'interHist' - an interactive visual interface for corpus exploration

Verena Lyding, Lionel Nicolas, Egon Stemle

European Academy of Bolzano/Bozen
viale Druso 1, 39100 Bolzano, Italy
{verena.lyding;lionel.nicolas;egon.stemle}@eurac.edu

Abstract

In this article, we present *interHist*, a compact visualization for the interactive exploration of results to complex corpus queries. Integrated with a search interface to the PAISÀ corpus of Italian web texts, *interHist* aims at facilitating the exploration of large results sets to linguistic corpus searches. This objective is approached by providing an interactive visual overview of the data, which supports the user-steered navigation by means of interactive filtering. It allows to dynamically switch between an overview on the data and a detailed view on results in their immediate textual context, thus helping to detect and inspect relevant hits more efficiently. We provide background information on corpus linguistics and related work on visualizations for language and linguistic data. We introduce the architecture of *interHist*, by detailing the data structure it relies on, describing the visualization design and providing technical details of the implementation and its integration with the corpus querying environment. Finally, we illustrate its usage by presenting a use case for the analysis of the composition of Italian noun phrases.

Keywords: visualization, corpus linguistics, language analysis

1. Introduction

Apart from building the basis for natural language processing tools, language resources can also be a source of knowledge directly explored by researchers. For example, language corpora are employed for retrieving information (e.g. by historians, market researchers, etc.) or as reference for authentic language in use in linguistic studies.

On the one hand, human users have informational needs that essentially differ from requirements arising in the context of corpus-based NLP research. On the other hand, they hold special sensory capabilities for perceiving and processing information. Thus, the importance and potential of providing adapted interfaces for the exploration of language resources should not be underestimated.

Research on advanced interfaces for language resources is yet another challenging task in the NLP context. In delivering resources to the user, concern with the development of powerful interfaces is a service to the human analyst comparable to the provision of e.g. annotation tools.[1]

In this paper, we are presenting *interHist*, an example visualization tool for analyzing language data from a corpus. In section 2. we give background information and information on related works. In section 3. we present its architecture and functionalities, and illustrate its usage for the analysis of Italian noun phrases.

2. Background and related work

2.1. Linguistic corpus analysis

Corpus linguistics is concerned with the study of language based on empirical data collected in electronic corpora (cf. e.g. Biber et al. (1998), McEnery and Wilson (2005), Lüdeling and Kytö (2009)). It provides a specific use case of human users directly accessing and interacting with language corpora.

Comparable use cases with respect to user interaction with language resources, include exploration tasks on *electronic lexica*, *word nets*, and analyses in *translatology*.

Corpus linguistics provides a systematic methodology that builds on quantitative information for deriving insights from language in use by exploration, or testing of linguistic hypotheses. Typical corpus linguistics' analysis tasks include the search and extraction of authentic text samples for specified linguistic phenomena, their inspection and analysis in context, and their quantification.

Corpus analyses procedures are usually assisted by query and analysis tools, such as concordancers (cf. e.g. Scott (2010)), or custom scripts for the extraction and processing of corpus data. In that way, examples matching a user-devised query are extracted automatically, while subsequently results are processed manually (e.g. irrelevant hits are removed, matching hits are categorised into distinct groups, etc.).

Despite the automatization of the data extraction step, the wealth of data that a user is confronted with can become an issue for corpus-based linguistic studies. Often several thousands of examples are retrieved for one query[2], which makes it challenging to carry out corpus analyses efficiently, that is reaching a high coverage with reasonable efforts.

[1] While e.g. part-of-speech tagging enables advanced modes for searching corpus data, clear and varied presentation formats facilitate navigation and data understanding.

[2] For example, even in a medium sized corpus like PAISÀ (250 million tokens; see section 3.3.1. below for details), the search for the word 'casa' (house) returns 97 948 results, while searching for the lemma 'vedere' (to see) yields 223 226 examples, and the search for the lemma 'pensare' (to think) followed by a preposition still gives 27 619 results.

2.2. Visualizations as interfaces

Visualization, and more specifically *information visualization* is concerned with "[t]he use of computer-supported, interactive, visual representations of abstract data to amplify cognition." (Card et al., 1999).

