work: the Big Data environment allows scholars to investigate trends and patterns that could not be studied in smaller corpora, e.g. the spreading of new words or new usages of existing words through social networks. Data mining techniques are indispensable to monitor this type of trends. As Word Space Models are already the principal technique for handling lexical semantics in Computational Linguistics, we suggest that they can also provide support for the lexicologist and the lexicographer in their traditional and new descriptive tasks: for the traditional task of describing individual words or lexical fields, support is needed for staying on top of the abundance of data, and for the new task of describing trends and developments, appropriate quantitative techniques need to be developed.

In this paper, we will focus on how a semantic space approach can support a lexicological analysis of polysemy in large corpora. It should be noted, though, that we are explicitly not presenting the Semantic Vector Space approach as a ready-made, stable technique. Rather, we will argue that in its current, computational linguistic implementation, the technique is

[1] Within the behavioural profile approach, there are also studies of lexical alternations, for which other multivariate methods like regression are available. However, since our focus is on the polysemy of 1 lexeme, rather than the alternation between multiple lexemes for the same concept, we do not go further into these.

too much of a black box to be suitable for in-depth lexical analysis, and that it needs be extended with more interactive features to become truly useful for lexicological and lexicographical purposes. As part of work in progress, we will demonstrate how such features might take the form of visual analytics tools.

In the next section we first give an informal discussion of the technique behind word spaces and how they model semantic structure and polysemy. We also discuss why these models are not transparent enough in their current form for lexicological research and how this can be remedied. Section 3 introduces a case study of a polysemous word in Dutch and demonstrates how word spaces in combination with visual analytics can be used to analyse the polysemy. Section 4 offers a general discussion and sets out a programme for future work. Section 5 wraps up with a summary and conclusion.

2. Word Space Models: implementations

Word Space Models were initially introduced in Cognitive Psychology to model lexical memory (Landauer and Dumais, 1997; Lund and Burgess, 1996) and then further developed in Computational Linguistics, where they are now the mainstay of semantic modelling in statistical Natural Language Processing (see Turney and Pantel, 2010 for an overview). Word Spaces Models, a.k.a. Semantic Vector Spaces or Distributional Models of Lexical Semantics are a family of approaches and different subtypes can be distinguished. Two dimensions are relevant here. First, models differ with respect to the type of context that is taken into account to model word meaning. The context definition can be either document-based or word-based. When studying word x, the basic observation in a document model is the occurrence of x in the context of a given document as a whole; more fine-grained word-based models by contrast take their observational starting-point in the occurrence of x in the context of another word, optionally taking into account the specific syntactic dependency relation between x and its context words. Document-based models have proven to be most suited for modelling syntagmatic and associative relations, for instance between *doctor* and *hospital* or *car* and *drive.* Word-based models, and especially syntactically informed ones, are better at capturing paradigmatic relations like the near-synonyms *hospital-clinic* (Sahlgren, 2006). Second, Semantic Vector Spaces allow words to be studied at type level or at token level, i.e. we can try to distinguish one word from the other, or we can try to distinguish one usage of a given word from another usage of the same word. In the former case, we are interested in synonymy and related lexical relations; in the latter, we are interested in polysemy. In the context of this paper, we are obviously interested in the latter approach, but we have to start by explaining type-level models, because the token-level models are methodologically speaking an extension of the type-level.

2.1. Type-level models

The type-level Semantic Vector Spaces can be most easily understood through an English toy example. Suppose we are interested in three target words: *dog*, *cat* and *coffee.* We have a corpus consisting of the following simple sentences:

1. The *dog* barked loud at the passerby.
2. The vet grabbed the *dog* by its neck.

3. The *cat* was purring loud.
4. The vet was scratched by the *cat* while grabbing it.

5. *Coffee* tastes better than tea.
6. The vet grabbed his cup and tasted his *coffee*.

We now turn this corpus into a small frequency table where the target words form the rows and the context features are represented in the columns. We use the simplest model possible: it ignores syntactic relations and merely counts raw frequencies of the nouns, verbs and adverbs. Disregarding the syntax in this way is called the 'Bag-of-Words approach' in Information Retrieval. There are also more complex models that do take into account syntactic dependencies, but for the sake of simplicity we will not go deeper into these. In a real world example with a large corpus, there will be far more context features (columns, typically several thousands) and the top frequencies will of course be much higher than 2 (Table 2).

The co-occurrence frequencies can be interpreted as coordinates that situate *dog, cat*, and *coffee* in a high-dimensional semantic space (hence the name Word Space Model), where each context word is a separate dimension. Developing this geometrical metaphor, we can also calculate the distance between two words in this high-dimensional space to measure how far their meaning is apart. In practice, vector algebra is used to compute the similarity between the three target words (or rather their vector representation) by taking the cosine of the angle between them, which is a

Table 2
Co-occurrence matrix for the target words dog, cat and coffee.

	bark	passerby	vet	grab	neck	purr	loud	scratch	taste	better	tea	cup
dog	1	1	1	1	1	0	1	0	0	0	0	0
cat	0	0	1	1	0	1	1	1	0	0	0	0
coffee	0	0	1	1	0	0	0	0	2	1	1	1

standard measure in Distributional Semantics (see e.g., Bullinaria and Levy, 2007). Intuitively, the angle between two similar concepts (i.e. cat and dog) is expected to be smaller than that between different words (i.e. dog and coffee). In other words: dissimilar word distributions result in a lower cosine similarity value:

cos(dog, cat) = 0.55
cos(dog, coffee) = 0.27
cos(cat, coffee) = 0.30

The result corresponds to basic intuition: 'dog' and 'cat' are most similar while 'dog' and 'coffee' share the least co-occurrences. We also note that 'dog' is slightly less similar (0.27) to 'coffee' than 'cat' (0.30) because 'dog' co-occurs with 'bark' and 'passerby' while 'cat' and 'coffee' do not.

The real world implementation of this technique will usually not use raw co-occurrence frequencies between the target word and its co-occurrences. High frequency words are not necessarily the most informative ones for the meaning of the target word. As in collocation studies, Word Space Models therefore use statistical measures of collocational strength to give a higher weight to context words that co-occur significantly more often than expected by chance. These high-weighted collocates are more informative for the meaning of the target word than others, irrespective of their raw frequency. In the example with the target word 'dog' for instance, the collocate 'vet' (as an abbreviation of veterinary), is semantically more closely related to the concept of ANIMAL than the word 'grab', even though it is less frequent. The set of collocational measures used in word space modelling is approximately the same as in collocational analysis. The weighting schemes we use for our present case study are based on Pointwise Mutual Information (PMI) and Log-Likelihood Ratio (LLR). The technical details of these measures are beyond the scope of this paper but see Thanopoulos et al. (2002) for a comparison of these and alternative collocational measures. Suffice to say that in general, weighting by collocational strength is a way to increment the importance of more informative context words and at the same time lower the influence of collocates that do not have discriminatory properties for the target's meaning. To calculate collocation strength and measure reliably whether two words co-occur more often than expected by chance, co-occurrence frequencies between a significant number of words must be collected from a corpus. In our toy example, we just impute the collocational weights calculated in a much larger corpus. Suppose *vet* gets a collocational weight of 4.3 if it co-occurs with *dog*, '3.5' with *cat* and only 0.8 for *coffee*. When the cosine similarity is then recalculated, *dog* and *cat* will be more similar to each other than in the unweighted calculation and less similar than before to *coffee*.

cos(dog, cat) = 0.87
cos(dog, coffee) = 0.31
cos(cat, coffee) = 0.32

Typically, weighted co-occurrence vectors are constructed for a sizeable part of a language's vocabulary as target words and a taking into account a few thousand mid-frequent context words. Table 3 shows how some of the weighted co-occurrence frequencies of our examples are situated in a matrix of 50.000 target words and 5000 context words. Calculating the cosine similarity between all pairs of target words results in similarity matrix with target words as both rows and columns and the cosine similarity between all target word pairs in the cells. The matrix has 1s on the diagonal because each target word is completely similar to itself. The matrix is also symmetrical, with the same values above and below the diagonal, because the cosine similarity between word A and B is the same as between word B and A. Table 4 shows the cosine similarities for our examples in the type-level similarity matrix of 50.000 target words by 50.000 target words. For each target word, it is now possible to look up the most similar word in the rest of the vocabulary. Assuming that very similar words are often near-synonyms, this type of word-by-word similarity matrices is typically used in computational linguistics for the task of automatic synonymy extraction.

The type-level model, as is self-evident from its name, models the distributions on a word type level. This means that the resulting semantic vector is an aggregation of the individual occurrences in the corpus, making it hard to use it to study

Table 3
Type-level matrix of PMI weighted co-occurrence frequencies.

	CW1	...	bark	cup	vet	...	$CW_{5.000}$
TW1	$w_{1,1}$						$w_{1,5k}$
...							
cat			1.2	0.2	3.5		
coffee			0.2	5.2	0.8		
dog			5.1	0.3	4.3		
...							
$TW_{50.000}$	$w_{50k,1}$						$w_{50k,5k}$

Table 4
Type-level matrix of PMI weighted co-occurrence frequencies.

