Highly interactive and natural user interfaces: Enabling visual analysis in historical lexicography

Roberto Therón
Department of Computer Science and Automation
University of Salamanca
theron@usal.es

Carlos Seguín
Department of Computer Science and Automation
University of Salamanca
seguin@usal.es

Laura de la Cruz
Department of Computer Science and Automation
University of Salamanca
laura_cs@usal.es

María Vaquero
Department of Computer Science and Automation
University of Salamanca
mvaquero@usal.es

ABSTRACT

Information technology, through the advances provided by computational linguistics and related disciplines, has opened the door to previously unthinkable possibilities of study in linguistics. The wealth and diversity of sources that is now available is fundamental to the understanding of language evolution and dictionary-making. However, these advancements are paired with a paradigm shift, in which both the user needs and the modes in which the users interact with technology have changed so much and so rapidly, that modern lexicography would need to resort to a new generation of tools to support its tasks. We present our work developed for the Nuevo Diccionario Histórico del Español (NDHE), in which the challenges of enabling deeper insight and supporting new user's tasks in diachronic linguistics have been approached from a human-computer interaction perspective. Thus, in contrast to what has happened in other disciplines in which visual analytics has focused its efforts since earlier, the analysis tools that are made now in the hands of the experts usually provide a volume of "raw" data so vast, that the data themselves can greatly hinder the work of experts. The linguistics community has already recognized the key importance of user-friendly interfaces. However, neither more powerful tools (in terms of automatic processing) nor user-friendliness alone are sufficient to support typical analytical tasks that take out the most from the multidimensional and ever-growing data stored in corpora and dictionaries. This paper discusses the benefits of producing corpus and dictionary analysis tools that go beyond user-friendliness and presents, interactive visual analysis tools produced for the NDHE and its sources.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: User/Machine Systems – *Human information processing.*

General Terms

Design, Human Factors, Languages.

Keywords

eLexicography, visual analysis, natural user interfaces

1. INTRODUCTION

The advancement and expansion of Internet and new technologies have facilitated access to a wealth of information hardly graspable. While this development is an exciting change, management, exploration and analysis of the flow of linguistic information is becoming a problem both on personal and social levels. Toffer coined this situation as "Information Overload" [13]. Too much information about a subject generates known difficulties to understand and to make decisions based on the data, so that too much information can be as damaging as a shortage of it. Despite the progress, automatic processing itself is not a solution to the problem as it lacks the basic and inherent skills to human reasoning, so it is important to combine it with the unique language and visual skills of humans. One research field, that aims to take advantage of these skills, is information visualization (InfoVis). In the words of Card et al. 0, information visualization is "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition."

McEnery and Hardie [15] point out user-friendliness as one of advantages of fourth generation corpus analysis tools and, in the conclusion of their textbook, they state that "if corpus techniques are to be used at all – let alone embraced – across a breadth of humanities subject areas, it is clearly important that the corpora should be made accessible via a user-friendly interface". Although "user-friendliness" in user interfaces is desirable, it is a vague notion and the actual challenge when it comes to the design of any advanced corpus analysis tool is to obtain a thorough understanding of the diverse community of users and the tasks that must be accomplished; thus, in getting beyond the vague quest for user-friendly systems, software designers should focus on specific goals that include well-defined system engineering and measurable human-factors objectives [20].

In recent years, visual analytics has emerged as a science that merges the intuition of humans with the power of automatic data processing, visualization and interactive environments, with the purpose of assisting researchers from various fields in solving multidimensional problems and examining complex systems [16], Keim et al. define visual analytics more precisely as "an interactive process that involves information gathering, data preprocessing, knowledge representation, interaction and decision making". We believe that visual analytics will play a key role in the future of digital access to textual cultural heritage in general, and in the advance or corpus linguistics in particular.

This research combines techniques from corpus and computational linguistics, data and text mining, information retrieval and information visualization, and human computer interaction, in order to cope with the information overload

problem associated with the analysis of linguistic corpora. This effort is part of a pioneering work that started a line of innovative research in the field of diachronic linguistics, more specifically in the analysis of historical dictionaries and corpora, through the development of interactive visual tools that allow the user to perform the analysis of phenomena related temporal evolution of the lexicon [21][22]. Although our research and developed tools are focused on the Spanish language, the visual solutions proposed can be extended to any language, provided that an appropriate historical corpus is available.