Fundamentally, visualizations are translations of data into graphics. By building on the humans' special capabilities for processing visual information (cf. Ware (2004)) selected characteristics of the data are represented by means of visual cues, also termed *visual variables* (Bertin, 1983), such as color, size, shape, etc.[3]. This translation process is guided by visualization principles which help to create information rich displays of large amounts of (complex) data, in a way so as to reveal patterns, trends and outliers in the data (Hearst, 2009).

Visualizations are increasingly adopted as tools for data analysis. Besides their general adaptedness to the human perceptual system, their use for data analysis is further enforced by interactive features that are incorporated in many computer-based visualizations and that allow for direct engagement with the data.

As tools for data analysis, visualizations are also a powerful means in interface design for language resources, which provide large amounts of semi-structured data and information connected to it. Visualizations create an added value by unlocking information that is provided by NLP tools to the understanding and analysis by the human user.

2.3. Language and linguistic visualization

The visualization of language and linguistics data is a specialized subfield of *information visualization*. Concern with visualizing abstract information related to social, cultural, geographical phenomena and all types of media started as early as the 90ies. The rising interest in visualizations at that time has been related to rapid improvements in advanced computer graphics.

Since then, for several years language-related visualizations primarily have focussed on representing document collections (cf. e.g. Widdows and Dorow (2002), DeCamp and Roy (2005), IN-SPIRE™[4], etc.). They typically do not visualize the text proper, but show how documents cluster according to topic or genre, what are the most prominent keywords and how the keywords distribute. As part of content analysis tools, document visualizations have also attained commercial relevance. The more recent Word and Tag Clouds (cf. e.g. Kaser and Lemire (2007), Hearst and Rosner (2008), etc.) based on lemma lists derived from text collections constitute a basic type of document visualization.

Another type of language visualizations loosely related to text, are visualizations of word nets, thesauri and dictionary data (cf. e.g. Visual Thesaurus[5] and VisuWords[6]).

Language visualizations with focus on linguistic specificities and language in its immediate textual context have started gaining wider attention only recently. This can particularly be observed from rising numbers of conferences and workshops on this topic over the past years (cf. ESSLLI 2009, AVML 2012, LINGVIS workshop at EACL 2012, Workshop on the visualization of linguistic patterns at DGfS 2013, etc.) and the increasing availability of related publications.

On the one hand, several visualizations of linguistic features have been developed (cf. e.g. Mayer et al. (2010), Collins (2007), Rohrdantz et al. (2010)).

On the other hand, we find a rising number of visualizations of textual data. Major works related to concordance data include Word Trees (Wattenberg and Viégas, 2008), Double Tree (Culy and Lyding, 2010), WORDGRAPH (Trenkmann et al., 2012), etc., while e.g. PhraseNets[7], Web Trigrams[8] and ConcGrams (Cheng et al., 2006) suggest visualizations for n-gram data.

Specific to corpora, the Corpus Clouds interface (Culy and Lyding, 2009) combines different views on query results and frequency information connected to it. On a more abstract level, the generic corpus search applications ANNIS (Zeldes et al., 2009) collects and integrates visualization modules for multilayer linguistic corpora.

3. The interHist visualization

With *interHist* we propose a visualization that abstracts over concordance data from text corpora. It provides a visual extension to search interfaces for linguistic analysis.

3.1. Aim and target group

interHist aims at supporting the analysis of language corpora for linguistic research purposes.

As explained in section 2.1. above, dealing with large sets of results is one of the major challenges that modern corpus linguistics is faced with. On the one hand, a thorough analysis of the results requires a close examination of all examples within their textual contexts. On the other hand, looking at examples one by one can get very time consuming and is often not feasible.

Thus, concerning analysis tools, there is a major demand for support and facilitation of data navigation and analysis. With *interHist* we provide a compact and interactive visual overview that complements the classic concordance view on the data. By allowing the user to browse an abstraction of the results set, it enables her to narrow down the data before starting the one-by-one analysis of the results. The drawback of abstractions, that is the loss of context information, is made up for by directly linking from the visualization overview to individual Keyword In Context lines.

