Structured Parallel Coordinates: a visualization for analyzing structured language data

Chris Culy
Verena Lyding
Henrik Dittmann

European Academy Bolzano-Bozen, ITALY
{christoper.culy,verena.lyding,henrik.dittmann}@eurac.edu

We present a visualization tool called Structured Parallel Coordinates (SPC), a specialization of Parallel Coordinates (cf., e.g., Inselberg, 2009), customized for the presentation and analysis of different types of structured language data, as found in corpora. Parallel Coordinates are a way of representing multidimensional data using a two-dimensional display. Interactive versions of Parallel Coordinates are flexible tools for data analysis, since selecting parts of the visualization allows for filtering the data (Inselberg, 2009). Language datasets often have dimensions which are interrelated or which have internal structure, a situation that is not accounted for by standard Parallel Coordinates. We describe the visual features and interactions provided by SPC to account for structured language data and give technical details of the tool. We present three sample applications of SPC that are closely linked to principle tasks in corpus analysis: (1) KWIC results as SPC, (2) ngrams and frequencies, and (3) ranking comparisons.
keywords: visualization, Parallel Coordinates, tools for analysis

Se presenta una herramienta de visualización llamada Structured Parallel Coordinates (SPC), una especialización de Parallel Coordinates (por ejemplo, Inselberg, 2009), adaptado para la presentación y análisis de diferentes tipos de datos de lenguaje estructurados. Parallel Coordinates son una forma de representar datos multidimensionales mediante una pantalla de dos dimensiones. Versiones interactivas de Parallel Coordinates son herramientas flexibles para analizar los datos, ya que seleccionando unas partes de la visualización se permite filtrar de los datos (Inselberg, 2009). Conjuntos de datos del lenguaje suelen tener dimensiones que están relacionadas entre sí o que tienen una estructura interna. Las características visuales y los interacciones preparadas por SPC son presentadas para explicar los datos y proporcionar los detalles técnicos. Se exponen tres ejemplos de SPC que están vinculados a las tareas principales en el análisis de corpus: (1) los resultados de KWIC como SPC, (2) ngrams y frecuencias, y (3) comparaciones de escala.
keywords: visualización, Parallel Coordinates, herramienta para el análisis

1. Introduction

Visualizations are a powerful means to support the processing, analysis and understanding of information by humans. The field of information visualization (InfoVis) is concerned with elaborating and evaluating possible ways to display different types of information and to create effective visualizations for it. Language data and any kind of linguistic information derived from it is often quite different from the types of information that InfoVis research has commonly focused on, like e.g. statistical and geospatial data. While InfoVis in general has matured, the specific concern with Linguistic Information Visualization (LInfoVis) is only recently getting more attention (see (Culy & Lyding, 2010a) and (Rohrdantz, Koch, Jochim, Heyer, Scheuermann, Ertl, *et al.*, 2010) for some examples), and applications targeted to language data are still scarce and not always linguistically informed (see (Wattenberg & Viégas, 2008) and (Hassan-Montero & Herrero-Solana, 2006) for some good general text-based examples).

In this paper we present *Structured Parallel Coordinates (SPC),* a visualization tool for the presentation and analysis of different types of structured language data, as found in corpora. It is targeted to use by language analysts ranging from linguists to language teachers and learners. We describe the visual features and interactions provided by the tool, and explain how they respond to requirements of the prospective users and the characteristics of the data we are dealing with. We demonstrate, based on three elaborated sample applications, how *SPC* can be customized for different tasks.

2. Related Work

Structured Parallel Coordinates are a specialization of the *Parallel Coordinates* visualization (cf. d'Ocagne (1885), Inselberg (2009)). *Parallel Coordinates* are a way of representing multidimensional data using a two-dimensional display. Each dimension is represented along a vertical axis, and the values for a piece of data are connected by a line (see Figure 1). Interactive versions of *Parallel Coordinates* are flexible tools for data analysis, since selecting points and lines in the *Parallel Coordinates* display is the same as filtering the data (Inselberg, 2009). They are typically used with data dimensions that are conceptually independent, such as car size, year of manufacture, and mileage. *Parallel Coordinates* have been applied in many different contexts (e.g. (Inselberg, 2009) or (Steed, Fitzpatrick, Swan, & Jankun-Kelly, 2009)), with few, if any, detailed applications to language. *Parallel Tag Clouds (PTC)* (Collins, Viégas, & Wattenberg, 2009; Lee, Henry Riche, Karlson, & Carpendale, 2010) is similar to *SPC* visually, but it uses the size of words to indicate their frequency, and is not a true *Parallel Coordinates* visualization, since multiple dimensions cannot be selected. We have also implemented *PTC* as an application of *SPC.*

