Visualization of sensory perception description

Kerren, Andreas; Pangrova, Mimi; Paradis, Carita

2011

Visualization of Sensory Perception Descriptions

Andreas Kerren, Mimi Prangova
School of Computer Science, Physics and Mathematics (DFM)
Linnaeus University
SE-351 95 Växjö, Sweden
Email (corresponding author): andreas.kerren@lnu.se

Carita Paradis
Centre for Languages and Literature
Lund University
SE-221 00 Lund, Sweden
Email: carita.paradis@englund.lu.se

***Abstract*—On the basis of a large corpus of wine reviews, this paper proposes a range of interactive visualization techniques that are useful for linguistic exploration and analysis of lexical, grammatical and discursive patterns in text. Our visualization tool allows linguists and others to make comparisons of visual, olfactory, gustatory and textual properties of different wines from different parts of the worlds, from different grape varieties, or from different vintages. It also supports the immediate creation of visual profiles for descriptions of sensory perceptions for exploratory purposes as well as for purposes of confirmatory investigations of linguistic patterns in text and discourse and their correlations to metadata variables.**

***Keywords*-multivariate visualization; interaction techniques; text visualization; scatter plot; dynamic queries; wine reviews;**

I. Introduction

In this paper, we present our work on the development of an interactive information visualization tool to be used on corpus data. The tool has been developed on the basis of 84,864 wine reviews, or tasting notes as they are also sometimes called, from the Wine Advocate[1] journal. Thanks to the capacity of the tool to handle large amounts of data and to its dynamic interface, it can be used for exploratory work as well as for confirmatory investigations in linguistics [1].

Wine reviews are descriptions and evaluations of wines written by professional wine tasters. They have a strict rhetorical structure and consist of three parts, starting with production facts and ending with an assessment and a recommendation of prime drinking time. The middle of the text, which is the most important part, is devoted to an iconic description of the wine tasting procedure from the taster's inspection of the wine's visual appearance through smelling, tasting and feeling its texture, i.e., from *vision* through *smell*, *taste*, and *mouthfeel (touch)* [2], cf. sample review (1).

> (1) "This great St.-Estephe estate has turned out a succession of brilliant wines. The 2005, a blend of 60% Cabernet Sauvignon and 40% Merlot, has put on weight over the last year. An opaque ruby/purple hue is accompanied by a sweet nose of earth, smoke, cassis, and cherries as well as a textured, full-bodied mouthfeel. While the tannin is high, there is beautifully sweet fruit underlying the wines structure. It will require 8-10 years of cellaring after release, and should drink well for three decades." (Wine Advocate 170, April 2007)

The visual appearance of the wine in (1) is described in terms of its clarity and color using the descriptors 'opaque ruby/purple'. The olfactory perceptions are primarily described through concrete objects, e.g., 'earth, smoke, cassis, and cherries', but also in terms of a gustatory property, 'sweet', while taste and mouthfeel are described through various gustatory and tactile properties ('high' (tannin), 'sweet' (fruit), 'textured, full-bodied'). Because almost all wine reviews describe the wines in terms of four different perceptual modalities, i.e., visual appearance, smell, taste and texture, they are a gold mine for linguistic explorations of descriptions of human sensory perceptions in discourse. Of particular interest are the descriptions of olfactory perception. There is no specific olfactory vocabulary, neither in English nor in (most) other languages of the world. Olfactory descriptions have to be made using words from other domains. In wine reviews, words for taste or words for objects such as fruit, herbs or flowers of different color are used. In general, dark objects are used in descriptions of red wines and pale objects describe white wines. In other words, olfactory descriptions are primarily made on the basis of the smell of objects and also their color and taste. Exploring patterns for perceptual descriptors and the context of their use in wine reviews provides useful information not only about the relations between descriptors of odor and other modalities, but also about language, perception and cognition in general [3].

Our tool supports the visual analysis of the corpus of wine reviews from the Wine Advocate. The wine reviews are available in the form of two databases that contain a large number of wines, metadata about the wines, and the actual reviews. In order for linguists to arrive at a better understanding of different text types, different discourses and their vocabularies, large corpora are of crucial importance. At the same time, it is also a challenge to identify linguistic patterns in large corpora, to organize the data, to make statistical calculations and to present the data to

[1] https://www.erobertparker.com/entrance.aspx

readers in intuitive and clear ways. The contribution of this paper is to find solutions to some of these challenges. The first challenge is that we have to be able to represent large amounts of multivariate data. For that purpose, advanced interaction techniques are essential, because they ensure the opportunity for selecting a subset of tasting notes and for getting detailed information about the tasting notes in order to proceed with further analyses. Secondly, we have to find an efficient way to interactively visualize the text of the individual wine reviews, which brings us to the field of interactive text visualizations. Thirdly, a number of compatible visualization approaches have to be combined in order to efficiently explore the language used in the descriptions of the wines.

The remainder of this paper is organized as follows. Section II gives a general overview of the advantages of our tool for linguists. Then, we discuss related approaches in Section III within the field of information and text visualization as well as in the field of linguistic analysis of sensory descriptions. In Section IV, we describe the wine database. Our own approaches to the visualization of wine tasting notes by using information visualization (InfoVis) techniques are presented in Section V. Initial results are briefly outlined in Section VI. We conclude in Section VII and suggest some investigatory paths for future work.

II. Linguistic Background

Advances in visualization offer important possibilities for organizing, presenting and analyzing linguistic data, in which case visualization techniques provide a way to view language in another formats than as linear stretches of letters. Visualization techniques offer the tools to capture lexico-semantic usage patterns and to represent interactions of different dimensions of language structure that characterize different texts and discourses. As demonstrated in the introduction, descriptions of wines in wine magazines are short texts with a very strict rhetorical structure. The language of such texts are of interest to linguists at various different discursive levels. Linguists want to know about what kind of words are used to describe the wines' visual properties, what kind of descriptors are used for olfactory, gustatory and tactile perceptions. They are interested in what words and expressions are used where in the texts. For instance, what kind of temporal expressions are used in different parts of a text, and what expressions of personal opinion, such as 'should', 'drinkable', 'recommend' are used where in the texts and why. More generally, linguists take an interest in how all linguistic patterns combine into what might be our understanding of the discourse beyond the text itself. In other words, visual imagery provides a way to represent things that would otherwise go unnoticed. The added value of the visualization tool presented in this paper is that it can be used interactively. The data can be easily explored, and because parameters and combinations of data and metadata can be changed, many questions regarding the potential of the data receive on-the-spot answers. As a result, new patterns emerge that can generate new research hypotheses about language use in different genres and text types.

