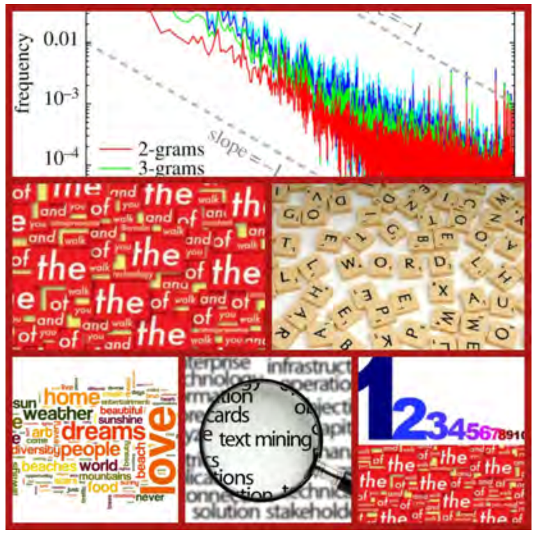
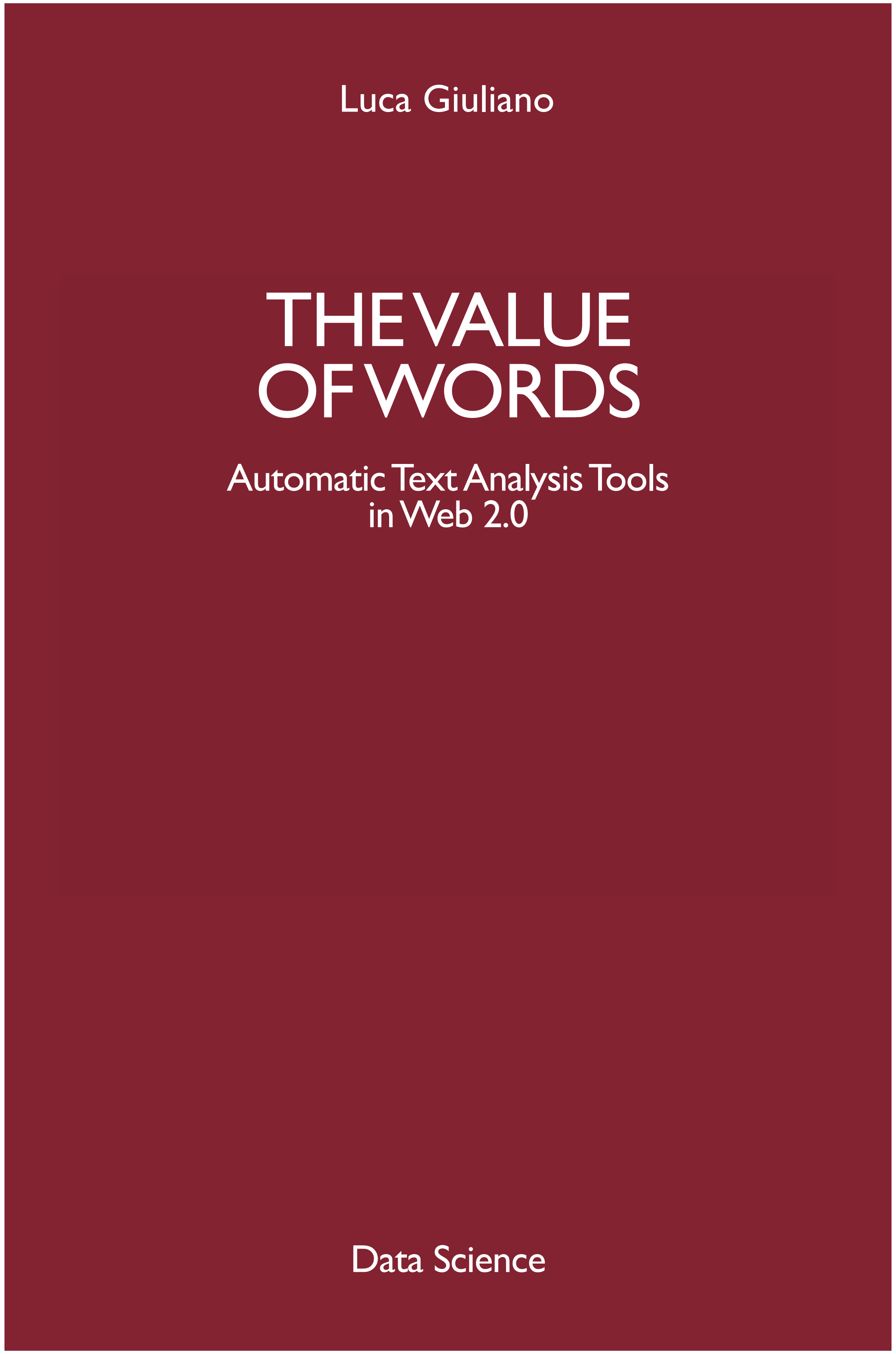
Luca Giuliano

THE VALUE OF WORDS

Automatic Text Analysis Tools in Web 2.0



Data Science



Luca Giuliano

THE VALUE OF WORDS

Automatic Text Analysis Tools in Web 2.0

Data Science

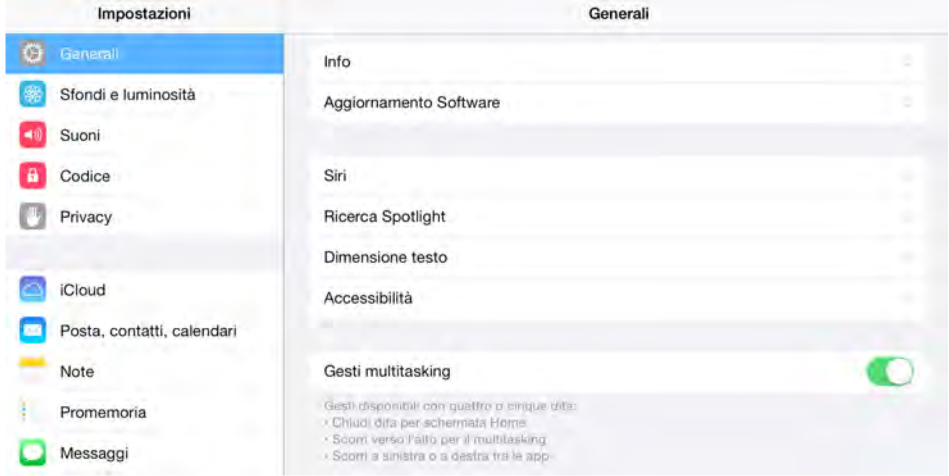
Luca Giuliano
The Value of Words. Automatic Text Anaysis Tools in Web 2.0.
Rome : Department of Statistical Sciences [2014] 112 p.
ISBN 978-88-908757-2-4

To view this book, you must have an iOS device with iBooks 3.0 or later and iOS 5.1 or later, or a Mac with iBooks 1.0 or later and OS X 10.9 or later. If you find errors or omissions, please contact the author: Luca Giuliano, luca.giuliano@uniroma1.it. A PDF version of this book is available at:

http://www.dss.uniroma1.it/it/ricerca/pubblicazioni

How To enable Multitasking Gestures on your iPad

By default the Multitasking Gestures is disabled on your iPad. To enable it, go to **Settings**>**General** and about halfway down the settings you will see Multitasking Gestures. Set the slider to **ON** and you will now have these new navigation methods available to you.



Generali

Info

Aggiornamento Software

Siri

Ricerca Spotlight

Dimensione testo

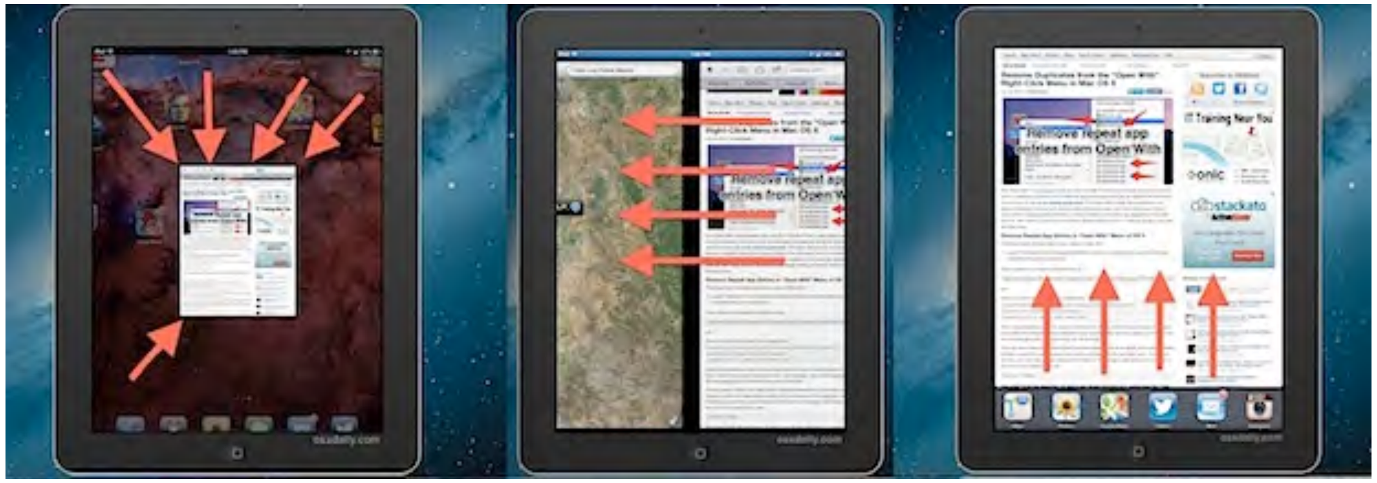
Accessibilità

Gesti multitasking ●

Gesti disponibili con quattro o cinque dita:
• Chiudi dita per schermata Home
• Scorri verso l'alto per il multitasking
• Scorri a sinistra o a destra tra le app

The concept is simple: By using 4 or 5 fingers, you can move between open applications on your iPad by swiping left or right. If you follow a link to a website, this gesture allows you to quickly go back to the book.

Furthermore, you can minimize an app by pinching your fingers together and you can view the multitasking bar by swiping up on your iPad.



Chapter 1

Worlds made of words

Accustomed, as we are, to multimedia communication, we tend not to give enough importance to words and writing. The Internet is first of all a world of words.

1

Computer science, Statistics, Linguistics

Everyday an immense amount of data and information transits on the Web and the advent of Web 2.0 has intensified these movements with Facebook and Twitter messages, blogs and comments, contributions on Wiki platforms, participations to forums and discussion groups, web pages and news. Meanwhile, the digitalization of bibliographical archives, of libraries and historical archives at all levels, including judicial, government acts and economic transactions, continues at great strides.

Fig. 1.1 The web transports every hour many billion words

According to EMC – Digital Universe Study 2014, the digital universe has grown to 4.4 zettabytes and “it is doubling in size every two years”. If we consider that 90% is made up of videos, the “words” (documents in txt, pdf and doc) moving on the Web could be estimated at 36 exabytes per month. This is equivalent to 8 CD-ROMs (640 megabytes) per person in the world. We are referring to an enormous amount of information about linguistic behaviours which was unthinkable until recently. It is a precious opportunity of analysis for the reading and understanding of social phenomena. An opportunity made more significant by Web 2.0, for the role of every user who autonomously puts on the Web the contents he/she is interested in, sharing with the other users judgments, opinions and choices.

Our management and calculus capabilities are not yet fully able to use the information extracted from the exponential growth of digital texts. However, the speed at which has improved in recent years the relationship between memorization and the management of great amounts of data, makes us hope well for the near future.

The analysis and textual data visualization tools we are about to explore in this application do not replace the commercial or research software used in computational linguistics and in linguistic statistics. For a more complete picture of the topic it is useful to consult a specific bibliography. A good starting point for those who are looking for a guide to the software available, is the portal BAMBOO DiRT – Digital Research Tools, and in particular the page: Analyze Texts.

Among the most important software developed in the research sphere, we can indicate: : Alceste, Alias-i, CATMA, DTM-VIC, Gate, IRaMuTeQ, Lexico3, TaLTaC2, TXM – Textométrie, T-LAB, WordStat.

The encounter between computer science and linguistics is strongly intertwined with statistics and mathematics. The synthesis takes place both at basic research and technological application levels, in particular in the fields of automatic translation, digital recognition, summary of spoken language and in the management of large information systems. The development of computer science since the 1950s has been appealing to linguists and, more in general, to scholars of similar subjects such as socio and psycholinguistics, bringing to the creation of computational linguistics (Jurafsky & Martin, 2000; Chiari, 2007).

Already before computer science had provided linguistic and text analysis with a quantitative approach, an important contribution from this point of view had been given by the inductive observations and the measurements of the Russian mathematician Viktor Bunyakowski : in 1848 he traced a sketch of the arithmetics of language taking into account word occurrences and length. Other similar contributions came later on from the inventors of stenography, like Jean-Baptiste Estoup (1916), and from Adolf Buseman's psycholinguistics with his study on childhood language (1925).

During the 1930s, George K. Zipf (1929) with his “principle of relative frequency” of graphic forms paved the way to lexicometrics and to real statistical linguistics. Between the '60s and the '70s, the French statistician Jean-Paul Benzécri began to apply a system of multivariate analysis (Correspondence Analysis) to linguistic data matrix which the words are points in a qualitative space.

Today, more than ever, the quantitative analysis of language is an extraordinary challenge for social science research methodology (Bolasco, 2013; Delahaye & Gauvrit, 2013; Mayer-Schönberger & Cukier, 2013). The analysis of communication processes cannot ignore the more and more widespread digitalization of information. Bibliographies, text and document classification, electronic publishing, the management of knowledge based on non-structured data, for example in the management of juridical documentation, continuously create new problems which can be solved only thanks to the intense collaboration between computer scientists, statisticians and the experts of substantive disciplines.

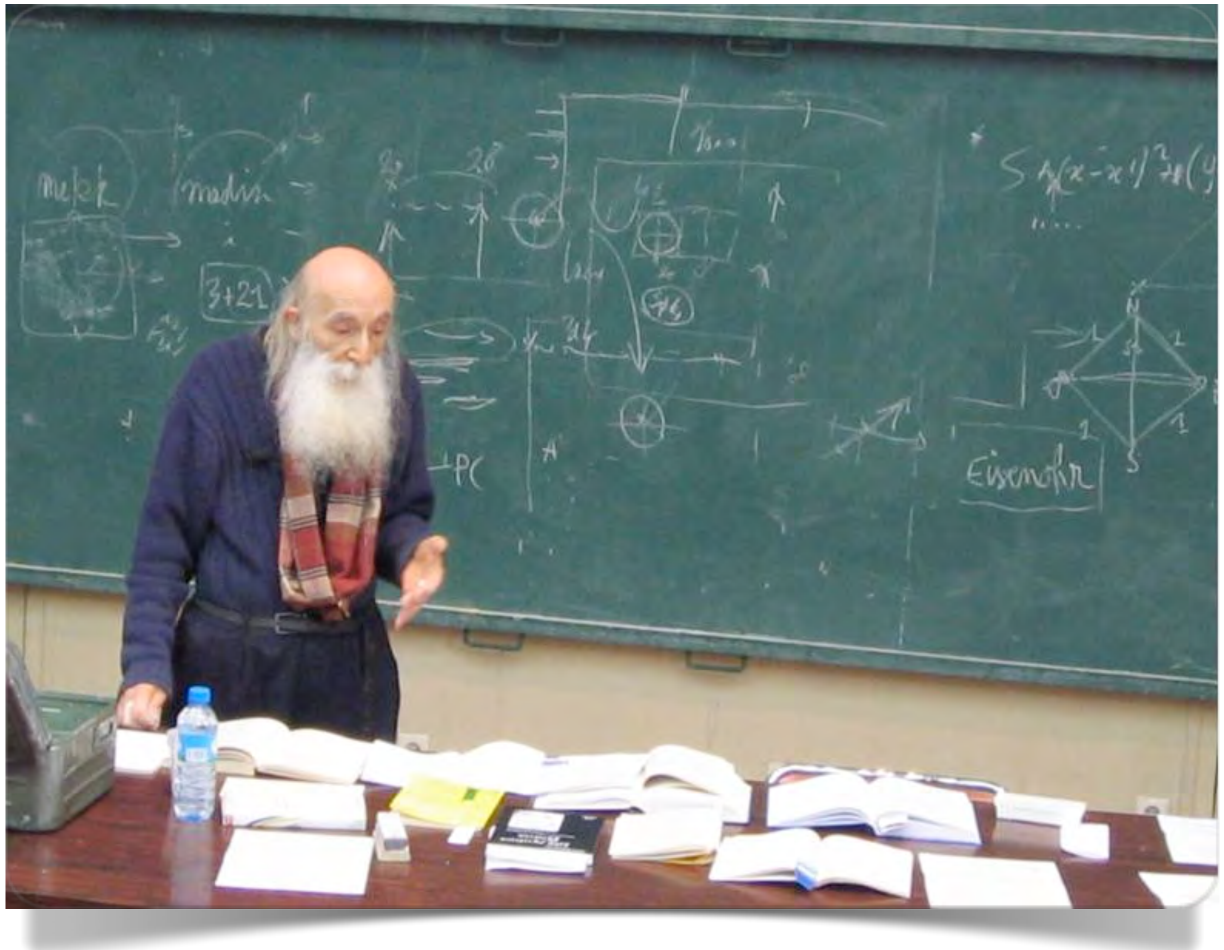


Fig. 1.2 Jean-Paul Benzécri

The intelligent management of memory

Document digitalization and communication via computer have made potentially available great amounts of information and data. However, this must not let us forget that, no matter how detailed and precise they may seem, they will always be incomplete in comparison to the reality and the complexity of the phenomena considered. Our knowledge is always intrinsically uncertain. For this reason, computer science and statistics have to be supported by the science of the uncertain: probability.

Text and language analysis cannot escape this limit. Every observation about lexicon and linguistic structures necessarily recalls the rules which characterize language as a social phenomenon and therefore creates problems of inference, which can be solved only with the support of the mathematical models of uncertainty. After all information itself – already in Shannon's traditional definition – is in inverse proportion to probability: An improbable event is more informative than a probable one.

In the interpretation of a text and in the analysis of its semantic content, information is expressed by a net of meanings of increasing and non-linear complexity which – to be efficient – has to adapt the models of our mind and our thought where originality, surprise and the unforeseen are inevitably features of intelligence and creativity.

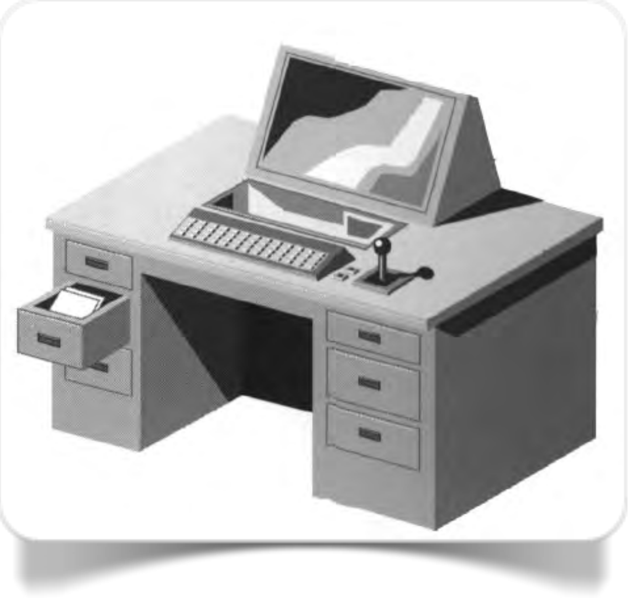


Fig. 1.3 Vannevar Bush's Memex

The first scholar to understand the advantages offered by technology in the “intelligent management of memory” was Vannevar Bush with his “memex” (memory extension), a tool for the consultation and indexation of archive papers invented during the 1930s. Memex should have reproduced the associative and non-linear paths of the mind (Bush, 1945).

The next step was made in 1960 by Ted Nelson with the project Xanadu, the prototype of a Literary Machine which was never finished (Nelson, 1981) , but which would have found its practical and revolutionary realisation some years later in Tim Berners Lee's World Wide Web (1991) and then in Ward Cunnigham's Wiki platform (1995).

The concept of hyperlink introduced by Ted Nelson has in little time become one of our natural ways of intellectual production to the point that we have actually forgotten its innovative content. A link is for us today a new sign of punctuation in written texts: It is a normative convention that helps to represent in a written text the reticular organization of our thought.

Research documents, if they have not been conceived in a structured form and with data classifiable a priori (surveying models, questionnaires, test), are irreducibly anchored to their original form: That of a letter, a newspaper article, a manifesto, a diary, a photograph, etc. Until the advent of digitalization, the “natural” documents produced for any reason in social life could not be subject to the empirical analysis of the sociologist if not with great fatigue and an enormous amount of time. W.I. Thomas and F. Znaniecki in 1920 ended their work *The Polish Peasant in Europe and America* analyzing in six years at least 1,000 letters and 8,000 various documents taken from the newspapers of the time, but nobody knows exactly how many they were, which of them were excluded and why. These problems emerged a long time after the publication, during a conference in 1938, when it was discovered that the majority of the original documentation had been destroyed (Madge, 1962).

A database constructed thanks to on the analysis of natural documents, typical of archives, is difficult to examine without reproducing the often singular and intuitive path of the scholar who has followed it. Today many non-structured documents are produced digitally “in a natural way”(Internet forums, emails, Facebook and Twitter text messages, blogs, online news) and can be therefore analysed both automatically with a lexicometric analysis and semi-automatically (CAQDAS - Computer Assisted Qualitative Data Analysis Softwares) using tools which guarantee the transparency of the researchers' methodological choices and follow step by step their conceptualisation, operativity and classification (Giuliano & La Rocca, 2008; Grbich, 2012).

Textual data information

As we have already pointed out, the web is a “living archive” with a great amount of information. One of the most interesting and efficacious results in the development of the Internet towards the interactivity of Web 2.0 is the possibility to have at one's disposal in addition to texts also on line tools capable of analysing the flows of information in order to identify the interpretative schemes and to reconstruct their meaning. We are obviously talking about first approach tools, which are rather simplified in some cases. The automatic analysis of textual data and text mining require more sophisticated approaches than the ones adopted by the tools so far described. However, as it will be possible to observe, they can be surprising, and as often happens with software, in some cases their potential goes beyond the objectives of those who have invented them, turning out to be tools open to the creative intervention of the user.

Abandoned to itself information is blurred by apparent disorder. Using some methodologies of automatic text analysis we can learn to manage better and with greater advantage the data at our disposal, selecting what we are interested in and synthetically showing the results.

Before proceeding with the analysis of some of these tools, which have been chosen for their simplicity, efficaciousness and reliability, we must understand some crucial concepts concerning the preparation of texts and the best way of using these in a mindful and proper way.

What do we mean when we talk about “on line texts”? In the Web we find a bit of everything: literary texts and poetry, newspaper articles, legislative documents, scientific papers, Wikipedia articles, social network messages, recipes, jokes, song lyrics. There is no limit to the variety of tests we can find on the Internet. Each of these texts, even if taken on its own, can be the subject of an automatic analysis. The result, however, will not be necessarily interesting: No more than it is to enter a library and count the books on the shelves or to enter a picture gallery and measure the frames of the paintings exhibited. Our interest towards one or more texts must be addressed by the questions we want to ask them or by the temporary answers which represent our work hypotheses. Texts, therefore, are analysable only if they are a corpus of texts; that is a series of texts which can be compared with certain points of view.

In many cases the corpus can be created easily: For instance, all the comments/remarks in a social network forum, chat or discussion group (closed corpus). Each comment/remark is a text or a text fragment. In some cases the comments/remarks could be grouped according to the characteristics of the writers (gender, age group, main interests, etc.) or simply according to the time in which they were posted. In other cases it can be necessary to create the corpus on the basis of certain representation and selection rules (sampled corpus). The corpus is always a consequence of the operative choices and decisions of who is carrying out the analysis. It is impossible to say beforehand if a corpus has been properly put together without considering the reason why it will be analysed.