3. Structured Parallel Coordinates

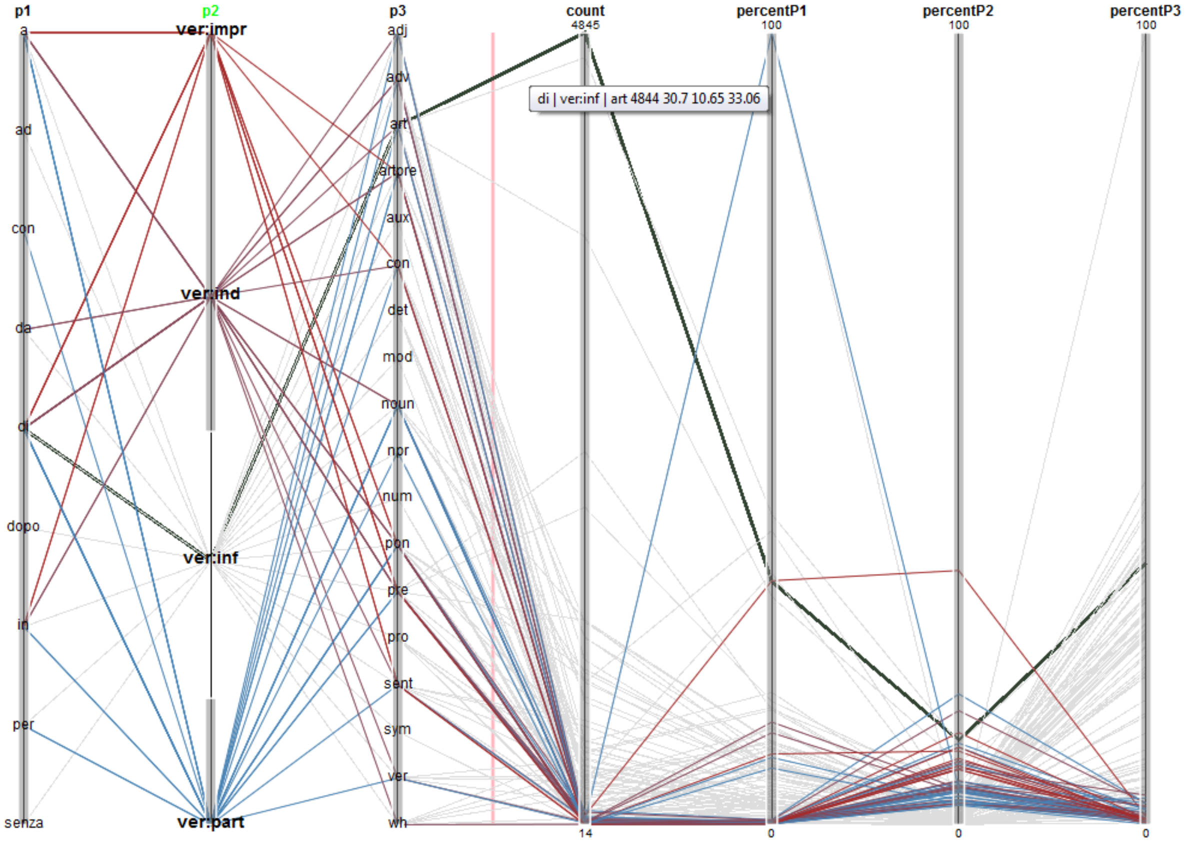
3.1. The corpus analysis context

With *Structured Parallel Coordinates* we aim at providing a tool that extends the concept and functionality of the standard *Parallel Coordinates* visualization to support the processing and analysis of language data derived from corpora. One particular problem in corpus linguistics is how to explore large result sets efficiently. A query might return thousands of examples from a corpus of millions of words and associated information (cf. the Corpus of Contemporary American English (COCA) (Davies, 2008), which has over 410 million words, or the recent freely available PAISÀ corpus of Italian (2011) with 500 million words). Our *Structured Parallel Coordinates* is a promising approach to visualizing corpus query results by providing the capability to explore and refine a set of query results without having to go back to the original data and redo a possibly complex query, in the same kind of spirit as, for example, Word Trees (Wattenberg & Viégas, 2008) or its linguistically specialized counterpart Double Tree (Culy & Lyding, 2010b). Other formats for corpus search results, such as KeyWord In Context (KWIC) lines and word lists, do not provide the flexibility and depth of information that we can provide by using *SPC*.