The rest of the paper is organised as follows: In section 2 we summarize the most relevant research related to our work. In the third section we discuss the challenges and opportunities of using visual interactive tools to enable a more appropriate access to the wealth of documents stored in corpora and dictionaries. Then, in the fourth section we present some results obtained by introducing natural user interfaces that can be coupled with the proposed visual interactive tools. Finally we expose the main conclusions gained from the experience of enabling visual interactive means for both analysts and regular users of corpora and dictionaries.

2. RELATED WORK

The use of visualization to gain a deeper understanding of abstract concepts, structures, relationships, etc. has a long history across domains. Within statistical approaches to linguistics, scatter plots, bar graphs, phylogenetic trees etc. can be regarded as standard visual tools that linguists have exploited extensively in order to obtain and to communicate their findings or exemplify particular hypothesis, models and theories. These valuable abstraction tools have benefited from computational and corpus linguistics, however, these fields could profit massively from the state-of-the-art in information visualization and visual analytics. Most of the visualisation techniques used in linguistics are static and print-oriented; the introduction of thoroughly designed interactive visual tools fully integrated with automatic data processing algorithms will help the aforementioned linguistic fields to take the big leap.

This approach has started to arise quite recently. The most innovative works in corpus linguistics regarding visualisation have often been done in collaboration with specialists in visualisation-related fields. That is the case of the recent work of Gregory and Hardie [11], which take advantage of standard geographical information systems and explore procedures by which historical corpora can be mined in order to obtain datasets that can be presented in visual form as maps. Despite the unquestionable value of this approach, the best value from interactive visualisations can be obtained when such visualisations are designed ad hoc, with the actual tasks and users in mind, rather than just finding ways of transforming data to fit into already existing systems and/or techniques.

In [23], a review of computational models for the integration of visual and linguistic information can be found. It is the first article that categorizes the research works that have dealt with the correspondence problem, namely how to associate visual events with words and vice versa. For a review of the main approaches to the problem we refer the reader to [4]. In spite of the great achievements reached in computational linguistics, the introduction of interactive visual tools in linguistics is a very recent tendency and is not yet popular, particularly among linguistics professionals. At present there are few works dedicated to the interactive visualization of linguistic data as compared to the well spread use of corpus tools that face such as allowing the exploration of syntactic patterns in personal corpora without doing manual annotation [17][18]. Specialized tools for language analysis still make little use of visualizations, and visualization tools for language related information (LInfoVis) [7]. This situation is changing, as data visualization has reached all knowledge domains; thus, for instance, Zhao et al. [27] have approached discourse analysis from a interactive visual perspective aimed at assisting computational linguistics researchers to explore, compare, evaluate and annotate the results of discourse parsers.

The advances are mainly related to document and text mining [24][9], with excellent and stunning results such as the ones available at Many eyes, a website devoted to popularization of data visualization, [25][26][5]. More directly related to the theme of the project, several research works can be mentioned here, such as [14], in which software for the visual exploration of a dictionary of Warlpiri is presented, or [8], in which a tool for the presentation and exploration of grammatical trees is proposed, and even successful business examples, such as VisualThesarus1, in which interactive visual maps (node-link diagrams) are used to visualize word relationships for various languages.

Word Clouds (and variations of this visualisation technique, [3]) have been used in recent years to present the most salient lexical items in a text, with the most salient words only shown, and with the size of a word increasing with its salience, thus presenting various ways to visualise a text such that the changing nature of syntactic patterns throughout the text become visible to the analyst.

In a related attempt to visualize lexical change, Kempken et al. described several treemap techniques used to visualize the productivity of rule sets in deriving nonstandard spellings in old German texts, among other aspects [12]. Rohrdantz et al. [19], present a new approach to detecting and tracking changes in word meaning by visually modelling and representing diachronic development in word contexts.