III. Related Work

Using both corpus methodologies including visualization of the data and experimental psychophysical techniques, Morrot et al. investigated the interaction between visual appearance and odor determination in wine description and wine tasting [3]. Their work presents the results of a study carried out with the help of a tool called ALCESTE. It is based on statistics about the distribution of words in a corpus of text to determine groups of words that co-occur in the same context. They found that the descriptors used to characterize white and red wines were different in terms of the colors of the objects used in the descriptions respectively (i.e., dark objects describe red wines and pale objects white wines). In addition to the corpus study, they also carried out a psychophysical experiment, which confirmed the corpus data, demonstrating the impact of vision on the human odor perception. In comparison to ALCESTE, our visualization tool gives users more possibilities to browse the text, to filter out uninteresting cases, and to interact with the visualizations. Thus, it does not afford pure statistical numbers only, but gives analysts an opportunity to explore the dataset and to get a better understanding of the texts' structure and content.

Another visualization approach, called Wine Fingerprints, has been discussed by Kerren [4]. In contrast to the tool presented in this paper, Wine Fingerprints focus on wine attributes, such as wine color, rating, grape type, price, or aroma, and not on the actual wine reviews. This data forms a multivariate data set, part of which can be hierarchically structured into a so-called aroma hierarchy. The Wine Fingerprints approach has various applications for business and industry in that it can create visual patterns of combinations of wine attributes and support comparisons of visual, olfactory and gustatory properties of different wines from different parts of the worlds, from different grapes, from different vintages etc. Both customers and companies can make visual comparisons of wines and select wines on a pictorial basis instead of on the basis of a list of multimodal perceptual attributes.

For the purpose of information visualization of complex textual data, we use different well-known techniques and interaction approaches. The general design of our visualization tool is based on standard coordinated and multiple view visualization techniques as described in Section V-A. An excellent starting point for related work of this kind of visualization techniques is the annual conference series on Coordinated & Multiple Views in Exploratory Visualization (CMV) or the work of Roberts [5].

In order to specify the layout of our tool and to define the functional requirements, we were inspired by the FilmFinder tool for exploring film databases [6]. It was one of the first tools, which integrated the concept of a two dimensional scatter plot with color coding, filtering, and details provided on demand (dynamic queries). The developers realized different encoding and interaction techniques for the representation of multivariate data.

The research project Many Eyes provides alternative methods for data analyses using innovative visualization techniques [7]. One of the approaches for supporting text analysis is the representation of a given text as a word tree [8]. The purpose of this visualization method is to afford an insight into the different contexts in which a word is encountered in an unstructured text. We used this concept in one of the text visualizations of our tool to facilitate rapid exploration of the wine tasting notes' content.

An approach for visual literary analysis, called Literature Fingerprinting, was presented by Keim and Oelke [9]. This work supports the visual comparison of texts by calculating features for different hierarchy levels and by creating characteristic fingerprints of the texts. Such features might be word length or measurement of vocabulary richness.

Salton and Singhal analyze the relationships between text documents according to different topics. They developed a tool called Text Theme [10] to represent such correlations visually. Single topics can then be identified and be compared with the help of textures or color coding. In contrast to this approach, our tool operates more on the syntactic level, i.e., higher-level themes cannot be compared directly.

Tag clouds provide information about the frequency of words used in a corpus of text [11]. This approach uses different font sizes for each word in the text to indicate how often this word is used by comparison with the others. Several extensions and approaches exists, such as Wordle or ManiWorlde [12], [13]. We use a simple tag cloud implementation to represent the word frequency in a group of tasting notes.

Stasko et al. developed a visualization tool for analyses of textual reports called Jigsaw [14]. The goal of their tool is to aid investigative analysts to faster understand the content of reports in order to predict possible threats and to prepare defensive plans accordingly. The main analysis unit in their work are pre-defined entities in the texts and the goal of the implemented visualizations is to represent relations and connections between these entities. As distinct from their work, our tool is not designed to focus on the significance of specific entities extracted from the texts but rather on the exploration of their content and linguistic constructions.

IV. Notes on the Data Set

The wine tasting notes are stored in two databases that contain information about different wines as well as the tasters' comments about them. In each database, the tasting notes are represented in different ways. The first database contains descriptive information about the wines, their unique ID number, their origin, vintages, wine ratings, dryness, color and the complete original wine review. The second database contains the same tasting notes including ID numbers, but they are segmented into words and word-class tags (so-called word tags, such as nouns (NN) or adjectives (JJ) [15]). The latter database was built from the former, the original database, by using the WineConverter tool, developed by Ekeklint and Nilsson from the computer linguistics group at Växjö University, Sweden. The result of this segmentation is a new structuring of the wine tasting notes where each word is described by additional information that accurately specifies its position in the text of the full tasting note. The location of each word in a tasting note is determined by the following information: ID number of the tasting note, number of the corresponding sentence in the tasting note, position of the word in this sentence, the word itself, and the word tag given to this word.

In order to get a better overview of appropriate visualization approaches for representing the tasting notes and their attributes, we had to take the great amount of analyzed data into consideration. Table I provides a list of substantial statistical numbers derived from the dataset to give an idea about the sheer quantity of the data to be visualized.

Table I
Statistical numbers derived from the wine databases.

