# Visualizations for exploratory corpus and text analysis

CHRIS CULY

VERENA LYDING

*Institute for Specialised Communication and Multilingualism, EURAC research*

*Abstract*

*In this paper we are concerned with how visualizations can support exploratory corpus and text analysis. We start by giving an overview of previous work in information visualization that is based on language data. We discuss how existing approaches differ from our approach . The core of the present article consists of a detailed presentation of five visualization components that we have designed for supporting the undirected exploratory analysis of text material. For each component we point out possible application contexts and motivations for the design choices and how they are related to established visualization principles. We conclude with a short discussion on future needs for the field of linguistic information visualization.*

*Keywords: corpus linguistics, visualization, exploratory analysis*

*Resumen*

*En este artículo nos ocupamos de cómo las visualizaciones pueden ayudar al análisis exploratorio del corpora y al análisis de texto. Comenzamos por dar una visión general de lo que se ha hecho previamente en visualización de información basada en datos del lenguaje. Discutimos cómo los enfoques existentes se diferencian del nuestro. El núcleo del articulo consiste en una presentación detallada de cinco componentes de visualización que hemos diseñado para ayudar al análisis exploratorio de datos de texto. Para cada componente señalamos posibles contextos donde pueden ser aplicados, las razones por las cuales tomamos las diferentes decisiones de diseño, y como se relacionan con los principios de visualización establecidos. Concluimos con una breve discusión de las necesidades futuras en el campo de visualización de información lingüística.*

*Palabras clave: lingüística de corpus, visualización, análisis exploratorio*

## 1. INTRODUCTION

The field of information visualization is concerned with "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" (Card et al., 1999), making use of the special capabilities for pattern recognition of the human visual system. The field of information visualization has been around for about 20 years, but it has recently started to mature. Major media outlets such as the New York Times regularly publish information visualizations, and basic tools to create visualizations easily are becoming increasingly available (e.g. Google Visualization API[1], Many Eyes[2]).

[1] http://code.google.com/apis/visualization/

[2] http://manyeyes.alphaworks.ibm.com/

At the same time, the multitude and extent of available corpora and text collections call for new methods to present and access this semi-structured data. Visualizations "use graphics to organize information, highlight important information, allow for visual comparisons, and reveal patterns, trends, and outliers in the data" (Hearst, 2009: ch. 10), and thus are particularly valuable for understanding the nature of a text collection in a broad way, without having a strong hypothesis in mind. These *exploratory* phases of text inspection can be considered a recurring part of most corpus-based studies (cf. e.g. Gilquin & Gries, 2009). While the directed search aspect of corpus and text analysis is well supported by common text analysis and query tools, there is currently little targeted support for exploratory search. Visualizations, especially interactive ones, are particularly suited for exploratory search. In fact, we see visualization as the future for exploratory linguistic analysis.

## 2. Previous work

Information visualization has mainly been concerned with numeric data. Until recently, language related visualizations had largely been concerned with presenting search results, especially for intelligence or business analysts (e.g. ThemeRiver in Havre et al., 2000). As well, visualizations of semantic information of document content, either at the lexical level (DocuBurst (Collins, 2007), Leximancer (Smith, 2000)), or at the document level (Rohrer et al., 1998) have been around for some time and have more recently been complemented by visualizations of thesauri (e.g. Visual Wordnet[3]).

Over the past few years, "clouds" have become a popular way to represent thematic or textual popularity, and have recently been included into corpus interfaces (cf. e.g. the beta interface to the DWDS corpus[4] and Monk[5]). Other word level visualizations include TileBars (Hearst, 1995), which presents document length and frequency for specific query terms together with their distribution across the text, and TextArc (Paley, 2002), an innovate alternative to standard concordances.

There have been some efforts to explore structure in texts. Arc Diagrams (Wattenberg, 2002) and work by Ruecker et al. (2008) are different visualization methods for representing patterns of repetition. Word Trees (Wattenberg & Viégas, 2008) are a technique for the visualization and interactive exploration of keyword in context lines as tree structures, and Phrase Nets (van Ham et al., 2009) visualize phrasal patterns.

[3] http://kylescholz.com/projects/wordnet/

[4] http://beta.dwds.de/

[5] http://www.monkproject.org/

Visualizations found in today's corpus query tools proper, are largely limited to dispersion plots showing the distribution of search terms over text (cf. WordSmith Tools, (Scott, 2004)), charts indicating frequency distributions of words over text types or over time, and networks for the display of co-occurrences. Additionally, corpus tools occasionally make use of color for highlighting, or size to indicate frequencies (in TAPoR[6]).

