Clustering for Semantic Purposes:
Exploration of Semantic Similarity in a Technical Corpus

Ann Bertels and Dirk Speelman

This paper presents an innovative approach, within the framework of distributional semantics, for the exploration of semantic similarity in a technical corpus. In complement to a previous quantitative semantic analysis conducted in the same domain of machining terminology, this paper sets out to discover fine-grained semantic distinctions in an attempt to explore the semantic heterogeneity of a number of technical items. Multidimensional scaling analysis (MDS) was carried out in order to cluster first-order co-occurrences of a technical node with respect to shared second-order and third-order co-occurrences. By taking into account the association values between relevant first and second-order co-occurrences, semantic similarities and dissimilarities between first-order co-occurrences could be determined, as well as proximities and distances on a graph. In our discussion of the methodology and results of statistical clustering techniques for semantic purposes, we pay special attention to the linguistic and terminological interpretation.

Keywords: specialized corpora, distributional semantics, Multidimensional scaling (MDS), semantic similarity, second-order and third-order co-occurrences.

1. Introduction and research objectives

This paper describes an exploratory co-occurrence analysis of the specialized language used in a technical corpus of French machining terminology. This corpus of 1.7 million tokens was lemmatised and tagged with Cordial Analyseur and contains technical journals, technical data sheets, ISO standards and textbooks (1996-2002). The overall research goal is to extract semantic information from this corpus in order to explore the semantic heterogeneity of

various technical items and refine a previous quantitative semantic analysis performed on the same data (Bertels et al. 2010; Bertels 2006 and 2011).

The research described in this paper is part of a larger project involving theoretical research on term classification, but it is not intended as a tool for term extraction. In previous studies, we attempted to determine whether a number of typical lexical items in a technical corpus were monosemous, in line with traditional terminology (Wüster 1931 and 1991), or whether some technical items were in fact polysemous, as suggested by descriptive terminology (Cabré 2000; Temmerman 2000; Gaudin 2003). Various experiments on specialized corpora, both from a distributional and contextual perspective, have confirmed the polysemy of certain lexical items, even within a specialized domain (Arntz and Picht 1989; Condamines and Rebeyrolle 1997; Temmerman 2000; Eriksen 2002; Ferrari 2002). In this paper, we want to go more deeply into this classification, especially into semantic heterogeneity, in order to get a clearer picture of different types of polysemy and vagueness. Recent studies tend to focus on the application of cognitive linguistics to specialized language and promote the use of specialized corpora for the retrieval and extraction of semantic information (Geeraerts 2010; Faber 2012).

A statistical regression analysis of about 5000 key items in our technical corpus put the traditional monosemy ideal (Wüster 1991) into question and showed that the most typical lexical items are not the most monosemous ones (Bertels et al. 2010; Bertels 2006 and 2011).
In order to quantify monosemy, we developed a monosemy measure based on the formal overlap of second-order co-occurrences of a technical node. Second-order co-occurrences are defined as co-occurrences of first-order co-occurrences of a node. The basic idea behind this measure is to assess monosemy in terms of “semantic homogeneity” (Habert et al. 2005). Monosemous words appear in semantically homogeneous contexts, which means that their co-occurrences belong to similar semantic fields. Polysemous words, on the other hand,

appear in semantically more heterogeneous contexts and their co-occurrences tend to belong to different semantic fields. Therefore, in order to gain insight into the semantics of the first-order co-occurrences of a node, we analyse their co-occurrences, i.e. the second-order co-occurrences of the node, more particularly the degree of formal overlap of these second-order co-occurrences.

- A higher degree of formal overlap indicates semantic homogeneity of the first-order co-occurrences. Indeed, if more second-order co-occurrences are shared by more first-order co-occurrences, the latter are semantically more closely related. As a consequence, the node is semantically more homogeneous and hence more monosemous.
- A lower degree of formal overlap of second-order co-occurrences reveals semantic heterogeneity of the first-order co-occurrences, which means they are semantically less closely related or not related at all. Consequently, the node is semantically more heterogeneous.

The results of our monosemy measure were validated by several specialized dictionaries and by a manual analysis of the most significant co-occurring items (Bertels et al., 2010). It proved to be capable of detecting semantic homogeneity of monosemous words, as well as semantic heterogeneity of polysemous words. Unfortunately, it was not successful in dissociating polysemous senses, nor in distinguishing between polysemy and vagueness, which are both considered “semantic heterogeneity”.

Figure 1 shows the results of the non-linear LOESS regression for the 4717 key items (black points) with respect to their typicality rank (X-axis) and predicted monosemy rank (Y-axis). The most typical lexical items of the technical corpus, displayed in the upper left part of the graph, prove to be less monosemous (e.g. *usinage*, *tour*). Their position on the Y-axis shows a lack of monosemy. Since technical items with 3 or 5 senses or homonymous items can

all have the same coordinate on the Y-axis, the research described in this article aims to split the multidimensionality of the Y-axis.

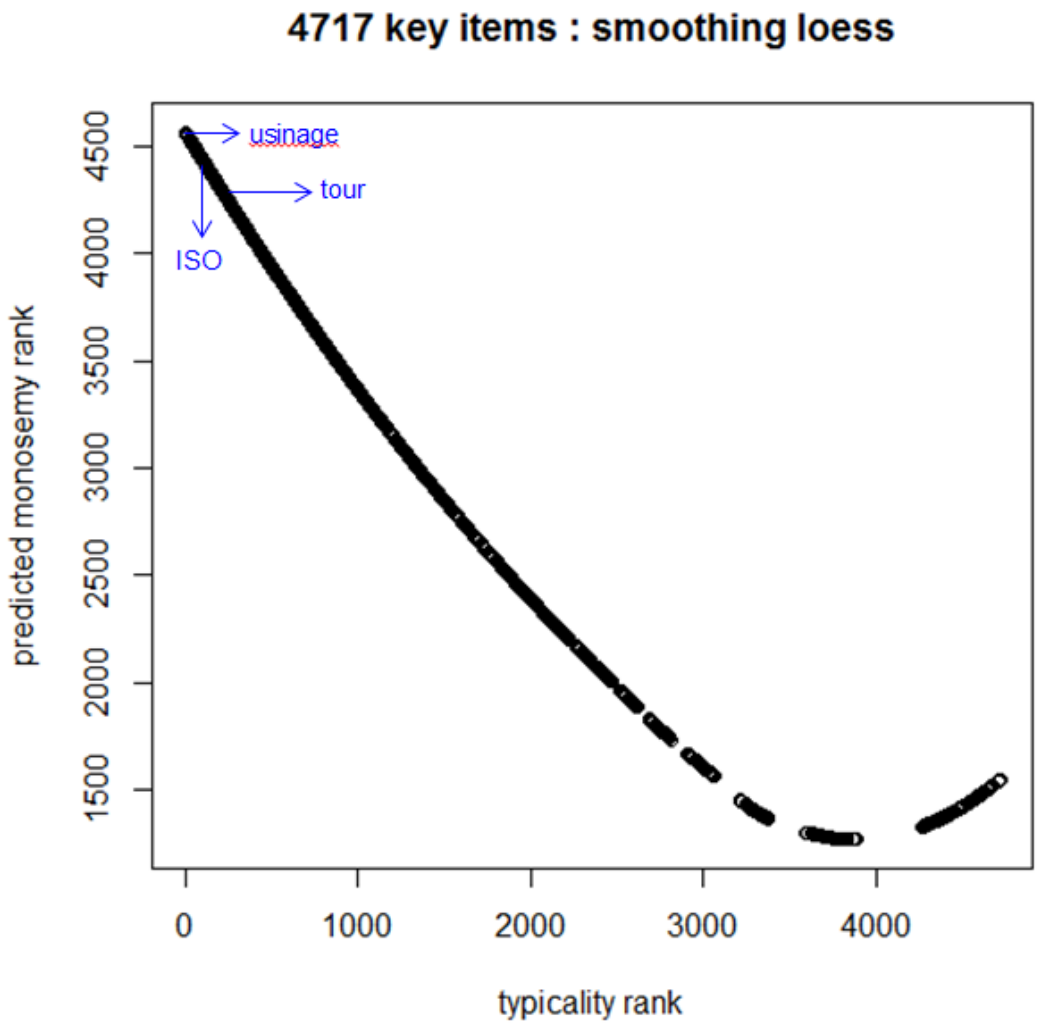

Figure 1. LOESS: non-linear regression (in the technical corpus)

In this study, we try to obtain a clearer picture of the semantic characteristics of a technical item by investigating the semantic distances between its first-order co-occurrences. Previously, our monosemy measure provided a quantification of the degree of semantic homogeneity or heterogeneity of the first-order co-occurrences of a technical node. Now, we want to plot the semantic similarities and dissimilarities between the first-order co-

occurrences on a plot, by means of clustering techniques. This means that first-order co-occurrences of a technical node are clustered with respect to the second-order and third-order co-occurrences they have in common. Third-order co-occurrences are defined as co-occurrences of second-order co-occurrences of a node. As a result, first-order co-occurrences which often appear with the same second-order co-occurrences are semantically related and will cluster together on a plot. If the resulting graph shows, for example, two clear-cut groups of first-order co-occurrences, this implies that the technical node has two different meanings. A graph showing just one group of first-order co-occurrences represents monosemy, whereas an expanding cloud reflects vagueness.

The objective of the research project is to obtain an accurate picture of the semantic distances and the corresponding semantic relations between significant first-order co-occurrences of a technical node, in order to assess its semantic heterogeneity (or homogeneity). We do not aim either at word sense disambiguation or at word sense induction. The goal of this article is not yet to differentiate between polysemy and vagueness, but to provide a visual tool, which helps us to find indications for the difference between polysemy and vagueness. This article focuses on how useful clustering and positioning techniques can be in determining distributional, and thus semantic, proximities between first-order co-occurrences, in order to refine our previous quantitative semantic analysis of these items.

The structure of this paper is as follows. First, we explain the methodological approach for clustering, more particularly Multidimensional scaling analysis (MDS), within the framework of distributional semantics (Section 2). Section 3 describes various configurations used to build up a co-occurrence matrix and focuses on the distinction between word forms and lemmas. The impact of these configurations on the clustering and visualisation is discussed in Section 4. Finally, Section 5 presents conclusions and suggestions for further research.