	TW_1	...	cat	coffee	dog	...	$TW_{50.000}$
TW_1	1						
...		1					
cat			1	0.32	0.87		
coffee			0.32	1	0.31		
dog			0.87	0.31	1		
...						1	
$TW_{50.000}$							1

polysemy. Suppose we have a polysemous word which has a balanced distribution over its different senses. On a type-level, this word would be modelled as the average meaning, smoothing out the different contexts in which these different meanings are (supposedly) embedded. Simplistically put, a type-level approach acts as if each word has only one meaning – not a satisfactory perspective for Lexical Semantics. Within Computational Linguistics, two types of solutions have been developed to overcome this limitation and model lexical polysemy.[2] The first type of solution stays within a type-level approach, but tries to find patterns within the type-level matrices that are indicative of different senses. These approaches first retrieve a set of semantically related words for a polysemous lexeme and within this set, using the same type-level methods, they identify groups of semantically similar words that are indicative of different senses (e.g. Pantel and Lin, 2002; Tomuro et al., 2007; Tamm and Sahlgren, 2014). However, any type-level approach necessarily aggregates over specific instances and abstracts away from the individual occurrences that realise the different meanings of a polysemous lexeme. Since these specific realizations are important in a qualitative, in-depth lexicological or lexicographical analysis, we turn to a second type of Word Space Models that have been developed to handle word's internal semantic structure and that do model polysemy on the level of individual tokens.

2.2. Token-level models

Token-level Semantic Vector Spaces allow more fine-grained modelling on the level of the individual occurrences of a word. While we lose the ability to look directly at the aggregate level, these token-level models allow a more detailed insight in the semantic properties of the individual occurrences.

Suppose we compile a second toy corpus with three sentences containing the word *cat*, but now in one of them *cat* has the specialised meaning of a type of sailing boat:

1. Blofeld was stroking the purring *cat* in his lap
2. The black dog barked at the *cat* in the tree
3. The cadet was sailing his *cat* against the wind

In a naive type-based approach, the infrequent meaning of sentence 3 would simply be lost and go unnoticed in the aggregation over all occurrences of *cat* and even a type-level approach that does take into account polysemy would probably not pick up this very infrequent meaning In a token-based model however, we will model each occurrence separately and the boat-meaning can be distinguished from the more frequent animal meaning.

[2] See Navigli (2009, 2012) and Turney and Pantel (2010) for a more in depth discussion of the two types of solutions.

Table 5
Co-occurrence matrix on token-level for the sentences in the toy corpus.

	bark	cadet	dog	lap	purr	sail	stroke	tree	wind
#1	0	0	0	1	1	0	1	0	0
#2	1	0	1	0	0	0	0	1	0
#3	0	1	0	0	0	1	0	0	1

Token-level models have been developed in Computational Linguistics since the mid 1990s for the task of automatic word sense disambiguation. By now a plethora of algorithms is available (see Agirre and Edmonds, 2006; Navigli, 2009, 2012 and Dinu et al., 2012 for an overview) but the method used in the our case study closely follows the seminal work by Hinrich Schütze (1998). Essentially, this approach models an individual token of a word by averaging over the type vectors of the context words. For example in sentence 1, the word 'cat' is modelled by averaging over the co-occurrence frequencies of the words 'stroking', 'purring' and 'lap'. Let us explain this method in more detail.

Suppose we create a matrix from our toy corpus of three sentences with a row for every occurrence of 'cat' (Table 5). Proceeding naively, we could still compute the cosine similarity between these tokens based on the raw frequencies in the table. However, note that token vectors #1 and #2 do not share a single co-occurring context word even though both usages refer the animal meaning. Consequently the similarity between all three tokens would be 0 and we would have no way of telling that token #1 and #2 share the same meaning whereas token #3 refers to another meaning of *cat.* In a real-world corpus this situation is even aggravated because the majority of informative context words will co-occur in only a limited amount of token-level observations. Suppose there are a 1.000 different context-words co-occurring with *cat* in the corpus and hence we have a matrix with the same amount of columns. If there are 10 words in the context window of each token (i.e. if each observation of *cat* is an utterance with 10 co-occurring words), this would mean that the remaining 990 columns contain a zero. This problem is called 'data sparsity' and makes it impossible to do any meaningful similarity calculation between tokens: most tokens will have zero similarity, even though they share the same meaning. Schütze's insight was that the problem of data sparsity can be solved by no longer looking directly at the first-order co-occurrences, but by replacing them with second-order co-occurrences, that is to say, the co-occurrences of the co-occurrences. We do not only model tokens through their co-occurring context words, we also in turn model the context words through their collocates on the type level. In essence, we enrich the sparse token matrix by projecting it in the denser type matrix.

To understand better how this works, let's look again at the *cat* sentences from our toy corpus. We now go through the steps in the construction of second-order token vectors.

STEP 1 We start from the sparse vector for Blofeld-sentence with three 1s for the co-occurring context words (*lap, purr* and *stroke*) (Table 6).

STEP 2 For each of these three context words (*lap, purr* and *stroke*), we look up the type vector from the large type-level PMI-weighted co-occurrence matrix that we constructed in the previous section (Table 7).

STEP 3 We add up the values for three vectors column by column and divide by three. We now have 1 vector that is an average of the type vectors for *stroke*, *purr* and *lap*. This vector can now said to represent the context in which *cat* occurred in the Blofeld-sentence. Because the token of *cat* is now modelled, not by its own co-occurring context words, but by the co-occurrences of the context words, we call this the *second-order co-occurrence vector*. We have now a token vector that is projected in the type-level word space (Table 8).

STEP 4 We repeat steps 1–3 for the two other tokens of *cat* in our toy corpus.

STEP 5 We now have matrix with a row for each token and the second-order co-occurrences as columns. Importantly, the token vectors in this matrix do not suffer from the sparsity problem any more because they have been construed in the much denser type-vector word space. Also the similarity between the context of token #1 and token #2 is now captured because the type vectors of the respective context words represent this similarity: the type vector of *purr* from sentence #1 and the type vector of *bark* from sentence #2 are similar because both strongly collocate with *pet* for example. This similarity has been propagated into the vectors of token #1 and token #2 (Table 9).

Table 6
The sparse token vector.

	bark	cadet	dog	lap	purr	sail	stroke	tree	wind
Blofeld was stroking the purring cat in his lap	0	0	0	1	1	0	1	0	0

Table 7
Selection of type-vectors for token #1's context words.

...	cup	blow	fur	pet	sea	...
TW1...						
...						
bark						
cadet						
dog						
lap	1.2	0.2	2.4	4.1	0.1	
purr	0.1	0.6	4.1	4.6	0.0	
sail						
stroke	0.4	1.2	3.2	4.5	0.2	
tree						
wind						
...						
$TW_{50.000}$						

Table 8
Averaging over type vectors of context words.

...	cup	blow	fur	pet	...
lap	1.2	0.2	2.4	4.1	
	+	+	+	+	
purr	0.1	0.6	4.1	4.6	
	+	+	+	+	
stroke	0.4	1.2	3.2	4.5	
SUM	1.7	2.0	9.7	13.2	
	÷3	÷3	÷3	÷3	
AVERAGE	0.6	0.7	3.2	4.4	

Table 9
Token vectors of second order co-occurrences.

	...	cup	blow	fur	pet	sea	...
Blofeld was stroking the purring CAT in his lap		0.6	0.7	3.2	4.4	0.1	
The black dog barked at the CAT in the tree		0.2	1.5	2.5	3.4	0.8	
The cadet was sailing his CAT against the wind		0.3	3.7	0.2	0.3	3.7	

STEP 6 Thanks to these denser representations, we can now calculate cosine similarities between the tokens. The final outcome is then a token-by-token similarity matrix that does show that token #1 and token #2 are very similar and probably express the same meaning, whereas token #3 is less similar and expresses another meaning (Table 10).

These six steps implement Schütze's (1998) original model of second order co-occurrences for token vectors. In our case study, we modified this approach slightly by integrating an additional weighting in step 3 that is inspired by the

Table 10
Token by token similarity matrix.

	Blofeld was stroking the purring CAT in his lap	The black dog barked at the CAT in the tree	The cadet was sailing his CAT against the wind
Blofeld was stroking the purring CAT in his lap	1	0.96	0.18
The black dog barked at the CAT in the tree		1	0.42
The cadet was sailing his CAT against the wind			1

Token by token similarity matrix.

Table 11

...	cup	blow	fur	pet	sea	...
lap w^{cat} = 2.2	1.2 ×**2.2**	0.2 ×**2.2**	2.4 ×**2.2**	4.1 ×**2.2**	0.1 ×**2.2**	
	+	+	+	+	+	
purr w^{cat} = 5.1	0.1 ×**5.1**	0.6 ×**5.1**	4.1 ×**5.1**	4.6 ×**5.1**	0.0 ×**5.1**	
	+	+	+	+	+	
stroke w^{cat} = 3.7	0.4 ×**3.7**	1.2 ×**3.7**	3.2 ×**3.7**	4.5 ×**3.7**	0.2 ×**3.7**	
SUM w^{TOT} = 11	4.6	7.9	38.0	49.1	1.0	
	÷11	÷11	÷11	÷11	÷11	
WEIGHTED AVERAGE	0.4	0.7	3.5	4.5	0.1	

inference process that a lexicologist (or any human interpreter for that matter) goes through when he or she disambiguates a polysemous item. Remember that in step 3 we averaged over the type-vectors of the context words of *cat*. In this averaging, each context word was treated on a par. The type-vector of *lap* had an equal share in the resulting second-order token vector as the type-vector of *purr*. However, it is clear that *purr* is a much better clue to the animal meaning of *cat* in sentence 1 than *lap*. After all, *lap* is polysemous itself and could also refer to rounds in a boat race. In other words, we would like to give *purr* a higher weight in the averaging over the context words' type vectors than *lap*. Conveniently, we can look up the collocational strength between *cat* and *purr* as well as between *cat* and *lap* in our large type-level matrix. This reflects how informative a context word is for modelling the meaning of *cat*. We now use this collocational strength measure (in this case Pointwise Mutual Information) as a weight in averaging over the type-vectors of the context words.