Summing up, a great part of the language related work is concerned with visualizing information derived from texts (e.g. automatically extracted key words) or visualizing entire documents. Much less consideration is given to visualizing linguistic features. Also, many of the language related visualizations lack a linguistic foundation. To give a concrete example, Word Trees provides an interesting and inspiring visualization of textual data, but does not make use of linguistic information, and thus lacks some of the options that it could provide (e.g. distinguishing words by their part of speech to treat homographs appropriately). Furthermore, visualizations for the linguistic analysis of textual data, and more specifically, visualizations targeted to the exploratory corpus/text analysis are extremely rare.

## 3. Visualizations in the near future

The trend in information visualization is to provide toolkits or components that are reusable in different contexts, rather than building visualizations that are application specific. While there have been recent calls for increasing the efforts to create visualization applications[7], we believe that the component approach is appropriate for linguistic visualizations, where we mainly find prototypical application examples, practically no toolkits and few components. In this section we present some of our visualization components, still under active development, to give an idea of the kinds of visualizations that are relevant to exploratory search.

When visualizing textual data, we have to decide what parts of the data are crucial to be displayed, how the data can be condensed, what abstractions are sensible and how different views of the data can be combined. Established visualization principles (cf. Card et al., 1999; Hearst, 2009) provide guidance on how best to create visualizations that meet the context-specific information aims. Thus, the starting point for constructing visualizations is understanding the task of the user. For each of our visualization components, we give the user's task and explain the general principles and techniques employed and how the component helps the user accomplish the task.

[6] http://portal.tapor.ca/

[7] E.g. Enrico Bertini: http://diuf.unifr.ch/people/bertinie/visuale/2009/06/im_sick_and_tired_so_many_libr.html