Number of tasting notes	84,864
Total number of words used in the tasting notes	8,332,666
Number of different words used in the tasting notes	46,000
Maximum length of the tasting notes	496
Number of word classes	43
Number of vintages	104
Range of wine rating values	1 to 100

V. Visualization Framework

In order to provide an overall perspective of the analyzed wine tasting notes, we follow Ben Shneiderman's mantra of information visualization: "overview first, zoom and filter, details on demand" [16]. This gives users an initial overview of the explored data and the possibility to proceed with investigation of its subsets. For this, we combined several visualization approaches to achieve our goals: scatter plots, tag clouds, word trees, bar charts / histograms, and a world map. The scatter plot is used to be the main entry point for using our tool as described in the following.

A. Visual Representations

1) Scatter Plot: The purpose of this visual representation is to give a first overview of the data. Because of the large number of tasting notes (cf. Table I), we decided to use a scatter plot for their initial display, i.e., each single tasting note is represented by a blue circle. This approach also

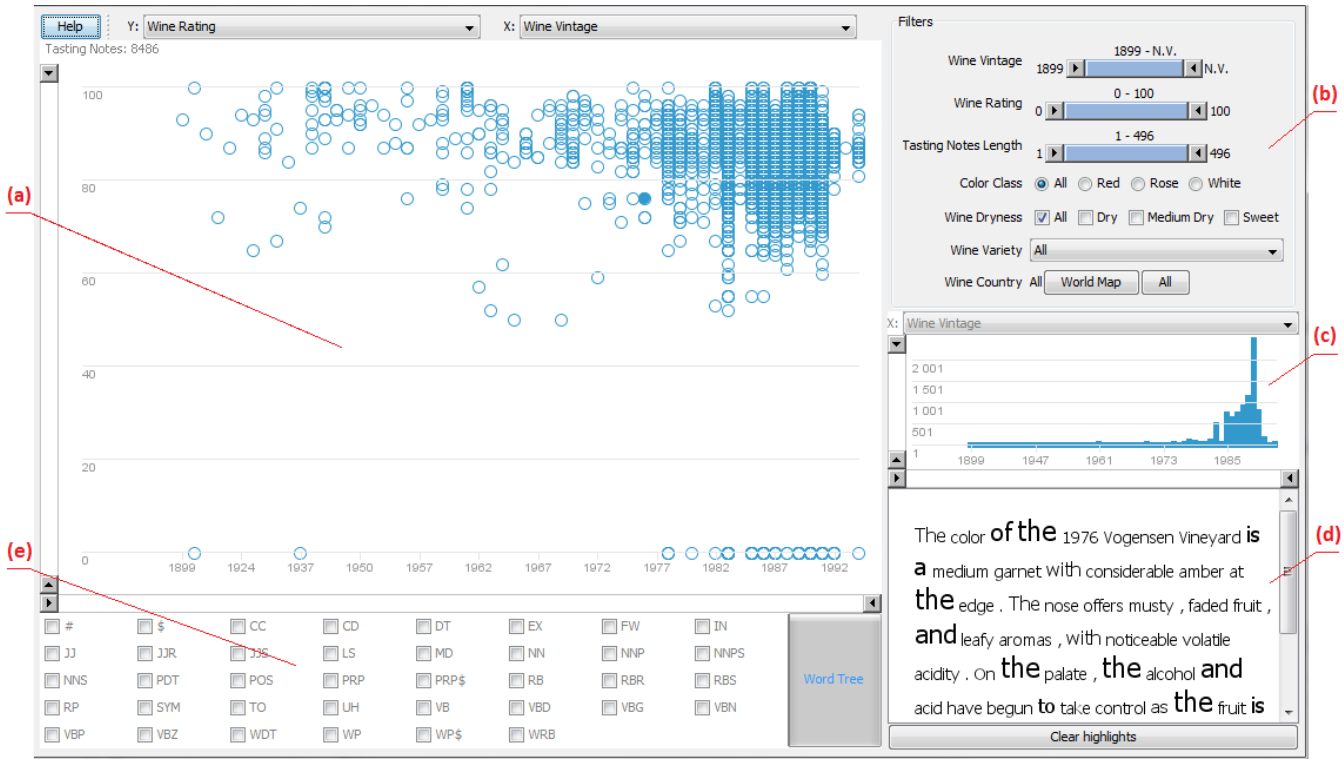


Figure 1. A snapshot of the main window of the application after starting. Note that one tasting note was selected in the scatter plot; its tag cloud is shown in the bottom right corner.

saves space and gives an idea about the distribution of the tasting notes on the basis of the values of two selected wine attributes, see Figure 1(a). Attributes currently supported by the scatter plot visualization are all possible pairs of Wine Rating, Wine Vintage, Color Class, Tasting Notes Length, and Wine Country.

2) Bar Charts and Histograms: Getting statistical information helps analysts to better understand the visualized data and finding the desired set of tasting notes. Bar chart diagrams are traditional approaches to statistical data visualization. In this work, they are supplied in order to show the number of tasting notes that correspond to the values of a specific wine attribute (Figure 1(c)).

3) Text Visualization: The visualization approaches that we apply for the representation tasting notes are word trees and tag clouds.

Word Tree: The word tree visualization facilitates rapid querying and exploration of text bodies [8]. In our tool, a word tree describes the sequence of words and phrases used in a group of tasting notes. The structure of the word tree is organized into two main groups of nodes: word tags and words. There are three prerequisites for proceeding with the word tree visualization:

- users need to select a group of tasting notes for further analyses,
- a specific tasting note for deriving the initial data (from now on referred as *root tasting note*), and
- the word classes of the words in this root tasting note that they would like to analyze.

The first three levels of the word tree contain data from the root tasting note. The other levels consist of data from the whole group. The root node of the tree is artificially added, and it contains the static text “Tags”, which suggests that the following level is composed of word tags. The second level contains the selected word tags that correspond to words in the root tasting note. The third level consists of all the words from the root tasting note that belong to the word tags on the previous level. Figure 2 gives an example of the word tree and the organization of its nodes. The levels of the tree alternate with each other to represent either word tags or words that correspond to the tags on the previous level. The children of each node representing a word class are the words from the analyzed group of tasting notes that belong to this word class. For instance, the word tree in Figure 2 displays two (selected) word tags of the root tasting note, i.e., “JJ” (adjectives) and “NN” (nouns, singular common). By looking at the children of “JJ”, the user can see that the root tasting note has four adjectives, e.g., “coarse”. Then, by looking at the next two deeper levels, the user can see that “coarse” has two successors: one noun (plural common; “NNS” $\rightarrow$ “flavors”) in another note from the analyzed notes group (black) and one determiner (“DT”$\rightarrow$ “, this”) in the root tasting note.