Finally, the research presented in this paper is part of a more ambitious project that includes several coordinated visualization techniques for a complete analysis in diachronic linguistics, so, in previous work, we have focused on the visual analysis of meaning evolution [21][22][10]. However, to our knowledge, few works have undertaken the task of addressing linguistic problems that have been traditionally studied with the aid of corpus linguistics, by means of visual analytics techniques and methodologies.

3. ENABLING VISUAL ANALYSIS AND CORPORA

Recently, the linguistic sources have been recognized as a key element of human cultural heritage. This fact is coupled with the need to integrate the ancient goals of philology with rapidly emerging methods from fields such as Corpus and Computational Linguistics [6]. However, the gold coin of such promise has two sides: on one the side, tools for exploring a corpus can be excellent aids to the linguist; on the other side they also, crucially, limit and define what we can do with a corpus [15].

At any rate, regarding the analysis of very large corpora, is impossible to avoid using these computer tools for practical reasons: the corpus alone solves few (if any) problems for a linguist. Tools that allow linguists to manipulate and interrogate the corpus data in linguistically meaningful ways unlock its potential [15].

On another level, nowadays many of this historical sources are being made available as public services, and researchers are

not the only type of user of linguistic corpora and specialized dictionaries anymore.

These are the main aspects that motivate our work: the use of automated processing of large amounts of data can greatly expand the range of research questions that can be addressed in many linguistics fields; but, if the user interfaces that are attached to the computational algorithms do not take into account basic human factors and the abilities, expertise and background of the users, many of its potential uses do still lie beyond the reach of what can be done with the computers.

In the following section an example of how a user interface can diminish the user's understanding of the data stored in a corpus. A visual interactive tool that solves this problem and enables the analysis of the same data source is also described.

3.1 Visual Analysis of RAE's Corpora

The Royal Spanish Academy (Spanish: *Real Academia Española*, *RAE*) is the official royal institution responsible for regulating the Spanish language. Among other corpora, the RAE has compiled two large corpora: a reference corpus of modern Spanish, called CREA (Corpus de Referencia del Español Actual), and a historical corpus, known as CORDE (Corpus Diacrónico del Español). Although we have developed tools for both corpora, in the following we refer only to CREA. It contains 200 million words of running text, providing an empirical basis for lexicographic and grammatical research. The corpus can be accessed online (http://corpus.rae.es/creanet.html) and its user interface can bee seen in Figure 1.

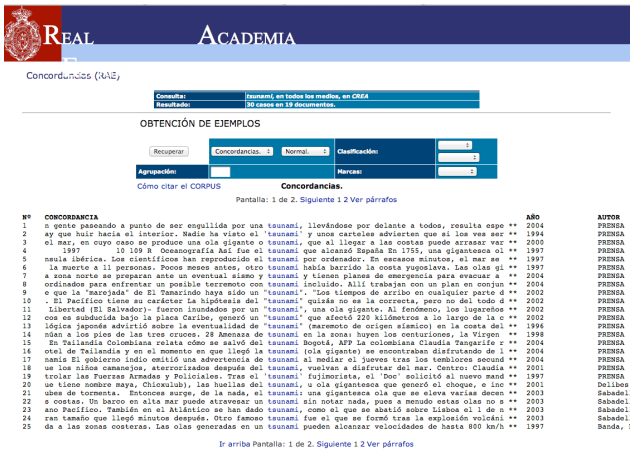

Figure 1. GUI for the exploration of CREA. The user is exploring the use of the word *tsunami* in Spanish in recent years.

Although the service indeed provides access to the very interesting content of the corpus (see for instance, the coucurrences of the word *tsunami* Figure 1) several issues related to the user interface, otherwise common in many other corpora interfaces. Due to the lack of space, we describe two of them:

1) the user queries the corpus by means of a combination of menu choices; while this may be valid for testing hypothesis, the fixed menu options prevent knowledge discovery (usually related to the data relationships, not to the raw data) in many cases.

2) while the corpus stores many documents relevant to a word use (30 cases in 19 documents in the example in Figure 1), only 25 cases are shown per page, so the users has to navigate the different pages, losing the context of the analysis.