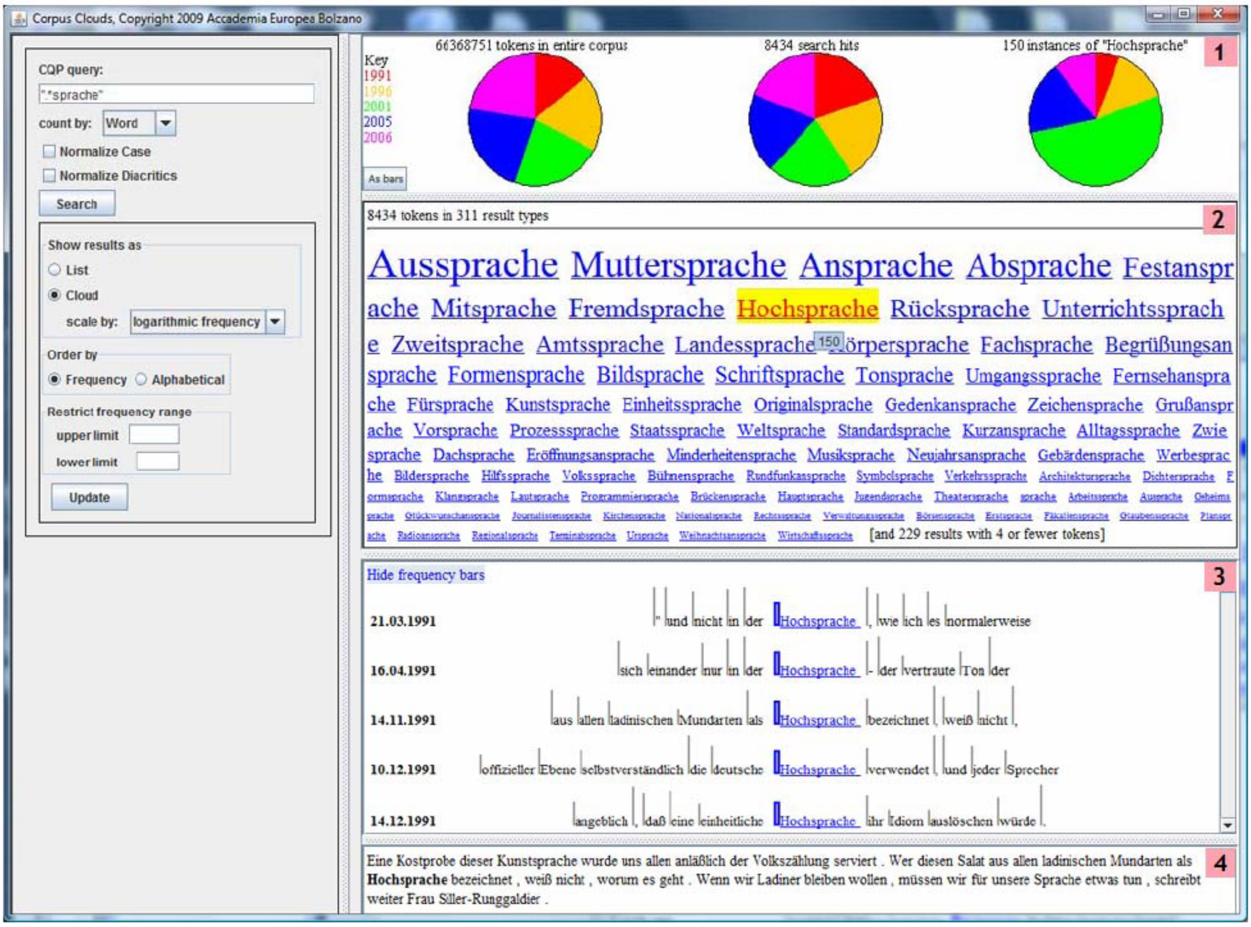
### *3.1. Corpus Clouds*

One aspect of exploratory corpus inquiry is getting an idea of what is frequent and how phenomena are distributed across the corpus. Corpus Clouds is a small program which provides visualizations of different types of frequency and distribution information for dynamic queries via a standard query system, integrated with a KWIC display. The overall design is inspired by Schneiderman's (1996) information visualization mantra overview first, zoom and filter, then details on demand, along with the idea of multiple views of the data, implemented in Corpus Clouds as four parallel views on query results [Fig. 1, from top]:

1. A distribution graph, showing the distribution of tokens and results over the corpus
2. A results pane displaying all strings that match the query
3. A KWIC display for a selected result type
4. A pane showing the extended context for one KWIC line

The different panels are coordinated by a technique called *brushing and linking*. Changes in one panel will update the other panels accordingly, for example selecting a specific result in the results pane, causes its distribution to be displayed as a graph in panel 1 and its concordance lines shown in the KWIC display panel. The KWIC display further provides for a view in which small bars indicate the frequency of each word in context, similar to the *sparklines* technique of Tufte (2006).

Clouds have been criticized in the human computer interface literature as not being good interfaces for web sites (Hearst & Rosner, 2008). However, the cloud view in Corpus Clouds is designed to meet the needs of corpus users, who are interested in frequency (of phrases as well as of words), by being interactive and flexible in its display order and scaling, as well as allowing for a simple list view instead of the cloud. By taking seriously the user's task (discovering information about frequency and distribution), we can repurpose a technique that is not optimal in other situations.


Corpus Clouds, Copyright 2009 Accademia Europea Bolzano
Key
1991
1996
2001
2005
2006
66368751 tokens in entire corpus
8434 search hits
150 instances of "Hochsprache"
As bars
1
2
3
4
Search
Update
Hide frequency bars
150

CQP query: ".*sprache"
count by: Word
☐ Normalize Case
☐ Normalize Diacritics

Show results as
○ List
● Cloud
scale by: logarithmic frequency

Order by
● Frequency ○ Alphabetical

Restrict frequency range
upper limit
lower limit

8434 tokens in 311 result types

Aussprache Muttersprache Ansprache Absprache Festansprache Mitsprache Fremdsprache Hochsprache Rücksprache Unterrichtssprache Zweitsprache Amtssprache Landessprache Körpersprache Fachsprache Begrüßungsansprache Formensprache Bildsprache Schriftsprache Tonsprache Umgangssprache Fernsehansprache Fürsprache Kunstsprache Einheitssprache Originalsprache Gedenkansprache Zeichensprache Grußansprache Vorsprache Prozesssprache Staatssprache Weltsprache Standardsprache Kurzansprache Alltagssprache Zwiesprache Dachsprache Eröffnungsansprache Minderheitensprache Musiksprache Neujahrsansprache Gebärdensprache Werbesprache Bildersprache Hilfssprache Volkssprache Bühnensprache Rundfunkansprache Symbolsprache Verkehrssprache Architektursprache Dichtersprache Formsprache Klangsprache Lautsprache Programmiersprache Brückensprache Hauptsprache Jugendsprache Theatersprache sprache Arbeitssprache Ausprache Geheimsprache Glückwunschansprache Journalistensprache Kirchensprache Nationalsprache Rechtssprache Verwaltungssprache Börsensprache Erstsprache Fäkaliensprache Glaubenssprache Plansprache Radioansprache Regionalsprache Terminabsprache Ursprache Weihnachtsansprache Wirtschaftssprache [and 229 results with 4 or fewer tokens]

21.03.1991 " und nicht in der Hochsprache , wie ich es normalerweise

16.04.1991 sich einander nur in der Hochsprache - der vertraute Ton der

14.11.1991 aus allen ladinischen Mundarten als Hochsprache bezeichnet , weiß nicht ,

10.12.1991 offizieller Ebene selbstverständlich die deutsche Hochsprache verwendet , und jeder Sprecher

14.12.1991 angeblich , daß eine einheitliche Hochsprache ihr Idiom auslöschen würde .

Eine Kostprobe dieser Kunstsprache wurde uns allen anläßlich der Volkszählung serviert . Wer diesen Salat aus allen ladinischen Mundarten als **Hochsprache** bezeichnet , weiß nicht , worum es geht . Wenn wir Ladiner bleiben wollen , müssen wir für unsere Sprache etwas tun , schreibt weiter Frau Siller-Runggaldier .