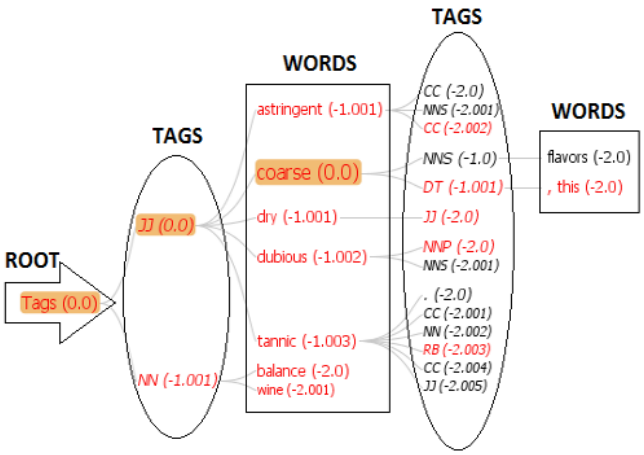


Figure 2. A word tree that shows the node organization into two main groups: word tags and words. Furthermore, there is another partition of the tree nodes as well: nodes that contain data from the root tasting note (colored in red) and nodes that contain data only from the other tasting notes (colored in black). The DOI value of each node is given in brackets on its right side.

Our word tree visualization represents a large data set of words and word tags. It is restricted by the size of the display and people's perceptive capabilities. To cope with these restrictions, our implementation applies the idea of Degree-Of-Interest (DOI) trees that provide a solution of these problems. They combine Focus&Context visualization techniques and degree-of-interest calculations to find a proper layout that fits within the bounds of the display. The technical idea is the use of a degree-of-interest function, which assigns a number value (DOI value) to each node indicating how interested the user is in this node. This value is then used as a criterion to determine, which of the nodes should be visible, which of them are in the focus and how they should be displayed [17], [18]. The nodes in the focus have the greatest DOI value and are slightly magnified. The size of all other nodes is directly proportional to their individual DOI value. An exception to this rule is the tree element that was selected last, which is the most magnified element, in spite of the fact that it has the same DOI value as the other focus nodes. Figure 2 demonstrates a degree-of-interest tree where the DOI value of the nodes are given in brackets on the right side of the node label.

Tag Clouds: The tag cloud visualization makes use of different font sizes for the words in a corpus of text to give a hint about the frequency of their usage. There are two prerequisites for the application of tag cloud visualizations to a tasting note's text. A group of tasting notes needs to be defined for further analyses, and one of them has to be selected for its text visualization. The text of the selected tasting note is then visualized by using the tag cloud metaphor, where each word has a different font size depending on how often this word occurs in the whole group of tasting notes, cp. Figure 1(d).

4) World map visualization: The origin of a wine is important information, visualized in a way that gives a rough overview of the wine-producing countries of the world. A natural approach for visualizing it is an interactive world map indicating the density of wine production in different countries. Figure 5 shows a world map representing information about the wines produced in different parts of the world that have been tasted and described in the tasting notes. The color saturation is directly proportional to the density of wines produced in each country.

B. Interaction and Coordinated Views

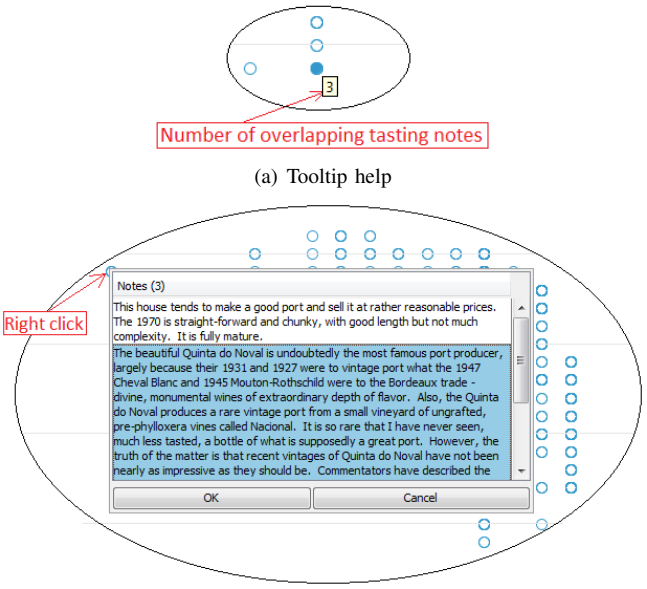
We combined the visualization approaches described in Section V-A with appropriate techniques for user interaction to build an efficient tool for analyzing wine tasting notes. The following subsections give a notion about the user interface and the overall layout of the application. In detail, there are five particular views intended to build an efficient overview visualization as displayed in Figure 1:

(a) a scatter plot (showing the distribution of tasting notes),
(b) filters (to reduce the complexity by filtering),
(c) bar charts and histograms (to show statistical data),
(d) tag clouds (for text visualization), and
(e) tag checkbox panel (to select specific word classes).

All aforementioned view are coordinated by standard highlighting and brushing techniques.

1) Distribution of Tasting Notes: The scatter plot axes on the left and bottom sides of the display correspond to one wine attribute each. Range sliders [19] are added to the axes in order to make it possible for the users to change the range of the wine attributes' values and therefore the scope of tasting notes visualized in the scatter plot. The number of visible tasting notes can be observed at the upper left corner of the scatter plot (8,486 in our screenshot example of Figure 1). Another possibility given to the user is to change the wine attributes plotted on the x-axis and y-axis by selecting other attributes from the combo boxes on the top of the display.