The *interHist* visualization is targeted at researchers, who carry out linguistic analyses on corpora in a semi-automated fashion, that is by directly engaging with the corpus data via query tools and manually analyzing the search results.[9]

[3] e.g. the quantity of a data item can be mirrored by the size of the representing visual element, or different types of data can be displayed with different colors

[4] http://in-spire.pnnl.gov/

[5] https://www.visualthesaurus.com/

[6] http://www.visuwords.com/

[7] van Ham et al., http://hint.fm/projects/phrasenet/

[8] http://chrisharrison.net/index.php/Visualizations

[9] As opposed to, for example, studies that are based on corpus derivates created in a fully automatic fashion, e.g. lists of collocations, clusters of keywords, etc.

3.2. Data structure

interHist draws on three types of information:

1. A *linguistic structure* specified by the corpus user by means of a formalized query
2. *KeyWord In Context* (KWIC) sequences matched by the query
3. *Occurrence frequencies* of the defined linguistic structure and its subsets

The *linguistic structure* is a sequence of linguistic items[10] as defined by the user.
It has to contain an *anchor* element, which is an obligatory token position with a unique value. In the example of Italian noun phrases, that will be laid out in detail below, the token with part-of-speech 'S' (*noun*) is the *anchor* item.
The *anchor* element splits the linguistic structure, specified as corpus query, into left and right context. Results patterns are recorded with reference to the *anchor* and their *occurrence frequencies* are calculated accordingly.
Finally, the *KWIC sequences* are word level representations of the query matches plus their surrounding context.
Summing up, the information that the visualization builds on is determined by the user's corpus query and completed with information extracted from the corpus.

3.3. Visualization design

The *interHist* visualization essentially provides a graphical representation of an intermediate layer between a *formalized corpus query* as provided by the user and the *exhaustive set of KWIC results* derived from the corpus. The intermediate layer builds on the query structure and frequency information derived from the results, and links to their KWIC representation.
interHist visualizes this information as a sequence of stacked histograms, where each histogram stands for a token position of the corpus query, or equally the words in the results' KWIC lines. A query that specifies six token positions will result in a visualization with a sequence of six stacked histograms.
Within each of the stacked histograms, the bar segments represent the distribution of the linguistic features per token position, as specified in the search query. Frequencies of the linguistic feature values per token position are indicated by the height of histogram segments. The linguistic feature values are encoded and distinguished by color. That is, color is used for associating corresponding feature values over different token positions. For example, in a corpus query that is defined as a sequence of part-of-speech values *adjectives* occurring in different positions with respect to the *anchor* would all be colored the same. Hovering over bar segments of the diagrams give the part-of-speech type as written text together with its number of occurrences.

[10]Typically this would be a sequence of closed class token-level attributes (e.g. parts-of-speech, semantic labels, etc.) or any listing of specific token values (e.g. all inflections of a certain verb, a list of prepositions, etc.). While theoretically also open class attributes could be used, the limited visual capacity of histograms requires token-level attributes with a restricted set of values.

The visualization allows for dynamic filtering of the results. By clicking on colored bar segments, the respective feature values (e.g. all parts-of-speech "adjective") are interactively activated as filters. This results in a second series of stacked histograms being created next to the initial series of histograms. The second series is representing the filtered results, while the first series shows the full results set.
A red border visually marks the restricting part-of-speech segment, as selected by the user, and second-level histograms accordingly. If filtering is active the brightness of first-level histograms is reduced.
Furthermore, the user can interactively adjust the layout of the data. The frequencies, represented by size of the stacked bars, can be viewed as relative values, as relative values with fixed minimum[11] or as log values.
Finally, from the *interHist* diagram users can directly access KWIC results matching the currently selected sequence of linguistic items. The selection is a type of filtering that is applied to the initial corpus query and creates a defined sub-set of the original results set.

3.3.1. Example visualization of 'Italian noun phrases'

We showcase *interHist* for the linguistic analysis of the composition of noun phrases in Italian. The analysis is carried out based on the PAISÀ corpus of Italian web texts (Lyding et al., to appear) which contains about 250 million tokens of contemporary Italian.
The investigation via *interHist*'s search interface is started by formulating a corpus query that approximates Italian noun phrases. It is formulated as a part-of-speech sequence which accommodates different structures of valid noun phrases of Italian (approximated to the detailed description in Renzi (1988)). An informal representation of the query is given in (a) below:[12]

(a) [predeterminer]? [determiner | pronoun]? [adjective]* [noun] [adjective | verb *ending in* ti|te|to|ta]?