Figure 5: Corpus Clouds

### *3.2. Double Tree*

Concordances with their KWIC display predate computers by centuries. They help the user in the task of discovering the linguistic context of words. One problem with the KWIC display is that it is not easy to make sense of a large number of results for a single term – it is difficult to detect regularities and differences in contexts. A second problem with standard KWIC displays is that they can only be sorted by left OR by right context, making it difficult to detect examples with a context of interest on both sides of the target term, without doing a new query. Wattenberg and Viégas (2008) provide Word Trees to help with the first problem. Word Trees collapse identical left or right contexts into a single line, giving a branching tree structure when contexts diverge. Two problems with Word Trees for corpus linguists is that infrequent results are suppressed without any notice, and only one side of context is visible at a time.

To overcome the problems with standard KWIC displays and to make up for the one-sidedness of Word Trees, we have made a new visualization, Double Tree (see Fig. 2), a two sided word tree, which displays both left and right contexts. Initially a Double Tree shows one word of context on each side. For each context word, color shade indicates the number of distinct words that precede/follow that node, while the total number of instances of the word (in that specific context) is shown when the mouse is over the node. Selecting a context word expands the context by one level and dynamically colors all the paths on the opposite side for the results containing the context word.

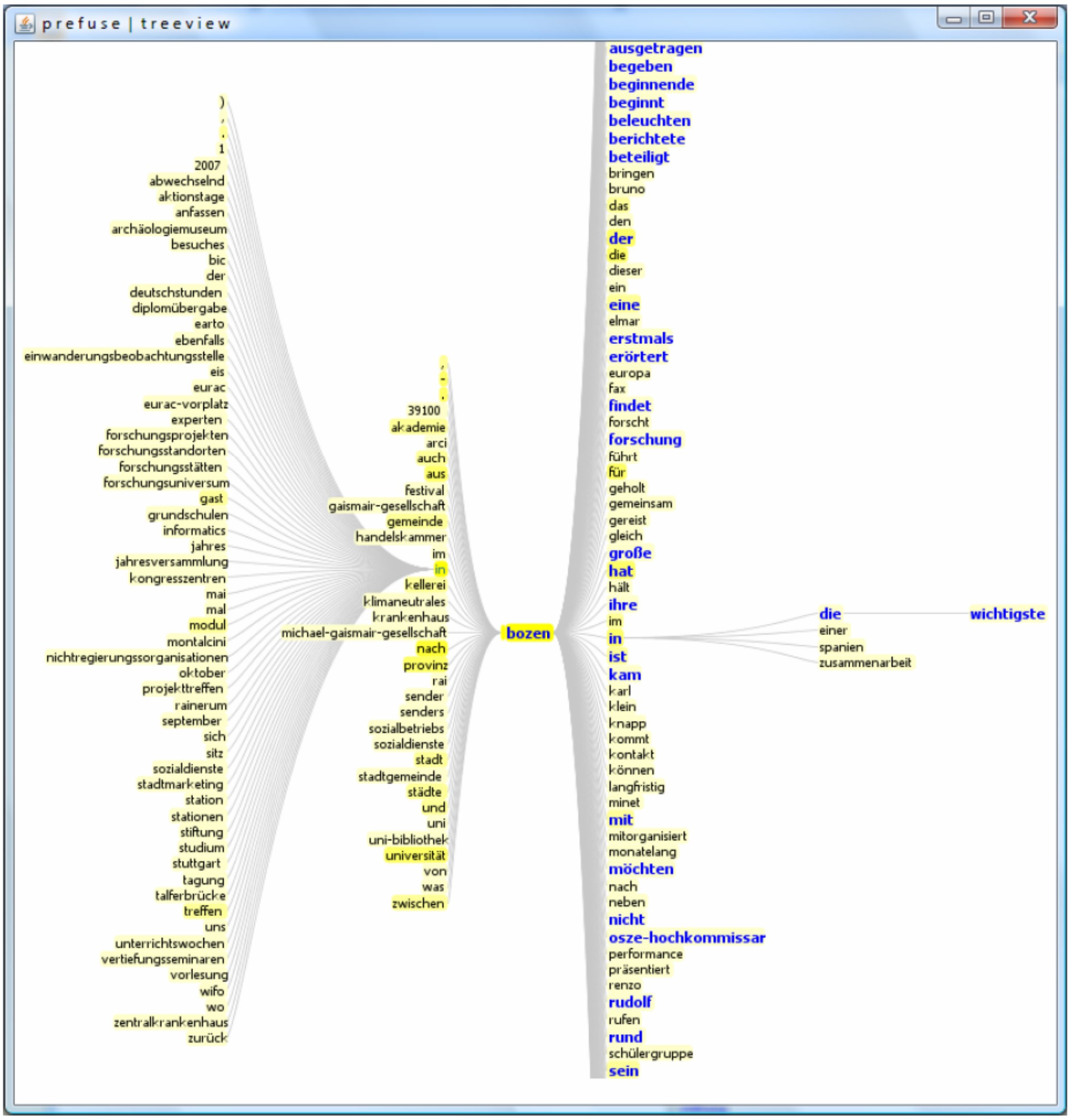
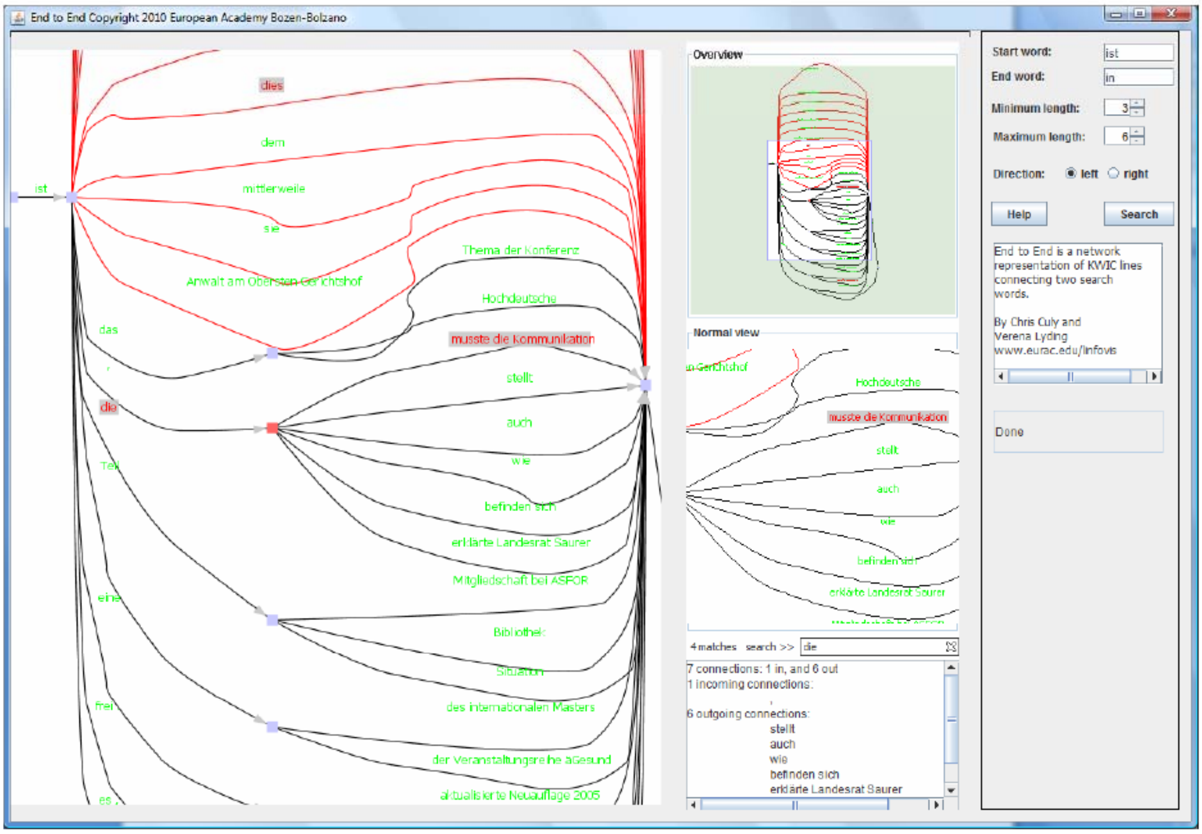