There is a drawback appearing as a consequence of the scatter plot concept and the data stored in the database: it might happen that more tasting notes share the same values for both of the wine attributes plotted on the axes. Such tasting notes overlap when they are visualized at the same spot in the scatter plot. This makes the selection of an element from the display more complicated. We added a tooltip to each element to give the user a hint about the number of overlapping tasting notes at the specific position (Figure 3(a)). Thus, an individual element can be selected from a popup list of the overlapping tasting notes, as shown in Figure 3(b). The selected tasting note differs from the others in its blue color in the scatter plot and in the popup list.



(a) Tooltip help

(b) Tasting note selection

Figure 3. Overlapping tasting notes in the scatter plot view.

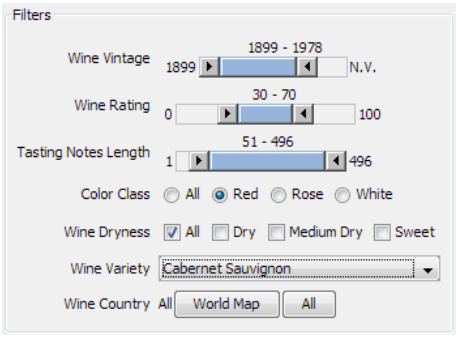


Figure 4. Types of filters implemented in the application.

2) Filtering: Filters are used to facilitate the task of the users to interact with the visualization and to find the best subset of elements to be further analyzed. Figure 4 shows a screenshot example of filters supported by our tool. By the current selection only those tasting notes are considered in the different views, which corresponding wines have a vintage between 1899 and 1978, a rating between 30 and 70 points, a length between 51 and 496 words, a red color, no specific dryness, geographical information, and made from Cabernet Sauvignon grapes.

The world map filter is a realization of the geographic visualization approach described in Section V-A4. It provides users with the opportunity to filter out tasting notes on the scatter plot depending on their origin (Figure 5). Different standard functionalities are supplied to assist working with the map like zooming in, zooming out and panning to a specific region of interest. Users have the possibility to select

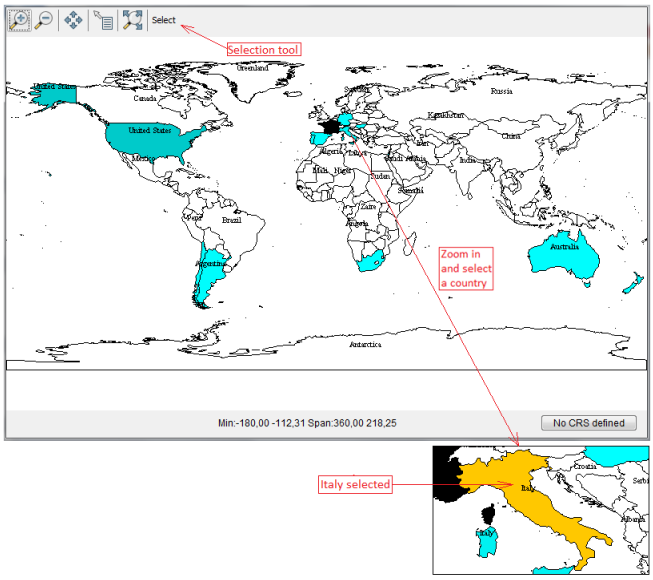


Figure 5. World map providing information about the tasted wines produced in different countries. It is also possible to use this view as an interactive filter for the specification of single countries.

a country on the map and our tool visualizes only those representations of tasting notes of wines produced in the specified region.

To provide a better software maintenance, we use a specific property file that contains a list of wine attributes and their required filter types. In this way, filters are dynamically created on the basis of this information and can be easily added or removed.

3) Statistical Information: The property file also contains a list of wine attributes that can be represented by bar chart diagrams. Figure 6 presents snapshots of histograms implemented in the application. An individual bar chart or histogram is created for each of the listed attributes showing the number of visible tasting notes corresponding to each of their values (Figure 6(a)). Only one of the diagrams is visualized at a time in order to save space. We added range sliders to the x- and y-axes to assist users in changing the range of visualized attribute values and to get a closer look at a specific section of the diagram, see Figures 6(b) and 6(c) where the vintage range was modified.

4) Word Frequency Analysis: After the selection of a tasting note in the scatter plot view, its text is visualized using a tag cloud approach (cf. Section V-A3). The font size of each word is estimated according to the frequency of its occurrence in all elements visible at the same time, including the selected one. Figure 7(a) shows a tag cloud example generated by our application.

The tag checkbox panel contains all word tags available. A coordinated interaction exists between the tag cloud view and the checkbox panel. On the one hand, when the user

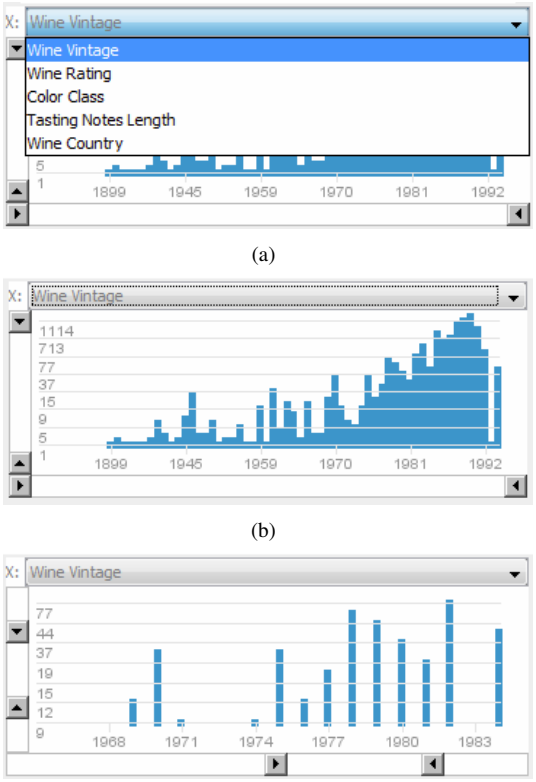


Figure 6. Screenshots of an interactive histogram for attribute "Wine Vintage".