The query matches any word sequence that contains an obligatory noun, which might be preceded by a sequence of one optional predeterminer, followed by one optional determiner or pronoun, followed by zero to any number of adjectives, and followed by an optional adjective or verb ending in 'ti', 'te', 'to' or 'ta, approximating the *participle* form (participio) which is often used as adjective, e.g. 'la frase *sbagliata*'.
Run on the PAISÀ corpus the query matches about 24 million examples of complex noun phrases, a noun together with at least one other element respecting the noun phrase specification of the query. In addition more than 500000 occurrences of nouns without context elements conforming to the query are found.

[11]Frequencies lower than a defined threshold are displayed by a bar with a fixed base size. This is to include categories in the display, that would be too small to be displayed according to the rendering of their frequency value.

[12]Each unit in square brackets indicates a token position. The sign following the brackets indicates how often the item might occur (with ? indicating 0 to 1 occurrences, * indicating 0 to many occurrences, and *no sign* indicating exactly one occurrence). The pipe symbol (|) reads as exclusive 'OR'.

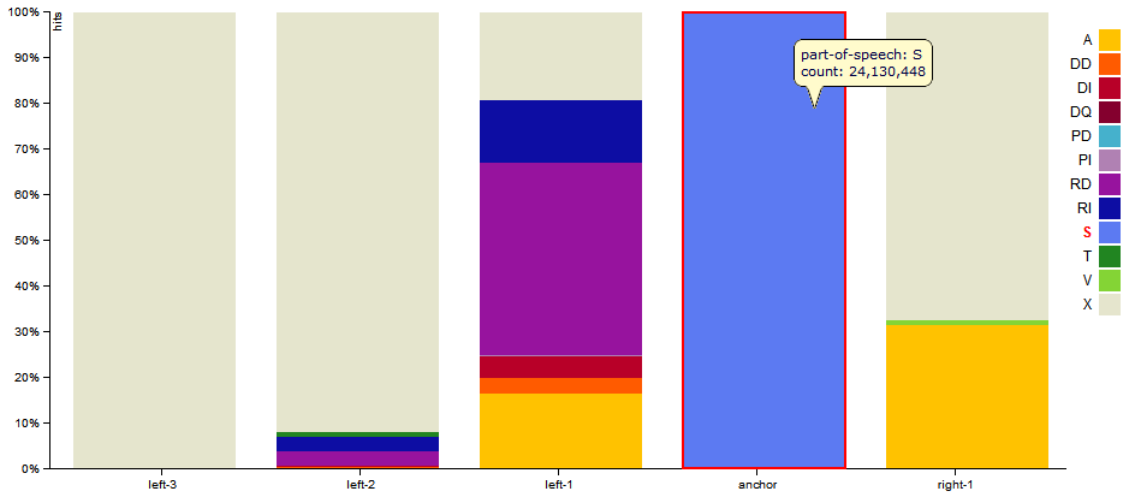

Figure 1: *interHist* visualization for Italian noun phrases

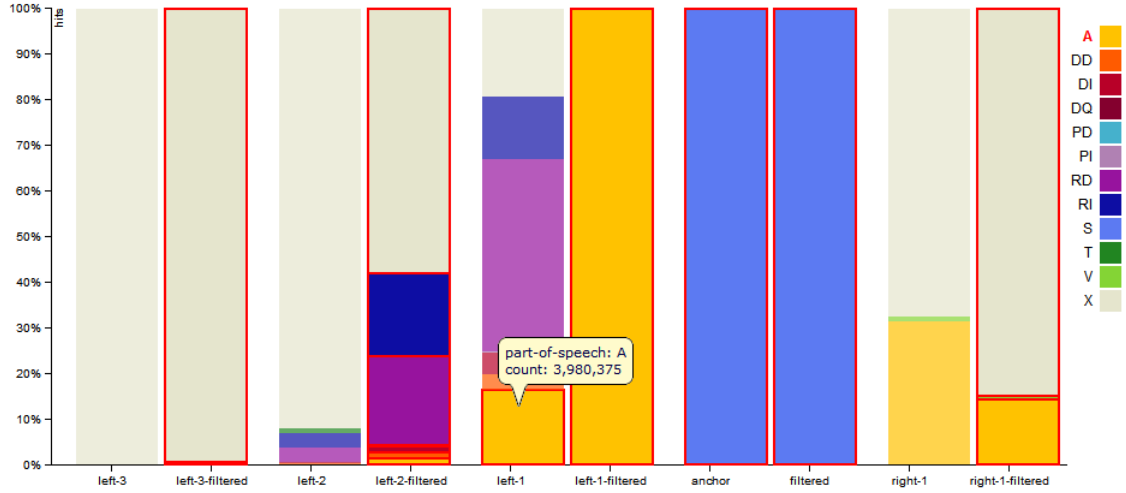

Figure 2: Filtered results by condition: *adjective* preceding the *noun*

Following the query description, the visualization is built on the token-level attribute *part-of-speech*, which provides for the abstraction over KWIC results.