At the same time, in analyzing corpus data we deal with information that has dimensions that are not necessarily conceptually independent, but can be interrelated or have internal structure, unlike the typical uses of *Parallel Coordinates*. One fundamental type of structure is the sequential order of linguistic units like words, phrases, or paragraphs, plus statistical information associated with it. Another type of structure comes from meta-information associated with corpus texts, e.g. dates, where the data for each point in time can be treated as a dimension, and these dimensions are ordered (chronologically) with respect to each other. Rank orderings of (co-)occurrences of linguistic units provide an example of dimensions that have an internal structure: the ranks. *SPC* is designed to specifically account for the special nature of structured language data such as these, unlike general *Parallel Coordinates* visualizations which are not tailored to either language data or data that is structured across dimensions

3.2. Visual features and interactions in SPC

As with other *Parallel Coordinates* implementations, *SPCs* place data of different dimensions on vertical axes (one axis for each dimension) that are lined up horizontally. The axes may represent purely textual data (e.g. words), or purely numeric data (e.g. frequencies), and a single instance of *SPC* can contain axes of both types. Figure 1 shows ngrams of ‘preposition’ + ‘verb’ + ‘any word class’ plus their counts, extracted from a subset of the Italian corpus PAISÀ. On the first axis words are displayed, the second and third axes show word classes, and on axis four, counts are given on a numerical scale.[1] The different data dimensions may have an order among each other, as is the case for sequences of words in KWIC data, with one dimension/axis for each word position in a KWIC. They may also be without order, or ordered and unordered dimensions may occur together, as is in fact the case in Figure 1. A light red line placed vertically between ordered and unordered axes visually indicates the separation of these differently interrelated dimensions; here it visually separates the ngrams from their counts.



di | ver:inf | art 4844 30.7 10.65 33.06

Figure 1. Ngrams with counts for sequences of: preposition + verb + any word

The colors of the lines shade from red to blue according to the order of the data on a determined axis (initially the left-most one). The specified axis, axis 2 in Figure 1, is indicated by a green label, and can be changed by clicking on the label of any other axis.

Data in SPC can be filtered by selecting data points (by dragging the mouse over it) on any axis. The connecting lines of data points which do not satisfy the resulting filters are rendered in gray.

[1] The remaining axes will be discussed in section 3.3.2. for Figure 3.

By looking at the connecting lines, in Figure 1 we can easily see that infinitive verbs (*ver:inf*) have the widest combinatory variety of all verbs. To inspect this situation more closely data is filtered by all verbs that are not infinitives. Looking at the *count* axis, we can see that ngrams containing infinitive verbs do not just have the greatest variety, but make up all ngrams with a frequency higher than about 100 occurrences. In Figure 1, details on the most frequent ngram are given in a pop-up. Here this is “di” + “*ver:inf*” + “*art*” (*article*) with a total of 4844 occurrences.

3.3. Three applications of SPC

3.3.1. SPC for KWIC analysis

Working with KWIC results that show search units in context is perhaps *the* principle approach to corpus-based linguistic research. As corpus searches often yield large numbers of results, to sort, filter and generally get an idea of the nature of the results is a necessary step of almost any KWIC based analysis. The strength of *SPC for KWIC analysis* is its compact representation format (for each position only one occurrence of every word type is shown), and the possibility to dynamically filter the data by selecting items.

Figure 2 shows an *SPC* for the KWIC results for the query for the lemma *vedere* (‘to see’) in a small corpus of Italian press releases (about 120.000 tokens), with two words of context to the left and right. The KWIC results are displayed with each axis representing a word position. Axes are ordered according to the sequential order of words in the KWIC. The words are displayed on each axis and presented in alphabetical order. KWIC sequences are represented by lines connecting the words. The results are filtered by position of the keyword, restricting the hits to future forms of the verb. The visualization shows that future forms make up the biggest part of the results. Having the results filtered by these forms, we can detect a particularity of the words preceding, directly or with a distance of 2, the future forms of *vedere*: they often represent events like here *incontro* (‘meeting’) and *conferenza* (‘conference’). Hovering over the topmost line of the filtered data marks the line in bold and shows the sequence of data points (here “*L’ incontro vedrà la partecipazione*” – ‘The meeting will see the participation’) in a pop-up box.