2. Methodological approach

2.1. Distributional semantics

The underlying idea of distributional semantics is that words with similar distributions have similar meanings. This is clearly expressed in Firth's (1968) adage, "You shall know a word by the company it keeps" and Harris' (1968) distributional hypothesis that words with similar meanings will occur in similar contexts (Clarke 2012).

Recent studies in distributional semantics determine semantic similarities between words on the basis of their distributional properties observed in corpora, mostly very large general language corpora. Two words are semantically similar if they appear in similar contexts, i.e. if they share syntactic contexts (Morlane-Hondère 2013; Morardo and Villemonte de La Clergerie 2013) or if they share first-order co-occurrences (Sahlgren 2006 and 2008; Peirsman and Geeraerts 2009; Ferret 2010; Heylen et al. 2012; Wielfaert et al. 2013). The latter studies generally determine semantic similarity by means of word space models (Semantic Vector Spaces or SVS) (Turney and Pantel 2010). Their identification of the most interesting and statistically significant co-occurring items relies on association measures. Words can then be clustered with respect to the (first-order) co-occurrences they have in common. Based on a dissimilarity matrix and a distance measure, words that very often appear with the same (first-order) co-occurrences cluster together in word space. Visualization of this word space shows groups of synonymous words (Ferret 2010) or semantically associated words (Peirsman and Geeraerts 2009). In case there is a problem of data sparseness on the level of first-order co-occurrences, this can be addressed by using second-order co-occurrences (Schütze 1998; Lemaire and Denhière 2006).

In this paper, however, the level of analysis is situated one level up in comparison with the studies mentioned above. Furthermore, the analysis is carried out on specialized language found in a technical corpus, rather than a general language corpus. Nevertheless, the most

innovative aspect of our study lies in the fact that it determines semantic similarity or proximity not between words themselves, but between first-order co-occurrences of a technical word. First-order co-occurrences are clustered and plotted on a visualisation, based on the second-order co-occurrences they have in common, in order to find groups of semantically related first-order co-occurrences of the technical node and determine how they relate to each other. In this analysis, we not only consider the fact that certain second-order co-occurrences of a node are shared (as in the monosemy measure), but we also take into account additional semantic information, in the form of the association score between first-order and second-order co-occurrences. If there is a problem of data sparseness on the level of second-order co-occurrences, we can still go one level up and look at the third-order co-occurrences related to the node.

These higher-order co-occurrences (Grefenstette 1994) are in fact second-order co-occurrences in relation to the first-order co-occurrences we attempt to cluster and plot in our analysis. It is interesting to note that lexical co-occurrence in specialized language on the first and second-order level has already been studied before, within the framework of distributional semantics, but the purpose was to construct taxonomies of specialized domains from noisy corpora based on hypernymy relations (Nazar et al. 2012).

2.2. Co-occurrence analysis

The first step in our exploratory co-occurrence analysis is the identification of statistically significant first-order co-occurrences of the technical nodes whose semantic heterogeneity (or homogeneity) we intend to examine. A recurrent co-occurrence analysis identifies all significant first-order and second-order co-occurrences and stores them in a co-occurrence matrix, which will serve as the starting point of the clustering experiments.

Several association measures can be used to assess the association strength between a node and its (first-order) co-occurrences, for example, the Log-Likelihood Ratio (LLR), Pointwise Mutual Information (PMI), Z-score, Dice Coefficient and Chi-squared (for a comprehensive overview, see Evert 2007). LLR (Dunning 1993) is a significance measure with its associated *p*-value. Although very often used as an association measure in co-occurrence analyses, the LLR tends to inflate the association score of co-occurring high-frequency words. PMI (Church and Hanks 1990) is an effect-size measure with a more intuitive interpretation. PMI is more reliable for high-frequency co-occurrences, but less reliable for co-occurrences involving low-frequency items, because it tends to overestimate the association score of infrequent items (Evert 2007).

In a preliminary study, on a small subset of the technical corpus (320,000 occurrences), we compared the impact of the LLR and PMI on the results of the clustering and on the plot (Bertels and Speelman 2013). This confirmed our expectation that the LLR inflates the association scores of high-frequency first-order co-occurrences: they were plotted at larger distances, at outlying positions even, despite their semantic relatedness. As a consequence, in this analysis, all co-occurrences were identified by using PMI as an association measure. Since there is no significance level or probability value associated to the PMI association measure and because PMI is less reliable for low-frequency items, we only consider co-occurrences which co-occur sufficiently often. If they have a co-frequency of at least 5, they are statistically reliable (Evert 2007). The lower the co-frequency threshold (but always at least 5), the higher the number of extracted co-occurrences. If the node is more frequent, this co-frequency threshold must be higher (e.g. 10 or 20), in order to avoid an overload of first-order co-occurrences on the plot.

Significant first-order co-occurrences were identified within a context window of 5 words to the left and right of a technical node. This span provides enough semantically

significant input, without adding too much noise, which might pose a problem when using a larger span. Significant second-order co-occurrences were identified within a span of 5 words to the left and right of the significant first-order co-occurrences, and significant third-order co-occurrences within the same span of the second-order co-occurrences. Co-occurring tokens can be extracted on the level of word forms or lemmas. Section 3 addresses these two distinct approaches and discusses various configurations to build up a co-occurrence matrix with either word forms or lemmas.

2.3. Clustering techniques

Different clustering techniques exist, e.g. Principal Components Analysis and Factor Analysis for tables with measurements, Correspondence Analysis for tables with counts and Multidimensional Scaling Analysis and Hierarchical Cluster Analysis for tables with distances (for a comprehensive overview, see Baayen (2008)). As will be discussed in the next section, we will use the technique of Multidimensional scaling, which involves dimension reduction prior to clustering. One example of an alternative approach to dimension reduction would have been the use of singular value decomposition as it is adopted in latent semantic analysis (Landauer and Dumais 1997). For a thorough discussion of the use of different dimension reduction techniques in distributional semantics, see Evert (2012).

2.3.1. Multidimensional scaling (MDS)

Since our aim is to explore the (semantic) distances between first-order co-occurrences on a plot, Multidimensional scaling (MDS) is an appropriate choice. We will use isoMDS, which is an implementation of non-metric MDS (Kruskal and Wish 1978; Cox and Cox 2001; Venables and Ripley 2002) in R, a freely available tool for statistical analysis. [1,2]

[1] IsoMDS is an implementation of “Kruskal's Non-metric Multidimensional Scaling”.

MDS takes as a starting point an item-by-item dissimilarity matrix, on the basis of which it attempts to visualise items in a low-dimensional space (typically in two dimensions) in such a way that the distances between the items in this low-dimensional space reflect the input dissimilarities as well as possible. The visual representation maximises the goodness-of-fit and minimises the distortion, created by reducing all the dimensions to the two dimensions displayed on the graph. The quality of this representation is determined by a stress percentage. Overall, an excellent stress percentage should be lower than 10% and a stress percentage above 15% is often considered as unacceptable (Clarke 1993; Borg and Groenen 2005).[3] However, this is not a clear boundary, because some studies manage it in a flexible way and consider a percentage between 15% and 20% as "less reliable", e.g. Borg and Groenen (2005, 47-48). Often, as is our case, the dissimilarity matrix that is the input for the MDS analysis is itself derived from an item-by-feature matrix, to which a distance measure is applied in order to obtain dissimilarities between the items with respect to their values in the columns. In our case this item-by-feature matrix has as its rows (=items) the first-order co-occurrences and as its columns (=features) the second-order or third-order co-occurrences.

2.3.2. Co-occurrence matrix

MDS allows us to cluster and visualise the proximities and distances, or the semantic similarities and dissimilarities, between significant first-order co-occurrences (*c*) of a technical node (extracted with PMI as an association measure), with respect to their association score (PMI) with significant second-order co-occurrences (*cc*). As a starting point, we used a co-occurrence matrix (used as a distance matrix) with all significant *c* of a technical node as rows, and all significant *cc* as columns. This *c* × *cc* co-occurrence matrix was generated by running Python scripts on the technical corpus of 1.7 million words. Since there is no probability value

[2] The R Project for Statistical Computing: www.r-project.org [accessed June 2014].
[3] http://geai.univ-brest.fr/~carpenti/2006-2007/Documents-R/MDS-non-metrique.html.

associated with PMI, a lower threshold will be applied for co-frequency of *c* with the node and for co-frequency of *cc* with *c*. At a lower co-frequency threshold above 5, the number of *c* and *cc* included in the *c* × *cc* matrix is higher, but they co-occur less often with the node and the *c,* respectively.

The more similar the association information in the columns is for the different rows, the more semantically similar the rows (or *c*) will be. Each entry in the *c* × *cc* matrix contains the association score between *c* and *cc* (see table 1). If there is no significant association between *c* and *cc*, the entry in the co-occurrence matrix is a very small number (0.00000001), which is nonetheless indispensable to calculate the values in the dissimilarity matrix (see Section 2.3.3). The main problem is that the *c* × *cc* matrix is often very sparse, because numerous *cc* are shared by very few *c*, which sometimes makes it impossible to generate graphs that are interesting from a linguistic point of view. However, this problem of data sparseness can be addressed by taking into account co-occurrences of a higher order, i.e. third-order co-occurrences of the node (or *ccc*) (Bieman et al. 2004; Lemaire and Denhière 2006).

Table 1. Simplified example of a *c* × *cc* co-occurrence matrix

	cc_1	cc_2	cc_3	cc_4	cc_5	cc_6
c_1	association score (PMI) between c_1 and cc_1					
c_2						
c_3						
c_4						

In the resulting *c* × *ccc* co-occurrence matrix for a technical node, the rows consist of all significant first-order co-occurrences (*c*) of the node, whereas the columns contain all

significant third-order co-occurrences (*ccc*) of all significant *c* and all significant *cc*. The value of an entry in the *c* × *ccc* matrix is not a PMI association score, as a *c* does not co-occur directly with a *ccc*. Instead, this value corresponds to the column sum of a new *cc* × *ccc* matrix for each *c* of a node, with the PMI association score between *cc* and *ccc* as the value of an entry. If there are *n ccc* for all significant *cc* of a *c*, the *cc* × *ccc* matrix for a *c* provides a sum for each column, thus generating a vector with *n* dimensions. The elements of the *n*-dimensional vector represent all the entries for the *c* row in the *c* × *ccc* matrix. This results in a new *c* × *ccc* matrix which is less sparse and semantically more rich and therefore better suited for our clustering experiments (see table 2).