STEP 3 WEIGHTED We multiply the values in the type vectors of each context word by the collocational strength of that context word with *cat*. We add up the values in the type vectors column by column and then normalise by dividing by the total weight. Compared to the unweighted second-order token vector, the second-order co-occurrences indicative of the animal-meaning will have a slightly higher value, whereas the other ones have slightly lower values (Table 11).

The rest of the process remains the same: we repeat steps 1–3 for all tokens resulting in a matrix of token vectors with second-order co-occurrences, which we can then turn into a token-by-token similarity matrix. The final output of the distributional modelling is a square similarity matrix with as values the cosine similarity between each pair of tokens. In the next section, we discuss the typical application for token-level Word Space Models in Computational Linguistics and why and how we develop a different approach that is more tailored to the needs of lexicologists and lexicographers.

2.3. Turning a computational technique into a lexicological tool

Token-level models have been developed in Computational Linguistics for the task of automatic word sense disambiguation (WSD). The aim of this task is to assign the correct sense label from a list of predefined senses to every occurrence of a polysemous item in a test set[3] that is made publicly available as a benchmark (see Agirre and Soroa, 2007 and Manandhar et al., 2010 for a description of the two most recent WSD & WSI SemEval tasks). Usually the test set has been annotated with sense labels from WordNet. For this type of evaluation, the outcome of the token-level distributional model, the token-by-token similarity matrix, is fed to an automatic clustering algorithm that classifies the occurrences into a number of groups (usually the number of predefined senses that are assumed to exist for a given polysemous item). A cluster quality measure then quantifies to what extent the tokens that had the same predefined sense label were also assigned to the same cluster. Typically, this sort of evaluation is also used during the development phase of a WSD system: all the different parameters of a token-level model we discussed above (e.g. the context window around a token, the number of first and second order context words, the collocational strength measure, the similarity measure etc.) are varied and optimised to maximise the cluster quality measure. This allows all WSD system developers to compare their models to the same benchmark and the assumption is that this, in time, will lead to the best possible WSD system with the

[3] These test sets are compiled by the computational linguistic community in the framework of regular WSD competitions, viz. the senseval en semeval series (http://aclweb.org/aclwiki/index.php?title=SemEval_Portal), at which the performance of different WSD systems is tested and compared.

best possible parameter settings. It is then assumed that such an optimal WSD system cannot only be applied to the benchmark data but to an unlimited amount of text data.

Productive as this joint research programme may have been in Computational Linguistics, we argue that there are two important reasons why this type of benchmark testing must be complemented with more qualitative assessment methods to fully exploit the possibilities of token-level Word Space Models for lexicology and lexicography. First, lexical semantic scholars cannot assume the existence of predefined senses, and, second, benchmark testing only evaluates the output of a Word Space Model and treats the actual identification process of semantic structure as a black box. Let us discuss these two points in further detail. First researchers of lexical semantics cannot assume they have a list of predefined senses at their disposal for obvious reasons: they are the specialists that have to identify senses and compile the sense inventory that computational linguists rely on. Moreover, it is not a priori clear that for any given polysemous items, there is a unique classification into distinct senses (Geeraerts, 1993; Kilgarriff, 1997; Hanks, 2000). Rather, the common view in lexicology is that word senses are not stable entities in abstracto but that actual semantic content of a word, when used, is highly context dependent. For lexicographers, who necessarily have to make a choice about which uses to lump together or split up in a dictionary, this context dependence means that the senses they distinguish are relative to the specific purpose or audience of a dictionary. In other words, each sense classification is only one of many possible perspectives on a lexeme's meaning. Interestingly, the assumption of predefined senses is not an inherent flaw of token-level Word Space Models. These models just calculate similarities between contextual usages and allow to group similar tokens bottom-up and independently of any a priori sense labels. Because of this property, token-level Semantic Vector Spaces are known in Computational Linguistics as automatic word sense *induction* techniques rather than automatic word sense *disambiguation* techniques. Nevertheless, the models are almost always benchmarked in a WSD task against predefined senses and the similarities are not assessed on their own merits nor on their ability to find interesting semantic patterns. The fact that they provide a means of bottom-up sense induction, is exactly why we think they provide a promising tool for in-depth lexical semantic analysis. Of course, a comparison of the semantic structure that these models capture with previous lexicological and lexicographical work is an interesting first step to assess whether the models ''make sense'' at all and can identify traditional sense distinctions. In our case study below, we will do such a first ''sanity-check'' with a set of manually disambiguated items. Furthermore, we will also argue in Section 4 that benchmark testing against traditional sense distinctions is crucial in the development phase to assess the effect of all the different possible parameter settings like context window size or collocational strength weighting. However, in the long term we foresee that different token-level models with different parameter settings, will offer different perspectives on a word's semantic structure which allows a lexical semantician to interactively discover new meanings and uses.

The second way in which benchmarking testing in Computational Linguistics does not fully meet the needs and aims of lexical semantic research, has to do with the non-transparency of the sense induction process. The benchmark testing paradigm only assesses the output of a token-level model, viz. the clustering of word occurrences into possible senses. How this structuring of tokens into similar groups comes about and how different parameter settings influence the structure, usually remains inside the black box of the algorithm. Typically, a new parameter setting is based on an intuitive idea (a simple example would be: *function words do not contribute much to identifying a word's meaning, so let's remove them from the context word list*). Whether this intuition proves correct is then evaluated in terms of an increase or decrease in the cluster quality measure for the WSD test set. An in-depth analysis of how a specific parameter setting was successful in some cases and not in others, is extremely rare in the computational linguistic literature. Yet, assessing different types of evidence for semantic structure in a large set of attestations of a lexical item, is exactly what lexical semantic research is all about. In other words, for token-level models to become a tool supporting lexical semantic analysis, a researcher must be able to look inside the black box and investigate whether e.g. syntactic patterns that lexical items occur in provide additional or different information about a word's meaning, compared to co-occurring words in a fixed context window. Lexical scholars must be able to interact with the models and analyse the influence of different parameter settings. Incidentally, also computational linguists themselves are aware of this need for more in-depth evaluation of Word Space Models. 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''. This seems to suggest a potential fruitful division of labour: lexical scholars, with their richer and theoretically more informed descriptive apparatus, can provide in-depth analyses of the semantic structure captured by Word Space Models, provided that computational linguists make intuitive interfaces available to investigate the computational models and thus offer lexical scholars a way to efficiently deal with ever larger data collections.

How do we propose then to turn token-level Semantic Vector Spaces into a tool that supports analysis of large datasets by a lexicographer or lexicologist? Remember that the output of our distributional model was a token-by-token similarity matrix, typically containing a few hundred tokens. It is quite evident that we cannot just hand over such a matrix to a lexicologist, in the sense that the matrix as such will not help the researcher tremendously with the interpretation of the data. We therefore turn to the paradigm of visual analytics which aims to capitalise on the human mind's ability to quickly

spot visual patterns and converts large quantitative data sets into a visual representation. Analysing this visual representation then hopefully allows to identify meaningful patterns in the data much easier and faster than pure number crunching. In the present study the quantitative dataset at hand is our token-by-token similarity matrix. The visualisation technique we will use, is Multidimensional Scaling (MDS, Cox and Cox, 1991), a dimension reduction technique that is widely used in cognitive psychology and that is explicitly developed to visualise and interpret similarity matrices. In this case, we use the technique to reduce the high-dimensional token by token similarity matrix to a two dimensional plot that tries to render the similarities as faithfully as possible: two tokens that are highly contextually similar will be close together in the plot, whereas dissimilar tokens will be far apart. This way, a lexical analyst can start exploring visually which semantic structure is captured by the distributional model. We will illustrate this technique in the case study in the next section: by additionally colour-coding the MDS plot and making it interactive, the user can not only explore the similarities between the tokens, but also look further into the context features that caused the distributional model to detect this specific semantic structure.

Let us finish this section by pointing out that we are not the first ones to apply Word Space Models to typical lexicological and lexicographic topics of interest: A number of studies have tried to identify lexical semantic change (Sagi et al., 2009; Cook and Stevenson, 2010; Gulordava and Baroni, 2011). Peirsman et al. (2010) study lexical variation between different varieties of the same language. However, all these studies stay within the benchmark testing paradigm and try to replicate previously identified semantic changes or lexical variants. The latter are a special type of test set. More genuinely linguistic is the work by Tamm and Sahlgren (2014), who show that Word Space Models can be fruitful applied to the analysis of semantic variation between register. However, they study polysemy on an aggregated level using a type-level word space. Also related to our approach is the work by Rohrdantz et al. (2011) and Rohrdantz et al. (2012), who use a Word Space Model to do a bottom-up, token-level analysis of lexical semantics in combination with visual analytics to bring the semantic patterns to light that the model captured. However, this work takes mainly targets a computational audience and wants to demonstrate the power of visual analytics. By contrast, starting with Heylen et al. (2012) and continued in Wielfaert et al. (2013) and the current paper, we have initiated a research programme that takes explicitly a lexicological perspective rather than a computational linguistic one. We want to turn Word Space Models into an analysis tool for lexicology (and lexicography), rather than turning lexicographic data into a test data set for Computational Linguistics. In this respect we link up with the tradition of turning quantitative corpus linguistic measures, like collocation and colligation, into usable tools for scholars of word meaning (e.g. Scott, 1996; Kilgarriff et al., 2004; Anthony et al., 2011).