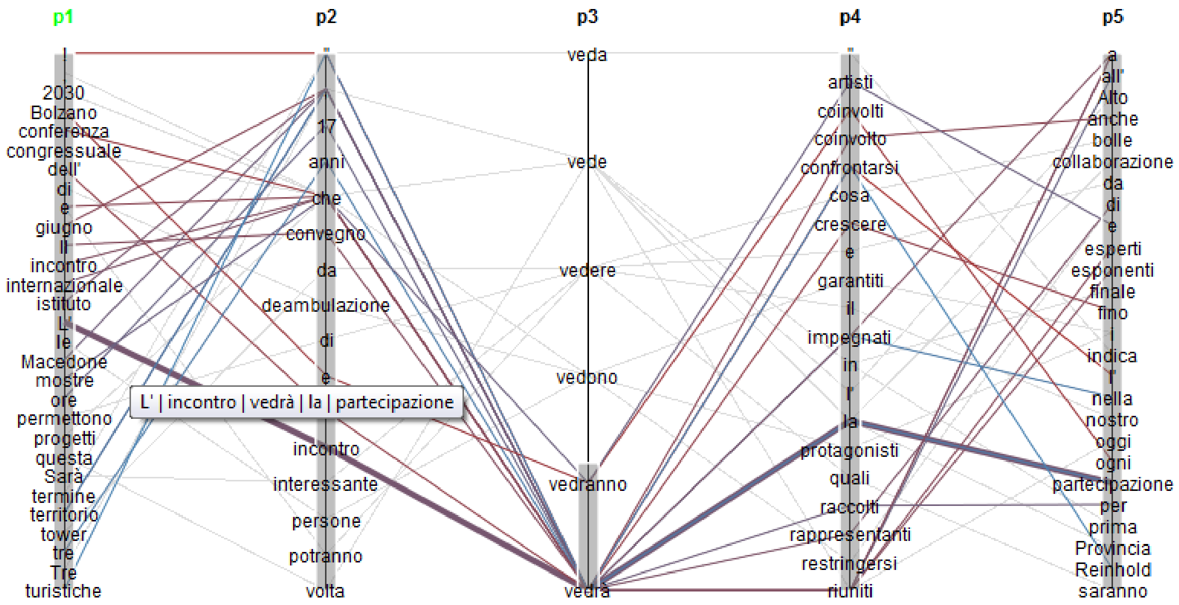


Figure 2. KWIC as SPC for the lemma *vedere* 'to see' in corpus of press releases

3.3.2. SPC for ngrams plus frequencies

Besides the question of which words (or other units of text) occur together, statistical information about their co-occurrence is of central relevance. *SPC for ngrams plus frequencies* is a visualization that combines these two types of information. Ordered data dimensions (the words of the ngrams) are presented together with unordered dimensions (statistical information related). In Figure 3 the first three dimensions show textual data—the results of a corpus query for pronouns followed by a form of the verb "to be" followed by "happy" or "sad" —, presented on the axes, in their original reading order, while the other four dimensions contain absolute and relative frequency information related to the ngrams, where *count* indicates the total number of the ngrams and *percentP1* indicating the percentage of a word in position *p1* occurring in the particular ngram compared to all occurrences of this word. In Figure 3, for example, "she" occurs in the ngram "she's happy" just 2 times and for 4.44% of all occurrences of "she" in the results set. On the axes with numerical data the upper values are indicated on top of the axes, with it being 165 for the absolute count of ngrams and 100 for the axes indicating percentages. In Figure 3, on axis *p1* "he" and "she" are selected, on *p2* "'s", and on *p3* "happy". The resulting visualization shows that "she's happy" is a lower percentage of all the uses of "she" (in this result set) than "he's happy" is of the uses of "he", with about 20%. Interestingly, the situation is inverse for ngrams for the unabbreviated form of "is". In fact "she is happy" makes up for 37.77% of all occurrences of "she", while "he is happy" makes up for only 28.91% of the total occurrences of “he”. The data is taken from a 100 million word Corpus of British English compiled from the web (ukWaC, (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008).