Table 2. Simplified example of a *c* × *ccc* co-occurrence matrix

	ccc_1	ccc_2	ccc_3	ccc_4	ccc_5	ccc_6
c_1	column sum for ccc_1 in *cc* × *ccc* matrix for c_1					
c_2	column sum for ccc_1 in *cc* × *ccc* matrix for c_2					
c_3						
c_4						

2.3.3. Dissimilarity matrix

Using the *c* × *ccc* co-occurrence matrix as a starting point, a dissimilarity matrix was created in R by calculating pairwise distances between the *c*. By default, isoMDS uses Euclidian distance to generate the dissimilarity matrix used to cluster and plot observations, in our case the first-order co-occurrences of a technical node. Euclidian distance is the spatial distance or straight line between two observations on a graph. Previous experiments, however, showed that Euclidian distance is less appropriate as a distance measure for a co-occurrence matrix with

both high and weak association scores, because on the plot, low scores lose all their weight compared to high scores (Bertels and Speelman 2013).

An alternative metric used to generate a dissimilarity matrix is cosine angle, a distance measure for observations represented as vectors.[4] Cosine angle determines (semantic) similarity by calculating the angle between the vectors. Observations with similar context vectors cluster together in multidimensional space, because the angle between them is small. Even if the magnitude of the values in the two vectors changes, the angle between them is not affected (van der Laan and Pollard 2003). Cosine angle is often used as a distance measure in distributional semantics (Padó and Lapata 2007; Sahlgren 2008), for example, to identify similarity between two text samples or between two words (Peirsman and Geeraerts 2009). In our previous experiments, cosine angle has yielded satisfactory results in plotting the MDS outcome for a technical node based on a small corpus (Bertels and Speelman 2013). The first-order co-occurrences in the rows of the co-occurrence matrix are considered to be vectors with one value in each column. Based on the association information in the columns, it is then easy to calculate similarity or dissimilarity.

3. Co-occurrences: word forms or lemmas?

As previously mentioned, our study aims to cluster first-order co-occurrences of several technical nodes, qualified by the monosemy measure as either semantically heterogeneous (*tour*, *usinage*) or semantically more homogeneous (*ISO*). The main goal is to identify semantic similarities and dissimilarities between first-order co-occurrences, displayed as proximities and distances on a plot, in order to understand the semantic characteristics of the technical nodes. In this paper, we specifically focus on the distinction between word form and lemma in the identification of significant co-occurrences, which is an interesting distinction from both a

[4] In R, cosine angle is implemented in the function distancematrix in the library hopach.

terminological and a semantic point of view. Moreover, this is an important issue in French, which is a moderately inflected language, more so than in languages like English.

As part of our quantitative semantic analysis, we developed a monosemy measure which considered technical nodes at the level of lemma or canonical form. All inflected word forms of a noun or adjective, for example, were grouped under one lemma (e.g. the word forms *tour* and *tours* for the noun *tour*). Similarly, one lemma encompassed all conjugated forms of a verb (e.g. *permettre*, *permet*, *permettent*, *permettant*, etc. for the verb *permettre*). Significant first-order and second-order co-occurrences, however, were identified at the level of word form. The main advantage of word forms, within the context of the current procedure, is that they are semantically more rich, especially in French. Identification of co-occurrences at the level of word form accounts, for example, for the semantic difference between *pièce usinée* (manufactured piece, i.e. "result") and *pièce à usiner* (piece to be manufactured, i.e. "intention"). If co-occurring items of *pièce* were extracted on the lemma-level (infinitive *usiner*), this semantic information would be lost. From a terminological point of view as well, word forms seem to be more interesting in co-occurrence identification, because they contain more precise information on how a term is formed. For the node *usinage*, for example, significant first-order co-occurrences reveal interesting multiword units such as *usinage grande vitesse* (high-speed machining) and *usinage haute précision* (high-precision machining). Although word forms are very useful for co-occurrences of some technical nodes (e.g. *usinage*), they do add considerable noise to the data. When considering co-occurrences at the level of word form, the number of co-occurrences is higher (because they are formally more diverse), but their frequency is lower. When co-occurrences are identified at the lemma level, the number of formally diverse co-occurrences is lower, but their frequency is higher. As a consequence, this is likely to have an impact on the plot of an MDS analysis, which will be discussed in Section 4. In this article, we want to look at the added value for both

configurations of co-occurrences as word forms and lemmas. As a starting point, we take the configuration at the word-form level because our monosemy measure was based on the word forms of the first and second-order co-occurrences and not on the lemmas. For some nodes, this configuration provides useful information. For other nodes, however, lemmatisation provides a clearer picture of the displayed first-order co-occurrences. We want to check every node again in order to find the most interesting configuration.

Note that for our monosemy measure, three configurations were compared in a subset of the technical corpus (320,000 occurrences) in order to assess the impact on monosemy degree and monosemy ranking. The three configurations were (1) lemma (node) – word form (*c*) – word form (*cc*), (2) lemma (node) – lemma (*c*) – word form (*cc*) and (3) lemma (node) – lemma (*c*) – lemma (*cc*) (for further details, see Bertels and Speelman 2012). When *c* and *cc* (configuration 3) or only *cc* (configuration 2) were considered on the lemma level, the degree of formal overlap of the *cc* of a node was higher, because the number of formally diverse *cc* was lower. As a result, the monosemy degree of all technical nodes was higher. Experiments showed, however, that semantically heterogeneous nodes were found to be semantically heterogeneous throughout the three configurations, showing a low degree of formal overlap in *cc*. Moreover, semantically homogeneous nodes were found to be semantically homogeneous in all three configurations as well. These forms, whether word forms or lemmas, therefore did not have a major impact on the outcome of the monosemy measure. Note that the quantitative analysis used monosemy rankings instead of monosemy degrees in order to determine the correlation between monosemy and typicality. Even when monosemy degree was higher for all considered nodes, this had no major impact either on their ranking or on the outcome of the regression analysis.

4. Discussion of the results

In order to evaluate the impact of the distinction between word forms and lemmas on the plot of the MDS analysis, several experiments were carried out, first on a subset of the technical corpus (Section 4.1) and then on the complete technical corpus (Section 4.2). In both sections, we focus on the technical node *tour*, not only because it is one of the most typical keywords in the technical corpus, but especially because it is semantically very heterogeneous and therefore particularly interesting for such comparative clustering experiments. The node *tour* in fact has two distinct technical meanings: "revolution" as in *dix mille tours par minute* (ENG. *ten thousands revolutions per minute)* and "tool for machining an object" as in *tour à commande numérique* (ENG. *CNC lathe*). It is used in several more general senses, depending on the context ("tower", "round", "lap", "turn", "tour or trip"). Sections 4.1 and 4.2 discuss the results for other technical nodes as well, i.e. *usinage* ("machining") and *ISO*. For the experiments described in this paper, we will only consider a *c* × *ccc* co-occurrence matrix, because of a significant data sparseness problem in the *c* × *cc* co-occurrence matrix. We focus here on words which we know have different senses. *Tour*, *usinage* and *ISO* turned out to be semantically heterogeneous in our previous quantitative analysis. Now, this "semantic heterogeneity" will be further explored, to see if we can also get it out of the data with an MDS analysis, and even go a bit further. The long-term objective is to generate a new monosemy measure which takes into account other parameters as well. This MDS analysis can be regarded as a first step in that direction.

4.1. MDS analysis of the technical journals

The subset of the technical corpus consisting of online technical journals (320,000 words) has previously been used for various evaluation experiments (Bertels and Speelman 2012 and 2013). As mentioned before, the nodes were considered at the lemma level. The first configuration discussed considers co-occurrences at the level of word form, while the second

configuration considers them at the lemma level. Note that a lower threshold for co-frequency is applied (minimum co-frequency of 5), in order to avoid overestimation problems in the PMI for low-frequency items. Function words are preserved in the list of significant second-order and third-order co-occurrences, as some of them provide useful semantic information: *pendant* (ENG. *during*), for example, indicates a process. However, these function words were filtered out of the list of significant first-order co-occurrences. It is not the likelihood of a semantic contribution, but rather the impact these have on the complexity of the analysis and representation that is different. Preservation of function words is manageable in the first case but not in the second, because they would make the plot too dense and therefore unreadable.

Table 3 shows a comparison of the characteristics of both configurations for the node *tour*, at a lower threshold of minimum co-frequency 5, without function words in the *c* rows. The number of significant *c* is similar in both configurations, although these *c* are not the same. For example, the list of significant *c* in the word form configuration contains both *broche* (ENG. *spindle*) and *broches*, whereas the list for lemmas only includes the lemma *broche*. On the other hand, some *c* are significant in the lemma configuration (higher co-frequency with node), but not in the word form configuration, for example *frontal*. As previously mentioned, a stress percentage lower than 15% is acceptable and useful for the interpretation of the resulting plot. Table 3 shows that the lemma configuration slightly outperforms the word form configuration.