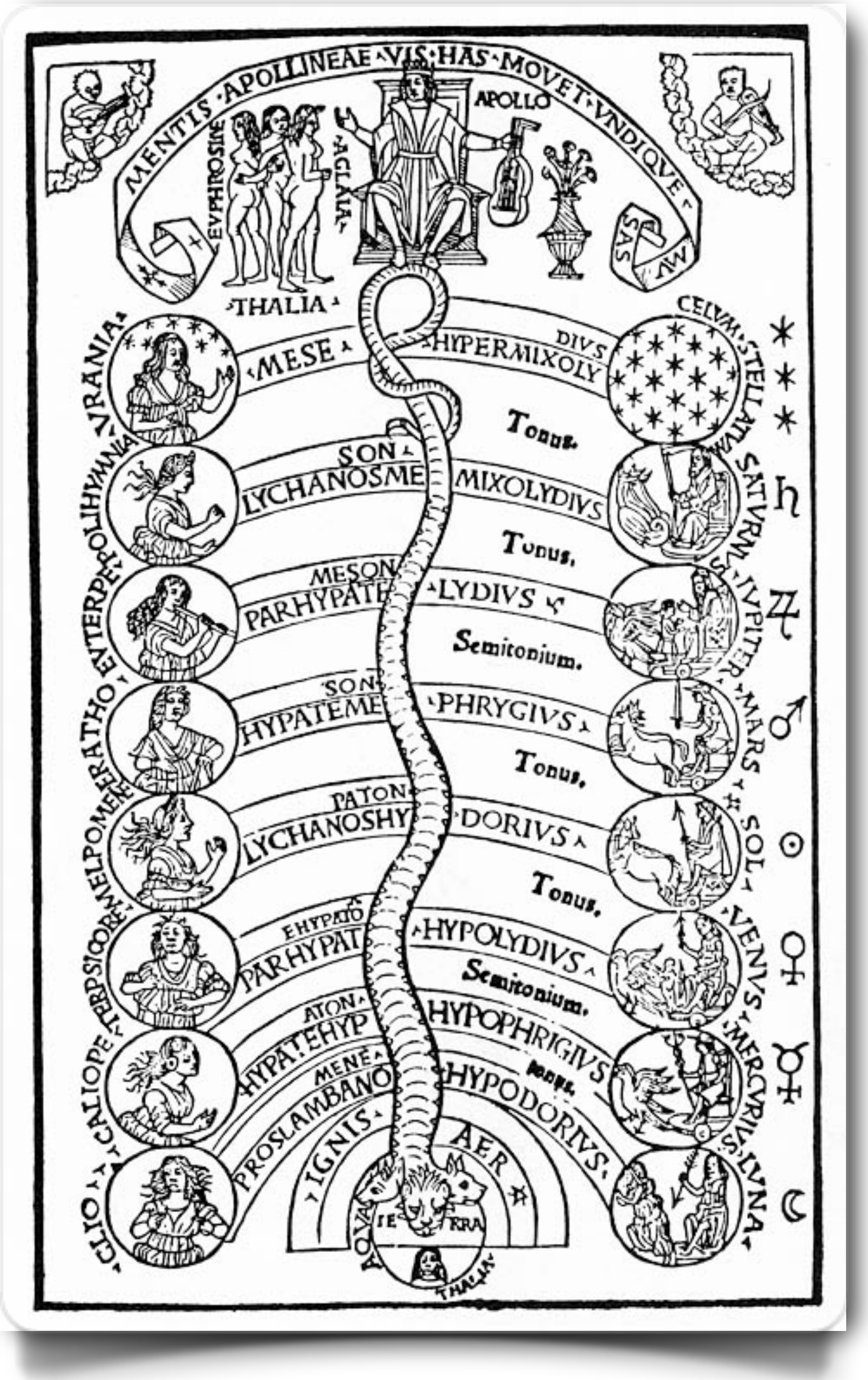
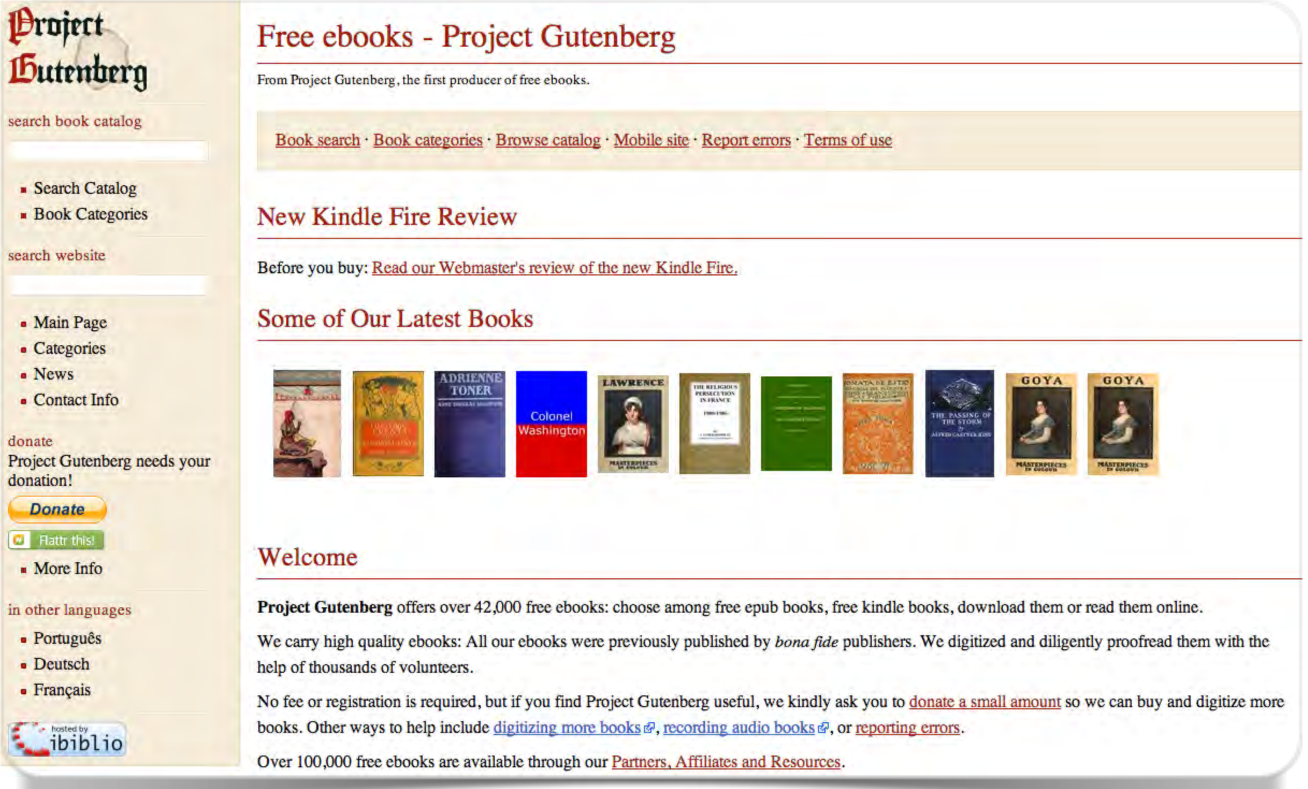


Fig. 1.4 Corpus Hermeticum

From on line texts to the corpus

As we have already said the on line sources for the identification of the texts to be used in the creation of a corpus can be various and most of them are known (social networks, Wikipedia, on line newspapers, blogs, RSS etc.). Less known might be some websites and portals useful when looking for digital publications and we therefore list some of the most interesting ones, especially those which cannot be forgotten in the map of the on line text analyst.



Project Gutenberg

search book catalog

- Search Catalog
- Book Categories

search website

- Main Page
- Categories
- News
- Contact Info

donate

Project Gutenberg needs your donation!

Donate

Flattr this!

- More Info

in other languages

- Português
- Deutsch
- Français

hosted by ibiblio

Free ebooks - Project Gutenberg

From Project Gutenberg, the first producer of free ebooks.

Book search · Book categories · Browse catalog · Mobile site · Report errors · Terms of use

New Kindle Fire Review

Before you buy: Read our Webmaster's review of the new Kindle Fire.

Some of Our Latest Books

Welcome

Project Gutenberg offers over 42,000 free ebooks: choose among free epub books, free kindle books, download them or read them online.

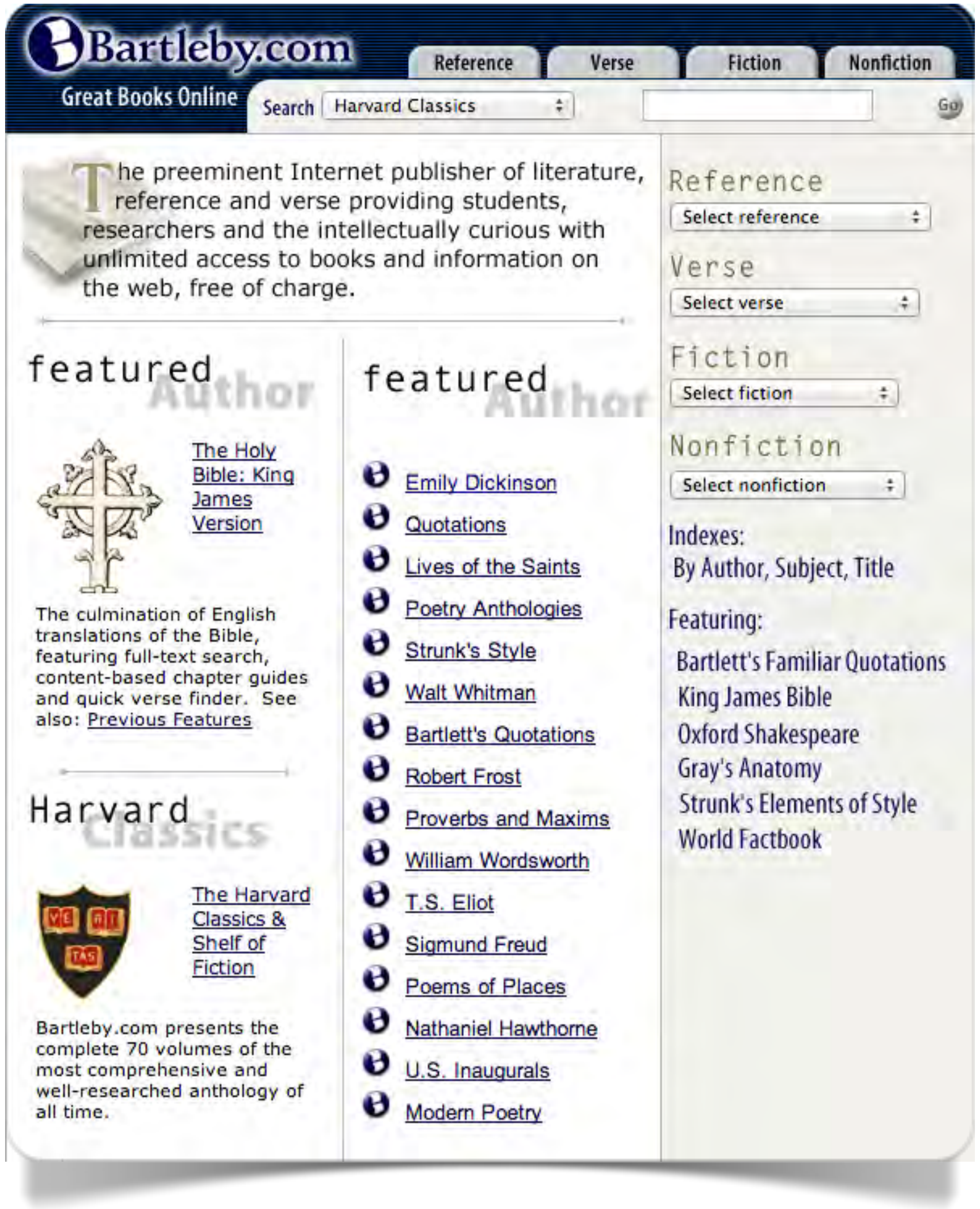
We carry high quality ebooks: All our ebooks were previously published by *bona fide* publishers. We digitized and diligently proofread them with the help of thousands of volunteers.

No fee or registration is required, but if you find Project Gutenberg useful, we kindly ask you to donate a small amount so we can buy and digitize more books. Other ways to help include digitizing more books, recording audio books, or reporting errors.

Over 100,000 free ebooks are available through our Partners, Affiliates and Resources.

Fig. 1.5 Project Gutenberg

Project Gutenberg can count on over 30.000 volumes in its catalog and it was certainly the first to work on the digitalization of printed books. It was created in 1971, well before the coming of the Internet, thanks to Michael Hart. It is considered the first e-book archive and it gathers texts belonging to public dominion or texts whose copyright has expired (in accordance with the laws of the USA). The texts, mainly written in English, are downloadable in different formats: From HTML to ASCII.



The preeminent Internet publisher of literature, reference and verse providing students, researchers and the intellectually curious with unlimited access to books and information on the web, free of charge.

featured Author

The Holy Bible: King James Version

The culmination of English translations of the Bible, featuring full-text search, content-based chapter guides and quick verse finder. See also: Previous Features

Harvard Classics

The Harvard Classics & Shelf of Fiction

Bartleby.com presents the complete 70 volumes of the most comprehensive and well-researched anthology of all time.

featured Author

- Emily Dickinson
- Quotations
- Lives of the Saints
- Poetry Anthologies
- Strunk's Style
- Walt Whitman
- Bartlett's Quotations
- Robert Frost
- Proverbs and Maxims
- William Wordsworth
- T.S. Eliot
- Sigmund Freud
- Poems of Places
- Nathaniel Hawthorne
- U.S. Inaugurals
- Modern Poetry

Reference

Verse

Fiction

Nonfiction

Indexes:
By Author, Subject, Title

Featuring:
Bartlett's Familiar Quotations
King James Bible
Oxford Shakespeare
Gray's Anatomy
Strunk's Elements of Style
World Factbook

Fig. 1.6 Bartleby.com

The Online Books Page is a digital library project directed by John Mark Ockerbloom, researcher at the University of Pennsylvania. It contains over 1 million volumes, having placed on the Web the sources of various universities and various projects, including the Gutemberg and Manuzio ones. It offers therefore volumes in all the languages of the world.

Internet Sacred Text Archive, created by John Bruno Hare, is the biggest archive of texts belonging to the sacred traditions, the mythology and the folklore of every place and every age.

Bartleby.com is a project which started being developed in 1993 at Columbia University. The name comes from a tale by Herman Melville (Bartleby, the scrivener. A story of Wall Street), one of the most famous in American literature. There are literary, scientific and political texts, such as the opening speeches of the Presidents of the USA.

PoliTxt is an archive of political texts. In addition to the opening speeches of the American Presidents, there are the transcriptions of television debates.

Virgo is the access point for University of Virginia Library digital texts and images. Formerly known as “Etext,” The Electronic Text Center.

Eserver.org is a digital humanities venture, founded in 1990 and based at Iowa State U, where writers, editors and scholars publish over 35,000 works free of charge, including materials in government.

WebCorp Live is a tool managed by Birmingham City University. It is used in language research, but it can also be consulted to to whittle down search engine results and to build thematic corpora. It is essentially a concordance generator that works using the best known web pages and search engines (Google, Bing and Yahoo!) with the possibility of selecting the news and the language one is interested in. The site offers different options, especially if an advanced research interface is used. The selection window ,choosing the sites according to their topics (for example, UK Tabloid Newspaper) or nationality (.it, .fr, .cn), allows to obtain strongly focused results. Also the the window used for the insertion of search words and phrases, helps to obtain flexible and sophisticated results by using simple search patterns based on wildcards and UNIX regular expressions.

Corpus preparation

The corpus preparation generally requires a preliminary phase of text conversion in ASCII (plain text) format and of normalisation. In some cases it is the analysis tool itself to carry out some simple cleaning and standardization operations. The straightforward “copy and paste” process is in most cases sufficient to eliminate, for instance, the format, the HTML markers or the images. Automatic text analysis pays particular attention to the “preparation of the corpus” and to the transcription and transcoding of stressed and special characters. Another important phase is the treatment of small letters. Capital letters make it possible to solve some problems of lexical ambiguity, but their presence duplicates the graphics because of the punctuation: The capital letter indicating the beginning of a sentence. In the analysis of on line texts all this has to be necessarily put aside because it would be the subject of an excessively specialized scientific treatment. One of the most drastic simplifications we must make use of in these cases is the transformation in small letters of all the letters present.

In the majority of on line applications the corpus consists of a collection of texts in order and it is impossible to make partitions. For example, a partition in a corpus made up of the swearing in speeches of the US presidents would be the subdivision of the speeches of the various presidents; or in a corpus made up of the text messages posted on Facebook, a partition could be the separation of the male and female messages.

Words, forms and occurrences

The constitutive elements of a text are words. In the automatic treatment of texts, words always appear in their graphic form, that is as a string of characters delimited by two separators. In the applications available on line that we are considering in this chapter, the user cannot choose the letters to consider as separators because it is the software itself that does this.

Graphic forms classified as count units are called word tokens. The automatic analysis of a text provides first of all the count of the graphic forms together with their word tokens. The list of graphic forms will be represented by distinctive word types. The graphic form inventory can be ordered in various ways. The most used are alphabetic order or decreasing order of the word tokens.

Not all the words of a text are the same from the semantic point of view. In the automatic analysis of language it is possible to distinguish the full words from the empty ones (in some cases these are indicated as stop words). Empty words are defined as words which do not have an interesting content for the analysis that is taking place (they are often grammatical words or a simple phrase connector), while full words are those which contribute significantly to the interpretation of the text.

In the automatic analysis of language text words are often filtered beforehand using a list of stop words. There is no univocal criteria in determining the list of words to insert as stop words. There are certainly grammatical words, as we have already said, but in some cases there can also be auxiliary verbs or words considered banal or not fundamental for particular purposes. In a list of stop words in English, French and Italian used by search engines, we can find common words such as the following, but which can certainly not be defined as empty words from the linguistic point of view:

English	Français	Italiano
able, about, accordance, actually, adopted, affected, affecting, afterwards, against, (…) wish, world, zero	Avoir, devrait, doit, droite, début, elle, encore, est, fait, (…) voie, vous, vu	Buono, comprare, consecutivo, dentro, deve, fine, fino, gente, indietro, (…) voi, volte, vostro

Fig. 1.7 Examples of stop words used by search engines (Source: Ranks.nl)

When it is possible, it is a good idea to check the list of stop words and if necessary, to change them, ignore them or apply personal criteria according to what are the analysis objectives.

Chapter 2

A tool to start: Textalyser

Textalyser is a good example of how it is possible to create with simplicity, but also with rigour a basic program for the automatic analysis of texts.

2

Textalyser

Textalyser offers different text analysis possibilities: The count of word tokens, clauses, syllables and n-grams. Furthermore, it uses some readability indicators (alternative tool with the same options: Text Analyzer).

The latest on line version works perfectly also with an iPad. The text can be inserted using copy and paste in a query window, specifying a webpage link or uploading the file directly from the computer.

In order to examine Textalyzer's options, let's consider as example the website with the swearing in speeches of the US presidents: Bartleby.com and let's choose Barak Obama's swearing in speech of the 20th of January 2009. We could obtain the text directly from the web address, but it is preferable to “copy and paste” in the specific window required because the webpage contains graphic forms which are not part of the speech (in other cases the text might be divided in various pages and therefore it would not be entirely uploaded using the link at the homepage). Textalyser performs by itself a very simplified normalisation, reducing all the text to small characters (for more sophisticated cleaning operations, one needs to use a “purely text” analyser like Notepad++).

In fig. 2.1 we see the results of the analysis using Textalyser's default options (minimum characters 3; number of words 10; stoplist English).

The output gives us immediately the following lexicometric measurements: The *complexity factor* (or *type-token ratio*) is the ratio between graphic forms or distinctive words (V) and the total of word tokens (N). It is one of the most important parameters to evaluate the suitability of a corpus for statistical analysis. The suitability threshold, according to which the lexical extension of the text as representative of the language used is considered adequate, is inferior or equal to 20%. Higher values can be found in small texts like the one in this example. The value is

Total word count (tokens, N)	1,309
Number of different words (types, V)	839
Complexity factor (Lexical Density, N/V)	64.1%
Readability (Gunning Fog Index: 6-easy 20 hard)	8.9
Total number of characters	16,637
Number of characters without spaces	7,837
Average Syllables per Word	1.58
Sentence count	137

Fig. 2.1 Lexicometric measurements of President Barak Obama's swearing in speech, the 20th of January 2009 (elaboration: Textalyser)

distorted also because in the default analysis have been considered only the words with a minimum of three characters and the *stop words* have been excluded. Selecting 1 character as the minimum word-length and no stoplist, the word tokens are N = 2,381 and the distinctive words are V = 903 with a type-token ratio of 37.9%. It is a value still very far from being adequate. However, the analysis is interesting the same if we make mainly qualitative evaluations.

The readability indicator (Gunning fog Index) is calculated according to the the length of clauses which make up the text and according to the length of the words divided in syllables. This is not relevant for the automatic analysis of the text, which aims at extracting its content, but it is useful the same for text comparison.

The most recurrent ten words (fig. 2.2) give us a first approximation of the rhetorical modalities present in Obama's speech, in particular if we compare them with the most recurrent ten words characterizing George W. Bush's second swearing in speech, the 20th of January 2005.

G.W. Bush	Occurrence	% Freq.	B. Obama	Occurence	% Freq.
our	50	4.4	our	67	5.1
freedom	24	2.1	you	14	1.1
liberty	14	1.2	nation	12	0.9
you	12	1.1	new	11	0.8
america	12	1.1	those	11	0.8
your	12	1.1	must	8	0.6
every	10	0.9	every	8	0.6
nation	9	0.9	what	8	0.6
own	9	0.8	these	8	0.6
country	8	0.7	less	7	0.5

Fig. 2.2 The ten most recurrent words in the swearing in speeches of the presidents G. W. Bush and Barak Obama of the 20th of January 2005 and 2009 (elaboration: Textalyser; stoplist active)

N-grams and keyword prominence

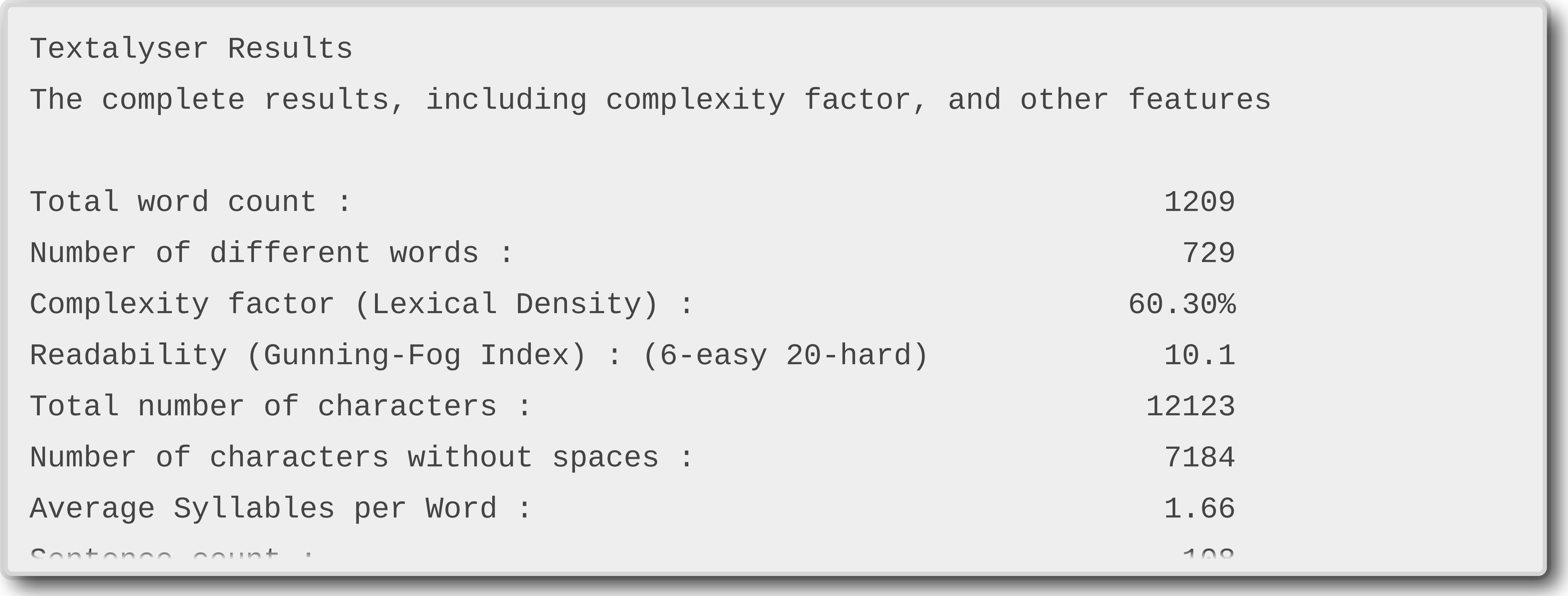
For a real extraction of the content 3-grams are more significant. These are text recurrences with three graphic forms, among which we immediately find expressions offering precise indications about the most relevant themes in the two speeches (fig. 2.3).