Furthermore, after several months using the service, we discovered an unexpected issue: the interface hides relevant information. Since the information is conveyed in tabular form, the fact that an horizontal scroll was needed in order to access further columns went unnoticed (we conducted informal usability studies and that was the case for almost all users). Figure 2 shows a mockup of the problem: the combination of the viewport and tabular layout hinders the analysis.

Figure 2. Due to both the technology available at the moment of development and the design of the user interface, a portion of the information available in the corpus is hidden to the user, which in many cases does not notice the need of an horizontal scroll in order to access all the relevant data.

To be able to gain a deep understanding of the enormous quantities of data stored in CREA, methodologies originated in fields such as information visualization, statistics, artificial intelligence, data mining, computational linguistics, graphic design, psychology of perception and human-computer interaction, are employed to solve parts of the specific problem.

Figure 3 shown the interface we developed for the visual interactive analysis of CREA. It features several interactive linked views: 1) coucur simplified geographic map rence tree (summarizes all the use cases found); 2) topic histogram (color coded, summarizes frequency of use across topics); 3) bubble map (simplified geographic map that conveys the use across countries); 4) timeline (conveys the use through time); 5) Quotation and document info; 6) View options for the coocurrences tree.

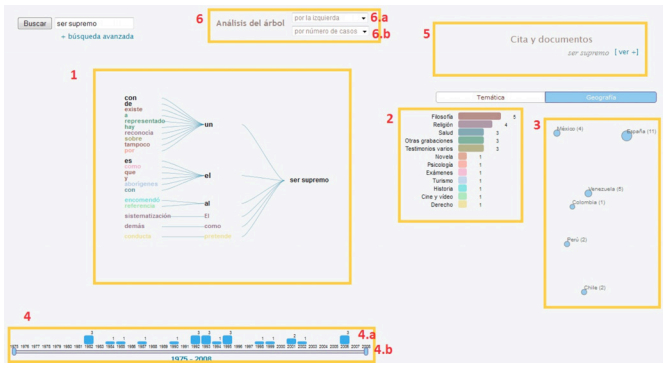

Figure 3. Overview of the visual interactive user interface developed for enabling a deeper understanding of the information of RAE's corpora.

Since all the relevant data to the case under analysis is provided, the presented visual interface enables exploratory analysis: the user can discover patterns or unexpected situations. All views are linked, so each interaction with a view is propagated to the other views. This way, as the user spends time exploring the data; she may filter some information out (select a particular

period of time, focus on a country or topic and so on). The combination of interactions creates a complex visual query that could not be reached with the conventional interface. Typically, the user will go on and new analysis questions triggered by the information conveyed will, in turn, trigger further interactions.

Another interesting feature is that although some data may be filtered out during an analysis session, this information is kept on the background in order for the user to not lose the problem context. For instance, see the example in Figure 4, where the user is exploring the word *Pearl*, she has decided to study the right coocurrences, the tree is ordered by frequency of use, and the color code refers to the topic. when the users hovers over the word *Jam* (in red color, that of the topic 'music'), the information related to the documented uses of the combination *Pearl Jam* is highlighted in all views, while keeping the context information. Thus, the user can easily see that *Pearl Jam* has been used in 25 of the 29 cases used under the 'music' topic; in 9 of the 18 cases in the year 1996, in 11 (45%) instances it was used in Spain, and so forth.

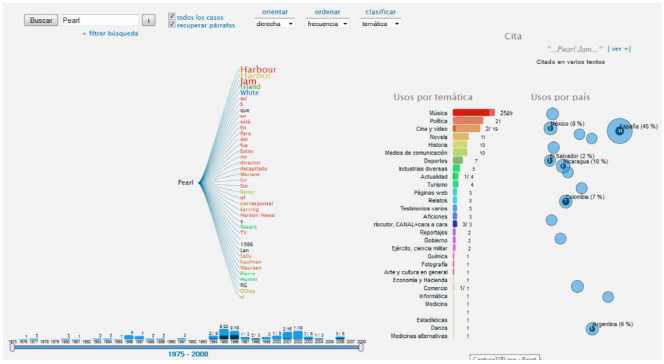

Figure 4. The user is interested in the use of the foreign word 'Pearl' in Spanish. By means of a series of interactions, the user can filter out the information that may answer her current analysis questions.