Figure 6: Double Tree

As with Word Trees, Double Trees implement a general goal of visualization to present more information in less space. Expanding only a given node is a limited example of a *fisheye view* (Furnas, 1981) according to the degree of interest. In addition, since the relatedness of left and right contexts is conveyed visually by assigning a unitary color to the words of one result phrase, Double Trees also implement Tufte's (2006: 70) principle of *sameness*.


End to End Copyright 2010 European Academy Bozen-Bolzano

Start word: ist

End word: in

Minimum length: 3

Maximum length: 6

Direction: ● left ○ right

End to End is a network representation of KWIC lines connecting two search words.

By Chris Culy and Verena Lyding www.eurac.edu/infovis

Done

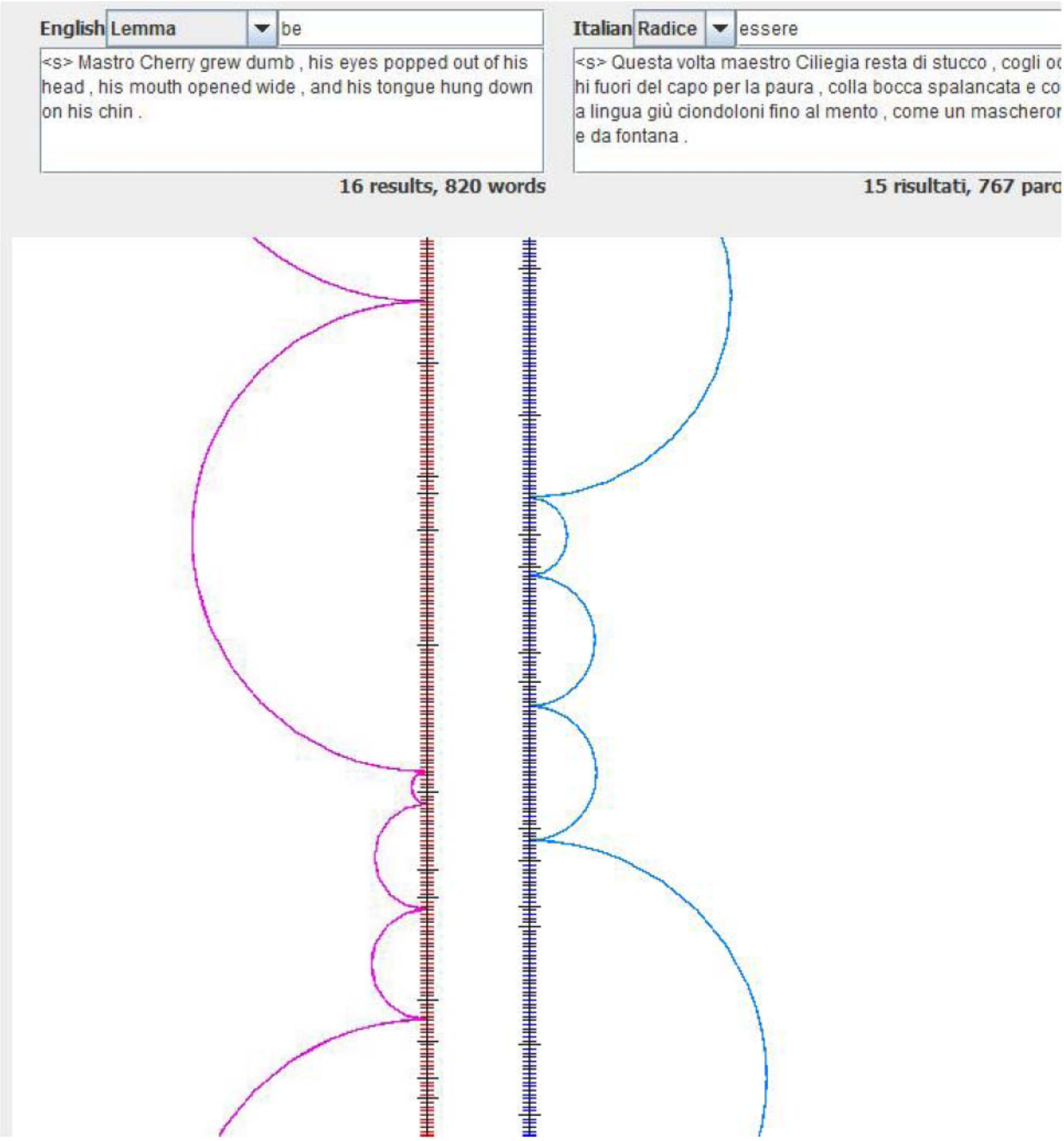
4 matches search >> die

7 connections: 1 in, and 6 out
1 incoming connections:
,
6 outgoing connections:
stellt
auch
wie
befinden sich
erklärte Landesrat Saurer

Figure 7: End-to-end for search "ist" "in"

### 3.3. End-to-end

Another type of exploratory search involves looking for variation and repetitions in connections between words, e.g. as collocations. End-to-end (see Fig. 3) is a component that creates a network of phrases connecting two search terms. In this way, the user can see at a glance all the connections between the terms. The user can also drill down by searching the network for particular words in between the search terms, or get a more detailed report of the context of a given word, i.e. providing details on demand. Other options allow the specification of the range of the lengths of the phrases, as well as optimizing the network from either a left to right perspective or a right to left perspective, thus providing multiple views of the data.



English Lemma be

<s> Mastro Cherry grew dumb , his eyes popped out of his head , his mouth opened wide , and his tongue hung down on his chin .

16 results, 820 words

Italian Radice essere

<s> Questa volta maestro Ciliegia resta di stucco , cogli oc
hi fuori del capo per la paura , colla bocca spalancata e co
a lingua giù ciondoloni fino al mento , come un mascheror
e da fontana .

15 risultati, 767 paro

Figure 8: Comparison Arcs

### *3.4. Comparison Arcs*

Another aspect of exploratory search is looking for positional patterns of occurrences. Wattenberg's (2002) Arc Diagrams provide a visualization of occurrence patterns by representing sequences of units as tokens on a line and connecting identical units with arcs. With Comparison Arcs (see Fig. 4), we expand the basic idea of Arc Diagrams in several ways. First, we allow the display of more than one sequence (in our case texts in the same or different languages) in parallel. Second, the visualization of Comparison Arcs is dynamic, visualizing the results of the user's search. Third, Comparison Arcs is user-selective, only showing the search results rather than showing all (automatically chosen) correspondences. Fourth, we incorporate linguistic information by allowing the user to search for lemmas and parts of speech in addition to words.