G.W. Bush	Occ.	Prominence	B. Obama	Occ.	Prominence
the united states	4	72.7	and we will	3	76.1
we have seen	3	**93.5**	a new era	2	43.0
of our time	2	56.5	who seek to	2	44.3
do not accept	2	60.7	of our nation	2	44.5
america will not	2	72.4	we will not	2	60.6
america’s influence is	2	72.8	that our power	2	63.1
of the world	2	75.7	of our economy	2	77.6
of ending tyranny	2	76.8	**say to you**	2	**83.3**
in our world	2	83.1	**a new age**	2	**85.8**
we have proclaimed	2	**84.6**	**time has come**	2	**87.2**
of liberty in	2	**87.5**	on this day	2	76.1

Fig. 2.3 3-grams of the swearing in speech of the presidents G. W. Bush and Barak Obama of the 20th of January 2005 and 2009 (elaboration: Textalyser)

The keyword prominence measures the average position of the segment in the text; the more the segment is near to the beginning of the text, the higher its value is. For example, the form *we have seen* in Bush's speech with 3 occurrences and a prominence of 93.5 is positioned in the first part of the speech and it has therefore a higher value.

Comparing the two speeches we can immediately observe how the 3-grams are very different. Obama's speech is characterized by a participatory tone (*and we will*; *our power*; *our economy*) and by affirmations which point out the novelty of his election and of the path to take: *time is come*; *a new age*; *a new era*, *say to you* are the most relevant segments. Instead, in Bush's swearing in speech there is a strong emphasis on the comparison between freedom and tyranny, on America's role in the present day and on its vulnerability: *we have seen* implicitly recalls the 11th of September 2001.



```
Textalyser Results
The complete results, including complexity factor, and other features


Total word count :                                             1209
Number of different words :                                     729
Complexity factor (Lexical Density) :                        60.30%
Readability (Gunning-Fog Index) : (6-easy 20-hard)             10.1
Total number of characters :                                  12123
Number of characters without spaces :                          7184
Average Syllables per Word :                                   1.66
Sentence count :                                                108
```

Fig. 2.4 Speech of the presidente Barak Obama of the 21th January 2013 (elaboration: Textalyser; min. char. 1; stoplist active - Source: Bartleby.com - scrolling text)

Chapter 3

Visualizing the weight of words: Wordle

Wordle - Beautiful Word with Clouds was created as a tool to produce aesthetically pleasant graphical representations of the most used words in a text, but used carefully, it allows us to obtain an efficacious synthesis of the content.

3

Wordle

Wordle was developed by Jonathan Feinberg as a Java applet in 2008 when he was still a researcher at IBM. We will find a simplified version among the visualization tools of the texts on Many Eyes. However, the most direct and funny way to use the “word cloud” is certainly from this website, where the application is free and where there are a variety of options which make it very flexible (Feinberg, 2010).

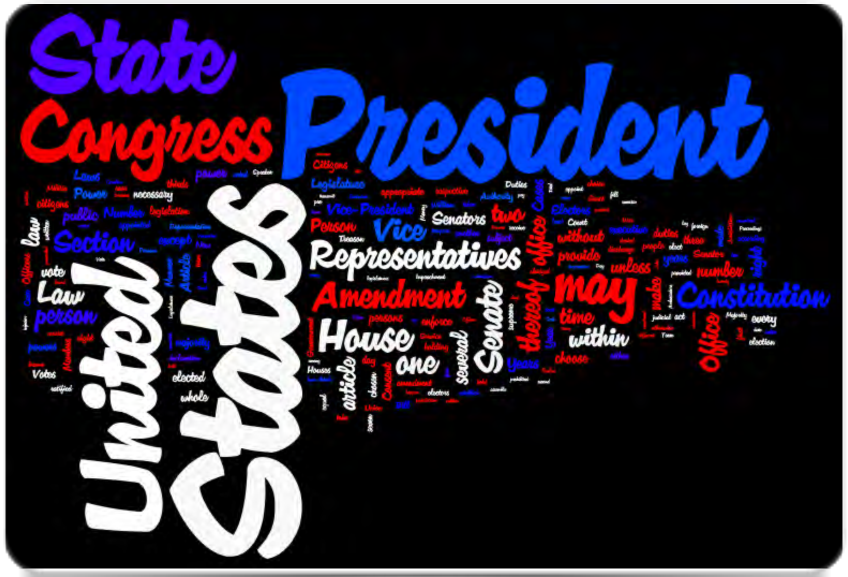
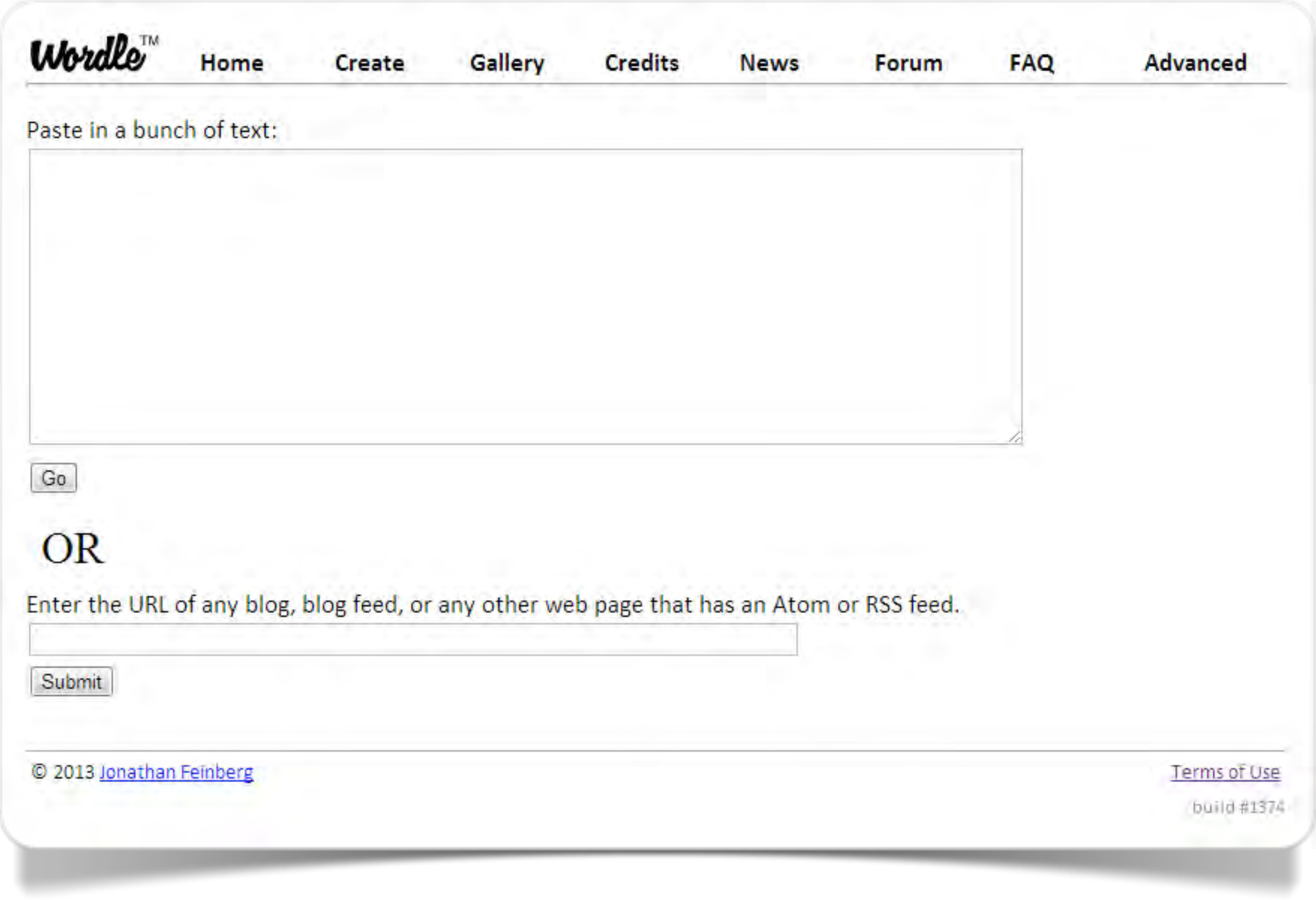


Fig. 3.1 US Constitution (elab. Wordle)

In Wordle the “heaviest” words (the ones with a greater number of occurrences) are represented proportionally in a bigger size. Its greatest limit is that – from the graphical point of view – size can be misleading, because some letters of the alphabet take up more space (for example the “o” and the “a”) than others (typically the “i” and the “l”).

Clicking on *Create* from the home page, Wordle opens a window for text acquisition by simply “copying and pasting”, by inserting the website address in the apposite place, or by using an RSS feed.

Clicking on *Go*, the program visualizes the word cloud with its predefined options. During the uploading phase it is important to decide immediately, in the preparation of the corpus, if to leave the words as they are in the text (Capital and Small letters) or to change all of them into capital or small letters in order to avoid the duplication of the words as a consequence of the different way in which they are written (Web/web). In any case this option can be activated after using the vertical menu *Language*.



Paste in a bunch of text:

OR

Enter the URL of any blog, blog feed, or any other web page that has an Atom or RSS feed.

© 2013 Jonathan Feinberg

Terms of Use

build #1374

Fig. 3.2 Wordle: Create your own

In this example (fig. 3.3) we have chosen *The Universal Declaration of Human Rights*, where the graphic characters have been made small. The first visualization is efficacious enough, also for the position taken by *right*, with *equal* on top and other fundamental key words such as *freedom*, *social*, *protection*, *law*, *everyone* around. Even though it is pleasant and efficacious, this result is purely casual (It is often necessary to make various attempts before obtaining a satisfactory image from the communicative point of view).

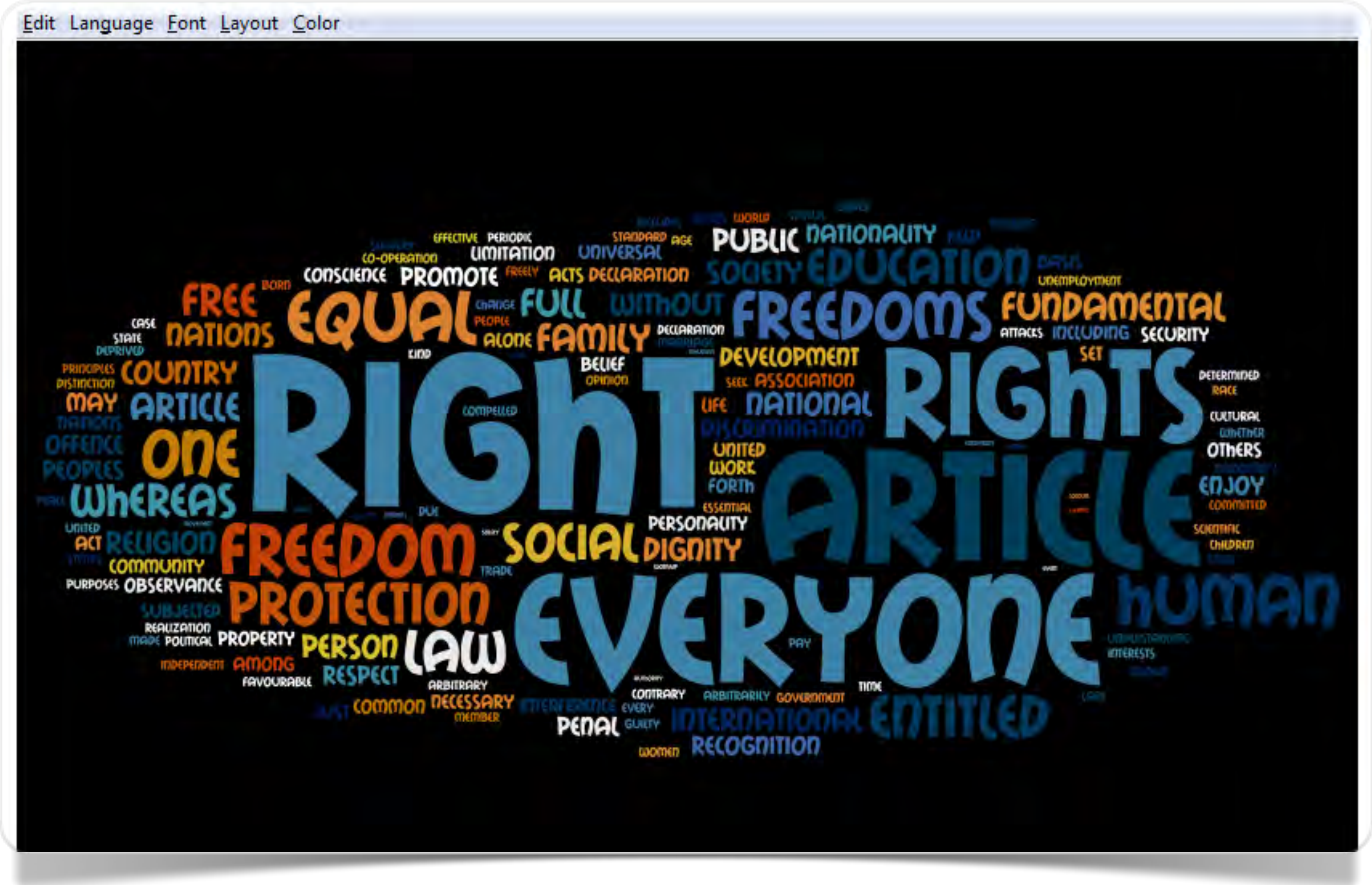


Fig. 3.3 *The Universal Declaration of Human Rights* (elab. Wordle with predefined options)

Options

The word cloud can be changed using the options of the vertical menu which contains different kinds of characters, colours and word orders, but these representation modalities do not have any meaning; They are only an aesthetic expedient.

Wordle default recognizes the language in which the text is written and it automatically eliminates the *stop words*. This option can be modified by the user from the vertical menu *Language*. The character ~ (tilde; 007E in Unicode or 126 in ASCII characters) is interpreted by the program as a joining character (*will~be*): the words are put together in the graph without the tilde itself being visualized (the same thing is valid for the character Unicode 00A0 or ASCII 160). The formation of "compound words" has to take place during the text preparation phase.

The menu *Layout* allows to intervene on the number of the words to represent (default: 150) and on their distribution: in alphabetic or casual order; all in horizontal; all in vertical or mixed; with the external bording of the cloud round or straight. A word can be removed from the graph by clicking the right button of the mouse. In this case it is useful to eliminate the word *article*, with 30 occurrences (one for every article of the declaration), but without content in the graphic representation of the document. In the menu Font and Color it is possible to select the graphic characters and the color palette. The visualization in figure 3.4 has been obtained eliminating the word *article* and using font *Teen*, a maximum of 80 words and a color palette *Firenze*.

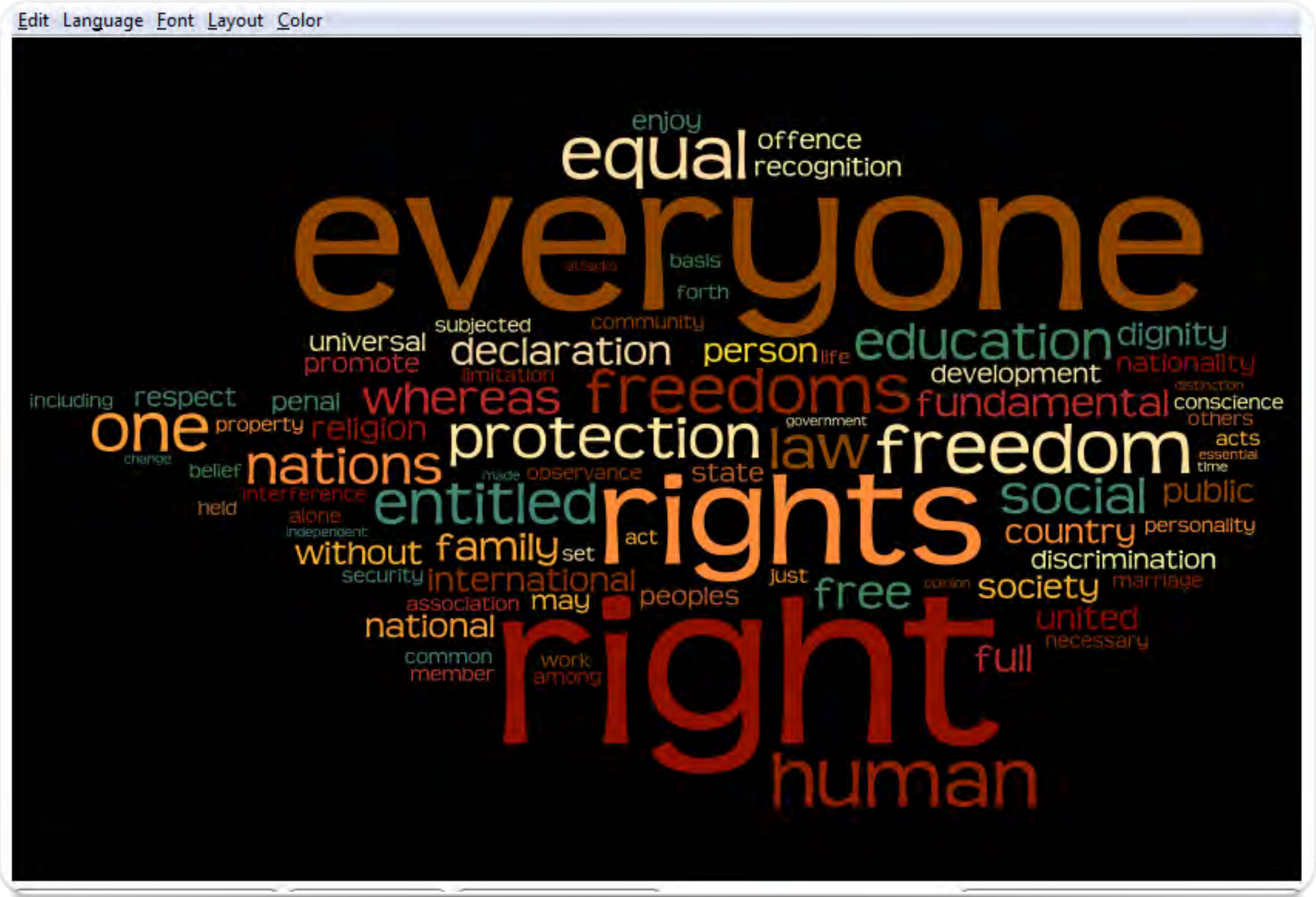


Fig. 3.4 *The Universal Declaration of Human Rights* (elab. Wordle with Font/Teen options, Maximum words 80, Color palette Firenze)

Integration of Textalyser with Wordle

From the uploading page *Advanced*, Wordle allows also to elaborate the textual data coming from a “table”or Excel format with two columns separated by the punctuation sign “colon”. In our example (fig. 3.5) a table has been inserted in Barak Obama's most significant 3-grams speech; The one already analysed with Textalyser (fig. 2.3). In this example the occurrences have been replaced by the indicator *keyword prominence*. The result obtained is very similar to a summary of the most relevant themes.

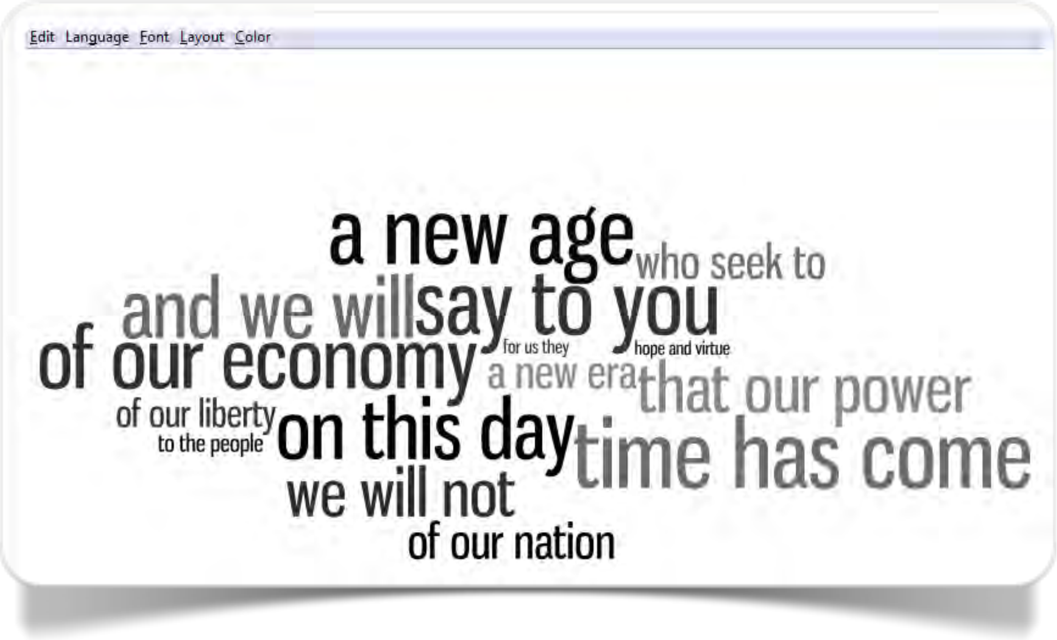


Fig. 3.5 3-grams of Barak Obama's swearing in speech of the 20th of January 2009 using the indicator *keyword prominence* (elab. Textlayser and Wordle)

Chapter 4

Tagxedo: The creative visualization

Tagxedo - Word Cloud with Styles was developed by Hardy Leung taking into account Wordle's experience, but with more artistic and playful intentions.

4

Tagxedo

The Tagxedo visualization is always based on a word cloud, but its graphic options and output management potential make it much more versatile. The first novelty introduced by Tagxedo, already when it came out, was the possibility for the user to choose the shape of the word cloud, with pleasant aesthetic results and a considerable communicative efficaciousness, as is shown by the examples available in the Gallery.