The dynamic nature of the proposed visual tool can be seen at http://www.youtube.com/watch?v=hK_aLTmCBGo.

4. NATURAL USER INTERFACES: ENRICHING THE ACCESS TO TEXTUAL CULTURAL HERITAGE

As mentioned before, another motivation of our work is the fact that the wealth of historical linguistic sources can be offered to other users beyond the academia.

The RAE has collected a vast amount of historical information and is currently developing different corpora and dictionary tools and services that will help in the editing of the Nuevo Diccionario Histórico del Español (NDHE). However, the online version of the RAE's dictionary is the service that receives almost all the traffic (over 40 million visits per month). Furthermore, most users do not explore the online dictionary but search for the meaning of a single word.

With the advent of touch screen mobile devices, tablets, and netbooks, motion sensor devices and speech recognition, and with the so called digital natives reaching puberty, we can expect the adoption of natural user interfaces (NUIs, user interfaces designed to feel as natural as possible to the user) in primary and secondary education within two or three years[1].

The term 'digital native' has prompted the debate over whether there is any adequate evidence for claims made about its implications for education [1]. However, there is an inherent potential in NUIs and students that are leveraging a large range of tools, applications and learning resources to make a personalized learning experience.

In order to enrich the digital access to all historical linguistic sources collected and made available by the RAE's, and go beyond the electronic imitation of the printed dictionary and the academic domain, we have developed a visual interactive tool that allow users to engage in virtual activities with movements similar to what they would use in the real world, manipulating complex textual content intuitively.

The user has access to the linguistic content using different entry points (CREA, CORDE, Diccionario Histórico del Español…) but the interaction links all the sources for a given word.

Different views for the each type of variables involved (diachronic, diatopic, diaphasic, etc.) are provided so the user is fostered to explore data. For instance, in Figure 5 the user is exploring a coocurrence tree.

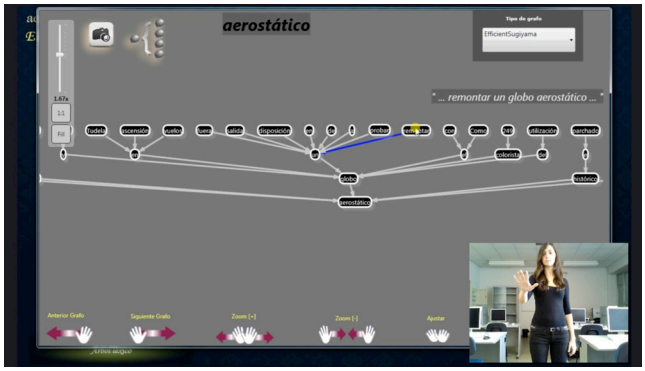

Figure 5. A user using gestures to navigate a coocurrence tree for the word *aeroestático* ('aerostatic'); the information is obtained from the CREA.

Opposite to the visual tool exposed in the previous section, the main purpose of the tool discussed here is to provide a fun environment and a fun experience rather than an effective means of analysis.

The system can be used by any person, but it has been designed with a student of secondary level as a target user.

Thus, several views and layouts are used to convey the information; for instance, the hierarchical information of the example in Figure 5 can be represented different graph layouts (force-directed, radial, etc..).

A similar approach has been followed for other type of information such as geographic data. Maps are used in different ways: small multiples to summarize (Figure 6), single maps with markers for location or choropleths to convey frequency of use.

[1] *NMC Horizon Report: 2012 (http://www.nmc.org/pdf/2012-horizon-report-K12.pdf).*

Figure 6. The user explores the geographic, diatopic and diachronic dimensions for the word *aeroestático*.

In order to enrich the experience, access to other type of available information is provided. In Figure 7 an image of an original lexicographic card is shown. The user can explore through all the cards related to a particular word.