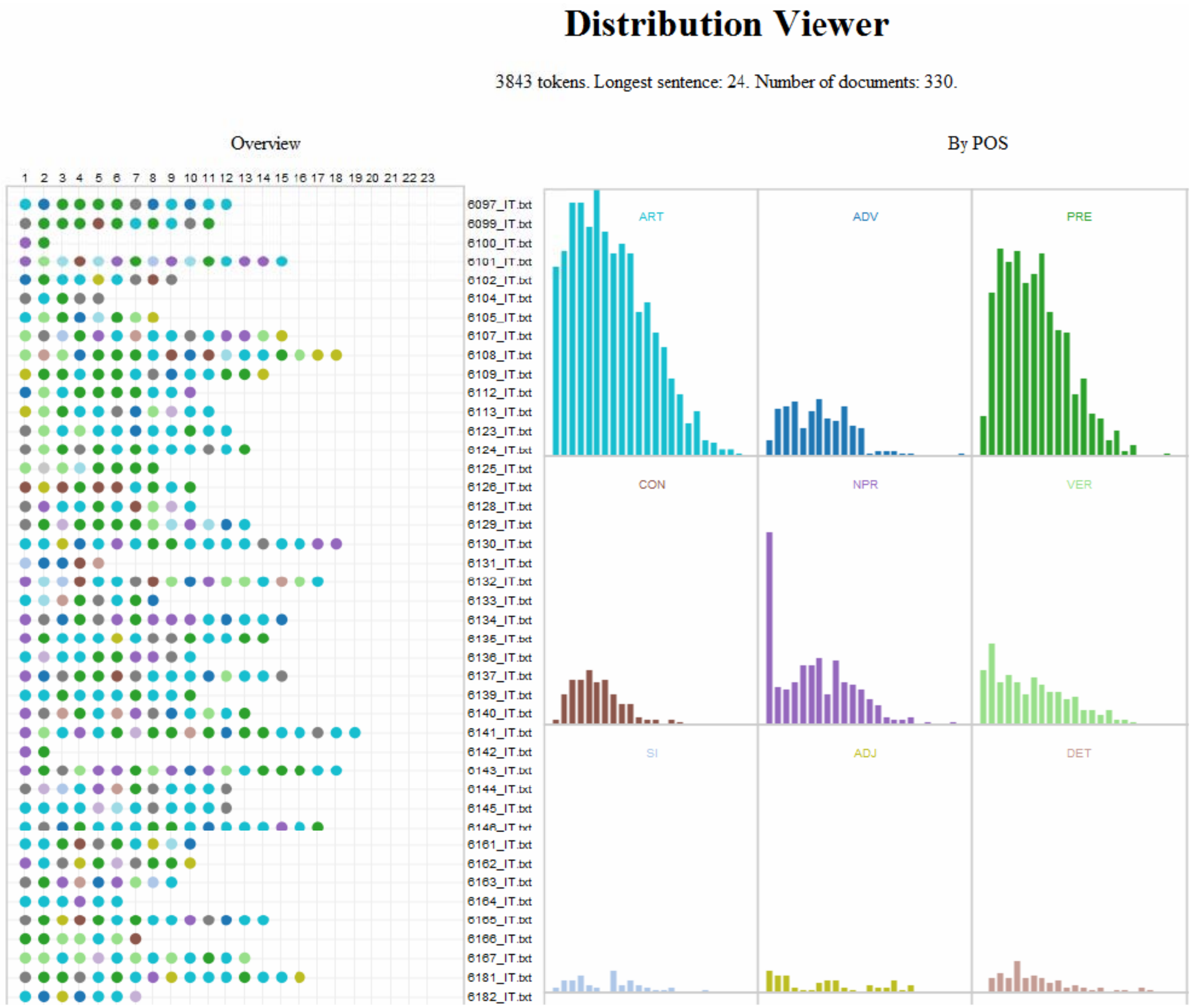
With respect to visualization principles, we give the user information on demand by providing information about tokens and sentence boundaries by moving the mouse over the diagram.

### *3.5. Distribution Viewer*

While Comparison Arcs visualizes co-occurrences of a particular type, Distributional Viewer (see Fig. 5) focuses on a different aspect of distribution by visualizing occurrences of a particular category. In our test example, we visualize the parts of speech of the initial words of each sentence in a small corpus. Notice that in contrast to Comparison Arcs, Distributional Viewer can handle corpora as well as individual texts.