3. Semasiological case study in Dutch

To illustrate the current possibilities of Word Space Models as introduced in the previous section, we present a case study of the Dutch noun *monitor*, which was one of 476 nouns included in Heylen et al. (2012) and which represent a clear case of polysemy. We collected a random sample of *monitor* tokens from our corpus and constructed a bottom-up, usage-based analysis of the senses that occur in the sample.

The corpus for our case study consists of Dutch newspaper materials from 1999 to 2005. For Netherlandic Dutch, we used the 500 million words *Twente Nieuws Corpus* (Ordelman, 2002), and for Belgian Dutch, the *Leuven Nieuws Corpus* (a.k.a. Mediargus corpus, 1.3 billion words). We randomly selected 1000 Belgian and 1000 Dutch newspaper issues from the corpus and extracted all occurrences of *monitor* as a noun, which resulted in 199 tokens. Both the type and token-level Word Space Models require that choices are made for the parameter settings discussed above (collocational strength measure, similarity measure, co-occurrence window around the tokens). In the models shown here, we opted for the settings that were shown to be optimal in our previous studies (Heylen et al., 2008; Peirsman et al., 2008, 2010); for a more technical account of the parameters that were used for the case studies, see Wielfaert et al. (2013).

The two main readings of *monitor* in Dutch, based on the van Dale dictionary (Den Boon and Geeraerts, 2008), are comparable to those in English and both developed from Latin *monitor* ''one who reminds, warns, admonishes, or checks''. Closest to this original meaning, is the use in Dutch of *monitor* as a person supervising and guiding a group of children in the context of organised playground activities, summer camps, sports activities and the like. In this reading, *monitor* is often found in compounds like *jeugdmonitor* ('youth monitor') or *speelpleinmonitor* ('playground monitor').[4] A second, metaphorically derived meaning, which is also present in English, is that of (computer) screen, as a 'visual warner' of the processes going on in a device. As we already, discussed in Section 2.3, token-level Word Space Models are still in a development phase, and as with any new instrument for analysis, it is probably a good first step to see whether they can

[4] Van Dale, the main dictionary of contemporary Dutch, mentions two other senses that are not present as such in our *monitor* data, namely *studiebegeleider* ('study supervisor') and *iemand die toezicht houdt; waarnemer* (someone who supervises; observer). http://vandale.nl/opzoeken?pattern=monitor&lang=nnhttp://vandale.nl/opzoeken?pattern=monitor&lang=nn.

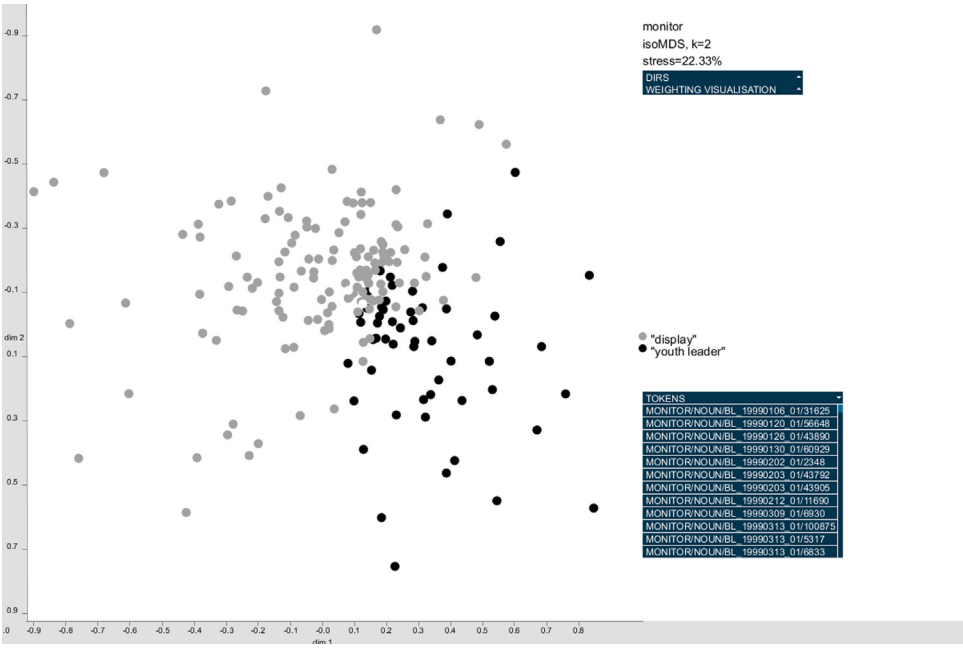

Fig. 1. Token cloud of the weighted model for *monitor*.

prove their metal by reproducing known knowledge before trusting them to deliver new insights. We therefore manually disambiguated the 199 tokens in our sample annotated them with the senses from the van Dale dictionary,[5] so that we can compare the token-level spaces with an existing sense classification.

We applied a token-level Word Space Model of the type described in Section 2.2 to the randomly selected sample of 199 tokens. The 199-dimensional similarity matrix was then reduced to just two dimensions and visualised in a scatter plot with Multidimensional Scaling. Theoretically speaking, the polysemy of *monitor* should be visible as two clearly separated clusters or 'token clouds', as we prefer to call them. The rationale behind this is that in the case of monosemy (i.e. the situation in which all the observations of a word represent the same meaning), the tokens should be relatively close to each other as their collocates are strongly related. However, in the case of polysemy (or homonymy, for that matter) distances should be larger and at least in theory large enough for separate clouds to appear.

The results that we obtain for *monitor*, as represented in Fig. 1 do not at first sight reveal clear clusters. However, a semantic structure is revealed if we colour-code the different senses from the manual disambiguation in the plot. The grey dots represent the screen sense and the black dots refer to the youth leader sense. Now it is more clear that the clustering technique does appear to be rather successful: each of the two senses is represented on the opposite side of the first dimension (the *x*-axis).

While this clearly establishes that a Semantic Vector Space approach can indeed be useful as a preliminary semantic classification tool for descriptive lexicology and lexicography, additional features and procedures may be implemented to enrich the use of the method.

As a first extension, we provided the option to easily switch between visualisations of models that use different parameter settings. Remember that in Section 2.2, we explained that we extended the original approach from Schütze (1998) by giving more weight to informative context words in the construction of a token vector. By comparing Fig. 1 (our weighted approach) to Fig. 2 (the original unweighted approach) we can now see that this adjustment succeeds better in telling two senses apart. In a similar way, the effect of other parameter settings can be visually explored.

In a plot of model with weighted context words, it also possible to give the data points a size that reflects the informational value of the collocates in the corresponding utterance. In Fig. 3 the size of the dots is relative to the maximum weight that was assigned to one of the context words of the respective token. In other words: if the data point is represented as a small dot in the plot, the context words of that specific *monitor* token have low weights, and there was not

[5] The manual disambiguation revealed meaning distinctions that were more fine grained than those covered by the dictionary (cf. discussion of token #4).

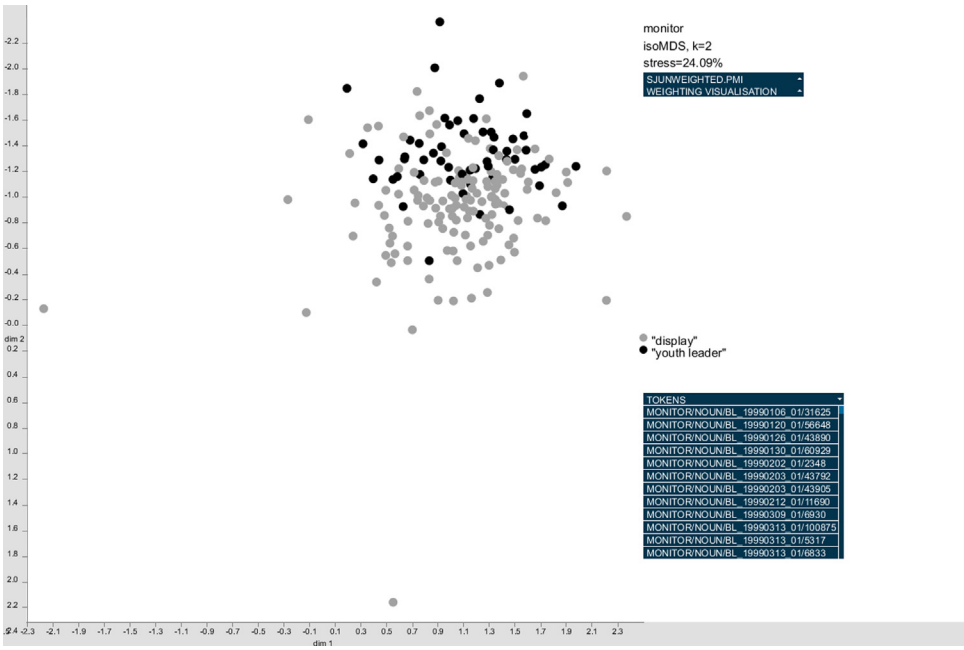

Fig. 2. Token cloud of the unweighted model for *monitor*.