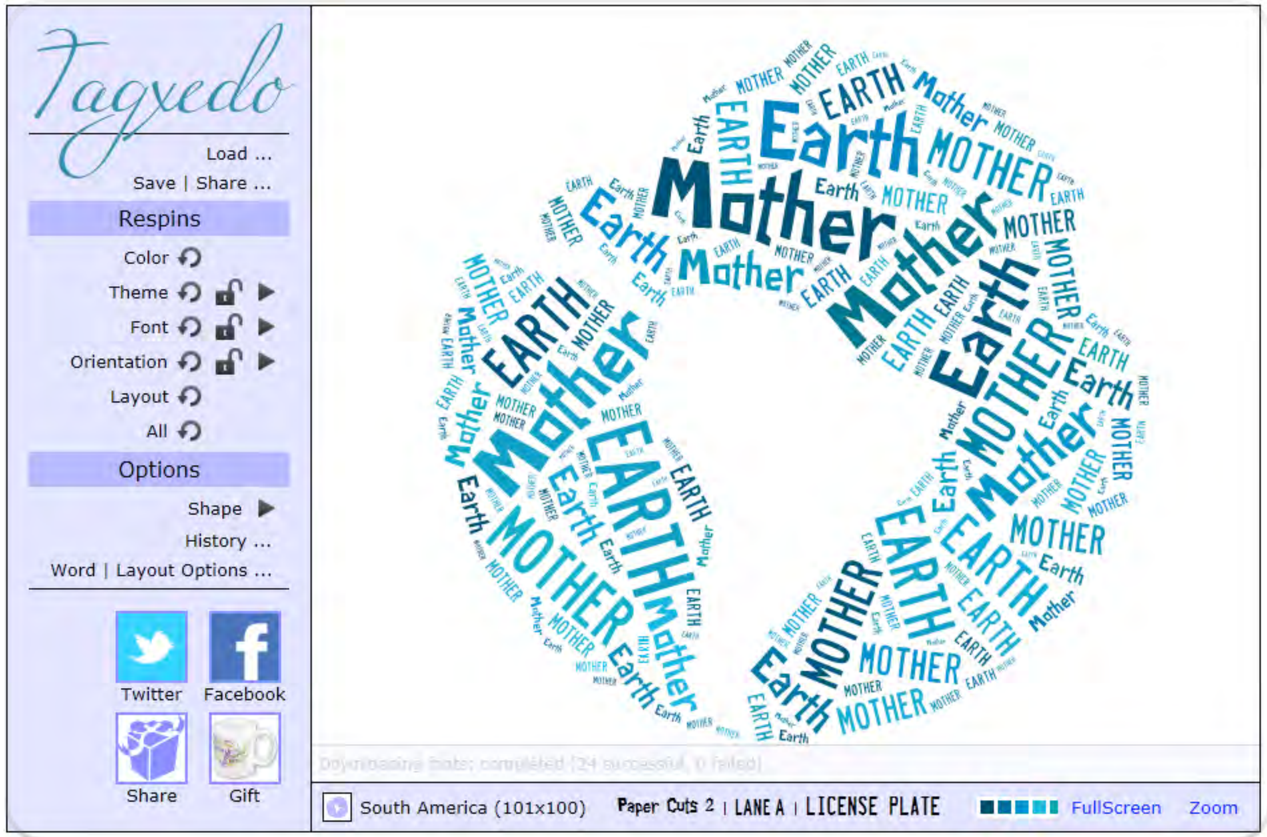


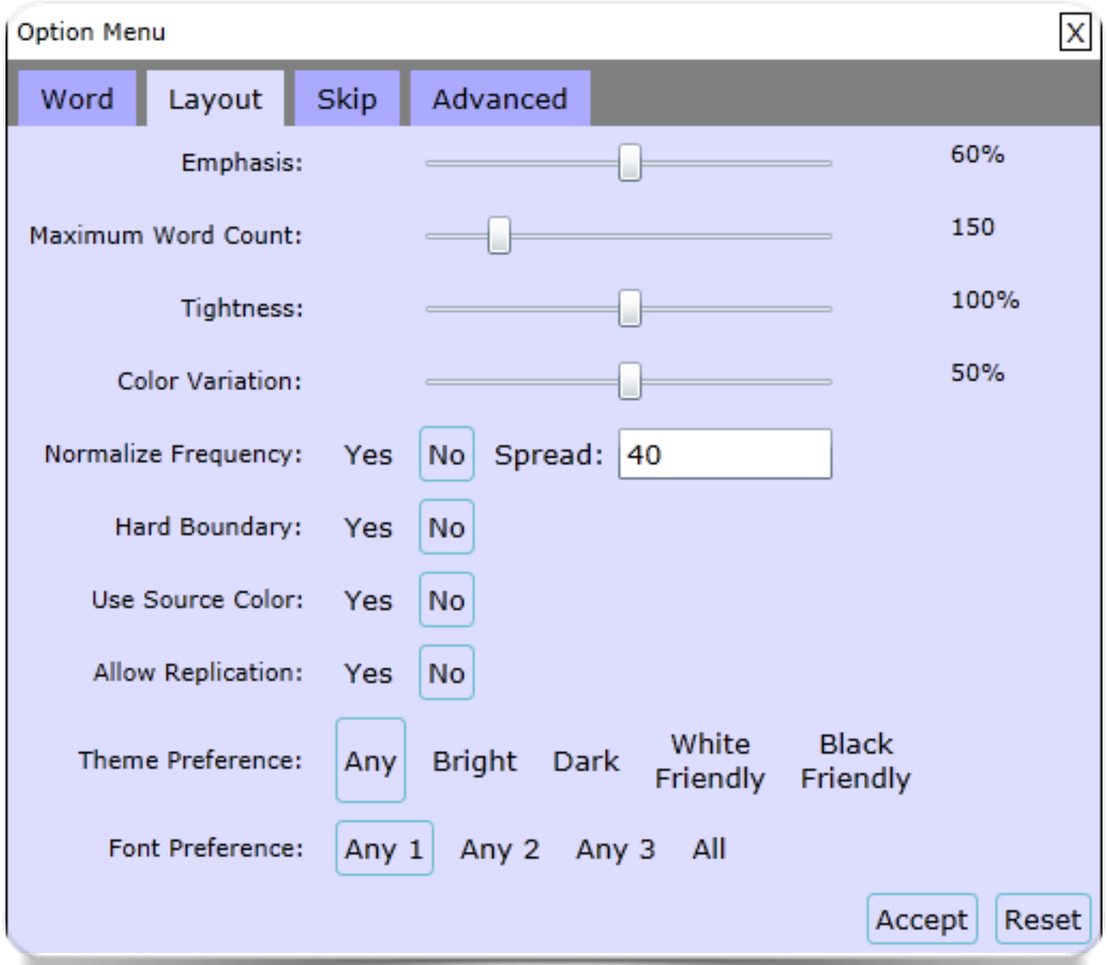
Fig. 4.1 Tagxedo Creator

In addition to the traditional copy and paste, file upload or URL insertion, the text acquisition modalities include access through Google News RSS feed or the latest 100 posts of a Twitter ID.

Layout options

The menu *Word - Layout options* (fig. 4.2) opens a series of cards with highly sophisticated intervention modalities. In the *Skip card* it is possible to select, among the first graphic forms in order of frequency, the ones to use for the final visualization. On the contrary of Wordle, Tagxedo only uses a list of English *stop words*, but with *Skip* it is always possible to remove from the vocabulary the words considered less significant by the user.

In the card *Layout* the option *Normalize Frequency* assigns a “theoretical” frequency to each word according to the order of occurrences. The Spread (with a default of 40) is the ratio between the highest and the lowest frequency. The higher the spread, the higher the dimension of the words with a greater frequency.



Emphasis: 60%

Maximum Word Count: 150

Tightness: 100%

Color Variation: 50%

Normalize Frequency: Yes [✓ No] Spread: 40

Hard Boundary: Yes [✓ No]

Use Source Color: Yes [✓ No]

Allow Replication: Yes [✓ No]

Theme Preference: [✓ Any] Bright Dark White Friendly Black Friendly

Font Preference: [✓ Any 1] Any 2 Any 3 All

Fig. 4.2 Tagxedo creator: Option menu

Tagxedo cannot count on a real tutorial, but Facebook page offers a lot of useful suggestions and explanations. In *101 Ways to Use Tagxedo* there are examples and “commemorative” visualizations created by Hardy Leung.

Figure 4.3 refers to the initiative of *New York Times online* that, for the 11 of September 10th anniversary, asked its readers to answer the question: *Where Were You on September 11, 2001?* The answers were 38 000.

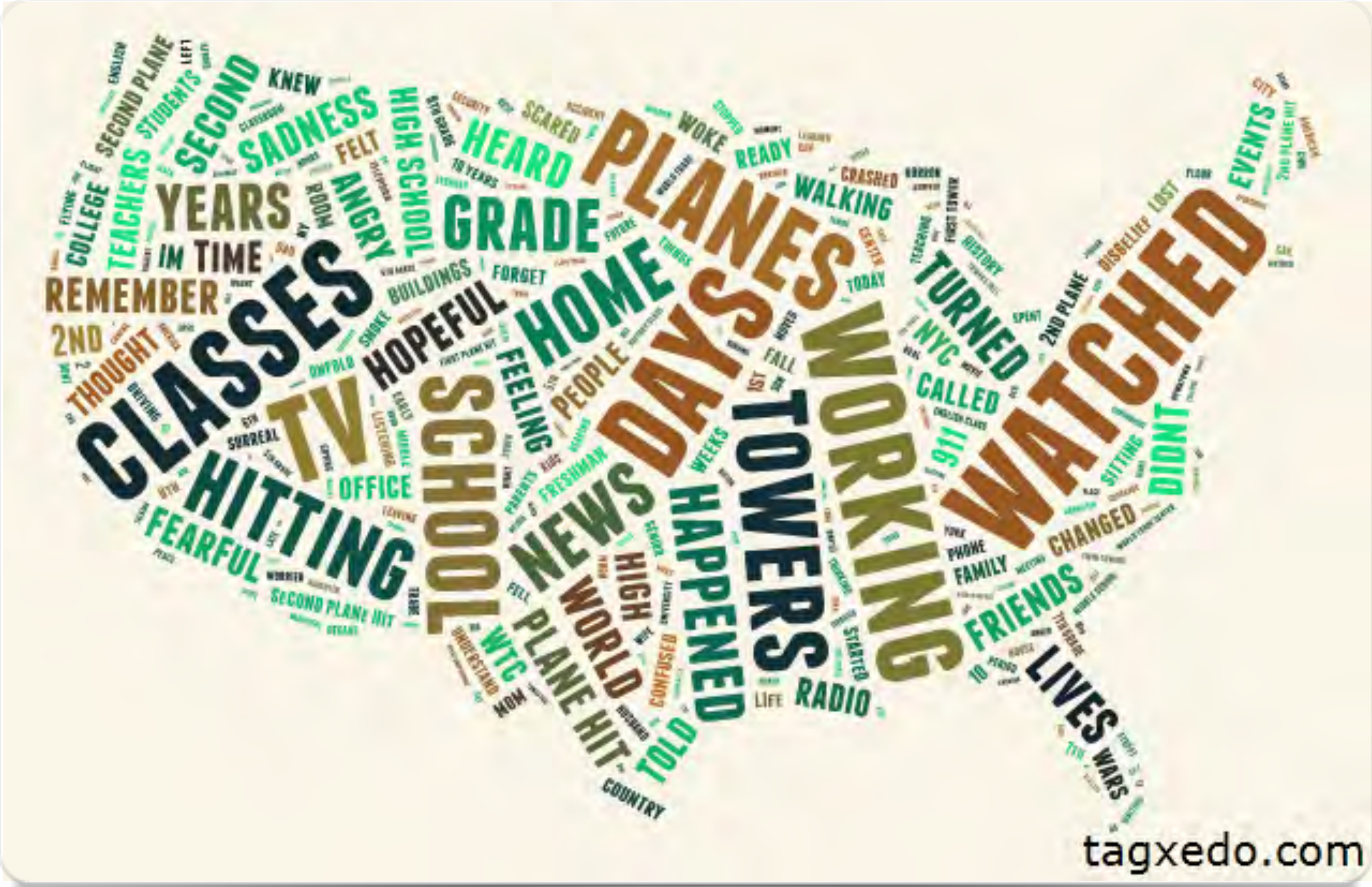


Fig. 4.3 *New York Times*: *Where Were You on September 11, 2001*? (elab. Tagxedo)

Chapter 5

Text visualization tools: Many Eyes

Many Eyes is an IBM on line laboratory capable of offering its users some powerful analysis and visualization tools both of textual and numerical data. Here we consider the visualization of textual data.

5

Many Eyes

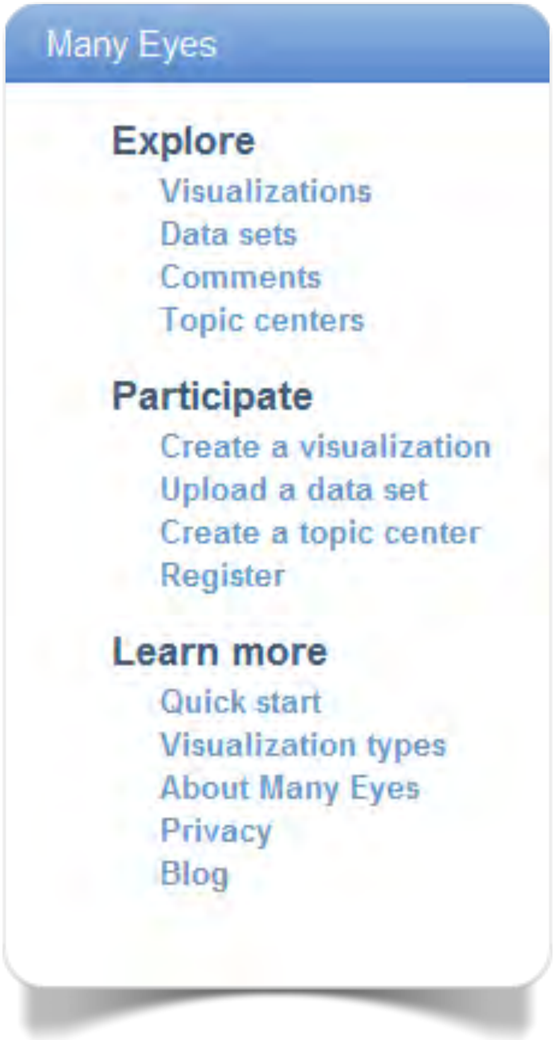


Fig. 5.1 Many Eyes:
Home Page Menu

With Many Eyes it is possible to freely use, elaborate and visualize the texts and the files already uploaded on the website database or, after subscribing for free, to upload a corpus previously prepared. The texts are uploaded from the menu *Participate – Upload a data set*.

The tools available are easy to access but, for proper use, it is advisable to read at least the documents from the menu *Learn more – Quick start*.

From *Explore* it is possible to see the visualizations created by website subscribers and saved on the site in chronological order. All the visualizations can be modified and saved (also by occasional users).

In the following examples we will follow the standard procedure for the creation of visualizations starting from the data set already uploaded on the website.

From the menu *Partecipate - Create a visualization* we follow the indications of the tutorial and we look for a corpus of sure interest: *Facebook Statuses Containing 'muslim', 'obama', and '9 11'.* It is a collection of personal statuses taken from Facebook coming up to the 10th anniversary of the Twin Towers attack containing references to the US president, to “muslim” and to the 11th of September.

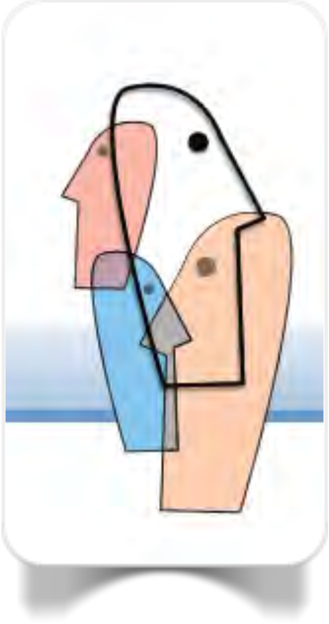
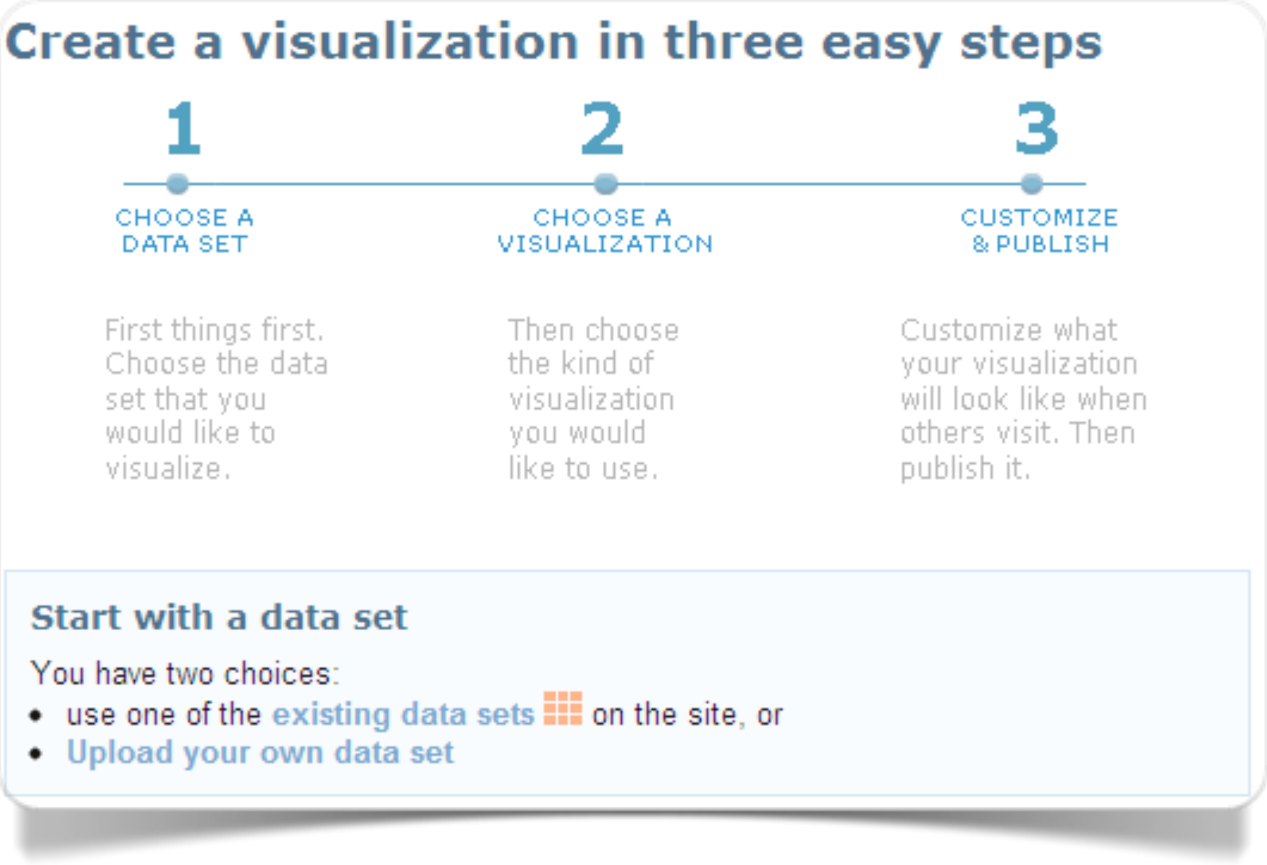


Fig. 5.2 Many Eyes: logo

From the visualization page of the file *Facebook Statuses Containing 'muslim', 'obama', and '9 11'*, clicking on *View as text* we can examine the characteristics of the entire corpus. Clicking on *Visualize us* different tools are offered according to the data set organization modalities. In our case we can refer simply to the test analysis tools: *Analyse a text*.



Create a visualization in three easy steps

1 CHOOSE A DATA SET

First things first. Choose the data set that you would like to visualize.

2 CHOOSE A VISUALIZATION

Then choose the kind of visualization you would like to use.

3 CUSTOMIZE & PUBLISH

Customize what your visualization will look like when others visit. Then publish it.

Start with a data set

You have two choices:

- use one of the existing data sets on the site, or
- Upload your own data set

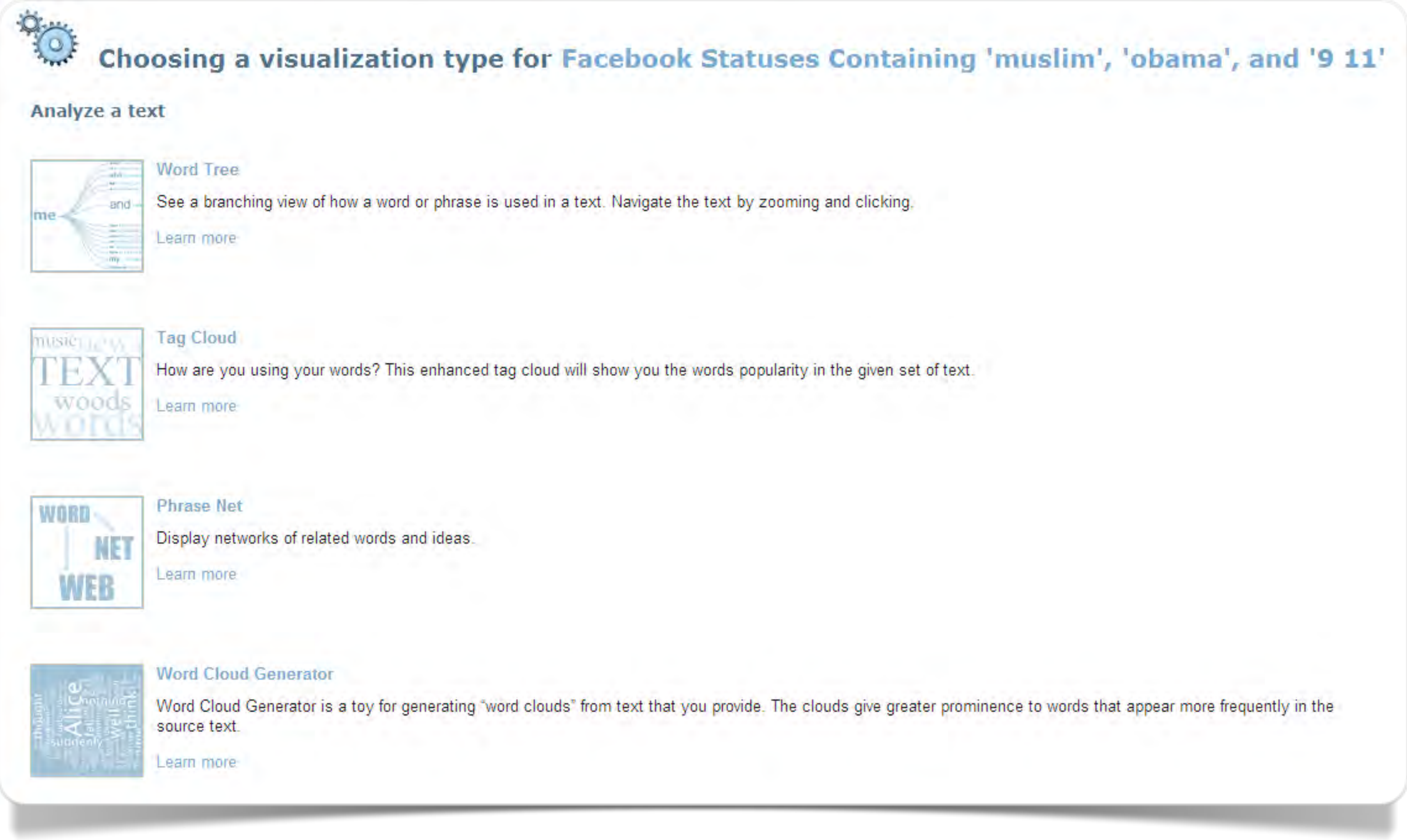
Fig. 5.3 Many Eyes: Visualization steps

Word Tree allows to analyse the text on the basis of a concordance classification tree, placing the selected word as pivot in the context of the words which come before and after.

Tag Cloud visualizes words and 2-grams (repeated segments of two words in a row) in a dimension which depends on the occurrences. It is similar to Wordle with a variation which allows to compare two texts (partitions) of the same corpus

Phrase Net s a diagram of the net of words related to one another through “bridge” or “link” words. The program is predefined for the English language, but it is adaptable also to other languages.

Word Cloud Generator is Wordle's original and simplified version, developed by Jonathan Feinberg when he was researcher at IBM.