Table 3. Word form and lemma configurations in the technical journals (node *tour*)

Configuration	**Number of significant *c***	**Stress of MDS analysis**
c – *cc* – *ccc* as word forms	38	14.81%
c – *cc* – *ccc* as lemmas	37	14.48%

Figure 2 shows the plot of the MDS analysis of word forms. The dispersion of the *c* reflects fairly clearly the semantic heterogeneity of *tour*. The most outlying *c*, *horizon*, refers to one of the general meanings of the node *tour* (as in *tour d'horizon*: "quick overview"). Furthermore, the outlying *c mille* indicates the more technical meaning "revolution" (as in *dix mille tours par minute,* "10,000 revolutions per minute"). *CNC*, *commande* and *numérique* cluster together and are clearly semantically related. In the upper right corner, the displayed *c* (*bibroches, inversés*) indicate a more specific meaning ("two-spindle machine"). The cloud of *c* in the bottom left corner represents the technical meaning "machine tool for machining an object or work piece". Several smaller clusters reveal the variety in usage contexts: a more specific usage (*tour (à) deux broches*) at the bottom left (*axes, broches, broche, outils, deux, trois, quatre, centres*), and more towards the centre of the plot (*vertical* and *verticaux*). The centre of the cloud contains *gamme, série, type* and a few less technical *c*. The more general *c* (*manutention, ligne, générale, production, conception*) are displayed in the upper left part of the cloud and reflect more general usage contexts of the node *tour*. What is striking is that some semantically identical word forms (e.g. the singular *vertical* and the plural *verticaux* or the adjectival forms *inversé*, *inversée* and *inversées*) are situated at quite a large distance from each other. This is due to the fact that they co-occur with other *cc*, that is to say, formally different *cc*. From a semantic point of view, this plot is not particularly useful and does not allow a coherent semantic interpretation.

As shown in Figure 3, the lemma configuration somewhat resolves this problem. The word forms *inversé*, *inversée* and *inversées* in Figure 2 are all displayed as *inverser*. The outlying *c horizon* still accounts for a more general meaning. The item *mille* is also in an isolated position and reflects a more particular meaning. Furthermore, *bibroche* seems to be more outlying, mainly because all inflected and conjugated forms of *inversé* are grouped under the lemma *inverser*. The more general *c* are displayed in the upper left section of the plot

(*manutention*, *constructeur*, *production*, *ligne*) and the more technical *c* in the lower section of the plot (see Figure 3).

For the node *usinage*, the dispersion of the lexical *c* (word forms) reflects the semantic vagueness of the node: there are no clear groupings of *c* (see Figure 4). Since the node *ISO* has only 3 lexical *c* in the subset of the technical journals, MDS-analysis in this subset is not useful.