Figure 1 shows the basic[13] *interHist* visualization corresponding to the study on Italian noun phrases, as approximated by the query in (a). Token positions are calculated relative to the anchor element 'noun' (S), displayed as light blue bar in fourth position. The anchor is the only obligatory element within the search query. It hence determines the results total, displayed on mousing-over the blue histogram bar. The *anchor* constitutes the central position of the token sequence, with a left context of up to three token positions and a right context of zero or one token expanding from it (marked in the visualization as "left-1, left-2, left-3" and "right-1").

A legend of the colors as assigned to part-of-speech values is provided on the right of the diagram. The descriptions correspond to the ISST-TANL part-of-speech tagset[14] as employed by the PAISÀ corpus. In addition, we introduced a category 'X' as dummy value designating positions not covered by all noun phrases.

The visualization helps to understand, that one position left of the noun, *determiners* ('RD'/'RI', determinative/indeterminative article) constitute the most frequently occurring word classes followed by *adjectives* ('A'). Two positions left predeterminers ('T') are more frequent than adjectives.

The visualization allows to filter the query results by interacting with the graphics. As explained above, filtering is done by clicking on a bar segment which stands for a part-of-speech value in a certain token position of the results set. For our sample analysis we selected the part-of-speech one position left of the noun to be of type adjective ('A').

Clicking on the yellow bar segment results in a display with two series of histograms for each token position, side by side. The updated visualization is shown in Figure 5.

In Figure 5 the second sequence of histograms displays the filtered subset of query results according to the filter-

[13] that is without any interaction applied

[14] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

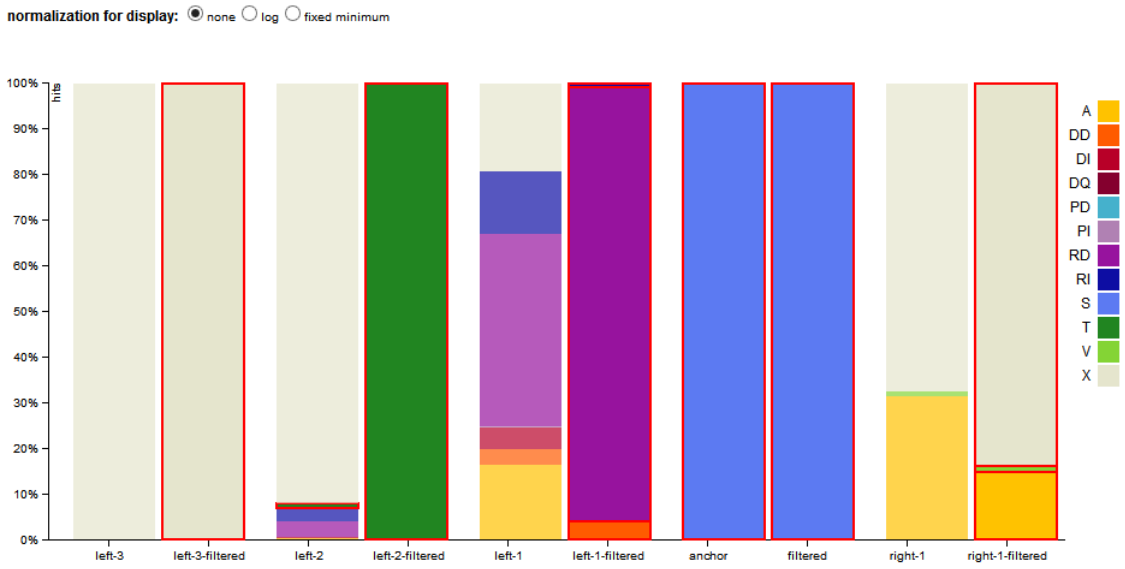

Figure 3: Filtered results by condition: *predeteriminer* in position left-2 of *noun*