Choosing a visualization type for Facebook Statuses Containing 'muslim', 'obama', and '9 11'

Analyze a text

Word Tree

See a branching view of how a word or phrase is used in a text. Navigate the text by zooming and clicking.

Learn more

Tag Cloud

How are you using your words? This enhanced tag cloud will show you the words popularity in the given set of text.

Learn more

Phrase Net

Display networks of related words and ideas.

Learn more

Word Cloud Generator

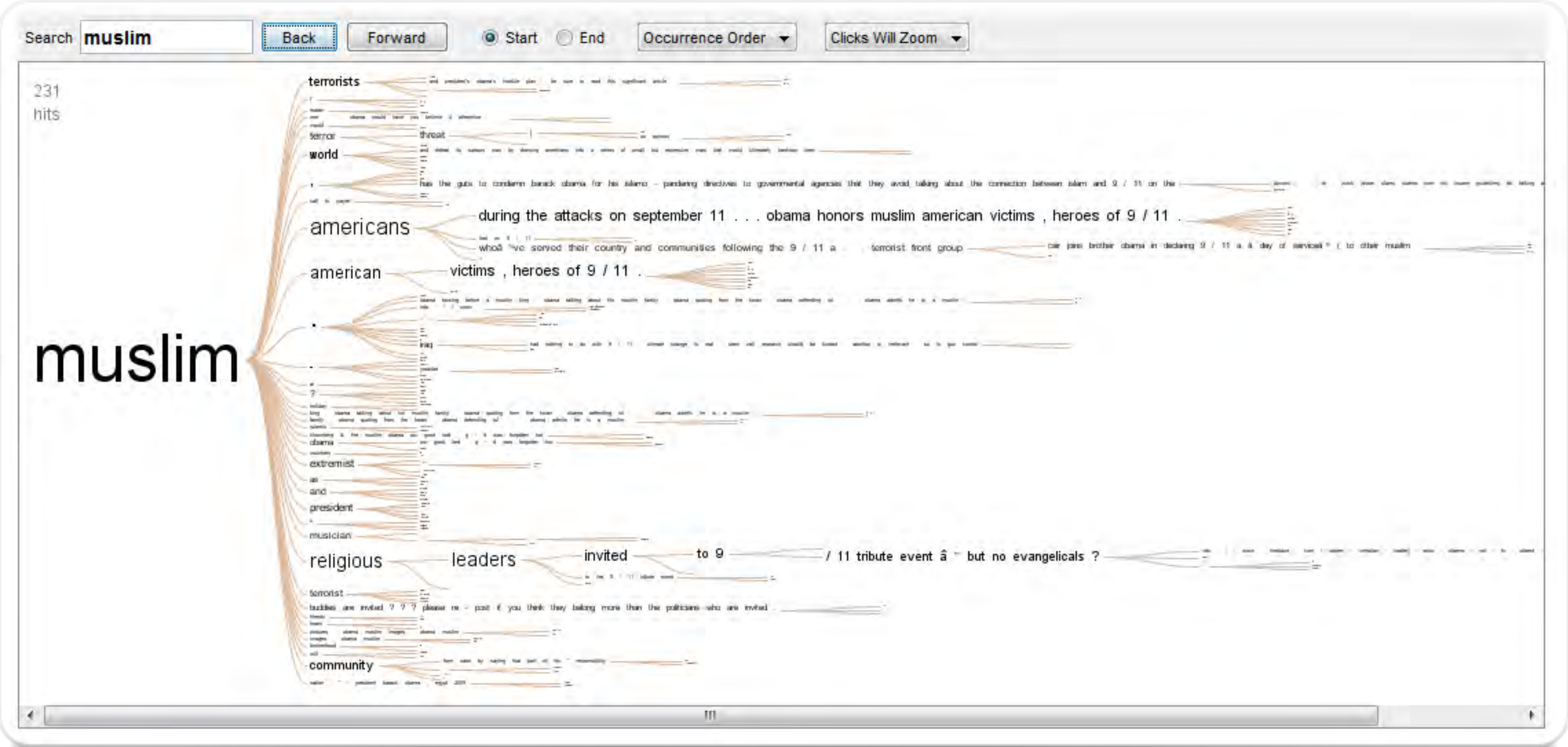
Word Cloud Generator is a toy for generating "word clouds" from text that you provide. The clouds give greater prominence to words that appear more frequently in the source text.

Learn more

Fig. 5.4 Choosing a visualization type

Word Tree

In Word Tree the concordances are classified in order to create a “word tree”: branches of sequences from a base word (pivot word) which represents its trunk, like in the leaves of a tree (if you insert in the field *Search* the selected word with the option *Start*), or ramifications of the sequences which come before the word, as happens in the roots of the tree (if you activate the option *End*).



Search muslim

● Start ○ End

Occurrence Order

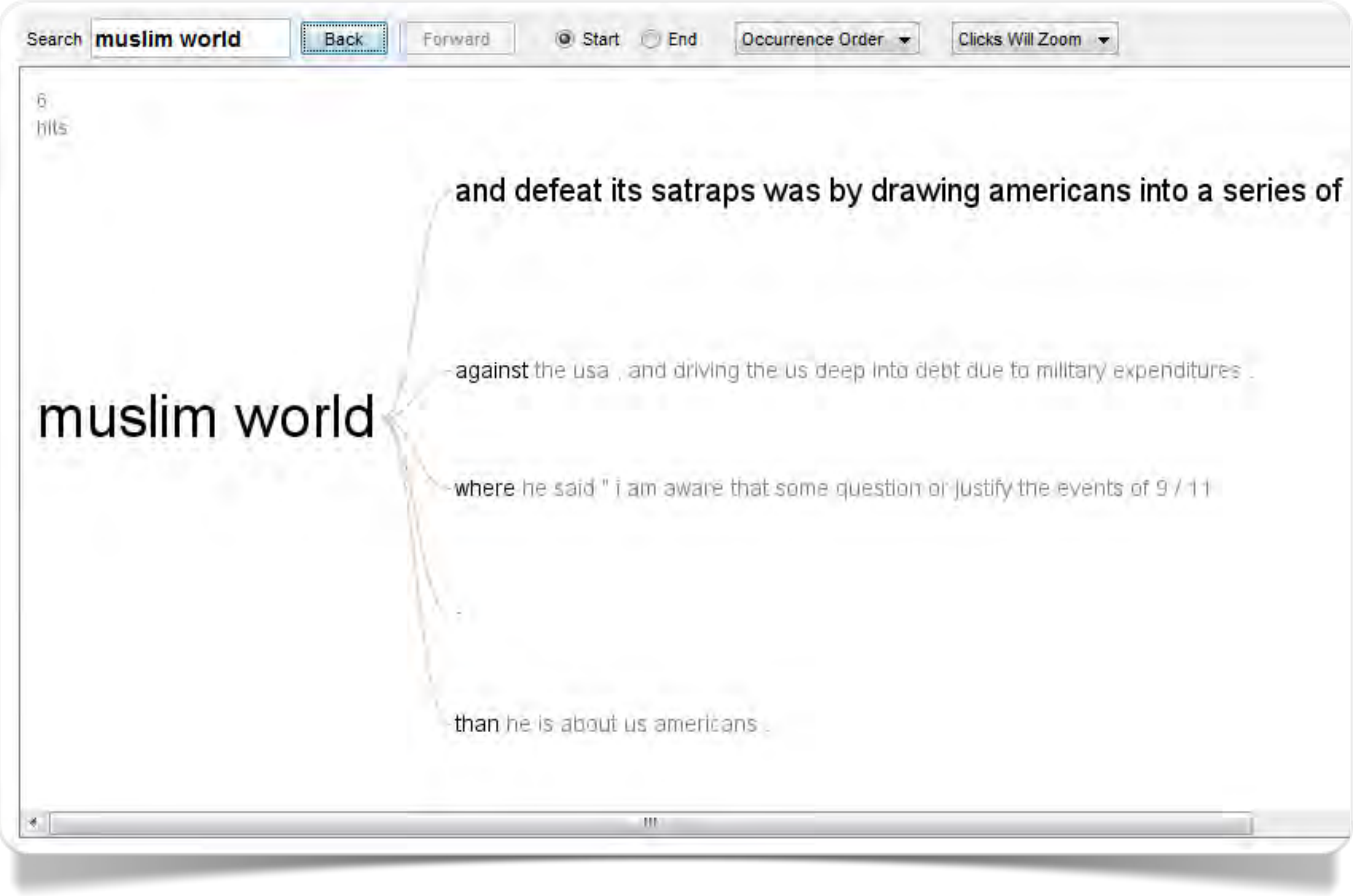
Clicks Will Zoom

231 hits

Back
Forward
muslim
terrorists
terror
threat
world
americans
during the attacks on september 11 . . . obama honors muslim american victims , heroes of 9 / 11 .
who&#x27;ve served their country and communities following the 9 / 11 a
terrorist front group
american
victims , heroes of 9 / 11 .
extremist
and
president
musician
religious
leaders
invited
to 9
/ 11 tribute event â ¨ but no evangelicals ?
terrorist
community

Fig. 5.5 Word Tree of the form *muslim* [Start] in the corpus 9-11 (elab. IBM’s Many Eyes)

Also in this case the words and the sequences are represented with dimensions proportional to the occurrences (*muslim* = 231 occurrences). The visualization is changeable ordering the sequences in alphabetical order or in order of frequency. In addition, all the sequences can be surfed and are dynamic with different modalities in order to highlight the semantic context. Figure 5.6, for example, was obtained from the visualization of figure 5.5 clicking on *world* (*muslim world* = 6 occurrences).



muslim world

and defeat its satraps was by drawing americans into a series of

against the usa , and driving the us deep into debt due to military expenditures .

where he said " i am aware that some question or justify the events of 9 / 11

.

than he is about us americans .

Fig. 5.6 Word Tree of the form *muslim world* [Start] nel corpus 9-11 (elab. IBM’s Many Eyes)

Tag Cloud

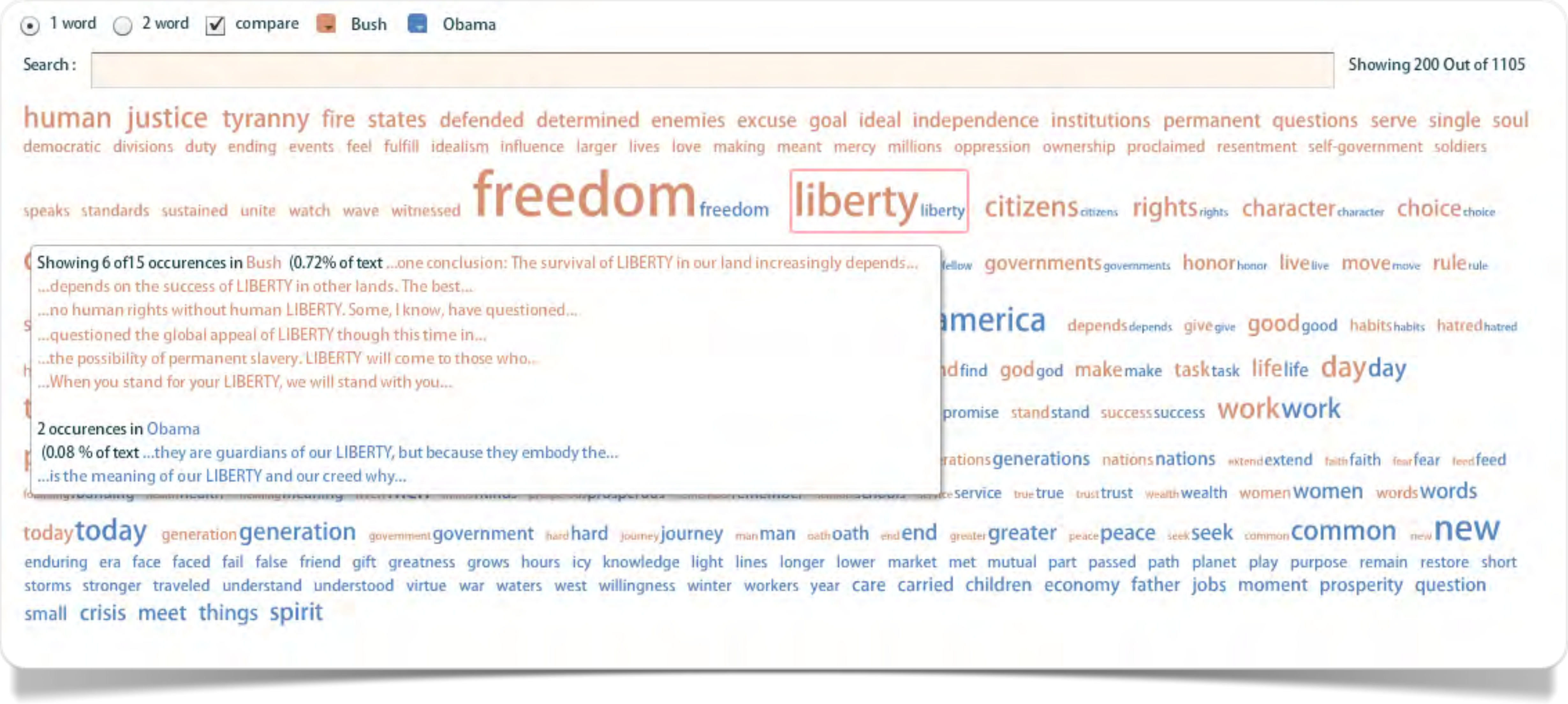
Tag Cloud visualizes the frequency of words and 2-grams (repeated segments of two words in a row). Also Tag Cloud eliminates the empty words from the text, but it does not allow the user to choose the language. Positioning the mouse on a word, it is possible to highlight the occurrences of the words in the context.

The most interesting novelty of Tag Cloud towards Cloud Generator, which is present among Many Eyes' tools, is the possibility to compare two texts within a corpus. The texts must be preceded by a marker of the following kind:

```
----------------Text title 1------------------

                   [text 1]

----------------Text title 2------------------

                   [text 2]
```

The corpus 9-11 is not suitable for this kind of application. Thus we select another corpus made up of the swearing in speeches of George W. Bush of the 20th of January 2005 and of Barak Obama of the 20th of January 2009.

The visualization of the words and their position (fig. 5.7) allows us to observe immediately the words which are present only in Bush's speech (in red) or in Obama's (in blue). The words in common are placed beside one another and the dimension of the word indicates its frequency in each text. Positioning the mouse on the word (for example, *liberty*) it is possible to observe the occurrences (15 in Bush and 2 in Obama) and the *concordances* for each text.



● 1 word ○ 2 word ☑ compare 🟧 Bush 🟦 Obama
Search :
Showing 200 Out of 1105
human justice tyranny fire states defended determined enemies excuse goal ideal independence institutions permanent questions serve single soul
democratic divisions duty ending events feel fulfill idealism influence larger lives love making meant mercy millions oppression ownership proclaimed resentment self-government soldiers
speaks standards sustained unite watch wave witnessed freedom freedom liberty liberty citizens citizens rights rights character character choice choice
fellow governments governments honor honor live live move move rule rule
america depends depends give give good good habits habits hatred hatred
find god god make make task task life life day day
promise stand stand success success work work
generations generations nations nations extend extend faith faith fear fear feed feed
service true true trust trust wealth wealth women women words words
today today generation generation government government hard hard journey journey man man oath oath end end greater greater peace peace seek seek common common new new
enduring era face faced fail false friend gift greatness grows hours icy knowledge light lines longer lower market met mutual part passed path planet play purpose remain restore short
storms stronger traveled understand understood virtue war waters west willingness winter workers year care carried children economy father jobs moment prosperity question
small crisis meet things spirit

Showing 6 of15 occurences in Bush (0.72% of text ...one conclusion: The survival of LIBERTY in our land increasingly depends...
...depends on the success of LIBERTY in other lands. The best...
...no human rights without human LIBERTY. Some, I know, have questioned...
...questioned the global appeal of LIBERTY though this time in...
...the possibility of permanent slavery. LIBERTY will come to those who...
...When you stand for your LIBERTY, we will stand with you...

2 occurences in Obama
(0.08 % of text ...they are guardians of our LIBERTY, but because they embody the...
...is the meaning of our LIBERTY and our creed why...

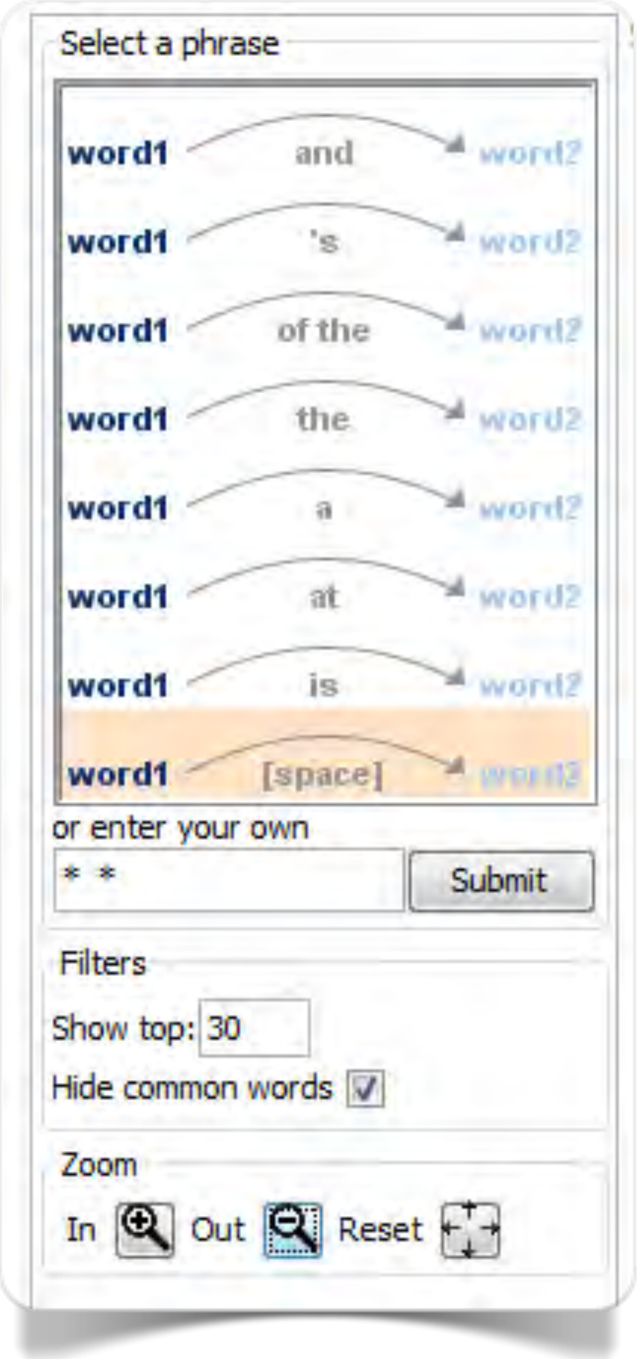
Fig. 5.7 Comparison between the 200 most frequent words of the swearing in speeches of the presidents G. W. Bush and B. Obama of the 20th of January 2005 and 2009 (Tag Cloud elaboration in IBM's Many Eyes).

Phrase Net

Phrase Net visualizes the diagram of the net of words related to one another selecting a word or a “bridge” character. The tool requires the selection of a predefined number of links, but with the option *Enter your own* it is possible to adapt the options to any language. The result in the figure has been obtained selecting space as a link between one word and another and visualising the 30 most frequent matches. The setting requires the exclusion of a list of common words (*stop words*).

Phrase Net is particularly efficacious in the identification of *repeated segments* with a crucial role in the reconstruction of the concepts representing the text content.

In this case (fig. 5.9) the central role of *muslim* (*muslim american/americans*, *muslim religious leaders*, *muslim terrorist*, *muslim islamic attack*) instead of Obama clearly emerges. The direction of the arrow indicates the word coming first and after in the segment (*9 11 memorial*). Positioning the cursor on the link, one is able to visualize the first ten occurrences of each segment.



or enter your own: * *

Show top: 30

Hide common words ☑

Fig. 5.8 Phrase Net: Selection options (IBM’s Many Eyes)

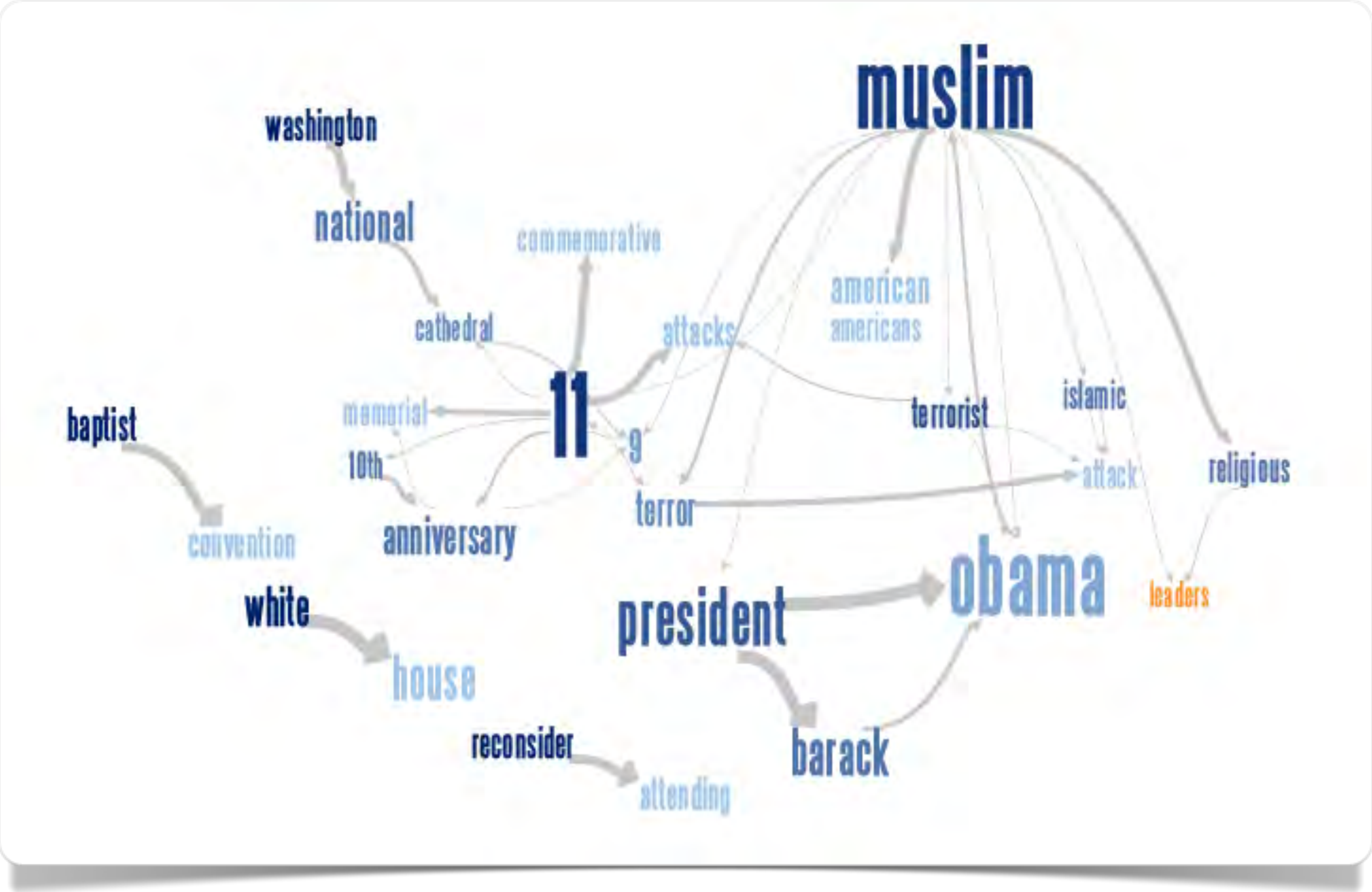


Fig. 5.9 Phrase Net del corpus 9-11 : option: [space] *30 top matches* (IBM's Many Eyes)

Word Cloud Generator

Word Cloud Generator s a simplified version of Wordle and it therefore does not require further explanations. In fig. 5.10 we can observe the visualization of the most frequent 100 words in the corpus corpus 9-11. The so called “banal” words, which were used to select the Facebook status (*Obama*, *muslime*, *9 11*),have been removed.