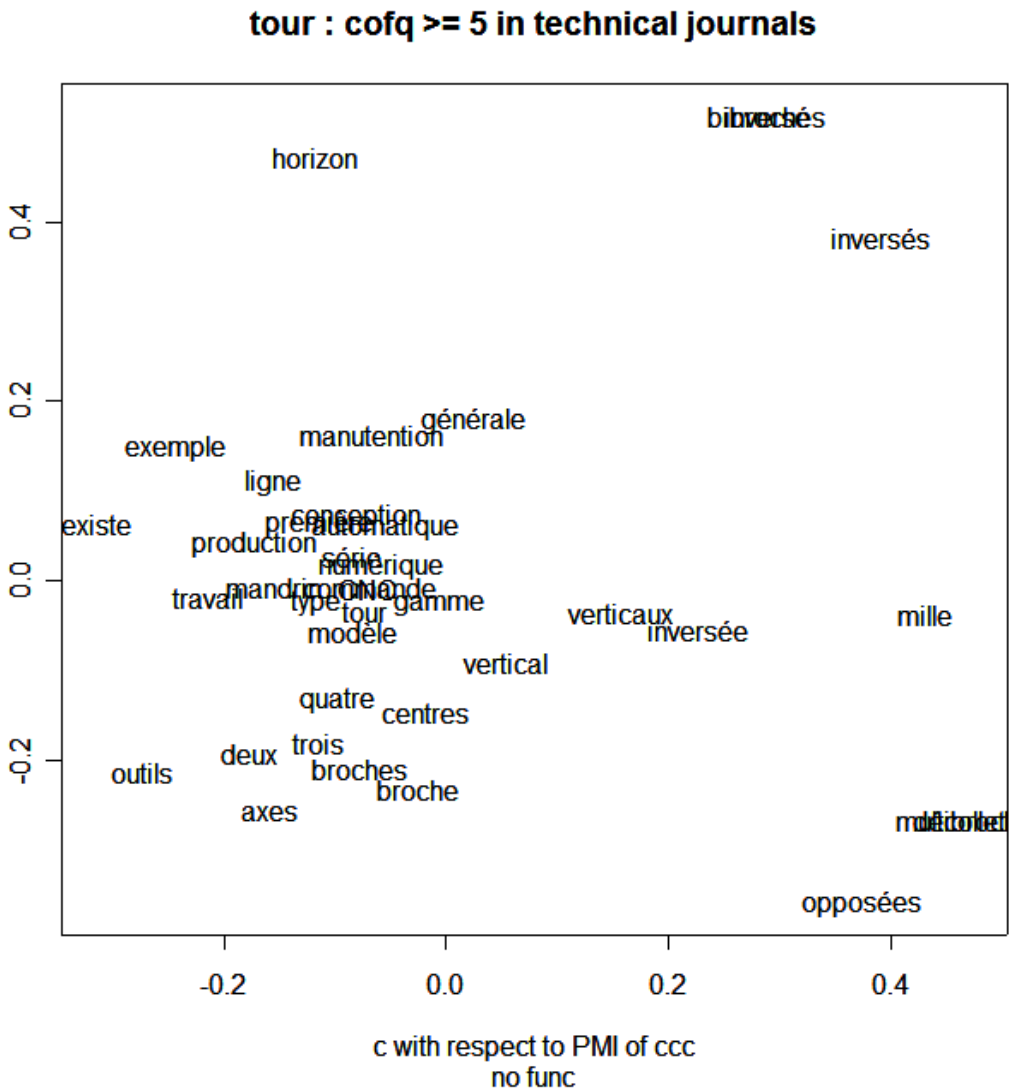

Figure 2. MDS of *c* of the node *tour* in the technical journals: word forms

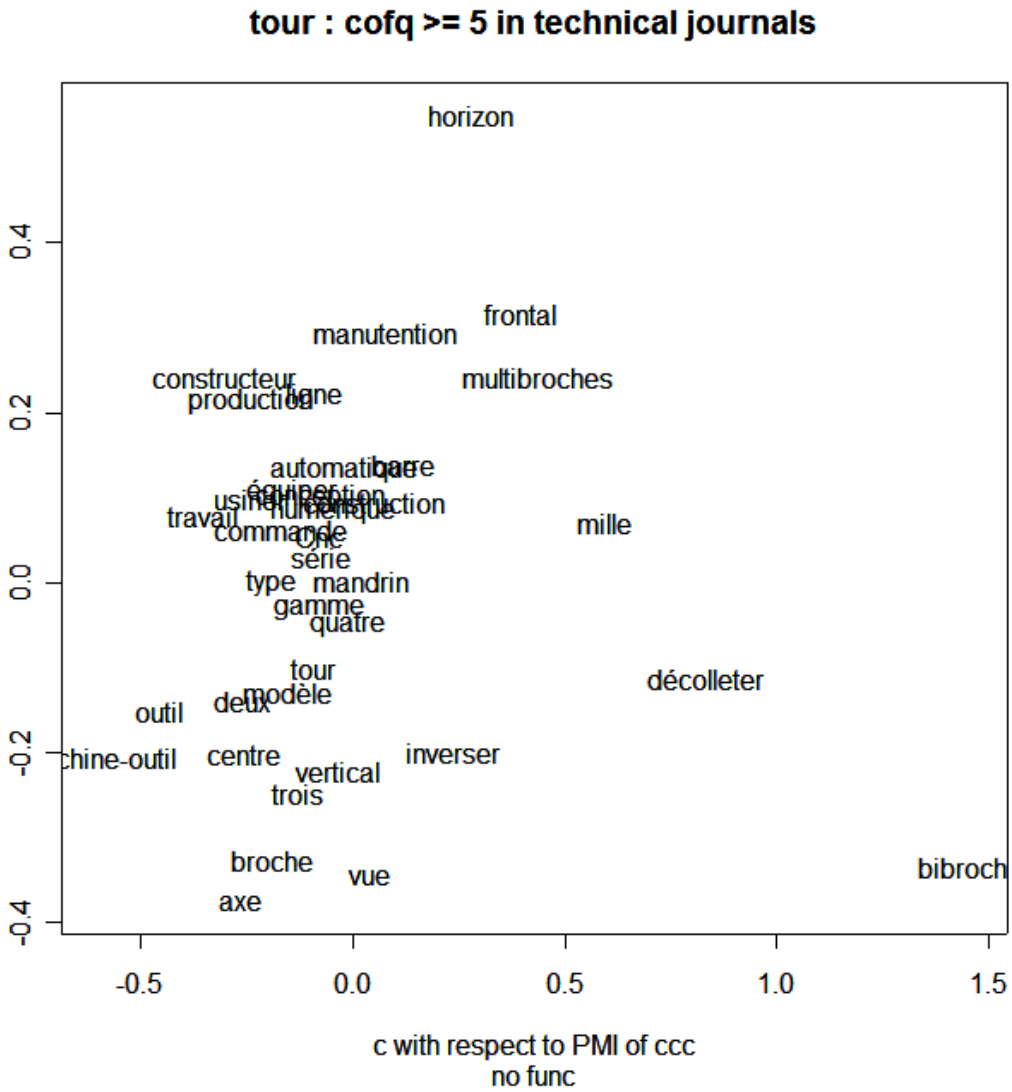

Figure 3. MDS of *c* of the node *tour* in the technical journals: lemmas

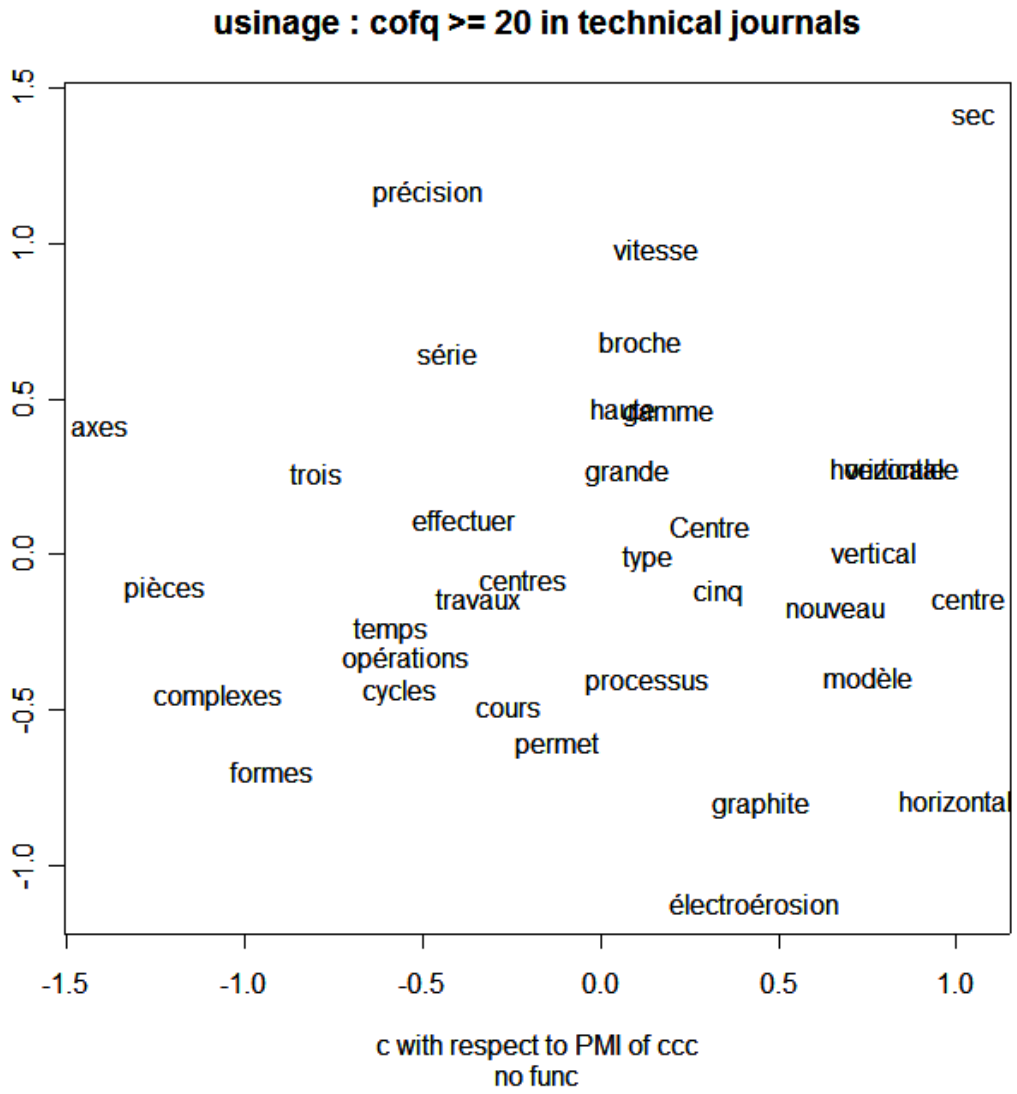

Figure 4. MDS of *c* of the node *usinage* in the technical journals: word forms

4.2. MDS analysis on the technical corpus

Similar experiments were carried out to detect semantic (dis)similarities between first-order co-occurrences in the complete technical corpus of 1.7 million words. The goal here is the comparison of the two variants of the approach (word forms and lemmas) rather than evaluating the method as a whole. Table 4 shows some interesting characteristics for several technical nodes in both configurations, i.e. co-occurrences as word forms and as lemmas.

The experiments used various lower thresholds for the technical corpus, i.e. a minimum co-frequency of 10, 20, 50 and 100. These lower thresholds had to be adopted, given that all frequencies and co-frequencies are (much) higher in the complete technical corpus than in a smaller subset. If, on the one hand, the number of significant *c* is too high, for example 100 or more, too many *c* will be clustered and displayed on the plot, making it too dense and therefore unreadable. The stress percentage will also be too high to obtain a reliable representation of the observed similarities and dissimilarities. On the other hand, if the number of significant *c* is too small, for example 5 or 6, these very few *c* are not likely to reveal interesting semantic information on the plot. Depending on the frequency of the node and the number of significant *c*, the lower threshold will be adapted.

Table 4. Configurations with word forms and lemmas in the technical corpus

Configuration			**Number of significant *c***	**Stress % of MDS analysis**
Word forms	*tour*	cofq ≥ 10	75	18.75%
Lemmas	*tour*	cofq ≥ 10	90	17.67%
Word forms	*usinage*	cofq ≥ 100	18	17.97%
Lemmas	*usinage*	cofq ≥ 100	23	18.68%
Word forms	*ISO*	cofq ≥ 10	32	12.40%
Lemmas	*ISO*	cofq ≥ 10	39	13.63%

For the node *tour*, the most interesting configurations have lower threshold 10. As shown in Figure 5 (for word forms) and, less convincingly, in Figure 6 (for lemmas), these configurations confirm the findings discussed in Section 4.1 for the technical journals. Again, *horizon* and *mille* are situated at outlying positions. Moreover, Figure 5 shows two other outliers (*monobroche* and *frontaux*), which have high values for some specific *ccc* in the columns of the *c* × *ccc*

matrix. Since these values represent the column sum of association scores between *cc* and *ccc*, it is unfortunately rather difficult to trace the origin of their outlying position.

However, the MDS analysis of these two configurations is only borderline reliable, as indicated by the rather high stress percentages of 18.75% and 17.67% (see Table 4). Furthermore, the semantic interpretation of the displayed *c* of *tour* extracted from the complete technical corpus is less clear and less convincing than those extracted from the subcorpus (see Section 4.1). The plot of the MDS analysis of the complete technical corpus could be considered as the superposition of 4 plots for each of the 4 subcorpora (technical journals, technical data sheets, ISO standards and textbooks). It therefore seems interesting to carry out an MDS analysis for each subcorpus. As suggested by Figures 2, 3, 5 and 6, significant first-order co-occurrences of *tour*, thematically slightly different in each subcorpus, seem to display a different co-occurrence behaviour, depending on the subcorpus.