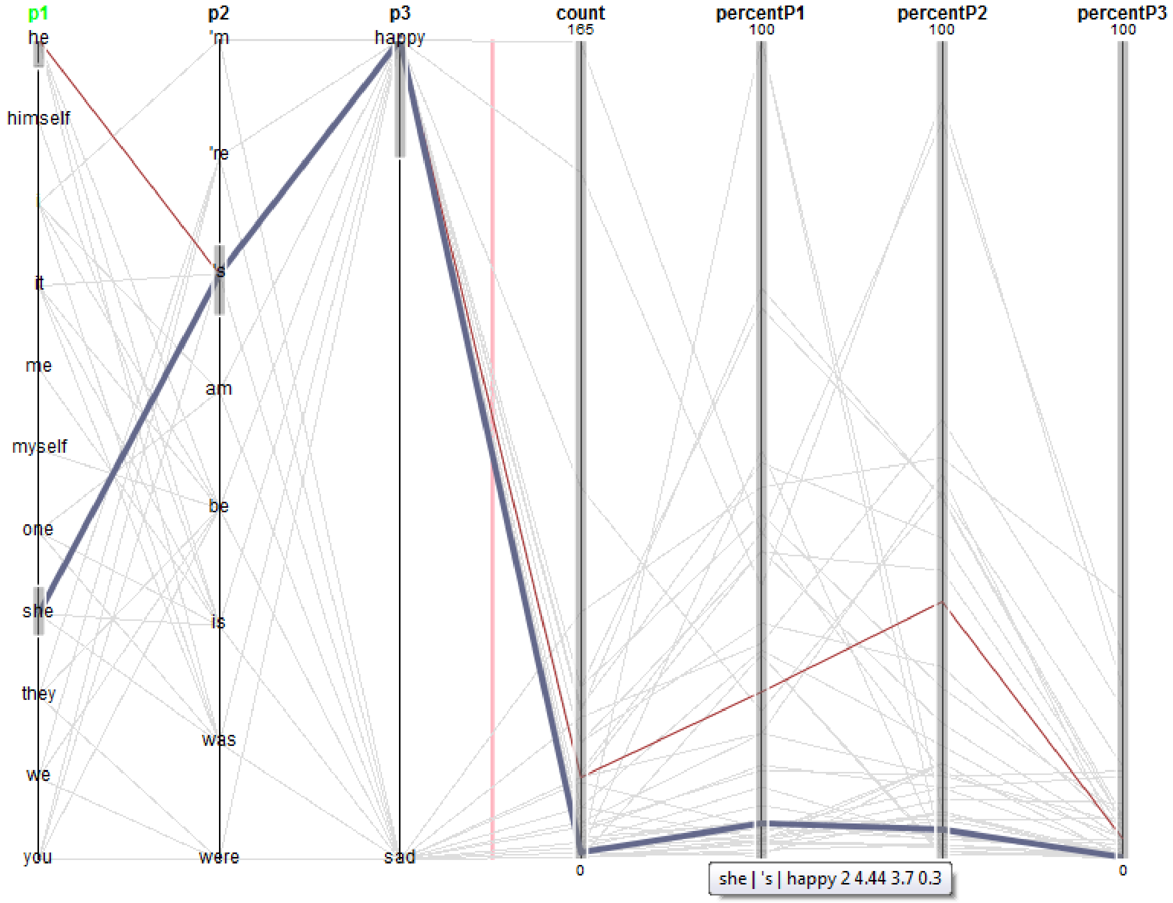


Figure 3. SPC for ngrams plus frequencies of *preposition* followed by lemma "to be" followed by "happy/sad"