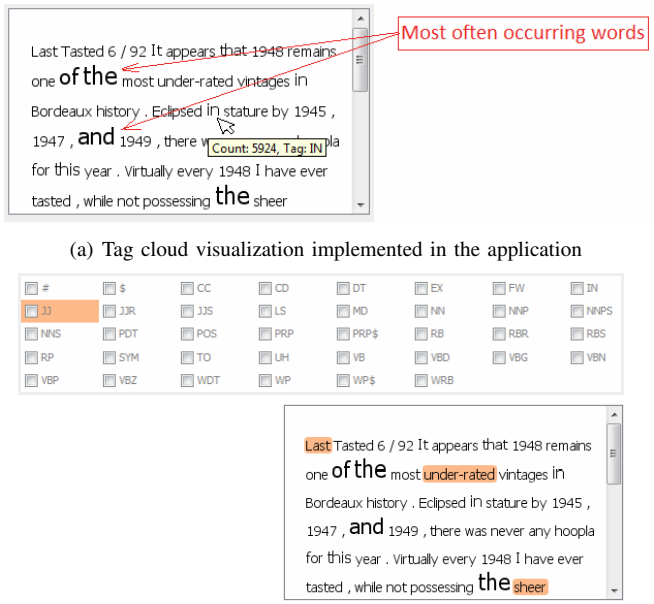


Figure 7. Word frequency analysis

selects a word from the tag cloud visualization, all words of the same class (i.e., with the same word tag) are highlighted in the tag cloud together with the tag itself in the tag checkbox panel. Figure 7(b) demonstrates this interaction after selecting the word "Last" in text of the tasting note. On the other hand, when a word tag is checked in the checkbox panel (such as "DT"), then it is highlighted together with the words corresponding to this tag in the text.

5) Sentence Structure Analysis: The basic concept and structure of the word tree visualization was already described in Section V-A3. Figure 8 presents an additional example of a word tree generated by our system. The visualization consists of three basic components: (a) a display containing the word tree, (b) a text area presenting the text of the root tasting note, and (c) a text area presenting the currently constructed sequence of words. All nodes that build a path from the root node to the currently selected node are in focus. Selecting a node from the tree changes the focus to the nodes contained by the path from the root to this node. A smooth animation is used to change the state of the tree to the newly selected focus [17]. The nodes in focus are highlighted with another background color and slightly enlarged. In the example, the node selected is "raspberries", and therefore, all nodes from the root to the node "raspberries" are in the focus. These nodes constitute a sequence of words which forms part of one or more tasting notes in the current scatter plot. This sequence is displayed at the bottom of the word tree, and it is also highlighted in the root tasting note if included there.

In Figure 8, the actual sequence of words is "glass, offering aromas of ripe raspberries.". The node labels are in red since they are contained in the root tasting note. Often, the tree depth and width exceed the display bounds. It is not possible all the nodes to be visualized in the space available. In order to surmount such problems, different techniques are integrated into the visualization, e.g., zooming in, zooming out, and panning controls [17].

There is a close relation between word tree visualization and the scatter plot. The word tree is constructed according to all combinations of words beginning with words from the root tasting note and followed by words from the whole group of tasting notes visualized in the scatter plot. This means that each sequence of words specified by the word tree exploration is contained in at least one tasting note of the current scatter plot selection. This relation is indicated by highlighting those tasting notes in the scatter plot (by a filled blue circle), which contain the sequence of words construed by the word tree (Figure 9(a)). The tool makes sure that highlighted elements are always visible. To distinguish them, their texts are in blue in the popup list of overlapping tasting notes. In the given example, there is only one tasting note that contains the sequence of words "glass, offering aromas of ripe raspberries" and its text is in blue in the popup list, see Figure 9(b).

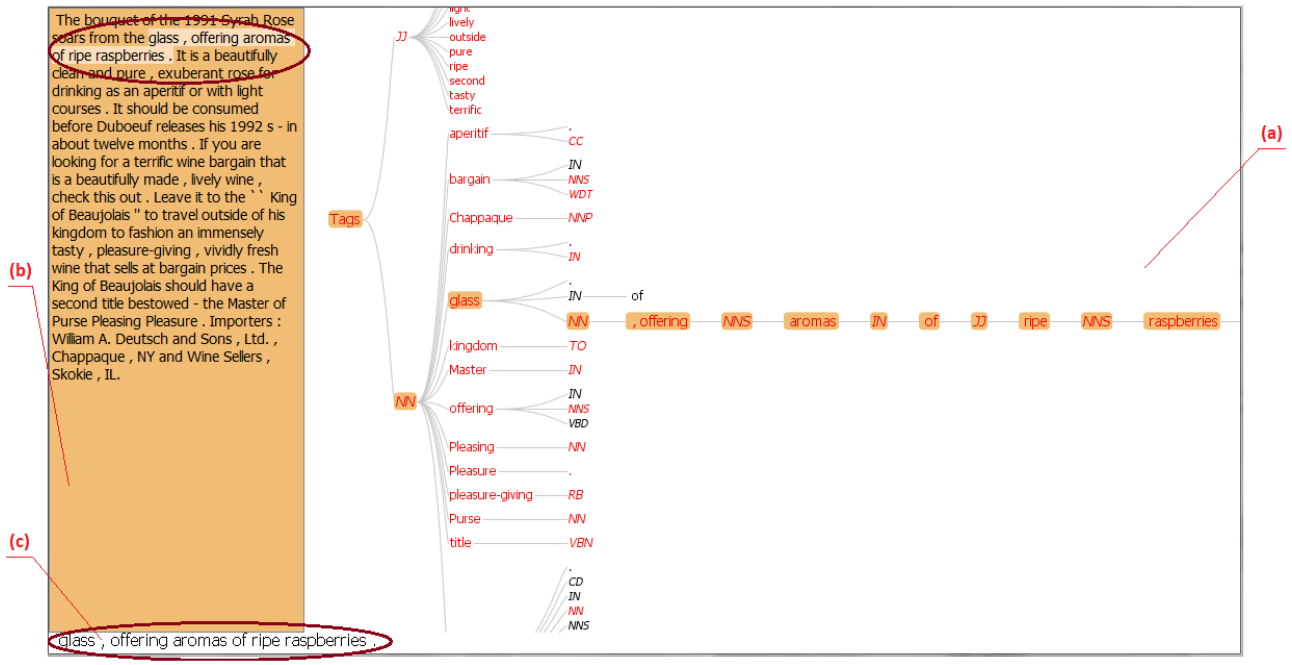
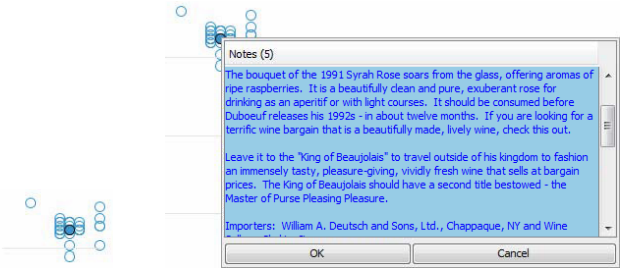