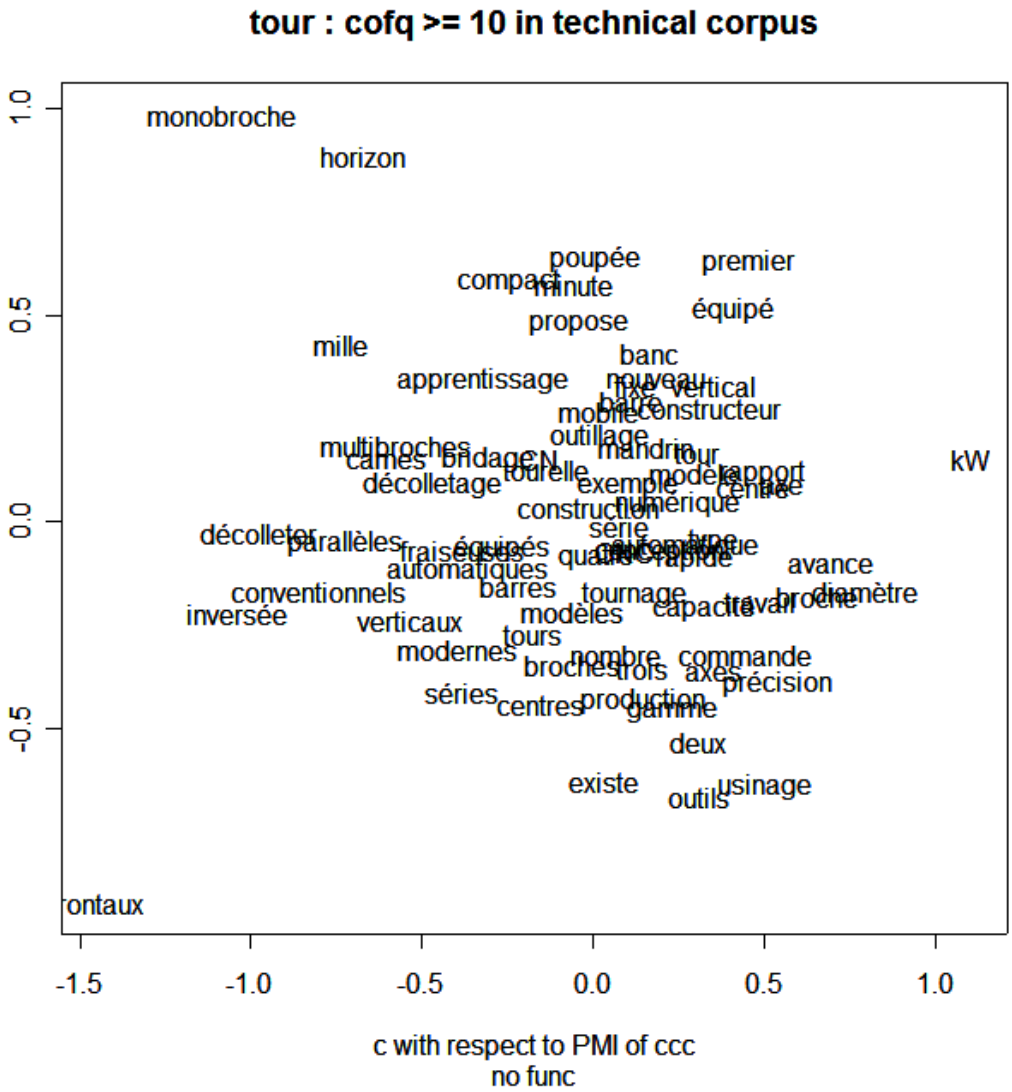

Figure 5. MDS of *c* of the node *tour* (in the technical corpus): word forms

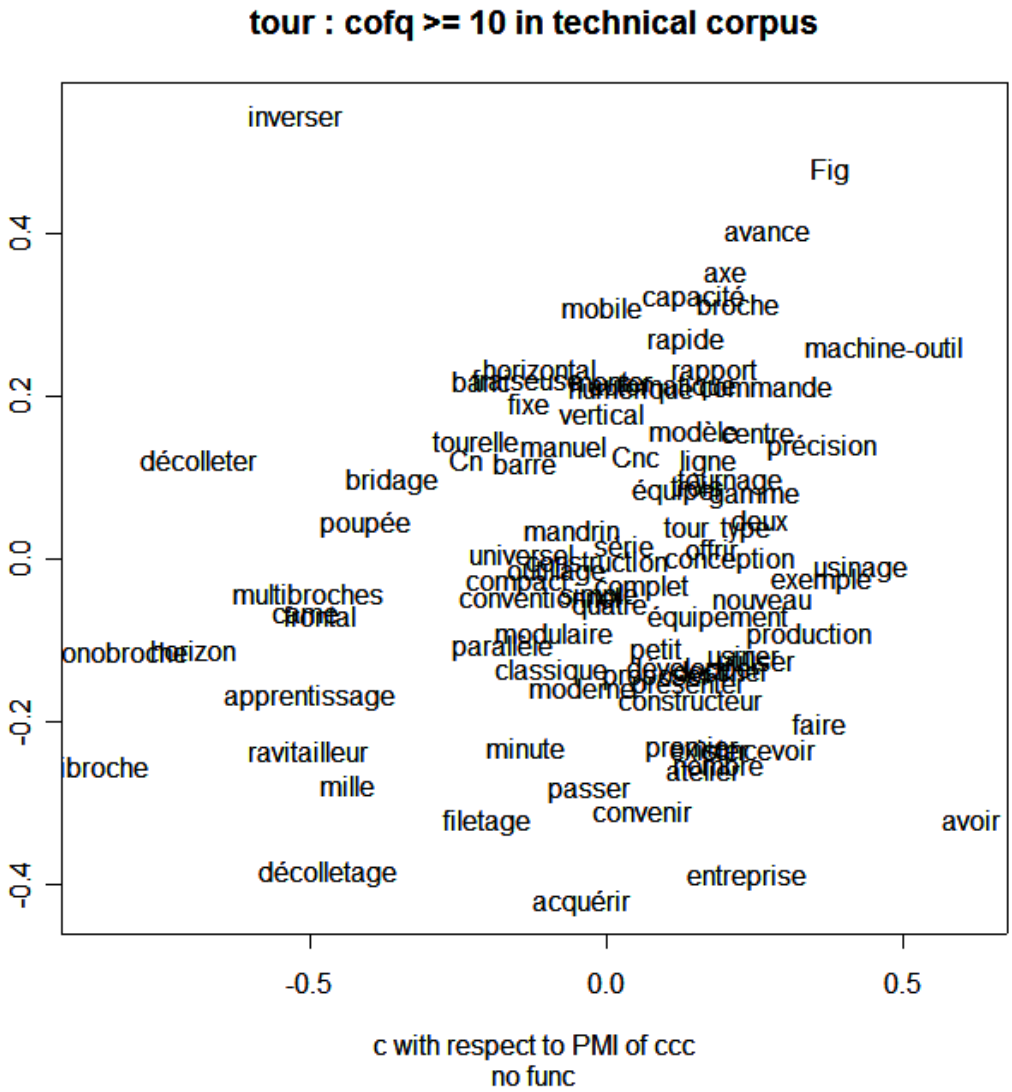

Figure 6. MDS of *c* of the node *tour* (in the technical corpus): lemmas

The technical node *usinage* was qualified by the monosemy measure as semantically heterogeneous. In fact, this deverbal noun is semantically vague, because it has a rather general meaning (“machining”), which is further specified by the complement, for example *usinage d’un trou* (“taraudage” or “tapping”) versus *usinage d’une pièce* (“fraisage, rectification” or “milling, rectification”). At a lower threshold of 100, the stress percentage is rather high in both configurations (see Table 4). For both configurations, Figures 7 and 8 show a cloud of *c* which clearly reflects the semantic vagueness of *usinage*. Furthermore, for the

word form configuration, Figure 7 shows several interesting groupings of *c*, in terms of the multiwords units that they constitute (*enlèvement de copeaux, haute précision, grande vitesse*). This suggests that an MDS analysis might be helpful for the identification and analysis of multiword units in future research. Note that we use MDS analysis as a way to find a robust method for discovering semantic clusters in all types of corpora, even in corpora without full sentences, as is the case in our technical corpus.