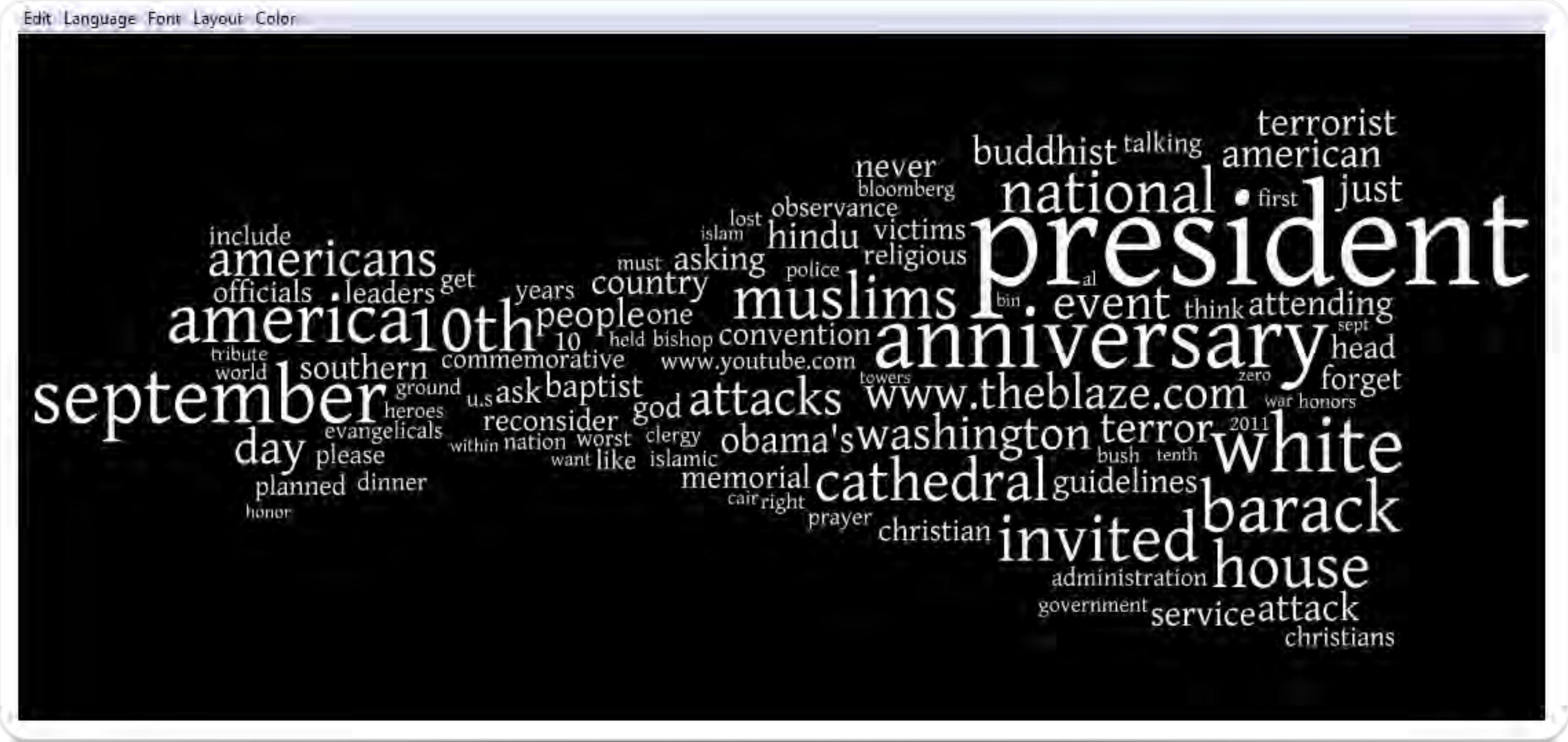


Fig. 5.10 Word Cloud Generator of the corpus 9-11 (elab. IBM’s Many Eyes)

Chapter 6

Netlytic Internet Community Text Analyzer

Netlytic is an online software created with the aim of providing a tool for the analysis of emails, newsgroups, blogs and, more in general, of the messages exchanged on the social networks.

6

Netlytic

Netlytic created by Anatoliy Gruzd (School of Information Management, Dalhousie University, Canada) has free access, but to elaborate texts, a subscription is required (Gruzd, 2010). You can start using by signing into the system with your existing Gmail or Yahoo! email account.

“Netlytic is a cloud-based text analyzer and social networks visualizer. Netlytic can automatically summarize large volumes of text and discover and visualize social networks from conversations on social media sites such as Twitter, Youtube, blog comments, online forums and chats. It is designed to help researchers and others to understand an online group’s operation, identify key and influential constituents, and discover how information and other resources flow in a network” (System Overview).

You can import a dataset using six options:

1. From RSS feeds.

2. Spreadsheets and other forms recorded in Google Drive or Dropbox.

3. From Twitter account, using keywords with operators, hashtags, or @usernames.

4. From YouTube: This option allows you to import comments data from any YouTube video.

5. From Instagram: Using this option, you can import Instagram comments using keywords or locations.

6. Text File in csv format or full-text transcripts with headers.

Using this last option, the simplest way to prepare the corpus which has to be analysed is to inserting *From* between one message and another. In the example which follows, the corpus consists in 529 messages (13th May – 7th July 2009) of a Facebook discussion group on the topic: *Fire fighters are heroes serving the community, soldiers are trained assassins serving the big corporation!*

```
From:

Claudia Green FIRE FIGHTERS ARE HEROES SERVING THE COMMUNITY, SOLDIERS ARE TRAINED ASSASSINS SERVING THE BIG
CORPORATIONS! WAKE UP NAIVE AND NARROW MINDED SUCKERS! PATRIOTISM TRANSLATES STUPIDITY!

From:

Kyle Eberhardt Lady we fight for eachother and your family not the government that employs us or thier
cronnies. All the BS is coming from within Washington DC your own government. The Us millitary is the last
half way honest thing this country has. Revolution is coming and most of us side with the PEOPLE that mean
you. So help us take back our country, because Washington is not going to help you put out the fire when the
time comes . Shure they'll take your money and promiss to do a good job but look how that's turn out

From:

Jerry Weed What an ignorant person you are. If you live in the US and think what we do for you is wrong then
get out of our country! Go live over in Iraq and Afghanistan and then tell us we're naive and narrow minded.

From:

Annabel Ward To be fair that's probably what the Iraqis think "get out of our country"-not really a great
line of defence for a nation that goes wherever it chooses.

[...]
```

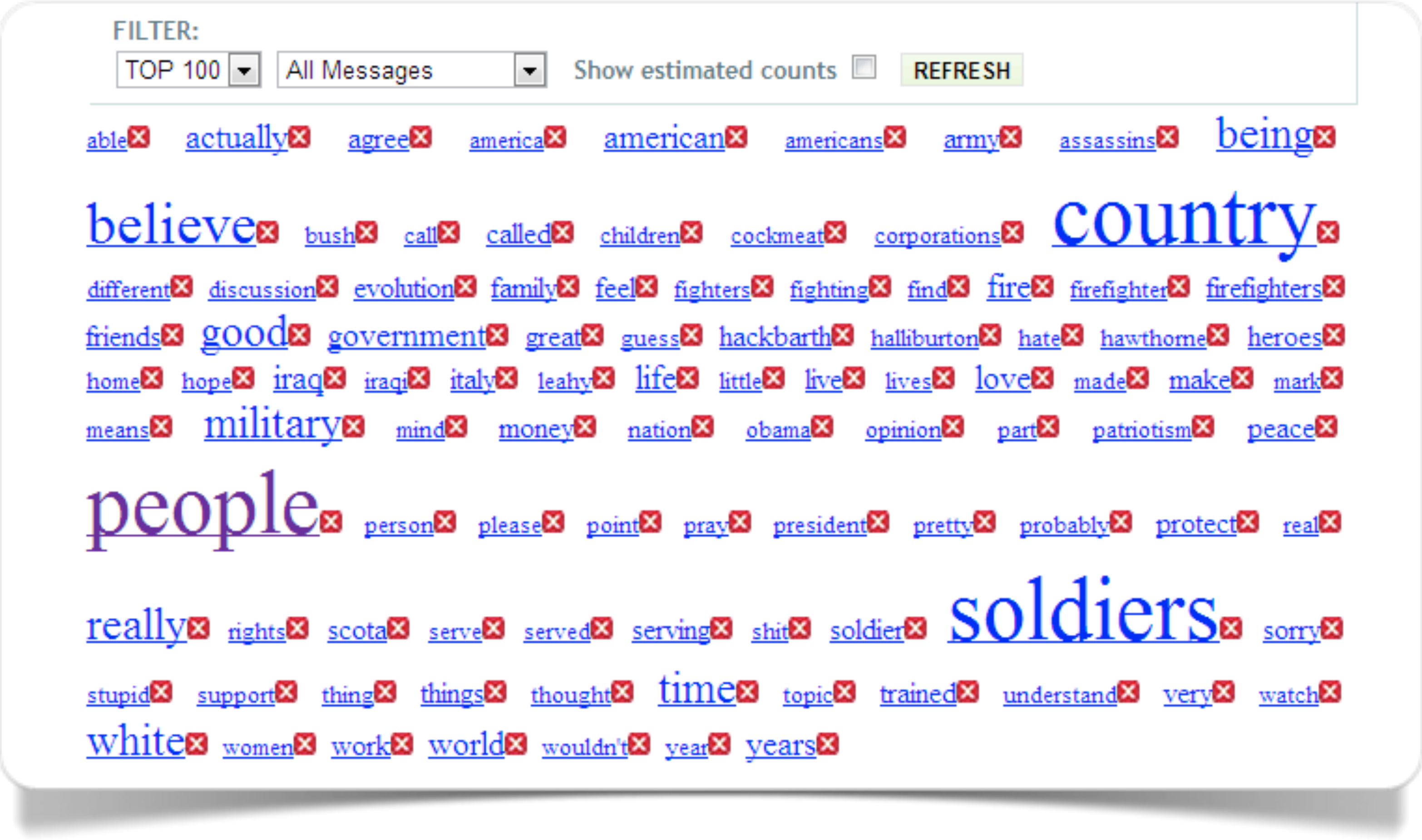
After the upload, the corpus can be analysed from the menu *Text Analysis* with two main tools:

1. *Keyword Extractor*: It identifies the most frequent graphic forms after eliminating the banal words and the stopwords, whose list of over 500 terms, can be visited from a link activated in the elaboration windows.
2. *Categories*: It classifies some graphic forms or text segments according to certain content criteria in order to guarantee a quick – also quantitative - visualisation.

Netlytic can count on various options for text normalisation and these allow the user to eliminate the causes of noise which are so common when communicating via computer. This operation can be carried out during the corpus upload phase or at the end of a first preliminary analysis using the tools described above. The application of Netlytic to our Facebook corpus will allow us to understand better the various stages.

Keyword Extractor

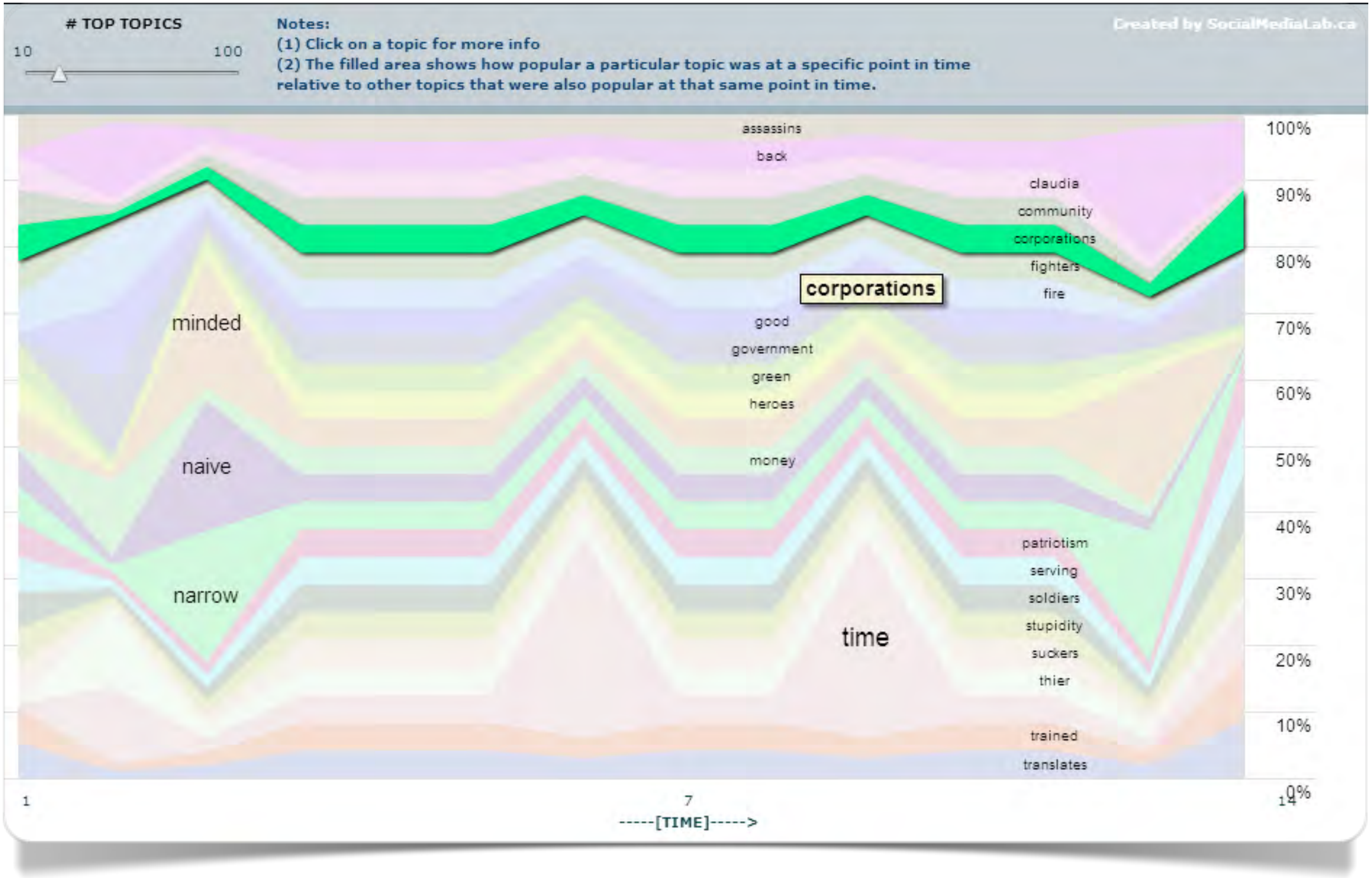
Thanks to the tool *Keyword Estractor* we obtain the results visualized in fig. 6.1 (very similar to a Many Eyes Tag Cloud) in which the most frequent words are represented in a proportionally bigger size. Names have been removed placing the cursor directly on the red cross on the right of each graphic form.



able actually agree america american americans army assassins being
believe bush call called children cockmeat corporations country
different discussion evolution family feel fighters fighting find fire firefighter firefighters
friends good government great guess hackbarth halliburton hate hawthorne heroes
home hope iraq iraqi italy leahy life little live lives love made make mark
means military mind money nation obama opinion part patriotism peace
people person please point pray president pretty probably protect real
really rights scota serve served serving shit soldier soldiers sorry
stupid support thing things thought time topic trained understand very watch
white women work world wouldn't year years

Fig. 6.1 Visualization of the top 100 extracted graphic forms with Keyword Extractor in the Facebook group “Fire Fighters are heroes”; 13th of May- 7th of July 2009 (elab. Netlytic)

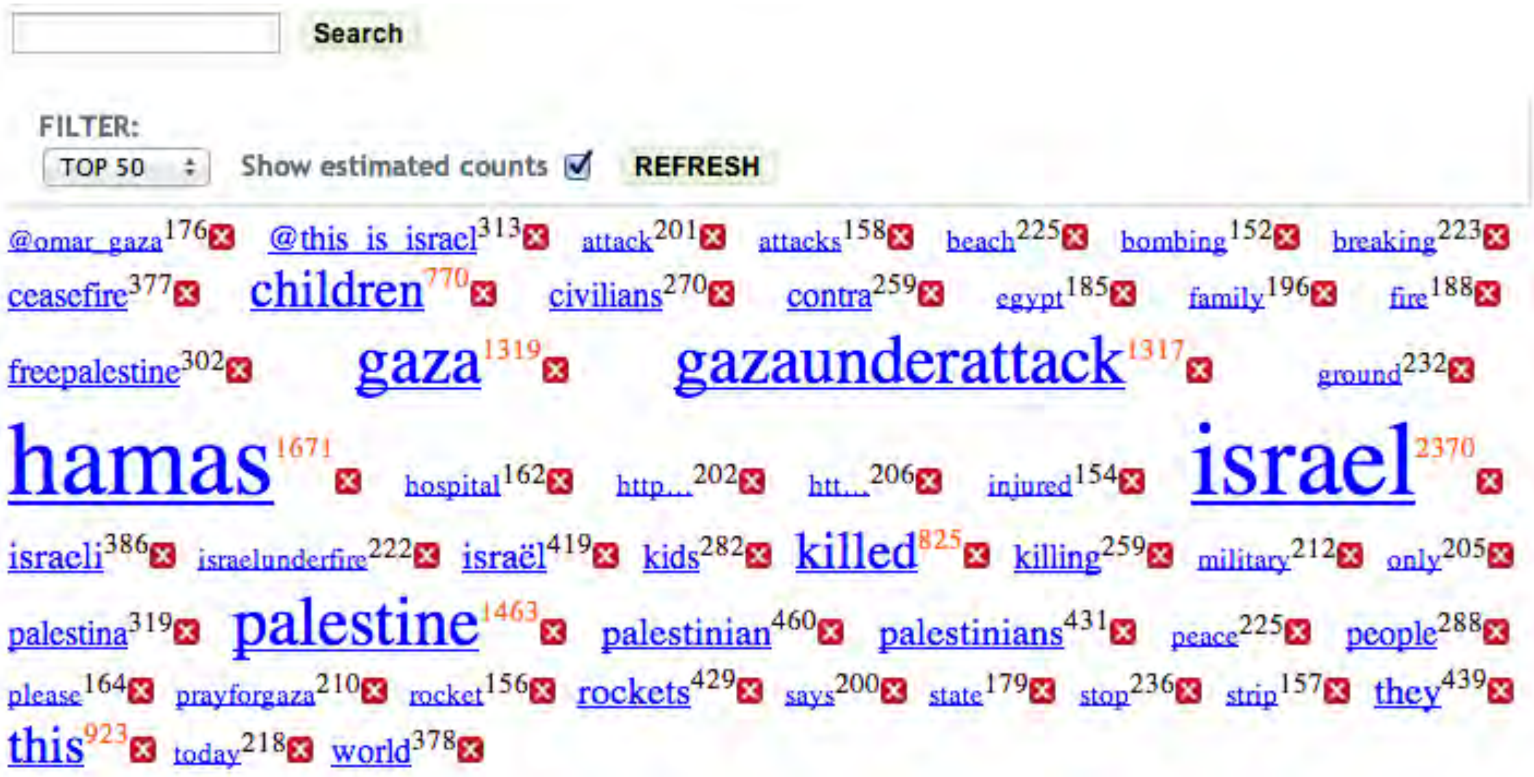
The graphical visualization of the result (fig. 6.2) allows us to appreciate the presence of the "concepts" extracted during the entire discussion period.



Notes:
(1) Click on a topic for more info
(2) The filled area shows how popular a particular topic was at a specific point in time relative to other topics that were also popular at that same point in time.

Fig. 6.2 Graphical visualization of the top 25 extracted graphic forms with Keyword Extractor in the Facebook group "Fire Fighters are heroes"; 13th of May- 7th of July 2009 (elab. Netlytic)

In the first part the participants mainly refer to the word *heroes* and in a polemic way to the activity of “worthless politicians”, while in the last part, they talk about the US presidents Bush and Obama (fig. 6.1), the vice-president Cheney and the American multinational Hulliburton (*corporations*) specialized in oilfield exploitation and strongly suspected of illegal practices during the war in Iraq. Significant is the presence of the words *fighter/s, money, narrow, stupidity* to express the indignation of most of the participants for the matching of the words *soldiers* and *assassins*. The percentage represents at a specific moment the distribution of the argumentations according to their frequency.

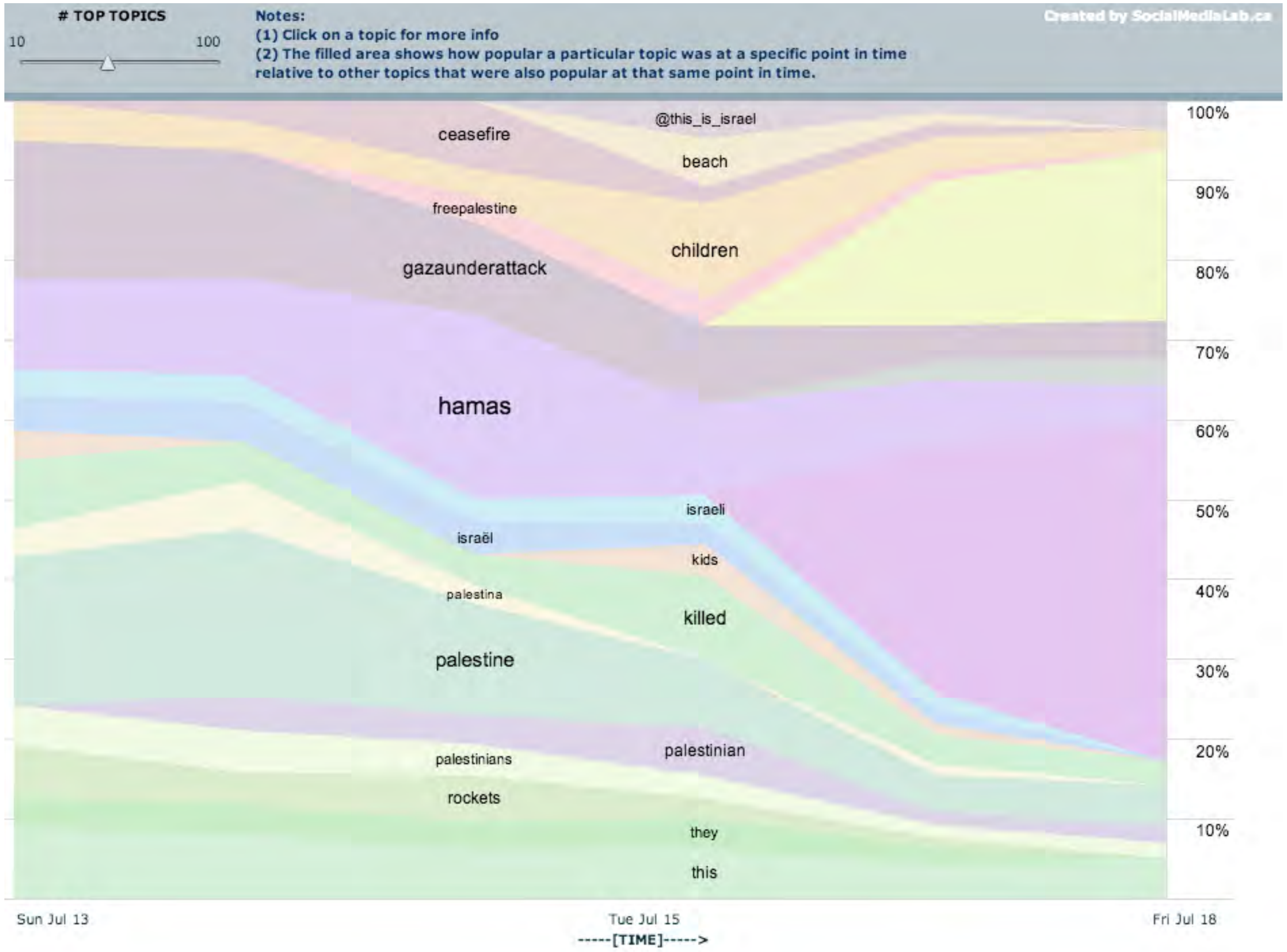


FILTER: TOP 50 Show estimated counts ☑

@omar_gaza 176 @this_is_israel 313 attack 201 attacks 158 beach 225 bombing 152 breaking 223
ceasefire 377 children 770 civilians 270 contra 259 egypt 185 family 196 fire 188
freepalestine 302 gaza 1319 gazaunderattack 1317 ground 232
hamas 1671 hospital 162 http... 202 htt... 206 injured 154 israel 2370
israeli 386 israelunderfire 222 israël 419 kids 282 killed 825 killing 259 military 212 only 205
palestina 319 palestine 1463 palestinian 460 palestinians 431 peace 225 people 288
please 164 prayforgaza 210 rocket 156 rockets 429 says 200 state 179 stop 236 strip 157 they 439
this 923 today 218 world 378

Fig. 6.3 Visualization of the top 50 extracted graphic forms with Keyword Extractor in Twitter: 13th - 18th of July 2014; keyword: #Istrael (elab. Netlytic)

In the second exemple (fig. 6.3 and fig. 6.4), we can observe the results of the analysis of 12,000 tweets (search keyword #*Israel*; 13th - 18th of July 2014).