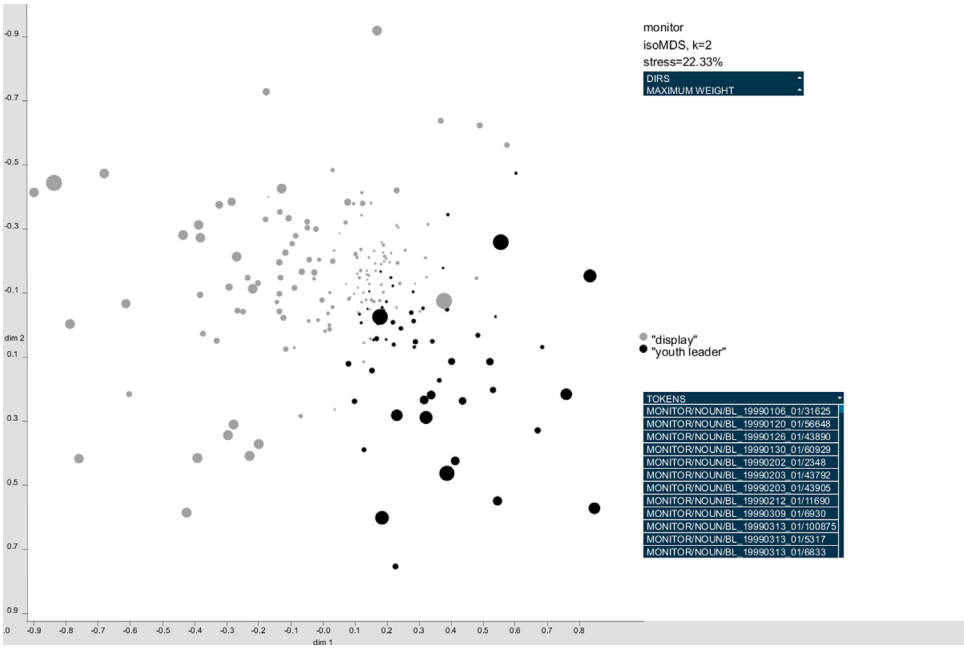

Fig. 3. Token cloud of *monitor* with weights as dot size.

much of a clue in the context to classify the token. Conversely, the bigger data points contain collocates that appear to have a high discriminative value. From a descriptive point of view, these tokens will be the primary candidates for manual scrutiny.

We also linked the dots in the representation to the actual examples. In this way, the initial attestations on which the representation is built can be readily accessed: a quick manual scanning of the data helps the researcher to interpret the

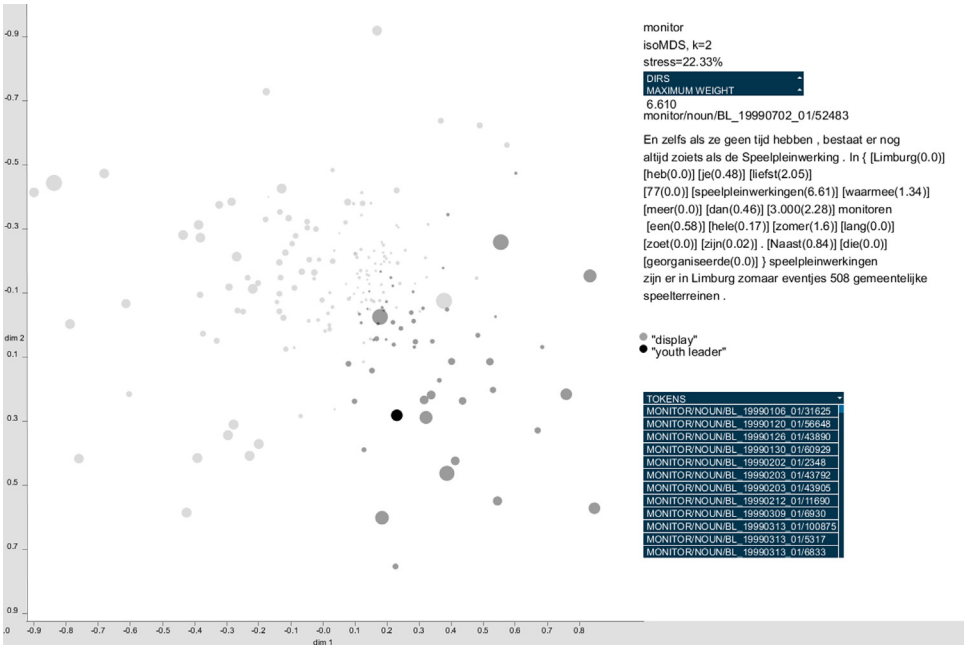

Fig. 4. Token cloud of *monitor* with highlighted concordance.

emerging structure. In Fig. 4, we see that one dot is highlighted by clicking and on the right-hand side, the concordance appears.

The weights of the context words are integrated in the concordance by putting them in brackets behind each context word. In the plot above the concordance of the highlighted token reads:

> In [Limburg(0.0)] [heb(0.0)] [je(0.48)] [liefst(2.05)] [77(0.0)] [speelpleinwerkingen(6.61)] [waarmee(1.34)] [meer(0.0)] [dan(0.45)] [3000(2.28)] monitoren [een(0.58)] [hele(0.17)] [zomer(1.6)] [lang(0.0)] [zoet(0.0)] [zijn(0.02)]. [Naast(0.84)] [die(0.0)] [georganiseerde(0.0)]. . ..
>
> In Limburg there are no less than 77 playground initiatives that keep more than 3000 youth leaders busy for a whole summer. Next to these organised. . .

This is an example of a well classified *monitor* token with the 'youth leader' meaning as it is surrounded by other 'youth leader' tokens. From the concordance, we can see that the weighting of context words also works as intended: the highly informative context word *speelpleinwerkingen* (playground initiatives) get the highest weight (6.61) and makes sure the token is appropriately positioned.

To illustrate the advantages of the interactive plot, we take a closer look at a few tokens that different context word weights and positions in the token cloud. The selected tokens have been marked with an arrow and a reference number in Fig. 5.

The first token, belonging to the screen sense, is situated at the top of the plot and could be considered as an outlier as it is not surrounded by any tokens of either sense. The full context of this token is the following, directly followed by its translation:

> **TOKEN #1**
>
> *Galerie Akinci is door de Zwitserse Emmanuelle Antille ingericht als woonkamer, waar je in een luie stoel [kunt (0.56)] [kijken(1.09)] [naar(0.57)] [haar(0.0)] [video-installatie(5.43)] Reflecting [Home(0.0)]. [Behalve(0.0)] [op (0.79)]* ***monitoren*** *[zijn(0.02)] [de(0.34)] videobeelden [ook(0.0)] [levensgroot(0.0)] op de [wand(0.0)] [geprojecteerd(0.0)]. De kunstenares, die zelf de hoofdrol speelt, lijkt hierdoor haast lijfelijk aanwezig.*
>
> Galery Akinci has been furnished like a living room by the Swiss Emmanuelle Antille, in which you can look from an easy chair to her video installation Reflecting Home. Apart from the displays, the video images are also projected life-size on the wall. The artist, who plays the lead role herself, seems to be physically present.

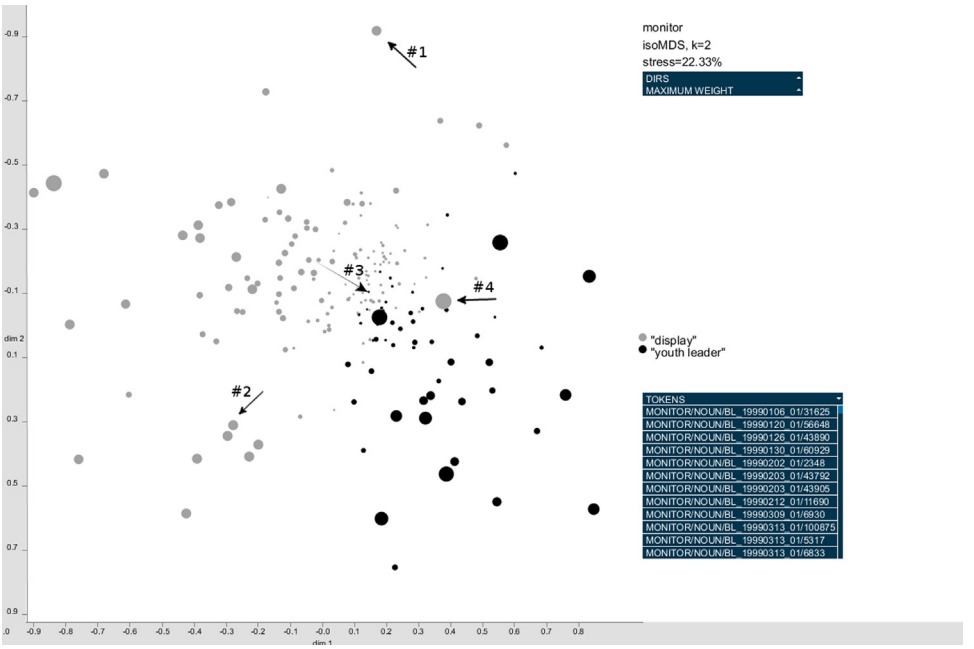

Fig. 5. Interactive perusal of the token cloud.