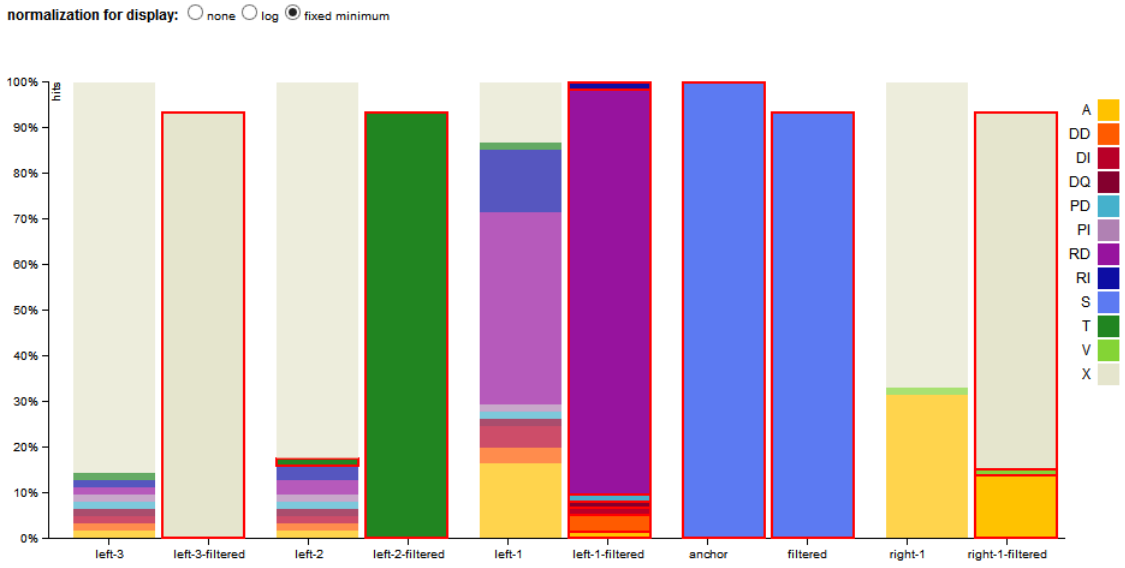

Figure 4: Filtering as in Figure 3 displayed with 'fixed minimum' values

ing condition applied to the results set. In this example the distribution of word classes over token positions got recalculated according to the condition: "tokens one position left (*left-1*) of the anchor noun ('S') are restricted to part-of-speech type *adjective* ('A')".[15] The first sequence of histograms still represents the distribution of word classes in the full results set. It is visually modified in to make transparent what filtering condition is active, by drawing a red line around the filtered element, and reducing the brightness of all other elements.

By coordinating views on a full results set and a subset of it, it can be observed how parts-of-speech in different token positions depend on each other.

The visualization in Figure 2 for example makes notice that under the condition of having an adjective preceding the noun in position left-1, the frequency of adjectives in position left-2 is increasing, while the occurrences of predeterminers in left-2 is decreasing.

To the contrary we could filter by *predeterminer* ('T') in position left-2. The resulting visualization is shown in Figure 3. We can see that under this condition, in left-1 the variation of part-of-speech types fundamentally reduces to *determinative article* ('RD') and *demonstrative determiner* ('DD'). To check if other parts-of-speech might be found with small frequencies, switching to 'fixed minimum' representation format is helpful. Figure 4 displays the respective visualization and shows that also *adjectives*, *indefinite* and *interrogative determiners* etc. occur in small numbers.