3.3.3 SPC for ranking comparisons

The comparison of occurrences of linguistic phenomena is what *SPC for ranking comparisons* is designed for. The comparison could be of the same phenomenon across different corpora or subcorpora, or it could be of different aspects of a single phenomenon. Each data axis contains the linguistic units under inspection ordered by frequency. Figure 4 shows a visualization of the top 20 most frequent words starting with *[Ss]elbst* (‘self’), counted by lemma, in 5 years of newspaper text, ranging from 1991 to 2006. The data is taken from a corpus of the German language South Tyrolean newspaper *Dolomiten*. In contrast to the KWIC and ngrams visualizations presented above, the axes are not ordered by the linear order of word sequences but with respect to the ordering of corpus metadata, the year of the originating text. Words are placed on the axes according to their rank ordering. “[NA]” indicates that a word that occurs in the top *n* words in another year is not among the top *n* in the year for that axis. In the case of multiple words with the same frequency, the words are given the same rank and the list of words is associated with the data point (rank). Since the list could be quite long (e.g. for hapax legomena), only one word is shown on the axis, and the others are shown when the user hovers the mouse over the word shown. E.g. in Figure 4 on the axis with data from 2001, *Selbstbestimmung* ('self-determination') shares its rank with another word (here *Selbstverständlichkeit* ('implicitness')), as is indicted by three dots following the term. This way of presenting words with the same frequency was decided upon in response to initial user feedback.

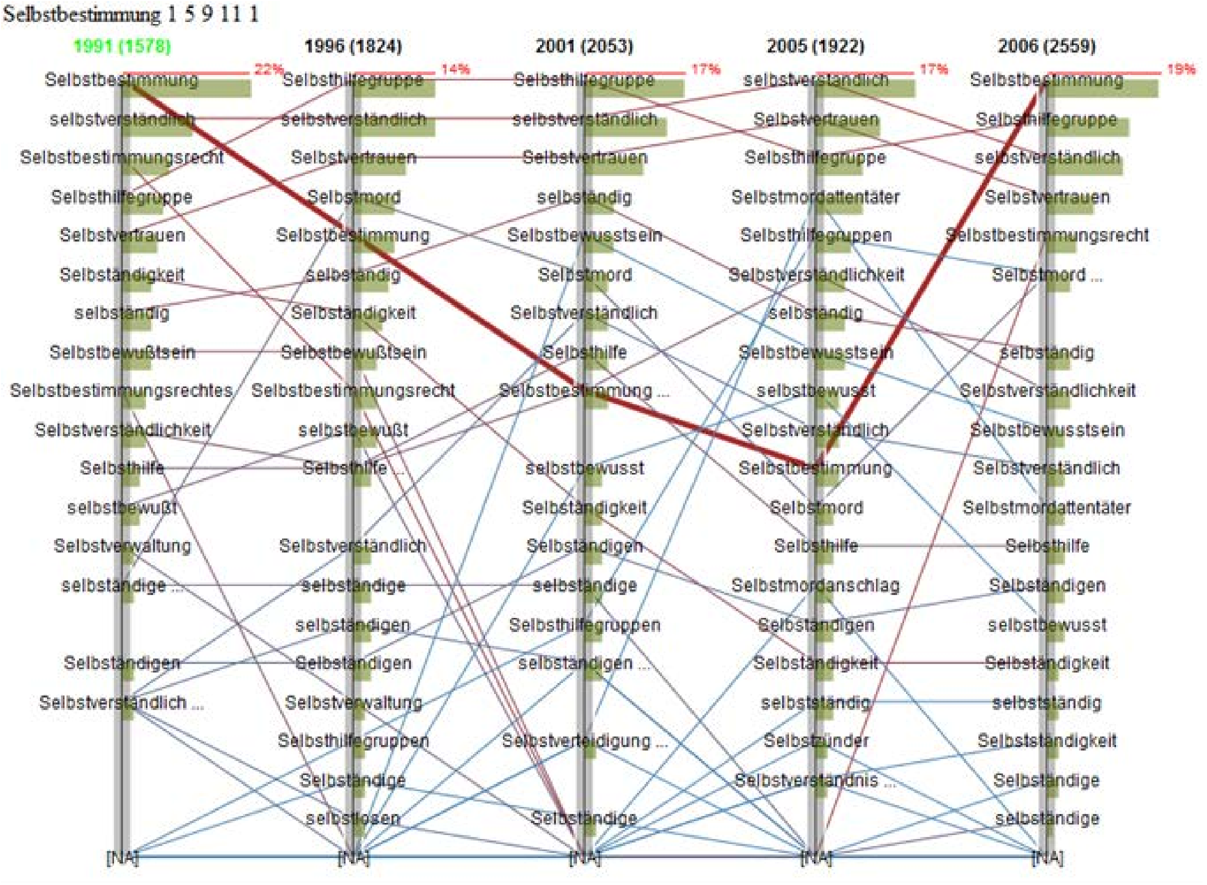


Figure 4. *SPC for rank orderings* of the top 20 words starting with "[sS]elbst" in 5 years of the Dolomiten