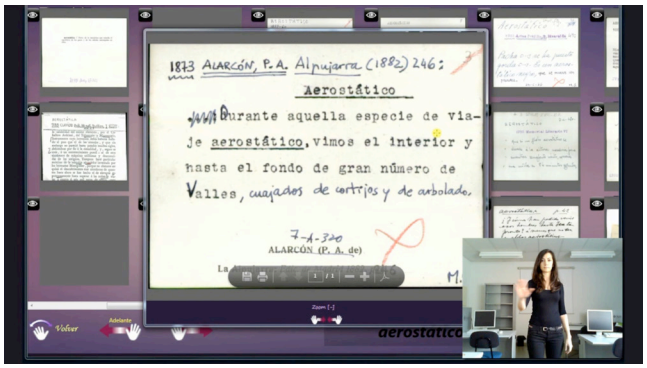

Figure 7. The user can explore other type of sources; in this case, she is exploring the lexicographic cards originally used by lexicographers for the first editions of the RAE's dictionary.

Other views include the entry of the word, if it exists, in the unfinished Diccionario Historíco del Español (see Figure 8), or the fragments of document to which the registered uses pertain (in CREA or CORDE).

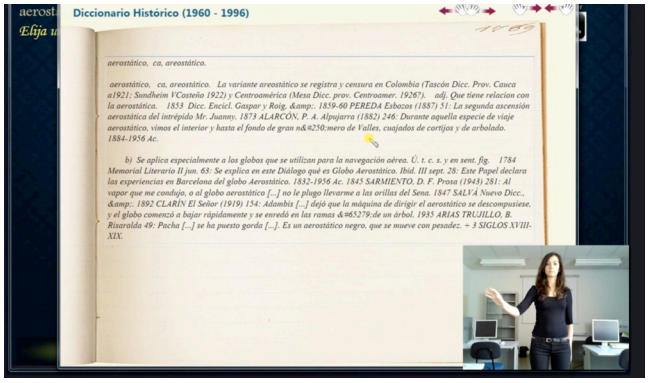

Figure 8. User reading the entry of *aeroestático* in the Diccionario Histórico del Español.

We have not conducted any formal usability studies for this gesture-based visual tool. However the user's feedback has been very positive and they reported an increased interest towards the language and its evolution, the geographic differences and the old lexicographic methodologies.

As it occurred in the previous section, the best way to assess the potential of our proposed tool is using the system. In the absence of this possibility, we refer the reader to this video: http://www.youtube.com/watch?v=o3wEHPV9wMI

5. CONCLUSIONS

We have discussed how using the right combination of novel representations and novel ways of interacting with them (i.e, gesture interfaces), it is possible to enhance the unique ability of expert analysis with the methods of automatic data processing. Firstly, it provides an overview of all data, and then, by enabling appropriate views for the different variables involved in the analysis (diachronic, diatopic, diaphasic, etc.), the analyst is fostered to explore data; as a result, through the intervention of her cognitive abilities (perception of patterns, atypical situations, etc..) and her experience, the discovery of aspects of the study that would otherwise be hidden or require a huge amount of time and effort, is enabled.

6. ACKNOWLEDGMENTS

The authors wish to thank Spanish Government project FI2010-16234, the participants of the usability evaluations and the insightful comments of the reviewers of this paper.

7. REFERENCES

[1] Bennett, Sue, Karl Maton, and Lisa Kervin. The 'digital natives' debate: A critical review of the evidence. *British journal of educational technology* 39.5 (2008): 775-786.

[2] Card, Stuart K., Jock D. Mackinlay, and Ben Schneiderman, eds. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[3] Chen, Ya-Xi, et al. Tagclusters: Semantic aggregation of collaborative tags beyond tagclouds. In *Proceedings of the* 10th *International Symposium on Smart Graphics*., SG 2009, Salamanca, Spain, May 28-30, 2009, Springer Berlin Heidelberg, 2009.

[4] C Collins, G Penn and S Carpendale. Interactive visualization for computational linguistics. In: *HLT '08: Proceedings of the46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, Morristown, NJ, USA, 2008. Association for Computational Linguistics, pp. 6–6.

[5] C Collins, S Carpendale and G Penn. Docuburst: visualizing document content using language structure. In: Proceedings of Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis'09), Eurographics Association, 2009, pp. 1039–1046.