The highest weighted collocates for this token are: *video-installatie* (5.43) 'video installation' and *kijken* (1.09) 'to look'. There are a number of things we can learn about how our distributional model works. First, despite the high weight of genuinely informative collocates for the meaning of this token, it has been plotted at a large distance of the tokens that are semantically related, namely the 'display' tokens. One hypothesis is that the outlier position can be attributed to the word *video-installatie* (video installation) because it is a low frequent word in our corpus and does not contribute enough relevant second-order collocates to the token vector in order to detect the similarity with the other 'display' tokens. Also, the high collocational association between *video-installatie* and *monitor* might be an artefact of the Pointwise Mutual Information measure which has been shown to have a bias towards low frequent events. Secondly, some informative context words like *videobeelden* (video images) and *geprojecteerd* (projected) get a zero weight. Clearly the weighting function is not optimal yet. In any case, the interactive inspection of the plot has given us valuable information about how to improve the distributional model, which would be hard to deduce from a single evaluation measure in a benchmarking testing procedure.

The second selected token also represents the 'display' sense, and it is situated with a handful of other 'display' tokens at the left side of the plot. The full concordance is as follows:

TOKEN #2

''*Daar stelden de daders vast dat er weinig te rapen viel,* ''*zegt zaakvoerder Louis.*'' *Ze pakten [dan(0.07)] [ook(0.0)] [alleen(0.37)] [wat(0.01)] [kabels(0.0)], [een(0.58)] [paar(0.71)] [klavieren(5.29)] [en(0.59)] monitoren [mee(0.49)]. [Het(0.0)] [gros(3.03)] [van(0.16)] [de(0.34)] [monitoren(5.58)] [lieten(0.0)] [ze(0.25)] [staan(0.9)].*

''There, the culprits realised there was little to steal,'' says business owner Louis. ''They took just some cables, a couple of keyboards and monitors with them. The majority of the monitors they left alone.

This token has a particularly interesting collocate, namely *monitor* itself, which has also been assigned the highest weight (5.58) among the target's collocates. As it turns out, we find pairs of self-co-occurring *monitor* tokens in this part of the plot. This might suggest choosing a different strategy for dealing with repetitions in the corpus. For a lexicologist, the main cue for the token's meaning would be *klavier* 'keyboard', which is correctly assigned a fairly high weight (5.29).

The third case is a so-called 'misclassifications', namely a 'youth leader' token which is in the middle of the plot, an area populated with 'display' tokens. Nevertheless, it is not hard to identify the main reason for this:

TOKEN #3

Als 14-jarige kwam [ik(0.0)] [voor(0.0)] [het(0.0)] [eerst(0.0)] [naar(0.57)] [Zwitserland(0.0)]. [Later(0.0)] [werd(0.23)] ik monitor [en(0.59)] [nu(0.0)] [coördineer(0.0)] ik [hier(0.46)] [de(0.34)] [hele(0.17)] [logistiek(0.0)]. ''*Het gigantische Palace Hotel, waarin bij ons bezoek 800 jongeren verblijven, werd meer dan honderd jaar geleden gebouwd door een Limburgse baron, maar ging 1 jaar na de opening al op de fles.*

As a 14-year-old I came for the first time to Switzerland. Later I became youth supervisor and now I coordinate the whole logistics. The gigantic Palace Hotel, where during our visit 800 youngsters reside, has built more than hundred years ago by a Limburgian baron, but went bankrupt just one year after the opening.

None of the informative collocates has been weighted. The best cues in this case would be *Zwitserland* (Switzerland), which is a country where traditionally many summer and winter camps for Belgian children are organised, and *coördineer* (coordinate). Again, this suggests changes to the weighting scheme.

The final example is a misclassification similar to #4, but for another reason:

TOKEN #4

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 geïntroduceerd door dokter Alfred Steinschneider.

For 30 years scientists have been looking for the solution. Risk babies are attached to a monitor to warn their parents in case of a breathing pause during the sleep of their child. That monitoring system has been introduced by doctor Alfred Steinschneider.

The word *risicobaby* ('risk baby') gets a particularly high weight while the other collocates within the context window have a fairly low weight. As a result, *risicobaby* is the collocate responsible for its odd position. *risicobaby* occurs only 12 times throughout our corpus and apparently its type vector is more associated to the 'youth leader' than to the 'display' sense. Interestingly, the *monitor* in this token does not refer to a prototypical screen or display, but rather to a medical device which monitors the physiological parameters of a sleeping infant. These parameters are usually visualised on a display part of the device, which makes that *monitor* is in this case a metonymical extension (more specifically: a *pars pro toto*) of the prototypical sense covered by our reference dictionary.

To conclude this section, let us point out that this type of analysis can be repeated for any lexeme of interest. Apart from the manual disambiguation, which is used in the current development stage, all other information visible in the plots is generated automatically and bottom-up from the corpus. To illustrate this, Fig. 6 shows a token cloud for the polysemous noun *piraterij*, which, like its English cognate *piracy*, is used both to refer to ''the act of attacking and stealing from a ship at sea'' and to ''the act of illegally copying someone's product or invention without permission''. Like with *monitor*, Fig. 6 shows that the token-level Word Space Model is fairly good at telling the two meanings apart when there is an informative (highly weighted) context word present in the token's context, but succeeds less well for tokens with uninformative contexts.

4. Discussion and future work

In the case study we looked at an example of polysemy in Dutch and used Word Space Models in the attempt to automatically apply collocational analysis on a large-scale corpus. The Word Space Models' output is made visually accessible by plotting it in two dimensions and adding colour-coding for the two manually disambiguated senses of *monitor* that occur in our sample. We argue that the technique is promising because it provides linguistically interpretable structures. Nevertheless, this application of visualised distributional modelling is work in progress and the technique will have to be further refined and enriched. In what follows, we discuss some of the possible avenues for future improvements that we consider and have experimented with.

First, the occurrences of *monitor* in the sense of 'display' we modelled are not a monolithic block, but rather show an internal structure that is not yet revealed in the analysis. In other words, we should try to increase the granularity of the analysis. Not only is *monitor* used to refer to screens which are attached to a computer or a video source, but also for more metonymical uses. This ranges from a military context where green screens are used to monitor a situation, to devices which keep an eye on the physiological parameters of a baby to prevent it from sudden death infant syndrome (see misclassified token #4 above). We also observed occurrences in which *monitor* still refers to the more prototypical display

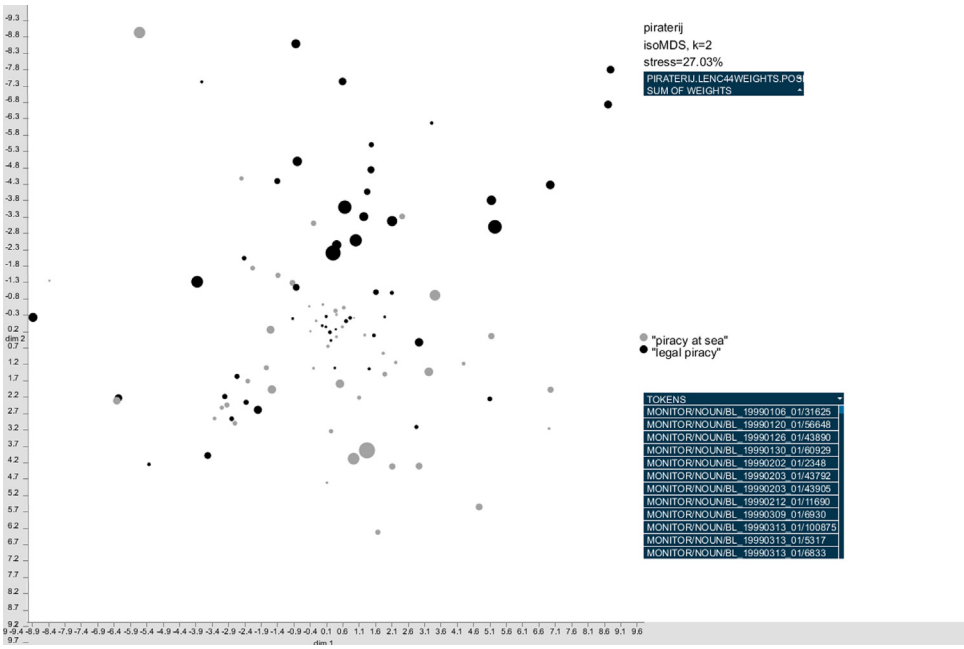

Fig. 6. Token cloud of *piraterij*.

which shows video images, but in the context of theatre, film and television, more specifically as a monitor on which the actors can see live what is happening on the theatre stage or directors who look at what is being filmed on a small screen. Nevertheless, these instances are all part of the broad 'display' sense. A sample of 200 tokens is simply not large enough to have enough tokens that represent such specific uses. Additionally, plotting more than 200 tokens simultaneously would also clutter the plot. Therefore, we should explore other heuristics to select a stratified sample which represents a specific use of the target word instead of naively drawing a random sample from the corpus. A possibility would be to apply a coarse grained clustering to all tokens first and then to the MDS visualisation per cluster in order to get a more fine-grained picture of set of related usages. This way the user can gradually zoom in on the different uses of the lexeme.