Moving from the filtered results sets to individual KWIC

[15] First-level and second-level histograms are calculated on totals for the full and restricted results set.

examples, we can inspect concrete uses of the patterns.
KWIC examples corresponding to the filtering in Figure 5 include:
molti argomenti trattati ('many tackled topics')
una certa familiarità ('a certain familiarity')
i saporiti pesci ('the tasteful fishes')
una sola volta ('a single time')
nuove aree usabili ('new usable areas')
KWIC examples corresponding to the filtering in Figure 3 and 4 include:
tutti i costi ('all (the) costs')
entrambi i piatti ('both (of the) plates')
tutte le attività umane ('all (of the) human activities')
tutto il mondo ('the entire world')
tutti gli impegni successivi ('all (of the) subsequent duties')

3.3.2. Implementation details and future extensions

The *interHist* visualization builds on the corpus query engine of the OpenCWB (Evert and Hardie, 2011). The query engine provides for the retrieval of corpus data on token level (i.e. word information and related annotations) and the tabulation of results according to token-level annotations. The calculation of frequency information on the results sets is done by custom scripts, prepared for *interHist*.
The *interHist* visualization is implemented in JavaScript using the D3 visualization toolkit[16], this means it is running on the client, inside the web browser, and is communicating with the server-side corpus engine via a web service.
An *interHist* demo is integrated with the PAISÀ corpus which is indexed for the OpenCWB and hosted at EURAC. The demo provides for custom searches that can be specified by means of the query language CQP. The demo has been tested for the Chrome, Mozilla Firefox and Opera web browser.
The primary objective for follow-up versions of interHist is to further extend the provided analysis functionalities. In particular, we foresee two tracks towards this end:
First, filtering options will be extended to allow for combined filters per token position, e.g. restricting the part-of-speech of the word in position left-1 to either definite ('RD') or indefinite determiner ('RI'), instead of allowing only for one value.
Second, subsequent filtering of filtered query results will be supported; that is second-level histograms (representing filtered results) will become clickable so as to create third-level histograms (representing a respective subset of the first selection), and eventually extending filtering to level N.
Furthermore, we aim to investigate how *interHist* can be connected to existing visualizations, which build on word information (e.g. Double Tree (Culy and Lyding, 2010) and others[17]) as opposed to categorical information (used by *interHist*).

4. Conclusion

In this article, we introduced the *interHist* visualization for the analysis of corpus data by linguists. It provides a showcase of how a visualization adapted for language data can serve as advanced interface component for language resources.
interHist's central innovation for the analysis of corpus data is the abstraction from token sequences to sequences of distributions of token characteristics (i.e. linguistic annotations). For the use case on Italian noun phrases, this allows for condensing 24 million query results into one display. The compact representation of query results based on their abstraction to part-of-speech sequences is integrated with a standard KWIC view on selected subsets of the results. *interHist* can equally operate on other token-level attributes, for examples lexical semantic classes or morphosyntactic features.
The visualization allows for the interactive exploration and user-steered switching between different views on the data under analysis.

[16] d3js.org
[17] see section 2.3.

5. References

Jacques Bertin. 1983. *Semiology of Graphics*. University of Wisconsin Press.

Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics : investigating language structure and use*. Cambridge Univ. Press, Cambridge [u.a.].

Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors. 1999. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Winnie Cheng, Chris Greaves, and Martin Warren. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11:411–433.

Christopher Collins. 2007. Docuburst: Radial space-filling visualization of document content. Technical report.

Chris Culy and Verena Lyding. 2009. Corpus clouds - facilitating text analysis by means of visualizations. In Zygmunt Vetulani, editor, *LTC*, volume 6562 of *Lecture Notes in Computer Science*, pages 351–360. Springer.

C. Culy and V. Lyding. 2010. Double tree: An advanced kwic visualization for expert users. In *Information Visualisation (IV), 2010 14th International Conference*, pages 98–103, July.

A. Frid-Jimenez J. Guiness DeCamp, P. and D. Roy. 2005. Gist icons: Seeing meaning in large bodies of literature. In *Proceedings of the IEEE Symposium on Information Visualization*.

S. Evert and A. Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proc. of the Corpus Linguistics 2011*, Birmingham, UK.

M. Hearst and D. Rosner. 2008. Tag clouds: Data analysis tool or social signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS)*, Hawaii, USA.

Marti A. Hearst. 2009. *Search User Interfaces*. Cambridge University Press, 1 edition.