Figure 8. Word tree visualization consisting of three basic components. The tree node labels corresponding to words that are contained by the root tasting note are colored in red.



(a) (b)

Figure 9. Highlighting of tasting notes in the scatter plot that contain the currently sequence of words constructed by the word tree.

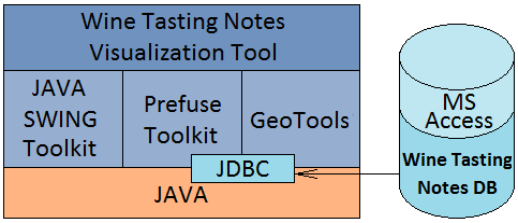


Figure 10. Software architecture of the tool.

C. Implementation Aspects

The tool's software architecture can be represented by four logical layers as shown in Figure 10. Because the original database containing the wine tasting notes was created using Microsoft Access®, we also decided to use this database management system (DBMS). The programming language that we decided to use for the implementation of the application is JAVA. We used four open source JAVA libraries to implement the required functionalities. The JDBC library was employed for establishing connectivity between the JAVA programming language and the database [20]. The graphical user interface was created with the aid of the JAVA Swing Toolkit [21]. We used the Prefuse Toolkit for the following interactive visualizations: the scatter plot, the bar chart diagrams and the word tree [22]. The world map visualization was created by the functionalities of the JAVA GIS Toolkit GeoTools [23].

Scalability Issues: The selected visualization approaches are appropriate for representing large amounts of data. From a theoretical point of view, there is almost no restriction placed upon the number of visualized tasting notes. The scatter plot and the bar chart diagrams provide different interaction techniques that give users the opportunity to focus on a subset of elements for further exploration. The tag cloud visualization represents the text of one tasting note using different font sizes for its words. With an increase in the number of tasting notes, the proportions between the font sizes of the words in the visualized text will be affected – not the number of the words. This feature makes the tag cloud visualization even more attractive. The word tree facilitates for users to explore the correlations beginning from a set of visible words and proceeding with other words which appear

as a result of previous choices. This approach allows the exploration of unbounded sets of words. However, increasing the number of tree nodes makes it more complicated to browse through the tree and to preserve the mental map. Because there is no upper limit for the number of tasted wines produced in the visualized countries, the world map is not restricted by the amount of visualized data either.

That said, our implementation currently imposes some restrictions on the functionality of the tool. The scatter plot together with the filters, the tag cloud visualization and the bar chart diagrams perform well for a number of up to 3,000 tasting notes. For more elements, the tool becomes slower and thus less interactive. One way of improving the application's performance is to migrate the database to another, more efficient DBMS. The current DBMS and the database schema are in fact the main bottleneck of our implementation. Because of the inappropriate design of the wine database, the word tree visualization cannot be efficiently built for more than 60 tasting notes. In order to overcome this restriction, the design of the database should be modified in such a way that it represents the tree structure of the words in the tasting notes. The response time of the world map view for standard user interactions takes about six seconds, which is a relatively long time. Here, we have to find out whether the GeoTools Toolkit API may provide a solution to this performance problem or if we have to move to another library.

VI. Results

Our tool offers possibilities for the exploration, the analysis and the presentation of large and complex amounts of data in ways so that linguists can make use of them in investigations of the structure of texts and discourses and of the lexical resources that languages have for the expression of meaning domains. Not only can visual images communicate concrete information, but they can also represent abstract information in the form of visual imagery, which is of particular significance in the case of wine descriptions of subjective sensory modal representations, which by nature are transitory and volatile. Through these techniques, textual data can be visually represented at a glance and can be interactively explored at the same time. This is clearly an innovation in linguistic research. The most essential part of wine descriptions is concerned with the description of passing sensory perceptions. They are captured by our visualization approaches in the form of scatter plots, tag clouds, word trees, and bar chart diagrams. Given the availability of tagged corpora, dynamic visualization techniques open up for linguistic advances through typological comparisons across different text types, different times and different languages. For instance, with the aid of the various filters of metadata we can explore linguistic patterns across subsets of tasting notes, subsets of ratings of wines, or subsets of grapes. And we can apply filters, such as only tasting notes containing more than 400 words, only sweet wines, only wines from Spain, etc., in various different combinations. The tool provides direct feedback in the form of interactive visualizations and is immediately able to answer questions such as: Do tasting notes have the same format across time? Or how do wines pattern that are described with the attribute 'sweaty saddle'? Thus it offers the possibility for linguists to play around with the variables and get a picture of the differences in the distribution of tasting notes in relation to the changes we make using the different filter settings straight away. We also obtain statistical information related to choices that we make in the form of bar chart diagrams and tag clouds. These functions are particularly important for the setting of parameters more accurately and for the subsequent formulation of new hypothesis for corpus investigations of text.