The ordering within the dimensions causes the positive or negative declination of the lines to have a meaning. Lines sloping upwards from left to right indicate an increase in rank ordering (though not necessarily of frequency, either absolute or relative to the series). Similarly, lines sloping downwards from left to right indicate a decrease in rank ordering. In Figure 4 the connecting line to the term "Selbstbestimmung" ('*self-determination*') is highlighted. It shows that from 1991 to 2005 it was continuously decreasing in rank, and from 2005 to 2006 got back to its original (first) ranking position. For the easy comparison of number of occurrences relative to total number of occurrences within each year, we added small horizontal bars to each data point on each axis to indicate percentages of occurrences. The visualization allows for a quick perception of the relative frequencies of items within and across dimensions. For the analysis in Figure 4 we can see that from 2001 to 2005 “Selbstbestimmung” decreased in rank, while its proportion in the results set did not change that much. To the contrary, the term holds the first rank both in 1991 and 2006, but in 2006 is only makes up 19% of the hits as opposed to 22% for 1991.

3.4. Technical details

SPC is designed to provide a core set of general visualization and interaction functionalities that can easily be customized and implemented for specific applications. It is implemented in JavaScript using the toolkit Protovis (Bostock & Heer, 2009) and runs in a browser. Functions and properties both on visual and interaction aspects of the visualization, as well as the fundamental aspects of the data ordering, can be easily customized for different applications of *SPC* as we have seen.

4. CONCLUSION

To sum up, *SPC* is a new type of *Parallel Coordinates* that is designed specifically for language data. It is a general tool that can be used to analyze different kinds of data with the current applications. For example, some of our colleagues are using *SPC* to analyze learner texts. *SPC* can also be extended to provide additional kinds of functionality, as we showed above in implementing *PTC* as an *SPC* application. *SPC* is an innovative tool for corpus analysis, which illustrates opportunities that are created when visualization techniques are adapted to the special needs of language information. *SPC* and the applications are freely available under an Open Source license.

REFERENCES

Bostock, M., & Heer, J. (2009). Protovis: A Graphical Toolkit for Visualization. *IEEE Transactions on Visualization and Computer Graphics, 15*(6), 1121-1128.

Collins, C., Viégas, F. B., & Wattenberg, M. (2009). Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST).* (pp. 91-98).

Culy, C., & Lyding, V. (2010a). Visualizations for exploratory corpus and text analysis. In *Proceedings of the 2nd International Conference on Corpus Linguistics (CILC-10).* A Coruña, Spain. (pp. 257-268).

Culy, C., & Lyding, V. (2010b). Double Tree: An Advanced KWIC Visualization for Expert Users. In *Proceedings of the 14th International Conference on Information Visualization (IV).* (pp. 98-103).

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present. http://www.americancorpus.org.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the Fourth Web as Corpus Workshop.* (pp. 47-54).

Hassan-Montero, Y., & Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. In *Proceedings of InSciT2006*, Mérida.

Inselberg, A. (2009). *Parallel Coordinates: VISUAL Multidimensional Geometry and its Applications*. New York: Springer.

Lee, B., Henry Riche, N., Karlson, A. K., & Carpendale, S. (2010). Spark Clouds: Visualizing Trends in Tag Clouds. *IEEE Tranactions on Visualization and Computer Graphics*, *16*(6), 1182-1189.

d'Ocagne, M. (1885). *Coordonnées Parallèles et Axiales: Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèlles*. Paris: Gauthier-Villars.

Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati (PAISÀ), http://www.corpusitaliano.it, 2011-

Rohrdantz, C., Koch, S., Jochim, C., Heyer, G., Scheuermann, G., Ertl, T. *et al.* (2010). Visuelle Textanalyse. *Informatik-Spektrum, 33*(6), 601-611.

Steed, C. A., Fitzpatrick, P. J., Swan, J. E., & Jankun-Kelly, T. J. (2009). Tropical Cyclone Trend Analysis Using Enhanced Parallel Coordinates and Statistical Analytics. *Cartography and Geographic Information Science, 36*(3), 251-265. doi:10.1559/152304009788988314.

Wattenberg, M., & Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics, 14*(6), 1221-1228.