Second, up until now we have looked at the Dutch word *monitor* from a semasiological point of view. In other words: we looked at the occurrences of *monitor* and discriminated the two main senses it has in our data. However, we could also follow the opposite direction (from concept to instantiation) and use an onomasiological profile to find the words that refer to a single concept, in casu BEELDSCHERM ('display'). Ruette et al. (2014) automatically detected a number of near-synonyms which all refer to displays in Dutch: *beeldscherm* (display), *computerscherm* (computer screen) and *monitor* (monitor) (Fig. 7).

In Heylen et al. (2012), we plotted a random sample of tokens belonging to these three lexical items in the same two-dimensional MDS representation of a token-level Semantic Vector Space. The *monitor* tokens have the lightest shade whereas instances of the near-synonyms *computerscherm* and *beeldscherm* are colour-coded in darker shades. In the right hand side of the plot we notice that tokens of the three near-synonyms are interspersed. This pattern corresponds to the foreseen structure of near-synonymy: words with the same meaning will occupy the same area of semantic space. However, on the left-hand side of the plot, we see there is an area with only (light-shaded) *monitor* tokens. The interactive exploration of these tokens and their linked concordances learns that these are indeed the *monitor* tokens with the 'youth leader' meaning. This illustrates how looking at a polysemous item from an onomasiological perspective can quickly reveal that the item has usages that do not fall under the concept of the onomasiological profile. This is an extra heuristic to identify polysemy (Fig. 8).

Third, next to onomasiological variation, Word Space Models also allow to investigate different sorts of lectal variation. This is extremely useful from a descriptive lexicographical or lexicological perspective, because it provides a way of assigning variety-specific labels to the meanings of a word. Our corpus consisted of a Belgian and Netherlandic newspaper material, and therefore we can look into regional variation. The figure above shows a plot of the same tokens as the previous one, but now the tokens are colour-coded by country with the darker shade for Belgium and the lighter one for the Netherlands. We can now see that the part of the plot that was previously exclusively populated by *monitor* tokens with the 'youth leader' meaning, is also exclusively Belgian. This correctly shows that the 'youth leader' meaning is indeed typically Belgian. Not shown here is that we can also track tokens back to the newspaper they were extracted from and

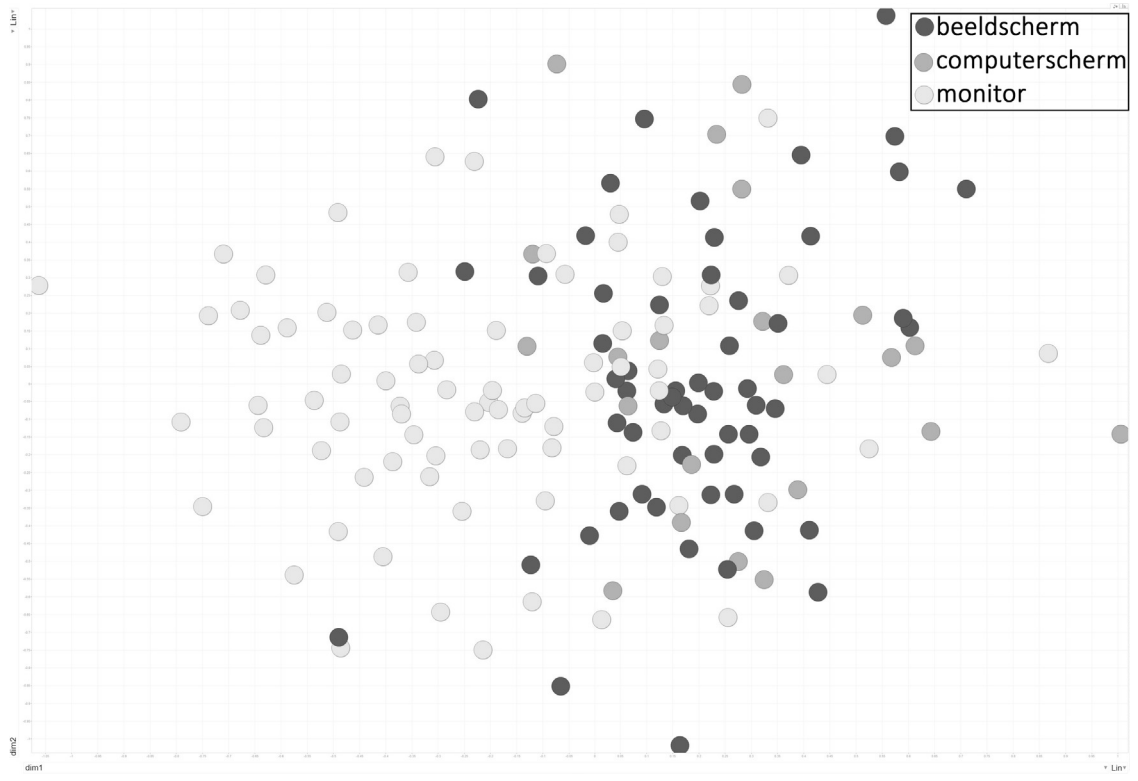

Fig. 7. Token cloud of Dutch near-synonyms for DISPLAY.

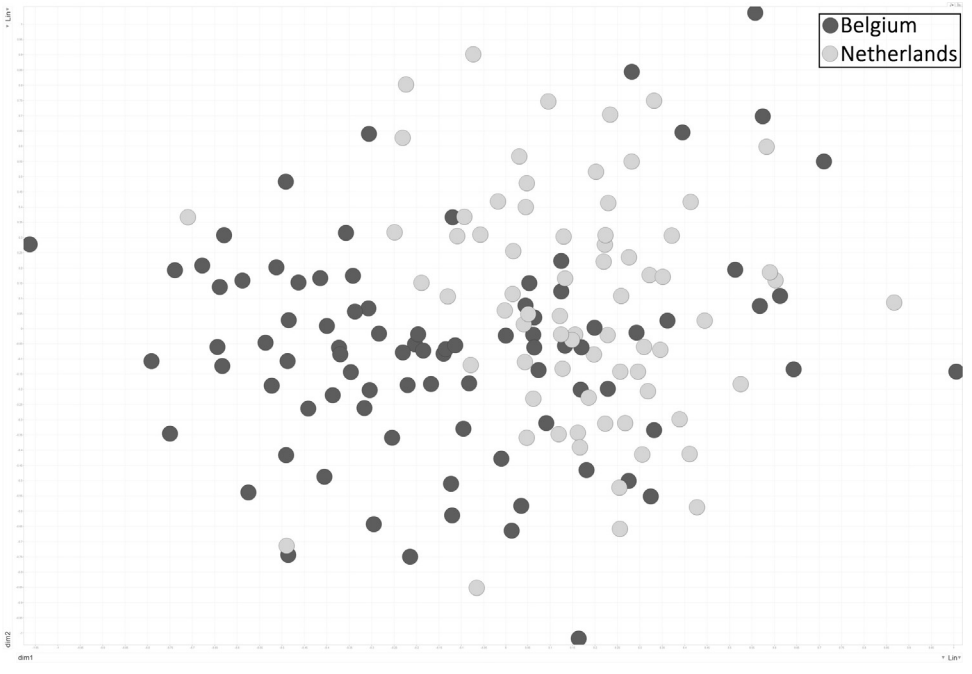

Fig. 8. Token cloud of Dutch near-synonyms for DISPLAY colour-coded by counrty.

that colour coding by newspaper shows that the 'youth leader' monitor tokens originate almost exclusively from the Belgian popular (tabloid-like) newspapers. These newspapers are in general more region-specific than the quality newspapers, which results in reports on local youth matters such as summer camps or youth movements. In a further stage, it could be interesting to look at the language register in the regional sections in contrast with the national articles.

Fourth and finally, we come back to the many different parameters that can be varied in Word Space Models for Lexical Semantics. On the one hand this is a strength of the model: the different parameter settings give different perspectives on the multifaceted phenomenon of Lexical Semantics. On the other hand, the parameter-richness also poses methodological challenges: we cannot expect a researcher to look for each lexical item at hundreds of different plots, all generated with slightly different parameter settings. Before token-level word spaces can become a bottom-up analysis tool for Lexical Semantics, we first need a better grasp of which type of semantic structure is captured by which combination of parameter settings. This can only be done by comparing models to manual analyses of lexical semantic structure by trained linguists. This means we need a lexicologically informed version of the benchmark testing paradigm prevalent in Computational Linguistics. On the one hand, this implies manually annotated datasets that cover a larger range of semantic relations than traditional WordNet senses (e.g. both denotational and connotational meaning distinctions, or metaphorical and metonymical extensions). In this respect, we are currently exploring token level models of the example instances linked to the fine-grained semantic distinctions in the Algemeen Nederlands Woordenboek,[6] an academic online dictionary of present day Dutch. On the other hand, lexicologically informed benchmarking also needs more fine-grained evaluation measures than the traditional precision, recall and F-measure to assess to what extent a Word Space Model captures the semantic structure annotated in a test dataset. Here, we are experimenting with measures like the McClain-Rao measure (McClain and Rao, 1975) that do not require the use of a hard clustering step to assess how well token-by-token similarity matrices capture a manually defined semantic structuring. Parameter sweep studies, like the one proposed in Lapesa and Evert (2013), could then help determine which specific parameter settings will highlight specific semantic relations between tokens. Eventually, lexicologists and lexicographers using the tool, can then be guided in the choice of parameters settings to investigate a specific type of semantic pattern like metonymy, metaphor or stylistic variation.