VII. Conclusion and Future Work

This work is concerned with various approaches for visualizing wine tasting notes that can be used to support linguistic analyses. Our data sources are large databases containing tasting notes and metadata related to the wines tasted. Linguists are interested in the language of such texts and the possibilities offered by the language to describe sensory perceptions to better understand descriptions of them. The purpose of our tool was to visualize this data in a way that would help linguists to get a better picture of wine descriptions. All solutions presented in this paper were carefully discussed with linguists during their development.

There are several improvements that can be made to enhance the visualization tool for wine tasting notes. In Subsection V-C, we discussed several problems with the current DBMS and the database schema. An improvement of this situation would be one of the fist candidates for the next software revision.

Another issue would be the tag clouds that can be improved. There are function words that occur very frequently in general language like "a", "the", "of", etc. They are visualized by the largest font sizes and therefore attract the attention of the users from other words that are more important and more interesting for linguistic analyses. An obvious solution to avoid this problem would be to create a user-defined black list of words that could be disregarded and excluded from the calculations. The tool could thereby avoid their overestimation.

The world map visualization and its performance could be improved too. It would be useful to add more interactive features to the map visualization. For example, the map could be extended by visualizing vintages, i.e., time-series data. Another idea would be to add an interactive control for tracing the wine production density in different countries on the basis of the wines' vintages. Range sliders or other controls could be integrated to change the time period of the data visualized on the map.

Finally, we recognize that our visual analyses are also related to tasks in the field of Sentiment Analysis. It would be very interesting to develop our tool also in this direction, see for example the handbook chapter [24].

ACKNOWLEDGMENTS

We would like to thank Ilir Jusufi for carefully proof-reading the final version of this paper.

REFERENCES

[1] M. Prangova, "Visualization of sensory perception descriptions," Master's thesis, Linnaeus University, School of Computer Science, Physics and Mathematics, Växjö, Sweden, 2010.

[2] C. Paradis, "A sweet nose of earth, smoke, cassis and cherries: Descriptions of sensory perceptions in wine tasting notes," in *Proceedings of the 7th AELCO International Conference*, Toledo, Spain, 2010.

[3] G. Morrot, F. Brochet, and D. Dubourdieu, "The color of odors," *Brain and Language*, vol. 79, no. 2, pp. 309–320, 2001.

[4] A. Kerren, "Visualization of workaday data clarified by means of wine fingerprints," in *Proceedings of the INTERACT '09 Workshop on Human Aspects of Visualization*, ser. LNCS, vol. 6431. Springer, 2011, pp. 92–107.

[5] J. C. Roberts, "Exploratory visualization with multiple linked views," in *Exploring Geovisualization*, A. MacEachren, M.-J. Kraak, and J. Dykes, Eds. Amsterdam: Elseviers, December 2004. [Online]. Available: http://www.cs.kent.ac.uk/pubs/2004/1822

[6] C. Ahlberg and B. Shneiderman, "Visual information seeking: tight coupling of dynamic query filters with starfield displays," in *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, ser. CHI '94. New York, NY, USA: ACM, 1994, pp. 313–317. [Online]. Available: http://doi.acm.org/10.1145/191666.191775

[7] IBM Research, "Many Eyes," http://manyeyes.alphaworks.ibm.com/manyeyes/, 2011.

[8] M. Wattenberg and F. B. Viégas, "The word tree, an interactive visual concordance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1221–1228, November 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1477066.1477418

[9] D. Keim and D. Oelke, "Literature Fingerprinting: A New Method for Visual Literary Analysis," in *IEEE Symposium on Visual Analytics Science and Technologie*, Sacramento, CA, USA, 2007, pp. 115–122.

[10] G. Salton and A. Singhal, "Automatic text theme generation and the analysis of text structure," Cornell University, Ithaca, NY, USA, Tech. Rep. TR94-1438, 1994.

[11] O. Kaser and D. Lemire, "Tag-Cloud Drawing: Algorithms for Cloud Visualization," in *Proceedings of Tagging and Metadata for Social Information Organization (WWW '07)*, Banff, Canada, 2007.

[12] F. B. Viegas, M. Wattenberg, and J. Feinberg, "Participatory visualization with wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, pp. 1137–1144, November 2009. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2009.171

[13] K. Koh, B. Lee, B. Kim, and J. Seo, "Maniwordle: Providing flexible control over wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 1190–1197, November 2010. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2010.175

[14] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information Visualization*, vol. 7, pp. 118–132, April 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1466620.1466622

[15] B. Santorini, "Part-of-speech tagging guidelines for the penn treebank project," University of Pennsylvania, Department of Computer and Information Science, Philadelphia, PA, USA, Tech. Rep. MS-CIS-90-47, 1990, (3rd Revision).

[16] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of the IEEE Symposium on Visual Languages (VL '96)*, 1996, pp. 336–343.

[17] J. Heer and S. K. Card, "Doitrees revisited: Scalable, space-constrained visualization of hierarchical data," in *Proceedings of the working conference on Advanced Visual Interfaces (AVI '04)*. New York, NY, USA: ACM, 2004, pp. 421–424.

[18] S. K. Card and D. Nation, "Degree-of-interest trees: A component of an attention-reactive user interface," in *Proceedings of the working conference on Advanced Visual Interfaces (AVI '02)*. New York, NY, USA: ACM, 2002, pp. 231–245.

[19] D. A. Carr, N. Jog, H. Prem Kumar, M. Teittinen, and C. Ahlberg, "Using interaction object graphs to specify graphical widgets," University of Maryland, Department of Computer Science, College Park, MD, USA, Tech. Rep. CS-TR-3344, 1994.

[20] Oracle, "The Java Database Connectivity," http://java.sun.com/products/jdbc/overview.html, 2011.

[21] ——, "JAVA Swing Toolkit," http://java.sun.com/docs/books/tutorial/ui/overview/intro.html, 2011.

[22] The Berkeley Institute of Design, "Prefuse Visualization Toolkit," http://prefuse.org/, 2009.

[23] GeoTools, "The Open Source Java GIS Toolkit," http://www.geotools.org/, 2010.

[24] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing*, 2nd ed., N. Indurkhya and F. J. Damerau, Eds. Chapman and Hall/CRC, 2010.