Notes:
(1) Click on a topic for more info
(2) The filled area shows how popular a particular topic was at a specific point in time relative to other topics that were also popular at that same point in time.

Fig. 6.4 Graphical visualization of the top 50 extracted graphic forms with Keyword Extractor in Twitter: 13th - 18th of July 2014; keyword: #Israel (elab. Netlytic)

Categories

As we have already said, the tool you can gain access to from the menu *Text Analysis*, allows you to classify words, compound words or sentences, following linguistic or semantic criteria till you obtain the synthesis of the contents considered or the general tone of the communication (Pennebaker e Graybeal, 2001).

The main predefined categories are: *agreement*, *certainly*, *disagreement*, *evaluation*, *opinion*, *positive*, *reference*, *self*, *uncertainty*, *us*. The user can add or change the pre-existent categories creating a personal list of words or classification criteria. This tool makes it possible to apply to the corpus a real semi-automatic content analysis procedure (Losito, 2002; Krippendorf, 2004).

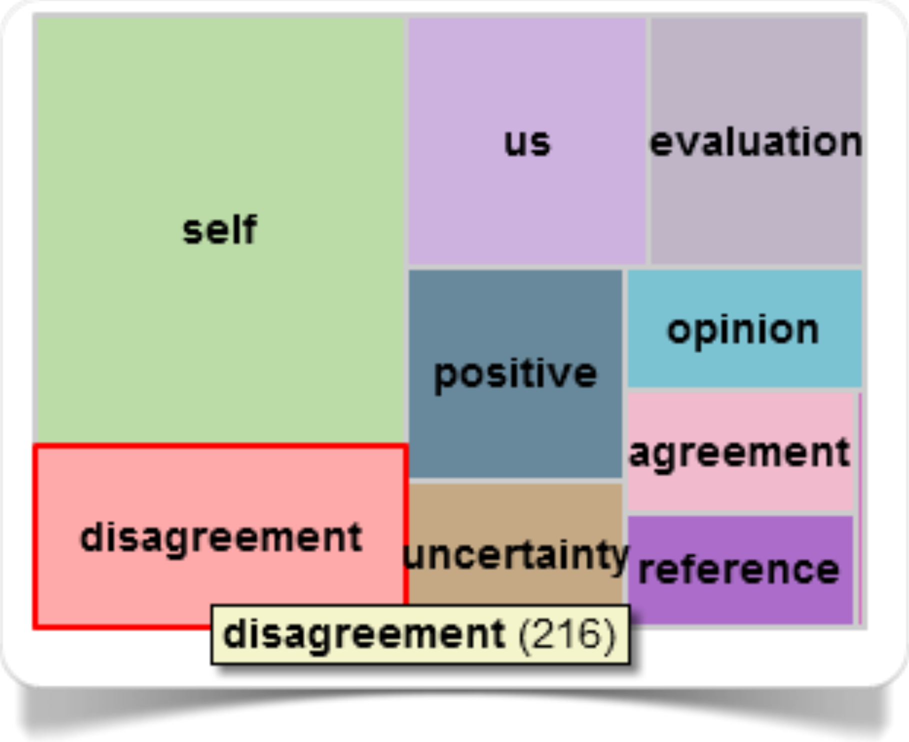


Fig. 6.5 Visualisation of the Cognitive & Social Categories in the Facebook group "Fire Fighters are heroes"; 13th of May - 7th of July 2009 (elab. Netlytic)

In this case the elaboration using the tool *Categories* points out that we are considering an exchange of opinions in which the elements of disagreement are more than those of agreement. In order to visualize the graph (fig. 6.5), it is necessary to click on the menu *Home*, select *Visualize* in correspondence of the corpus under elaboration and then choose *Cognitive & Social Categories*.

The Gallery 6.1 shows the main categories with reference to the tweets of the conflict between Israel and Hamas and some screenshots of depth in the categories *Condition* and *Feelings (bad)*.

Gallery 6.1 Visualization of the Cognitive & Social Categories in Twitter: 13th - 18th of July 2014; keyword: #Israel (elab. Netlytic)

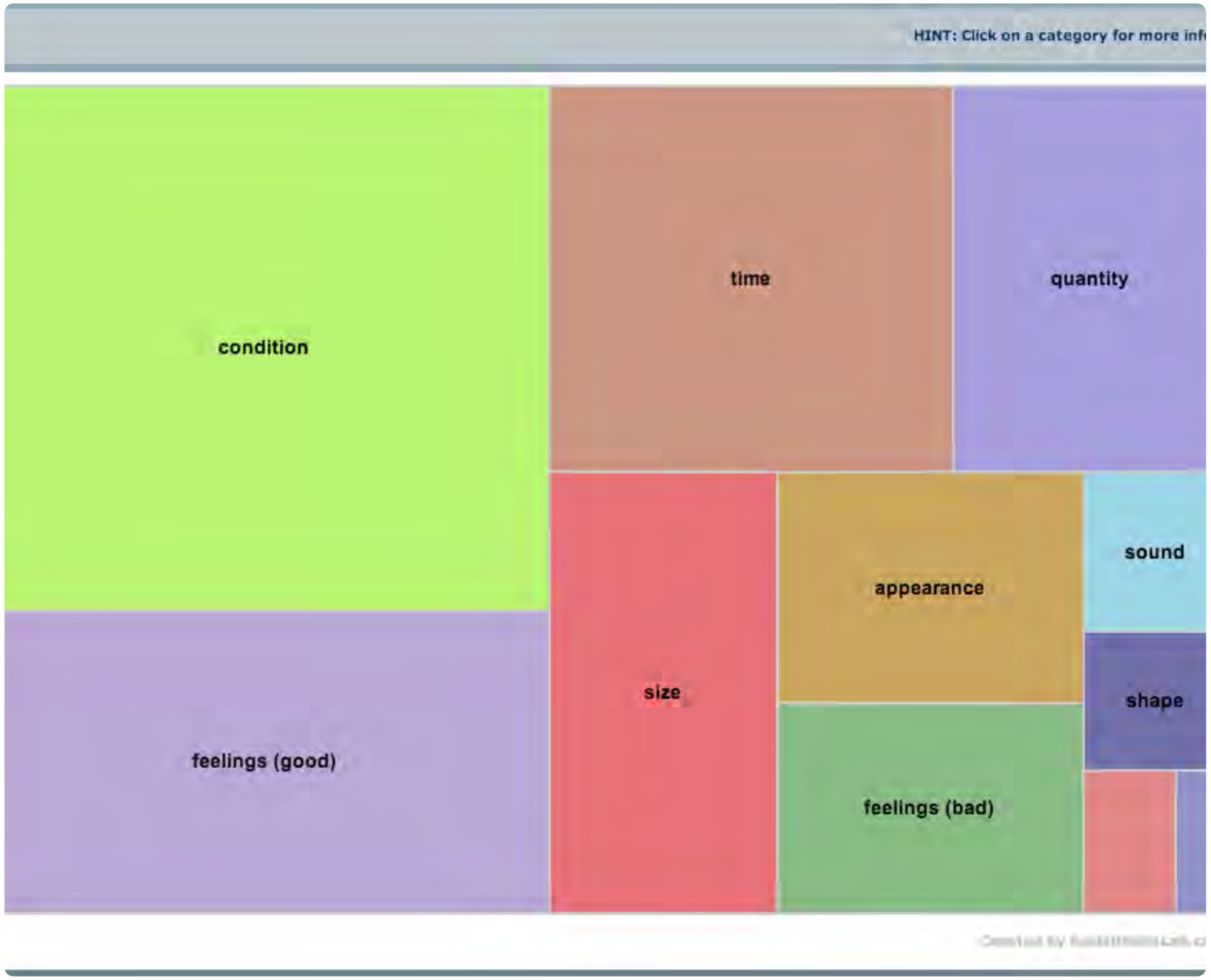
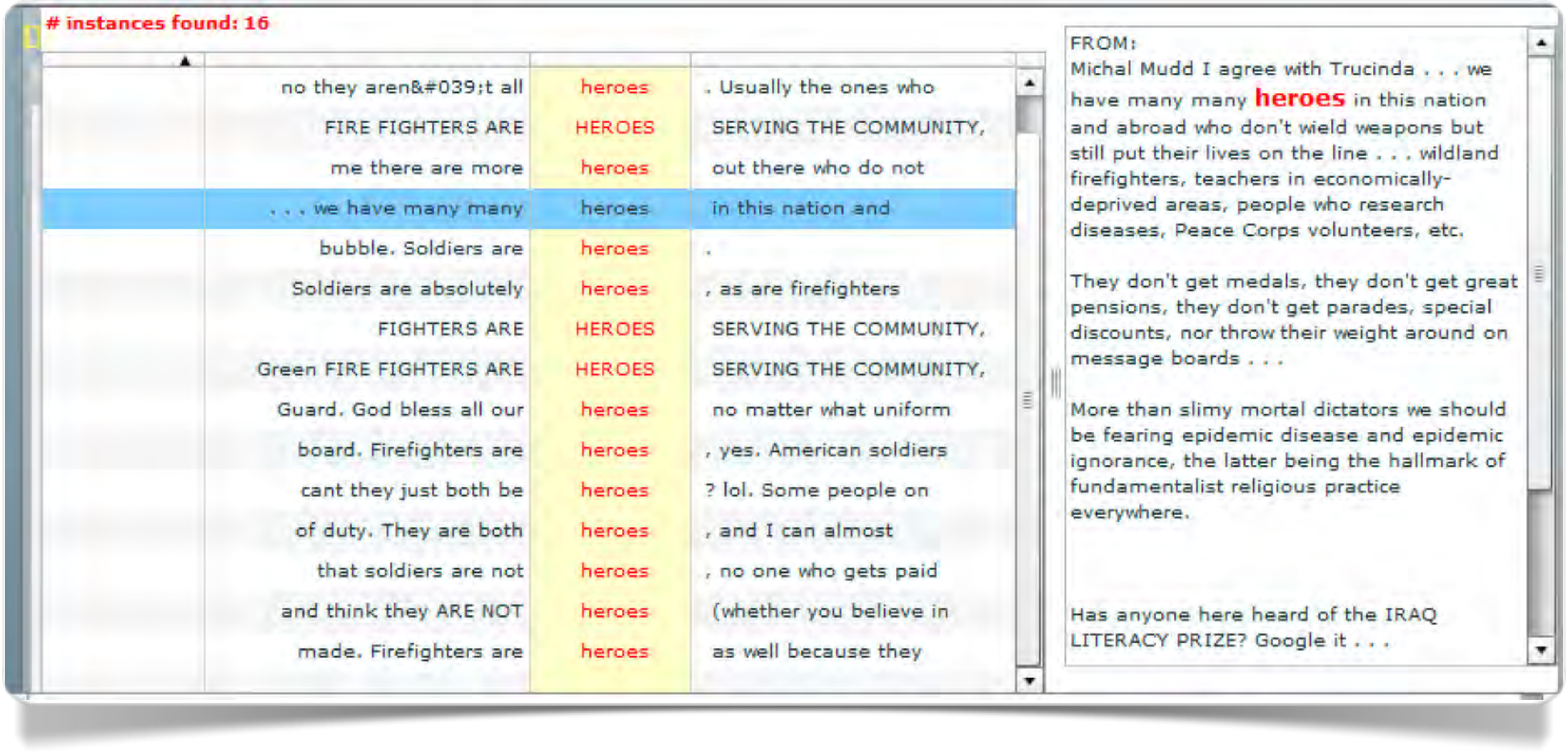

HINT: Click on a category for more inf

Fig. 6.1.1 Categories

Concordance visualization

In all the tools presented, positioning the cursor on one of the categories or on the words themselves, it is always possible to obtain the identification of the concordances (fig. 6.6) and then the overall context (message or text) in which one can find the word capable of facilitating the interpretation of the results and the general meaning of the visualizations (Haythornthwaite & Gruzd, 2007).



instances found: 16

no they aren&#039;t all	heroes	. Usually the ones who
FIRE FIGHTERS ARE	HEROES	SERVING THE COMMUNITY,
me there are more	heroes	out there who do not
. . . we have many many	heroes	in this nation and
bubble. Soldiers are	heroes	.
Soldiers are absolutely	heroes	, as are firefighters
FIGHTERS ARE	HEROES	SERVING THE COMMUNITY,
Green FIRE FIGHTERS ARE	HEROES	SERVING THE COMMUNITY,
Guard. God bless all our	heroes	no matter what uniform
board. Firefighters are	heroes	, yes. American soldiers
cant they just both be	heroes	? lol. Some people on
of duty. They are both	heroes	, and I can almost
that soldiers are not	heroes	, no one who gets paid
and think they ARE NOT	heroes	(whether you believe in
made. Firefighters are	heroes	as well because they

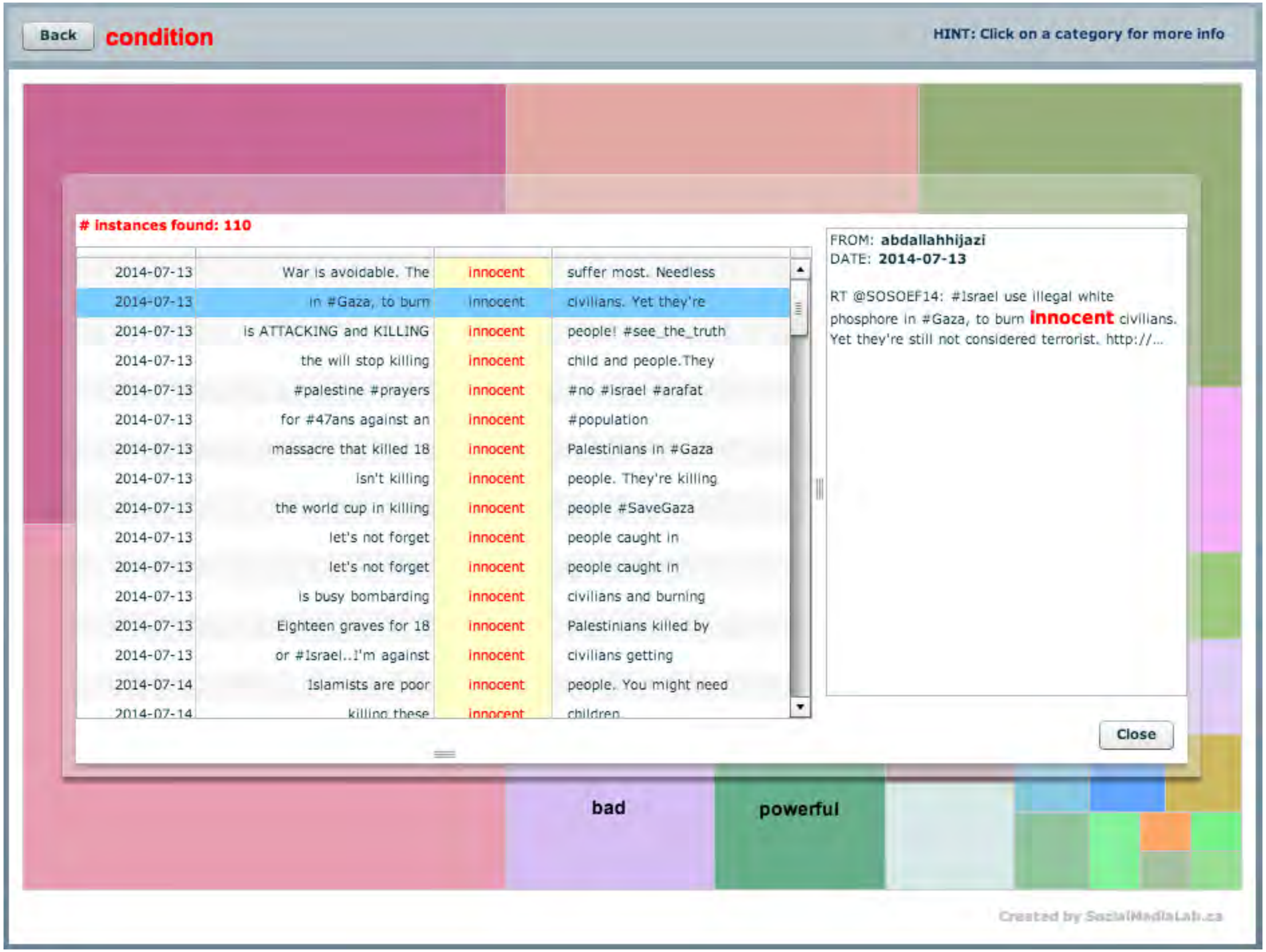
FROM:
Michal Mudd I agree with Trucinda . . . we have many many **heroes** in this nation and abroad who don't wield weapons but still put their lives on the line . . . wildland firefighters, teachers in economically-deprived areas, people who research diseases, Peace Corps volunteers, etc.

They don't get medals, they don't get great pensions, they don't get parades, special discounts, nor throw their weight around on message boards . . .

More than slimy mortal dictators we should be fearing epidemic disease and epidemic ignorance, the latter being the hallmark of fundamentalist religious practice everywhere.

Has anyone here heard of the IRAQ LITERACY PRIZE? Google it . . .

Fig. 6.6 Visualization of the concordances and of the context of the word *heroes* in the Facebook group “Fire Fighters are heroes”; 13th of May - 7th of July 2009 (elab. Netlytic)


Back
condition
HINT: Click on a category for more info
Close
bad
powerful
Created by SocialMediaLab.ca

instances found: 110

2014-07-13	War is avoidable. The	innocent	suffer most. Needless
2014-07-13	in #Gaza, to burn	innocent	civilians. Yet they're
2014-07-13	is ATTACKING and KILLING	innocent	people! #see_the_truth
2014-07-13	the will stop killing	innocent	child and people.They
2014-07-13	#palestine #prayers	innocent	#no #israel #arafat
2014-07-13	for #47ans against an	innocent	#population
2014-07-13	massacre that killed 18	innocent	Palestinians in #Gaza
2014-07-13	isn't killing	innocent	people. They're killing
2014-07-13	the world cup in killing	innocent	people #SaveGaza
2014-07-13	let's not forget	innocent	people caught in
2014-07-13	let's not forget	innocent	people caught in
2014-07-13	is busy bombarding	innocent	civilians and burning
2014-07-13	Eighteen graves for 18	innocent	Palestinians killed by
2014-07-13	or #Israel..I'm against	innocent	civilians getting
2014-07-14	Islamists are poor	innocent	people. You might need
2014-07-14	killing these	innocent	children

FROM: **abdallahhijazi**
DATE: **2014-07-13**

RT @SOSOEF14: #Israel use illegal white phosphore in #Gaza, to burn **innocent** civilians. Yet they're still not considered terrorist. http://...

Fig. 6.7 Visualization of the concordances and the context of the category *Innocent* in Twitter: 13th - 18th of July 2014; keyword: #Israel (elab. Netlytic)

Network Analysis

From the menu *Text Analysis* you gain access to the tool *Analysis*, and even though it is not closely inherent to the automatic analysis of textual data, it allows you to visualize the exchange of opinions among the users, or simply, the co-occurrences of the names within the messages or the segments in which the corpus is divided.