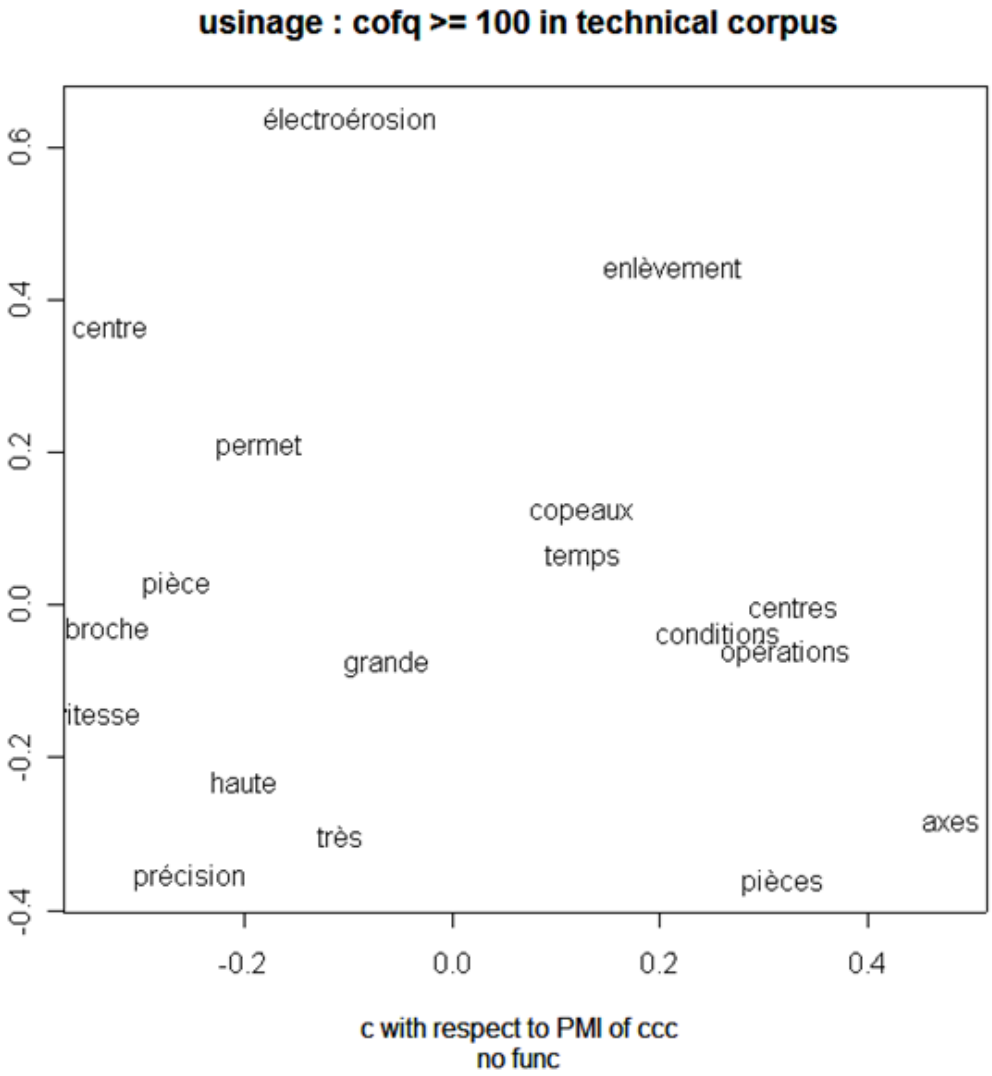

Figure 7. MDS of *c* of the node *usinage* (in the technical corpus): word forms

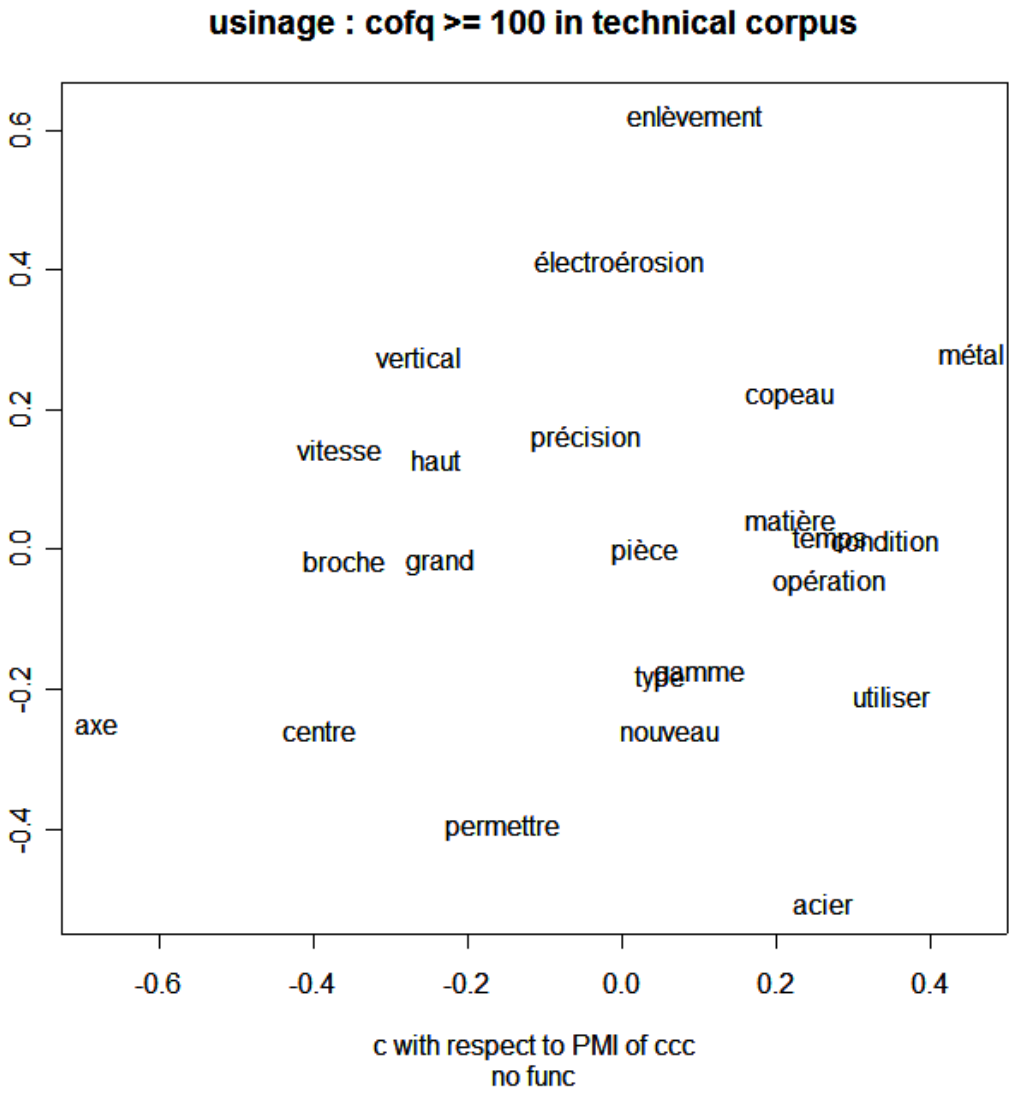

Figure 8. MDS of *c* of the node *usinage* (in the technical corpus): lemmas

Finally, for the technical node *ISO*, which is semantically rather homogeneous, both configurations at lower threshold 10 are reliable: they have acceptable stress percentages of 12.40% and 13.63%, respectively. Function words were filtered out of the rows in the co-occurrence matrix, but numbers were kept in, because of the particular characteristics of the node *ISO*. In the configuration on word forms, numbers and references to the standards cluster together to the left of the plot (see Figure 9). The bottom section shows a specific use of the node *ISO* with the indications *9000*, *9001*, *9002*, *certifiée* and *certification*: these *c* all

refer to the well-known quality standards and certifications. In the plot for the lemma configuration (see Figure 10), numbers are displayed on the left section of the plot and *c* related to quality standards and certifications here too cluster together in the upper left section of the plot.

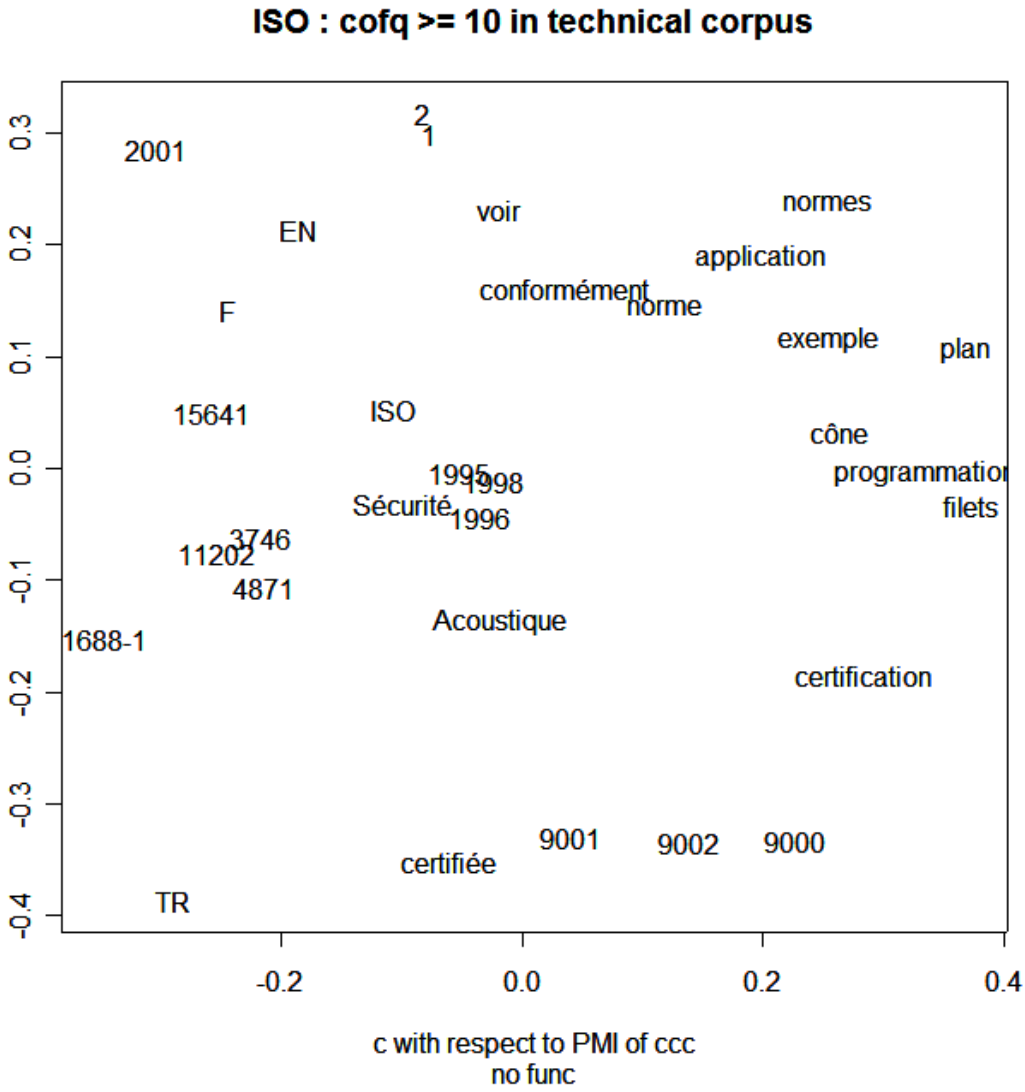

Figure 9. MDS of *c* of the node *ISO* (in the technical corpus): word forms

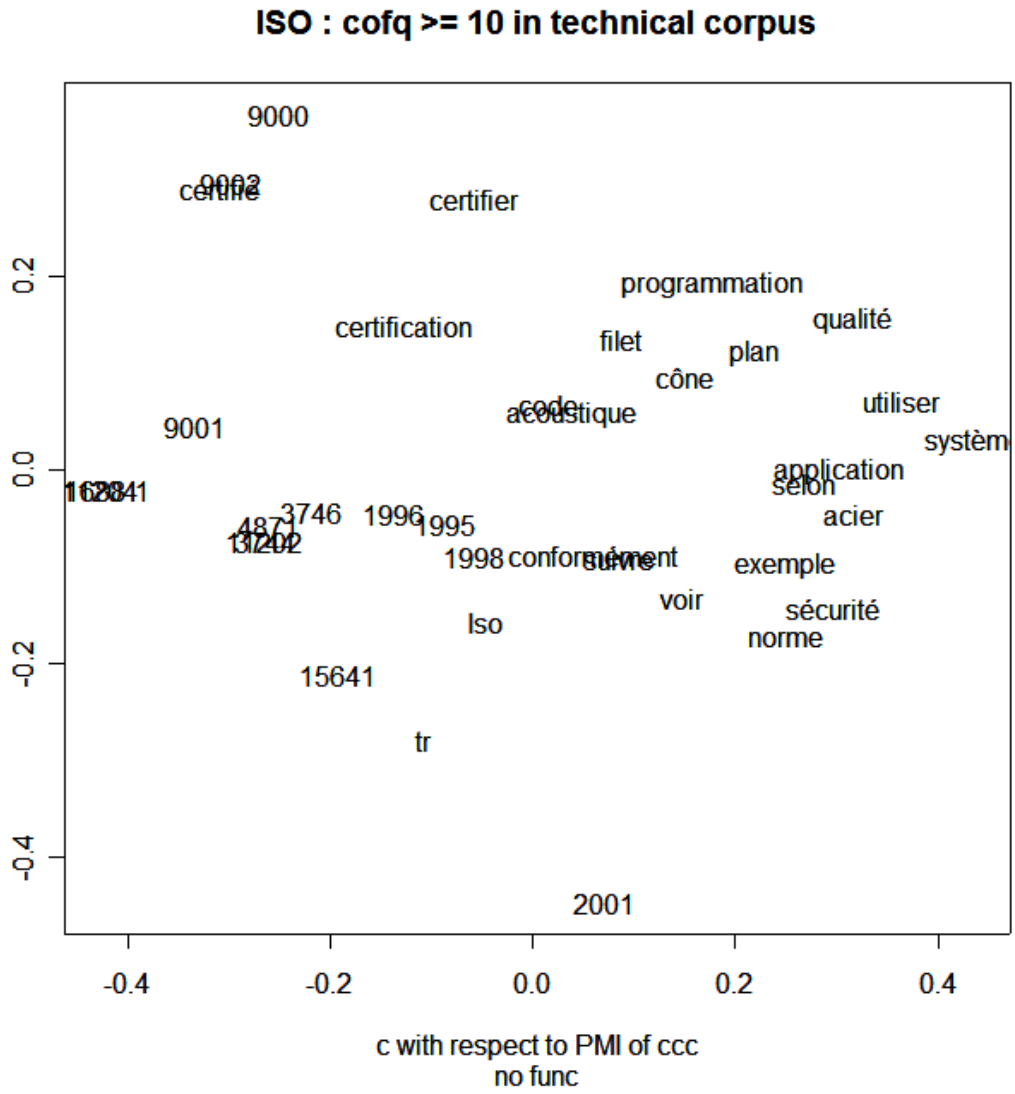

Figure 10. MDS of *c* of the node *ISO* (in the technical corpus): lemmas

5. Conclusions and further research

This paper presented the methodology of clustering for semantic purposes, and discussed its visual results both from a terminological and a semantic point of view. In several experiments, first-order co-occurrences of technical nodes were clustered based on shared-order second and third-order co-occurrences and by taking into account their respective association strength. The results of the statistical clustering techniques showed semantic similarities and

dissimilarities between first-order co-occurrences by means of proximities and distances on a plot.

MDS analysis was carried out on a subset of technical journals (320,000 words) as well as on the complete technical corpus (1.7 million words). Two configurations for the co-occurrence matrix were generated, for word forms and for lemmas, in order to assess the impact of the identification of co-occurrences at the level of word form and lemma. For the node *tour*, the plots for the subcorpus showed, in both configurations, some groupings of co-occurrences and some outliers, which indicated different meanings. For the complete technical corpus, the semantic interpretation of the plots was less convincing, and suggested that MDS analysis might be necessary for each specific subcorpus. The four subcorpora seem to have their own stylistic and thematic characteristics, even within the same technical domain of machining terminology.