We use two different visualization techniques, which together provide multiple views of the data. One visualization technique is essentially a *starfield* (Ahlberg & Shneiderman, 1994), where each part of speech is given a different color and plotted on a grid with sentence position on the horizontal access and text on the vertical axis. This allows the user to see broad patterns in the distribution.

To allow the user to follow up on initial observations, we provide two other views, one which shows for each part of speech its distribution across sentence document position, and the other which shows for each sentence document position, the distribution of parts of speech in that position. In both cases, we use Tufte's (1999) technique of *small multiples*, which shows separate bar graphs for each case (part of speech or sentence document position). As well, as with the other visualizations, appropriate additional information is provided by moving the mouse over the diagrams.



# Distribution Viewer

3843 tokens. Longest sentence: 24. Number of documents: 330.

Figure 9: Distribution Viewer

Both Comparison Arcs and Distribution Viewer emphasize the point that we can visualize not only words, but lemmas, parts of speech and more. Almost all language-related visualizations that we are aware of visualize only words.

## 4. Discussion

In this paper we have presented a collection of visualization components for the exploratory analysis of corpus and textual data. All these visualizations build on established visualization principles to best integrate textual data and connected linguistic information into concise displays.

To extend these initial efforts, we need more experience with linguistic information visualization in general, and the user's demands with respect to particular tasks. On the theoretical side, we need to determine what visualization techniques are applicable to language data. On the practical side, we need to evaluate what kinds of visualizations can benefit what users in what kinds of tasks, and what visualization alternatives are favored over others in specific usage contexts.  Furthermore, we need to find ways to guarantee efficient and flexible interoperability of all tools and components that aid the work of the language analyst especially since we see visualization tools becoming a central aspect of the next generations of corpus and text analysis tools.

## References

Ahlberg, C., & Shneiderman, B. (1994). Visual information seeking: tight coupling of dynamic query filters with starfield displays. In B. Adelson, S. Dumais, & J. Olson (Eds.). *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating interdependence*. CHI '94. ACM, New York. (pp. 313-317).

Card, S., Mackinlay, J., & Shneiderman, B. (Eds.). (1999). *Readings in Information Visualization: Using Vision to Think.*. San Diego: Academic Press.

Collins, C. (2007). Docuburst: Radial space-filling visualization of document content. Knowledge Media Design Institute, University of Toronto, Technical Report KMDI-TR-2007-1.

Furnas, G. (1981). The FISHEYE view: A new look at structured files. Bell Laboratories Technical Memorandum. No 81-11221-9.

Gilquin, G., & Gries, S. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*(1). (pp. 1-26).

Havre, S., Hetzler, B., & Nowell, L. (2000). ThemeRiver: Visualizing Theme Changes over Time. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*. INFOVIS. IEEE Computer Society, Washington, DC, 115.

Hearst, M. (1995). Tilebars: Visualization of term distribution information in full text information access. In *Proceedings CHI'95*, Denver, Colorado. (pp 56–66).

Hearst, M., & Rosner, D. (2008). Tag Clouds: Data Analysis Tool or Social Signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*. HICSS. IEEE Computer Society, Washington, DC, 160.

Hearst, M. (2009). *Search User Interfaces*. Cambridge: Cambridge University Press.

van Ham, F., Wattenberg, M., & Viégas, F. (2009). Mapping Text with Phrase Nets. *IEEE Transactions on Visualization and Computer Graphics* 15, 6. (pp. 1169-1176).

Paley, W. (2002). TextArc: Revealing Word Associations, Distributions and Frequency. Interactive Poster presented at the IEEE INFOVIS'02.

Rohrer, R., Sibert, J. & Ebert, D. (1998). The shape of Shakespeare: Visualizing text using implicit surfaces. In *Proceedings of the IEEE Symposium on Information Visualization,* Washington: IEEE Computer Society Press. (pp. 121–129).

Ruecker, S., Radzikowska, M., Michura, P., Fiorentino, C., & Clement, T. (2008). Visualizing Repetition in Text. CHWP. Retrieved from http://www.chass.utoronto.ca /epc/chwp/CHC2007/Ruecker_etal/Ruecker_etal.htm.

Scott, M. (2004). WordSmith Tools. Liverpool: Lexical Analysis Software.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, Washington: IEEE Computer Society Press. (pp. 336–343).

Smith, A. (2000). Machine Mapping of Document Collections – the Leximancer System. In *Proceedings of the Fifth Australasian Document Computing Symposium, DSTC, Sunshine Coast, Australia*.

Tufte, E. (2006). *Beautiful Evidence*. Cheshire, Connecticut: Graphics Press LLC.

Tufte, E. (1999). *Envisioning Information*. Cheshire, Connecticut: Graphics Press LLC.

Wattenberg, M. (2002). Arc diagrams: visualizing structure in strings. In *Proceedings of the IEEE Symposium on Information Visualization,* Washington: IEEE Computer Society Press. (pp. 110–116).

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