Owen Kaser and Daniel Lemire. 2007. Tag-cloud drawing: Algorithms for cloud visualization. In *Proc. WWW 2007 Workshop on Tagging and Metadata for Social Information Organization*, Banff, Canada, May.

A. Lüdeling and M. Kytö. 2009. *Lüdeling, Anke; Kytö, Merja: Corpus Linguistics*, volume 1/2 of *Lüdeling,*

Anke; Kytö, Merja: Corpus Linguistics. Mouton de Gruyter, Berlin.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. to appear. The paisà corpus of italian web texts. In *Proceedings of the Web as Corpus Workshop at EACL 2014*.

Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank, and Daniel A. Keim. 2010. Visualizing Vowel Harmony. *Linguistic Issues in Language Technology*, 4(Issue 2):1–33, December.

Tony McEnery and Andrew Wilson. 2005. *Corpus linguistics : an introduction*. Edinburgh Univ. Press, Edinburgh.

Lorenzo Renzi. 1988. *Grande grammatica italiana di consultazione. La frase. I sintagmi nominale e preposizionale.*, volume 1. Il Mulino, Bologna.

Christian Rohrdantz, Steffen Koch, Charles Jochim, Gerhard Heyer, Gerik Scheuermann, Thomas Ertl, Hinrich Schütze, and Daniel A. Keim. 2010. Visuelle textanalyse. *Informatik-Spektrum*, 33:601–611.

Mike Scott. 2010. What can corpus software do? In Anne O'Keeffe and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, Routledge Handbooks in Applied Linguistics, pages 136–151. Routledge, London.

M. Trenkmann, M. Potthast, H. Gruendl, P. Riehmann, B. Stein, and B. Froehlich. 2012. Wordgraph: Keyword-in-context visualization for netspeak's wildcard search. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1411–1423.

Colin Ware. 2004. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Martin Wattenberg and A B. Viégas. 2008. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228.

Scott Cederberg Dominic Widdows and Beate Dorow. 2002. Visualisation techniques for analyzing meaning. In *Proceedings of the Fifth International Conference on Text, Speech and Dialogue*, pages 107–115.

Amir Zeldes, Julia Ritz, Anke Lüdeling, and Christian Chiarcos. 2009. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*.

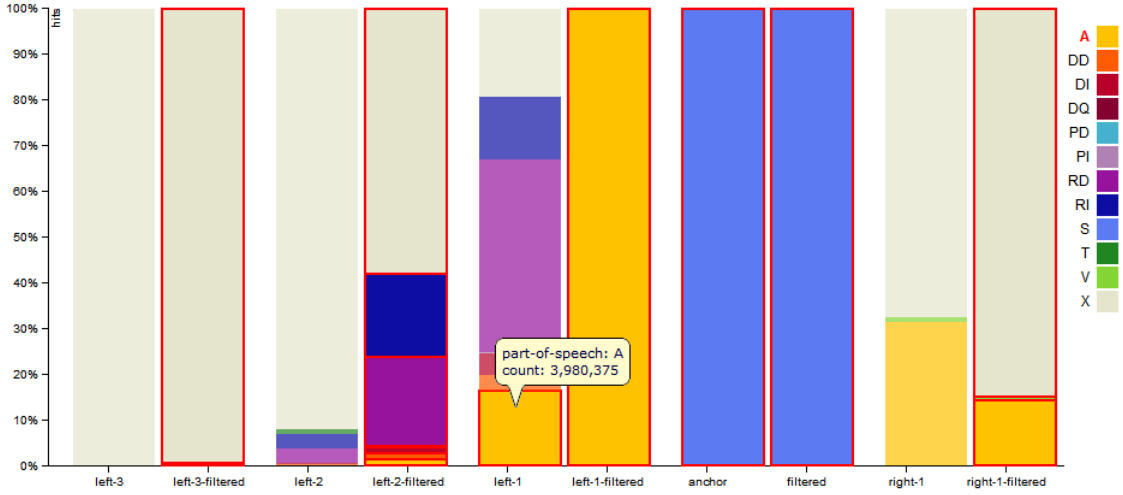

Figure 5: Figure 2: interHist visualization of Italian noun phrases together with results filtered by ”adjectives“ in position left-1