[6] Crane, Gregory, and Anke Lüdeling. Introduction to the special issue on corpus and computational linguistics, philology, and the linguistic heritage of humanity. *Journal on Computing and Cultural Heritage (JOCCH)* 5.1 (2012): 1.

[7] C Culy and V Lyding. Double tree: an advanced KWIC visualization for expert users. *In: 14th International Conference Information Visualisation*, 2010, pp. 98–103.

[8] D Derrick and D Archambault. TreeForm: Explaining and exploring grammar through syntax trees. *Lit Linguist Computing,* 2009; pp. fqp031.

[9] A Don, E Zheleva, M Gregory, S Tarkan, L Auvil, T Clement, B. Shneiderman and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. *In CIKM '07: Proceedings of the sixteenth ACM conference on information and knowledge management*, New York, NY, USA, 2007. pp. 213–222.

[10] Esteban, Andrés, and Roberto Therón. Corpusexplorer: supporting a deeper understanding of linguistic corpora. In *Proceedings of the* 11th *International Symposium on Smart Graphics*. Springer Berlin Heidelberg, 2011.

[11] Gregory, Ian N., and Andrew Hardie. Visual GISting: bringing together corpus linguistics and Geographical Information Systems. *Literary and linguistic computing* 26.3 (2011): 297-314.

[12] S Kempken, T Pilz and W Luther. Visualization of rule productivity in deriving nonstandard spellings. In: *Proc. of SPIE-IST Electronic Imaging (VDA '07)*, vol. 6495, 2007.

[13] Levy, David M. Information Overload. *The Handbook of Information and Computer Ethics* (2008): 497.

[14] C D Manning, K Jansz and N Indurkhya. Kirrkirr: Software for Browsing and Visual Exploration of a Structured Warlpiri Dictionary. *Lit Linguist Computing* 2001; 16(2):135–151.

[15] McEnery, Tony, and Andrew Hardie. *Corpus linguistics: method, theory and practice*. Cambridge University Press, 2011.

[16] A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler (2008). Visual Analytics: Scope and Challenges. In S. Simoff, M. Böhlen, & A. Mazeika (Eds.), *Visual Data Mining* (pp. 76–90). LNCS 4404. Berlin: Springer-Verlag.

[17] O'Donnell, M. 2008. Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of the ACL-08:HLT Demo Session (CompanionVolume),* Columbus, Ohio, June 2008. Association for Computational Linguistics. pages 13–16.

[18] O'Donell, M 2011. Visualising patterns in text, *Keynote Talk at the 29th Annual Conference of the Spanish Association of Applied Linguistics (AESLA*). Salamanca.

[19] Rohrdantz, Christian Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim and Frans Plank. Towards Tracking Semantic Change via Visual Analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT).* 2011. Portland, Oregon

[20] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, and Steven Jacobs. 2009. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (5th ed.). Addison-Wesley Publishing Company, , USA

[21] Theron, Roberto, and Laura Fontanillo. Diachronic-information visualization in historical dictionaries. *Information Visualization* (2013): 1473871613495844.

[22] R. Theron, L. F. Fontanillo, A. E. Marcos, and C. S. Herrero. Visual analytics: A novel approach in corpus linguistics and the Nuevo Diccionario Histórico del Español. *In Proc. of III Congreso Internacional de Lingüística de Corpus*, 2011.

[23] R Srihari. Computational models for integrating linguistic and visual information: A survey. *Artif. Intell. Rev* 1995, 8(5-6):349–369.

[24] F van Ham, M Wattenberg and F. B. Viegas. Mapping text with phrase nets. *IEEE Transactions on Visualization and Computer Graphics* 2009; 15:1169–1176.

[25] F B Viegas, M Wattenberg, F van Ham, J Kriss, and M McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics* 2007; 13(6):1121–1128.

[26] M Wattenberg and F B Viégas. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics* 2008; 14(6):1221–1228.

[27] D Zhao, Jian, et al. Facilitating Discourse Analysis with Interactive Visualization. *Visualization and Computer Graphics, IEEE Transactions on* 18.12 (2012): 2639-2648.