5. Conclusion

In this paper, we gave a non-technical introduction to Word Space Models, and more specifically, to token-level models based on second order co-occurrences. We argued that these models are a logical extension to the set of corpus-linguistic tools available to lexicologists and lexicographers, because they allow for a systematic combination of two existing traditions in quantitative corpus-based lexicology: on the one hand, statistical methods for finding contextual clues to word meaning (collocation analysis), and on the other hand statistical methods for grouping attestations into senses ('behavioural profile' analysis). We suggested that such an extension of the lexicological tool set is necessary for at least two reasons: to manage the increasing amounts of data to be addressed by traditional descriptive lexicology, and to implement the new types of trend analysis that are made possible by 'Big Data'. However, we also pointed out that a number of adaptations are necessary before token-level Word Space Models (a technique that was first formulated in Computational Linguistics) can be fruitfully applied to in-depth analyses of Lexical Semantics. As part of an ongoing research programme that explores bottom-up statistical approaches to lexical semantics and that is geared precisely to customising the computational linguistic methods to the needs of corpus-based lexicology and lexicography, we presented a case study of the Dutch polysemous word 'monitor'. With the case study, we demonstrated how token-level Word Space Models can be usefully combined with visual analytics techniques; we showed how this combination of techniques yields a first overview of the different usages that are present in a set of attestations, and how it can lead to a better insight into the underlying mechanisms that allow distributional models to capture semantic structure. Although token-level Word Space Models are currently far from providing a ready-made technique for lexical analysis, this case study tries to deliver a proof of concept that such models constitute a promising path for quantitative corpus-based semantics to pursue.

Acknowledgements

This research was funded by KU Leuven BOF grant 3H110243 for the OT project *From lexical to semantic sociolectometry: New methods for the corpus-based analysis of variation in lexical categorization*.

[6] http://anw.inl.nl/.

References

Agirre, E., Edmonds, P., 2006. Word Sense Disambiguation: Algorithms and applications Text, Speech and Language Technology Series, vol. 33. Springer, Dordrecht.

Agirre, E., Soroa, A., 2007. SemEval-2007 Task 02: evaluating word sense induction and discrimination systems. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, pp. 7–12.

Anthony, L., Chujo, K., Oghigian, K., 2011. A novel, web-based, parallel concordancer for use in the ESL/EFL classroom. In: Newman, J., Baayen, H., Rice, S. (Eds.), Corpus-based Studies in Language Use, Language Learning, and Language Documentation. Rodopi, New York, pp. 123–138.

Baroni, M., Lenci, A., 2011. How we blessed distributional semantic evaluation. In: Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–10.

Bullinaria, J.A., Levy, J.P., 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. Behav. Res. Methods 39, 510–526.

Church, K.W., Hanks, P., 1989. Word association norms, mutual information and lexicography. In: Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, pp. 76–83.

Cook, P., Stevenson, S., 2010. Automatically identifying changes in the semantic orientation of words. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp. 28–34.

Cox, T.F., Cox, M.A.A., 1991. Multidimensional scaling on a sphere. Commun. Stat. Theor. Methods 20, 2943–2953.

Den Boon, T., Geeraerts, D., 2008. Van Dale Groot Woordenboek van de Nederlandse Taal, 14th ed. Van Dale Lexicografie, Utrecht/Antwerpen.

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

Evert, S., 2004. The statistical analysis of morphosyntactic distributions. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp. 1539–1542.

Firth, J.R., 1957. Papers in Linguistics, 1934–1951. Oxford University Press, London.

Geeraerts, D., 1993. Vagueness's puzzles, polysemy's vagaries. Cognit. Linguist. 4, 223–272.

Geeraerts, D., 2010. Theories of Lexical Semantics. Oxford University Press, London.

Glynn, D., 2010. Testing the hypothesis. Objectivity and verification in usage-based cognitive semantics. In: Glynn, D., Fischer, K. (Eds.), Quantitative Methods in Cognitive Semantics. Corpus-driven Approaches. Mouton de Gruyter, Berlin & New York, pp. 239–270.

Gries, S.Th., 2006. Corpus-based methods and cognitive semantics: the many meanings of to run. In: Gries, S., Stefanowitsch, A.T. (Eds.), Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis. Mouton de Gruyter, Berlin & New York, pp. 57–99.

Gulordava, K., 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, pp. 67–71.

Hanks, P., 2000. Do word meanings exist? Comput. Hum. 34, 205–215.

Harris, Z., 1954. Distributional structure. Word 10 (23), 146–162.

Heylen, K., Peirsman, Y., Geeraerts, D., Speelman, D., 2008. Modelling word similarity. An evaluation of automatic synonymy extraction algorithms.In: Proceedings of the Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco, pp. 3243–3249.

Heylen, K., Speelman, D., 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: Visualization of Language Patters and Uncovering Language History from Multilingual Resources, Avignon, France, pp. 16–24.

Kilgarriff, A., 1997. I don't believe in word senses. Comput. Hum. 31 (2), 91–113.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D., 2004. The sketch engine. In: Williams, Vessier, (Eds.), Proceedings of the Eleventh Euralex Congress. UBS Lorient, France, pp. 105–116.

Landauer, T.K., Dumais, S.T., 1997. A solution to plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev. 104 (2), 211–240.

Lapesa, G., Evert, S., 2013. Evaluating neighbor rank and distance measures as predictors of semantic priming. In: Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013), Sofia, Bulgaria.

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods Instrum. Comput. 28, 203–208.

Manandhar, S., Klapaftis, I., Dligach, D., Pradhan, S., 2010. SemEval-2010 Task 14: word sense induction & disambiguation. In: Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, pp. 63–68.

McClain, J.O., Rao, V.R., 1975. CLUSTISZ: a program to test for the quality of clustering of a set of objects. J. Market. Res. 12, 456–460.

Navigli, R., 2009. Word sense disambiguation: a survey. ACM Comput. Surv. 41(2). ACM Press, pp. 1–69.

Navigli, R., 2012. A quick tour of word sense disambiguation, induction and related approaches. In: Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM), Spindleruv Mlyn, Czech Republic, pp. 115–129.

Ordelman, R., 2002. Spoken Document Retrieval for Historical Video Archives – Dutch Speech Recognition in the Echo Project. Technical Report. University of Twente, Parlevink Group.

Pantel, P., Lin, D., 2002. Discovering word senses from text. In: Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02), Edmonton, Canada, pp. 613–619.

Peirsman, Y., Heylen, K., 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, Hamburg, pp. 34–41.

Peirsman, Y., Geeraerts, D., Speelman, D., 2010. The automatic identification of lexical variation between language varieties. Nat. Lang. Eng. 16 (4), 469–491.

Rohrdantz, C., Hautli, A., Mayer, T., Butt, M., Keim, D.A., Plank, F., 2011. Towards tracking semantic change by visual analytics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA, pp. 305–310.

Rohrdantz, C., Niekler, A., Hautli, A., Butt, M., Keim, D.A., 2012. Lexical semantics and distribution of suffixes: a visual analysis. In: Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, Avignon, France, pp. 7–15.

Ruette, T., Geeraerts, D., Peirsman, Y., Speelman, D., 2014. Aggregating dialectology and typology: linguistic variation in text and speech, within and across languages. In: Szmrecsanyi, B., Wälchli, B. (Eds.), Linguistic Variation in Text and Speech, Within and Across Languages. Mouton de Gruyter, Berlin & New York, p. 205.

Sagi, E., Kaufmann, S., Clark, B., 2009. Semantic density analysis: comparing word meaning across time and phonetic space. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics, Athens, Greece, pp. 104–111.

Sahlgren, M., 2006. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Spaces (Ph.D. dissertation). Department of Linguistics, Stockholm University.

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

Scott, M., 1996. WordSmith Tools. Oxford University Press, Oxford.

Sinclair, J.M., 1991. Corpus, Concordance, Collocation. Oxford University Press, Oxford.

Tamm, M.K., Sahlgren, M., 2014. Temperature in the word space: sense exploration of temperature expressions using word-space modelling. In: Szmrecsanyi, B., Wälchli, B. (Eds.), Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech. De Gruyter, Berlin, pp. 231–267.

Thanopoulos, A., Fakotakis, N., Kokkinakis, G., 2002. Comparative evaluation of collocation extraction metrics. In: Proceedings of the 3rd Language Resources Evaluation Conference, Las Palmas, pp. 620–625.

Tomuro, N., Lytinen, S.L., Kanzaki, K., Isahara, H., 2007. Clustering using feature domain similarity to discover word senses for adjectives. In: Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing, pp. 370–377.

Turney, P.D., Pantel, P., 2010. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. 37, 141–188.

Weaver, W., 1955. Translation. In: Locke, W.N., Booth, A.D. (Eds.), Machine Translation of Languages. MIT Press, Cambridge, MA, pp. 15–23.

Wiechmann, D., 2008. On the computation of collostruction strength. Corpus Linguist. Linguist. Theory 4 (2), 253–290.

Wielfaert, T., Heylen, K., Speelman, D., 2013. Visualisations interactives des espaces vectoriels sémantiques pour l'analyse lexicologique. In: Actes de SemDis 2013: Enjeux actuels de la sémantique distributionnelle, Sables d'Olonne, pp. 154–166.