For the technical nodes discussed in this paper, the lemma configuration yields a more coherent semantic interpretation, as the various inflected and conjugated word forms are not all displayed on the resulting plot. Of course, the MDS analysis needs to be expanded to more numerous technical nodes in order to explore their semantic heterogeneity and confirm the findings so far.

Finally, future MDS analysis should take into account word class information or POS-tags for first-order co-occurrences intended to be clustered and plotted. This would also be relevant for second-order and third-order co-occurrences, which have an important disambiguation function in the underlying $c \times ccc$ co-occurrence matrix. This integration of statistical co-occurrence information with syntactic co-occurrence information could be a significant step towards the identification and analysis of multiword units, which are very important in specialized language.

References

Arntz, Reiner, and Heribert Picht. 1989. *Einführung in die Terminologiearbeit.* Hildesheim: Georg Olms Verlag.

Baayen, Rolf H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics using R.* Cambridge: Cambridge University Press.

Bertels, Ann, and Dirk Speelman. 2012. "La contribution des cooccurrences de deuxième ordre à l'analyse sémantique." *Corpus* (11): 147–165.

Bertels, Ann, and Dirk Speelman. 2013. "Exploration sémantique visuelle à partir des cooccurrences de deuxième et troisième ordre." In *Actes de Traitement Automatique des Langues Naturelles (TALN 2013) Atelier Sémantique Distributionnelle (SemDis).* 126–139. Sables d'Olonne, France.

Bertels, Ann, Dirk Speelman, and Dirk Geeraerts. 2010. "La corrélation entre la spécificité et la sémantique dans un corpus spécialisé." *Revue de Sémantique et de Pragmatique* 27: 79–102.

Bertels, Ann. 2006. *La polysémie du vocabulaire technique. Une étude quantitative.* PhD thesis. University of Leuven.

Bertels, Ann. 2011. "The Dynamics of Terms and Meaning in the Domain of Machining Terminology." *Terminology* 17(1): 94–112.

Biemann, C., Bordag, S. and Quasthoff, U. (2004). "Automatic acquisition of paradigmatic relations using iterated co-occurrences." In *Proceedings of Language Resources and Evaluation (LREC 2004)*, 967–970. Lisboa, Portugal.

Borg, Ingwer, and Patrick Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications.* New York: Springer-Verlag.

Cabré, Maria Teresa. 2000. "Terminologie et linguistique: la théorie des portes." *Terminologies nouvelles* 2: 10–15.

Church, Kenneth W., and Patrick Hanks. 1990. “Word Association Norms, Mutual Information, and Lexicography.” *Computational Linguistics* 16(1): 22–29.

Clarke, Daoud. 2012. “A Context-Theoretic Framework for Compositionality in Distributional Semantics.” *Computational Linguistics* 38(1): 41-71.

Clarke, K.R. 1993. “Non-parametric Multivariate Analyses of Change in Community Structure.” *Australian Journal of Ecology* 18: 117–143.

Condamines, Anne, and Josette Rebeyrolle. 1997. “Point de vue en langue spécialisée.” *Meta* 42(1): 174–184.

Cox, Trevor F., and Michael A.A. Cox. 2001. *Multidimensional Scaling*. Boca Raton: FL. Chapman & Hall.

Dunning, Ted. 1993. “Accurate Methods for the Statistics of Surprise and Coincidence.” *Computational Linguistics* 19(1): 61–74.

Eriksen, Lars. 2002. “Die Polysemie in der Allgemeinsprache und in der juristischen Fachsprache. Oder: Zur Terminologie der “Sache” im Deutschen.” *Hermes* 28: 211–222.

Evert, Stefan. 2007. *Corpora and Collocations*. Extended Manuscript of Chapter 58 of Lüdeling A. & M. Kytö, 2008, *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter. http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf. [accessed June 2014].

Evert, Stefan. 2012. “*The Role of Dimensionality Reduction in Distributional Semantics*.” Presentation at Leuven Statistics Days. Leuven, 8 June 2012.

Faber, Pamela (ed.). 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter.

Ferrari, Laura. 2002. “Un caso de polisemia en el discurso jurídico?” *Terminology* 8(2): 221–244.

Ferret, Olivier. 2010. “Similarité sémantique et extraction de synonymes à partir de corpus.” In *Actes de Traitement Automatique des Langues Naturelles (TALN 2010)*. Montréal, Canada.

Firth, John R. 1968. “A Synopsis of Linguistic Theory, 1930–1955.” In *Selected Papers of JR Firth, 1952–59,* ed. by John R. Firth, 168–205. Bloomington, Indiana University Press.

Gaudin, François. 2003. *Socioterminologie: une approche sociolinguistique de la terminologie.* Bruxelles: Duculot.

Geeraerts, Dirk. 2010. *Theories of Lexical Semantics.* Oxford: University Press.

Grefenstette, Gregory. 1994. “Corpus-derived First, Second and Third-order Word Affinities.” In *Proceedings of Euralex 1994. International Congress on Lexicography*. 279–290. Amsterdam, the Netherlands.

Habert, Benoît, Gabriel Illouz, and Helka Folch. 2005. “Des décalages de distribution aux divergences d’acception.” In *Sémantique et corpus.* ed. by Anne Condamines, 277–314. Paris: Hermes-Science.

Harris, Zellig. 1968. *Mathematical Structures of Language*. New York: Wiley.

Heylen, Kris, Dirk Speelman, and Dirk Geeraerts. 2012. “Looking at Word Meaning. An Interactive Visualization of Semantic Vector Spaces for Dutch Synsets.” In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2012).* 16–24. Avignon, France.

Kruskal, Joseph B., and Myron Wish. 1978. *Multidimensional Scaling. Sage University Paper series on Quantitative Applications in the Social Sciences*, number 07-011. Newbury Park, CA: Sage Publications.

Landauer, Thomas K., and Susan T. Dumais. 1997. “A Solution to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Introduction and Representation of Knowledge.” *Psychological Review* 104(2): 211-240.

Lemaire, Benoît, and Guy Denhière. 2006. “Effects of High-Order Co-occurrences on Word Semantic Similarity.” *Current Psychology Letters* 18(1). http://cpl.revues.org/index471.html. [accessed June 2014].

Morardo, Mikaël, and Eric Villemonte de La Clergerie. 2013. “Vers un environnement de production et de validation de ressources lexicales sémantiques.” In *Actes de Traitement Automatique des Langues Naturelles (TALN 2013) Atelier Sémantique Distributionnelle (SemDis).* 167–180. Sables d’Olonne, France.

Morlane-Hondère, François. 2013. “Utiliser une base distributionnelle pour filtrer un dictionnaire de synonymes.” In *Actes de Traitement Automatique des Langues Naturelles (TALN 2013) Atelier Sémantique Distributionnelle (SemDis).* 112–125. Sables d’Olonne, France.

Nazar, Rogelio, Jorge Vivaldi, and Leo Wanner. 2012. “Automatic Taxonomy Extraction for Specialized Domains Using Distributional Semantics.” *Terminology* 18(2): 188–225.

Padó, Sebastian, and Mirella Lapata. 2007. “Dependency-based Construction of Semantic Space Models.” *Computational Linguistics* 33(2): 161–199.

Peirsman, Yves, and Dirk Geeraerts. 2009. “Predicting Strong Associations on the Basis of Corpus Data.” In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2009).* 648-656. Athens, Greece.

Sahlgren, Magnus. 2006. *The Word-Space Model*. PhD thesis, Stockholm University, Sweden.

Sahlgren, Magnus. 2008. “The Distributional Hypothesis.” *Rivista di Linguistica* 20(1): 33–53.

Schütze, Hinrich. 1998. “Automatic Word Sense Discrimination.” *Computational Linguistics* 24(1): 97–123.

Temmerman, Rita. 2000. *Towards New Ways of Terminology Description. The Sociocognitive Approach.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Turney, Peter D., and Patrick Pantel. 2010. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research* 37: 141–188.

van der Laan, Mark J., and Katherine S. Pollard. 2003. “A New Algorithm for Hybrid Hierarchical Clustering with Visualization and the Bootstrap.” *Journal of Statistical Planning and Inference* 117: 275–303.

Venables, William N., and Brian D. Ripley. 2002. *Modern Applied Statistics with S.* New York: Springer-Verlag.

Wielfaert, Thomas, Kris Heylen, and Dirk Speelman. 2013. “Interactive Visualizations of Semantic Vector Spaces for Lexicological Analysis.” In *Actes de Traitement Automatique des Langues Naturelles (TALN 2013) Atelier Sémantique Distributionnelle (SemDis).* 154–166. Sables d’Olonne, France.

Wüster, Eugen. 1931. *Internationale Sprachnormung in der Technik: besonders in der Elektrotechnik.* Berlin: VDI-Verlag.

Wüster, Eugen. 1991. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie. 3. Aufl.* Bonn: Romanistischer Verlag.

Authors’ addresses

Ann Bertels
Leuven Language Institute
& QLVL – Faculty of Arts
University of Leuven
Dekenstraat 6 – b 5302
B-3000 Leuven (Belgium)
ann.bertels@ilt.kuleuven.be

Dirk Speelman
QLVL – Faculty of Arts
University of Leuven
Blijde-Inkomststraat 21 – b 3308
B-3000 Leuven (Belgium)
dirk.speelman@arts.kuleuven.be

About the authors

Ann Bertels is Assistant Professor (tenure-track) at the KU Leuven (Belgium), where she teaches French for Specific Purposes at the Leuven Language Institute (ILT). Her teaching and research interests include quantitative semantics, quantitative lexicology, terminology and language for specific purposes, specialised corpora, pedagogical lexicography and business communication.

Dirk Speelman is Associate Professor at the Department of Linguistics at the KU Leuven. He teaches courses on information science, statistics for the humanities, linguistic methodology (especially corpus linguistics) and usage-based model of language. His main research interest lies in the fields of corpus linguistics, computational lexicology and variationist linguistics.