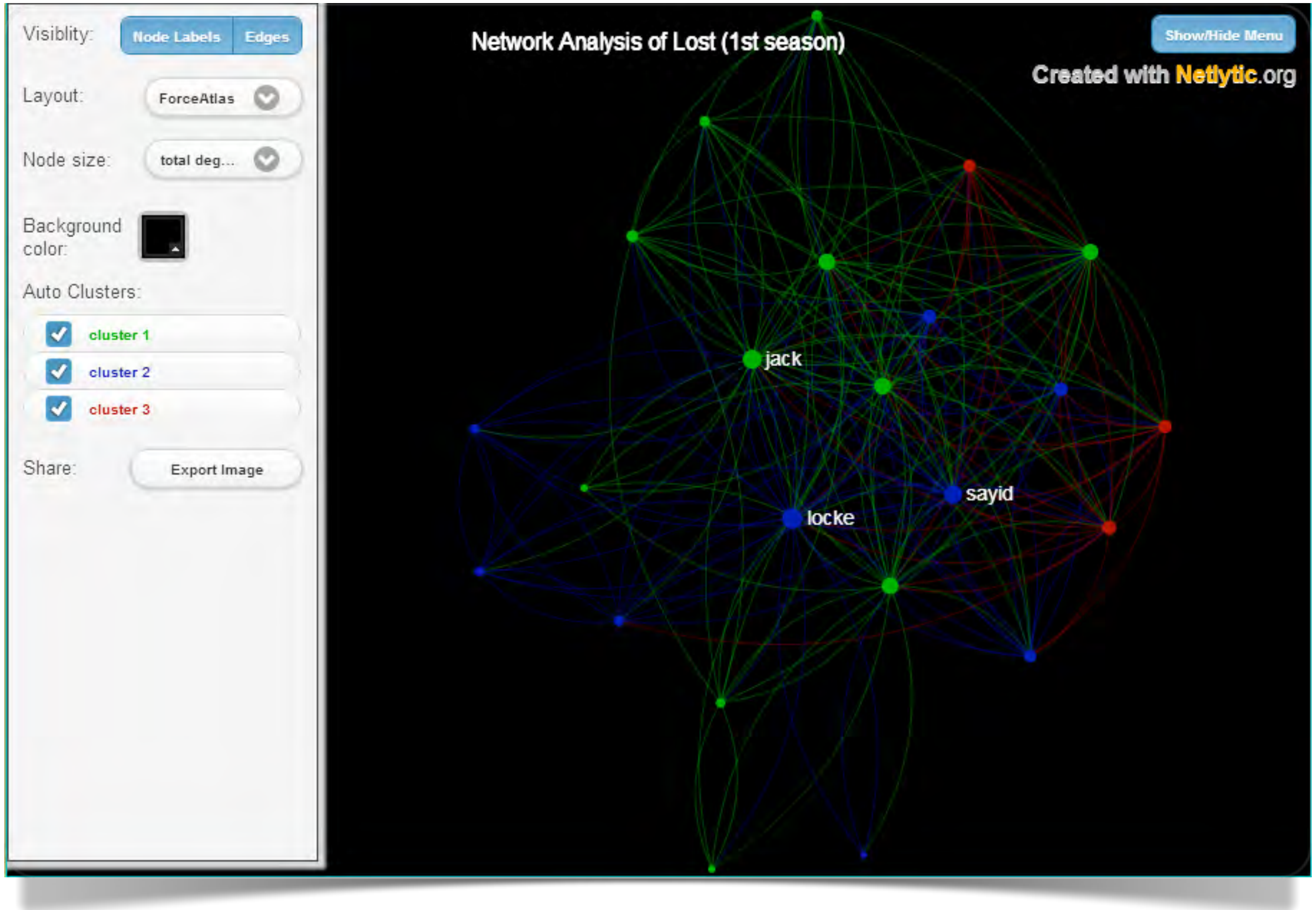
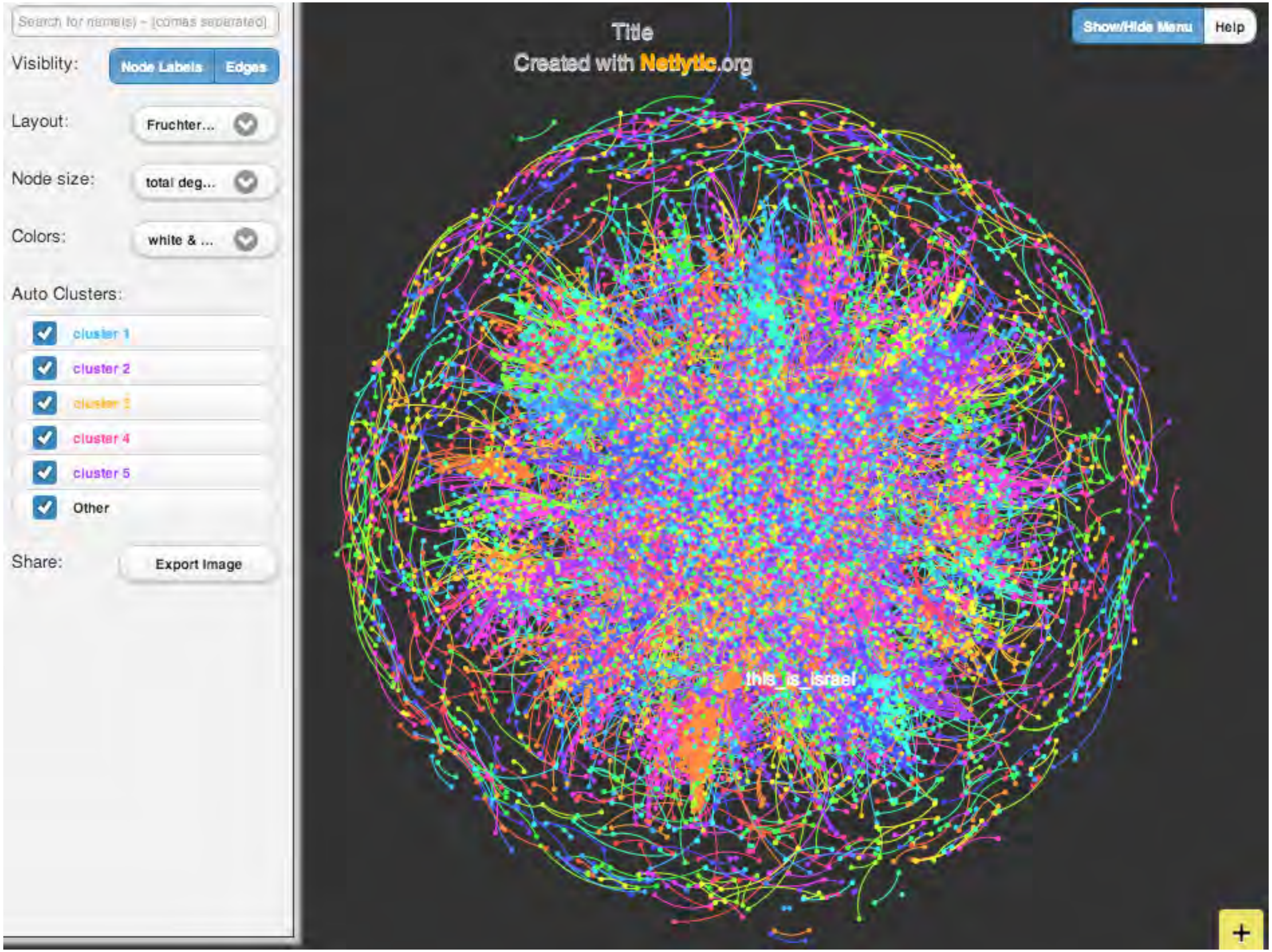


Fig. 6.8 The network of *Lost* (ABC Studios - elab. Netlytic)

In Figure 6.8, we can observe the network of the first season of the television series Lost (ABC Studios) where it is clear the centrality of the interactions around three characters: Jack, Sayid, and Locke. In Figure 6.9 is shown the network of tweets on the conflict between Israel and Hamas.



Visiblity: Node Labels Edges

Layout: Fruchter...

Node size: total deg...

Colors: white & ...

Auto Clusters:

- ☑ cluster 1
- ☑ cluster 2
- ☑ cluster 3
- ☑ cluster 4
- ☑ cluster 5
- ☑ Other

Share: Export Image

Search for name(s) - (comas separated)
Title
Created with Netlytic.org
Show/Hide Menu
Help
this_is_israel
+

Fig. 6.9 Network Analysis of Twitter: 13th - 18th of July 2014; keyword: #Israel; Who mentions whom (elab. Netlytic)

Chapter 7

TAPoR Text Analysis Portal of Research

TAPoR is a project sponsored by six Canadian universities (Montréal, Alberta, Toronto, New Brunswick, Hamilton).

7

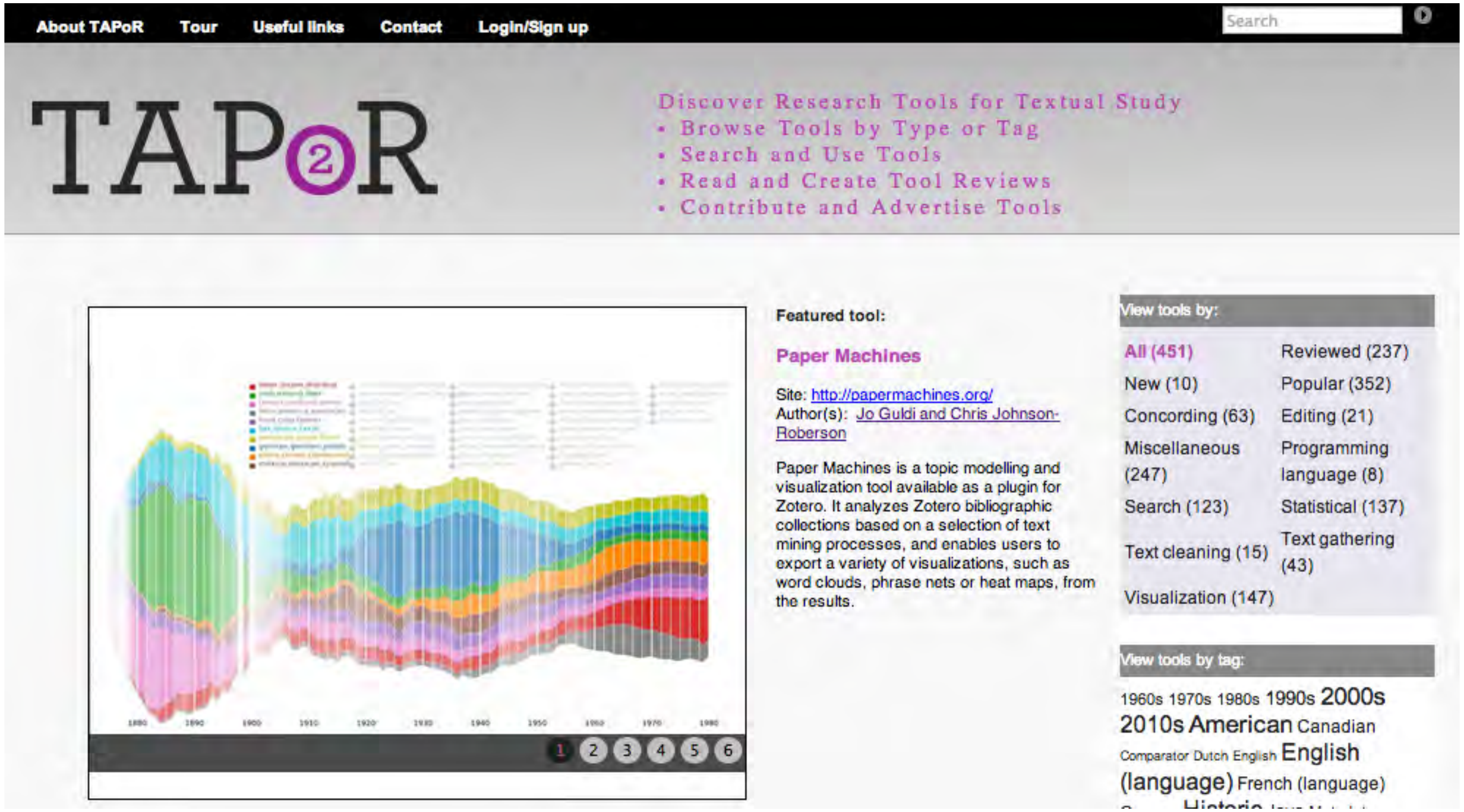
TAPoR & TAPoRwere

TAPoR – Text Analysis Portal of Research, started in 2002 and coordinated by Geoffrey Rockwell, is in full development (Tapor 2) and aims at becoming an experimental portal for the analysis of digital texts (Rockwell, 2003).

The portal is rich of documents and updatings on the state of research of computational linguistics and automatic text analysis. In the section *TAPoR Text* are gathered sample texts and a selection of didactic material. The user can subscribe for free and create with a personal account, a section of personal texts; if he/she thinks it is the case, he/she can let them be publicly used by everyone. Logging in, one enters two main work environments - TAPoRware and Voyant – which interact inside the portal. Some options are, however, more complex and unstable and is therefore preferable to gain access to the most reliable visualisation and analysis tools from outside. A guide to the main functions is available in English, French and Italian.

Fig. 7.1 TAPoR project - 2002 logo

TAPoRwere – Prototype of Text Analysis Tools has been developed by Geoffrey Rockwell, Lian Yan, Andrew Macdonald and Matt Patey. The tools are compatible with HTML, XML and Plain Text files and they can be elaborated directly in URL or uploading the file from a personal computer. In general, in this first environment you can find introductory tools, but also experimental applications still to be fully developed, such as Raw Grep (a concordance generator that uses as pivot text strings instead of words), Keyords Finder (a keyword identifier based on the principle of the maximum frequency of content words and n-grams), Word Brush (a visualizer of "brush" words with applications, which for now are only aesthetic). Lately it has been added a viewer of frequencies in which the words are represented by drops of water: Voyant Term Fountain (click on *Open*, select *Shakespeare's play* and *OK*).



Discover Research Tools for Textual Study

- Browse Tools by Type or Tag
- Search and Use Tools
- Read and Create Tool Reviews
- Contribute and Advertise Tools

Featured tool:

Paper Machines

Site: http://papermachines.org/
Author(s): Jo Guldi and Chris Johnson-Roberson

Paper Machines is a topic modelling and visualization tool available as a plugin for Zotero. It analyzes Zotero bibliographic collections based on a selection of text mining processes, and enables users to export a variety of visualizations, such as word clouds, phrase nets or heat maps, from the results.

View tools by:

All (451)	Reviewed (237)
New (10)	Popular (352)
Concording (63)	Editing (21)
Miscellaneous (247)	Programming language (8)
Search (123)	Statistical (137)
Text cleaning (15)	Text gathering (43)
Visualization (147)	

View tools by tag:

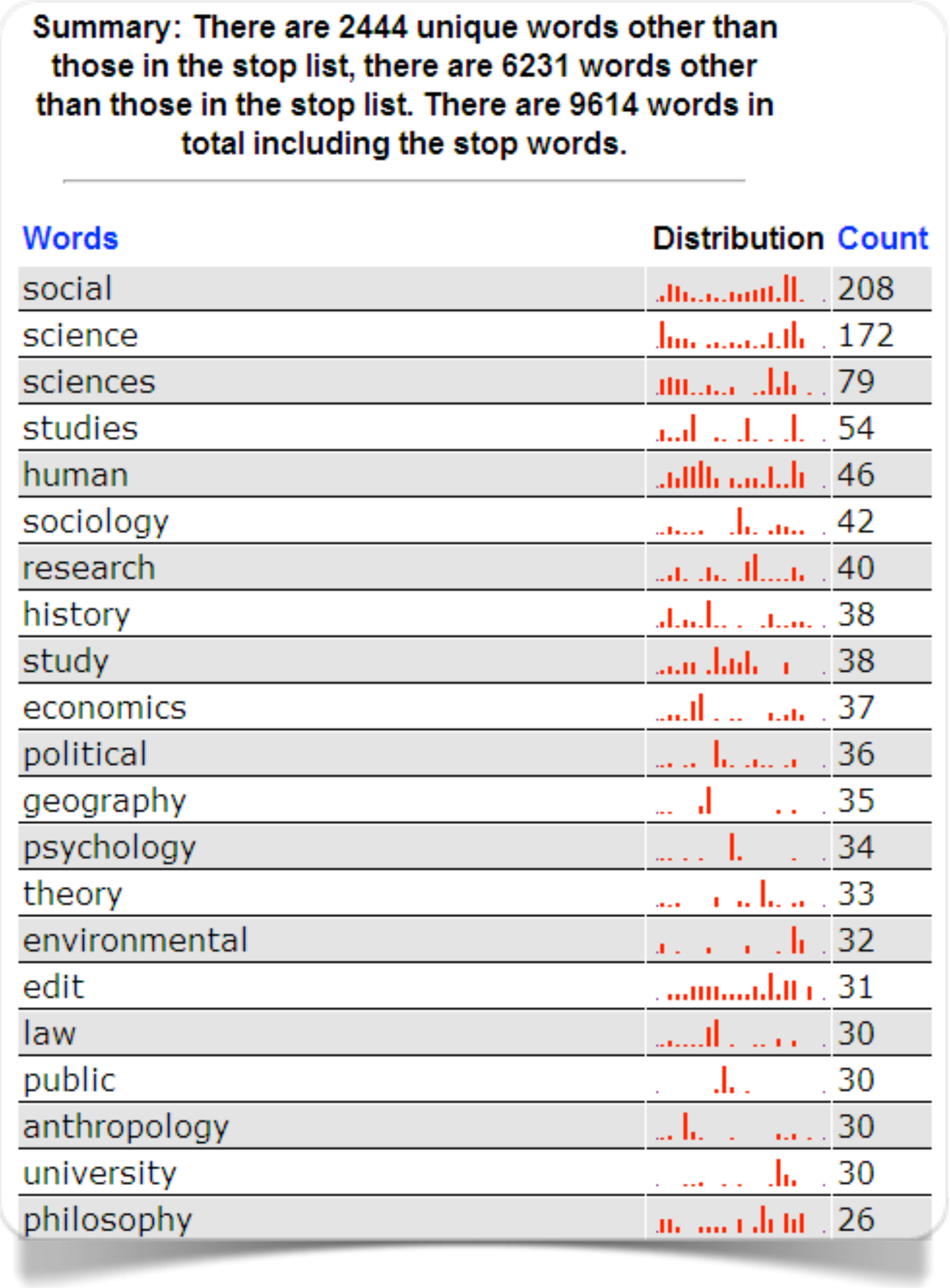
1960s 1970s 1980s 1990s 2000s 2010s American Canadian Comparator Dutch English English (language) French (language) Historic

Fig. 7.2 TAPOR 2 - Main Page

List Words

List Words counts the occurrences and the frequency of the vocabulary. It is possible to apply the *Glasgow Stop Words* list generated by automatic methods whose origin dates back to *Text Mining and Information Retrieval* studies (Lo Tsz-Wai *et al*., 2005) or a list of words to exclude from the analysis prepared autonomously by the user. There are different output options, among which the vocabulary in decreasing or increasing order of frequency, in alphabetical order or in order of appearance in the text.

In the example we can see the output of *Social Science* in Wikipedia. In the second column of fig. 7.3 we can see the distribution of a certain number of chosen words (usually the five or twenty most frequent words) in all the corpus divided into 20 fragments (5% of the corpus per fragment).



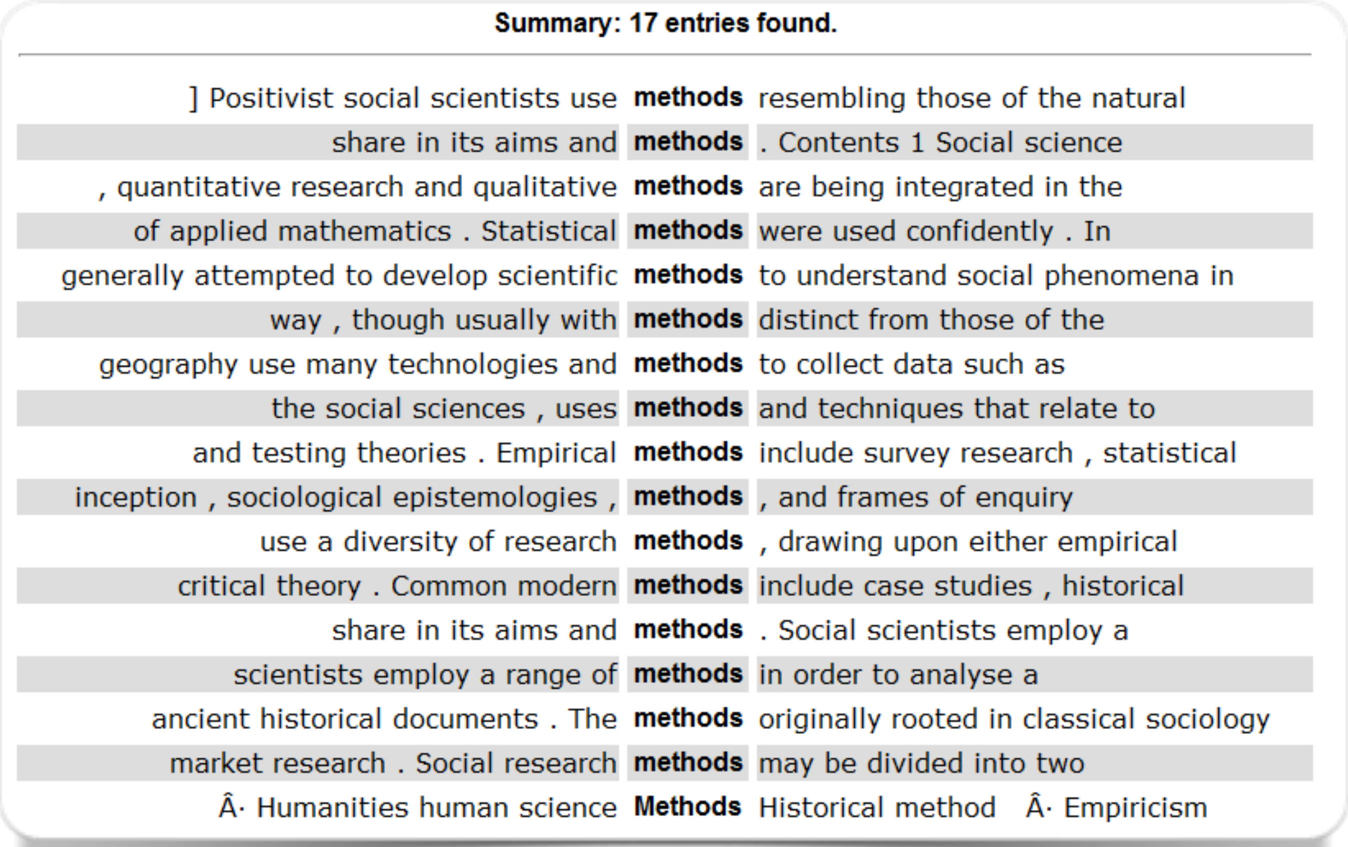
Summary: There are 2444 unique words other than those in the stop list, there are 6231 words other than those in the stop list. There are 9614 words in total including the stop words.

Words	Distribution	Count
social		208
science		172
sciences		79
studies		54
human		46
sociology		42
research		40
history		38
study		38
economics		37
political		36
geography		35
psychology		34
theory		33
environmental		32
edit		31
law		30
public		30
anthropology		30
university		30
philosophy		26

Fig. 7.3 The twenty most frequent forms in the Wikipedia entry *Social Science* (elab. Taporware)

Concordance

The tool Concordance produces as a risults le concordances with reference to a certain pivot word or a regular Unix expression (word/pattern). Among the options there is the possibility to choose the context in which the concordances can be placed, the paragraphs and the indication of the number of the surrounding words to be considered. In the example (fig. 7.4) we can see the output of the concordances of the word *methods* in the Wikipedia entry “Social Science”.

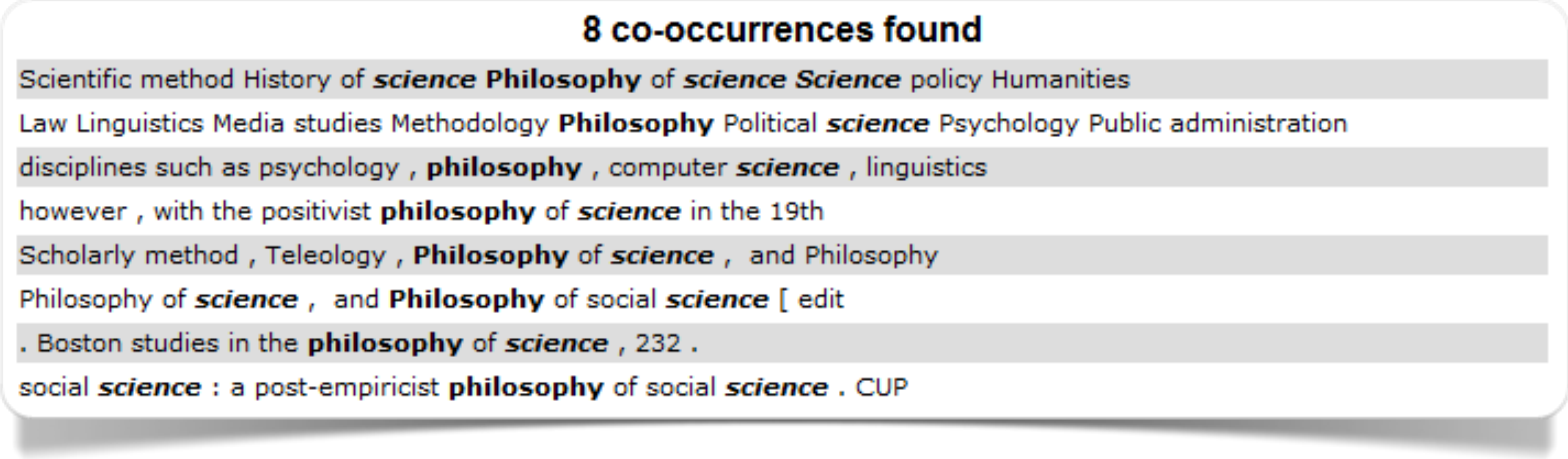


```
                   Summary: 17 entries found.

       ] Positivist social scientists use methods resembling those of the natural
                    share in its aims and methods . Contents 1 Social science
  , quantitative research and qualitative methods are being integrated in the
     of applied mathematics . Statistical methods were used confidently . In
generally attempted to develop scientific methods to understand social phenomena in
                way , though usually with methods distinct from those of the
      geography use many technologies and methods to collect data such as
               the social sciences , uses methods and techniques that relate to
         and testing theories . Empirical methods include survey research , statistical
inception , sociological epistemologies , methods , and frames of enquiry
              use a diversity of research methods , drawing upon either empirical
          critical theory . Common modern methods include case studies , historical
                    share in its aims and methods . Social scientists employ a
             scientists employ a range of methods in order to analyse a
       ancient historical documents . The methods originally rooted in classical sociology
        market research . Social research methods may be divided into two
              Â· Humanities human science Methods Historical method   Â· Empiricism
```

Fig. 7.4 Concordances of *methods*;
Wikipedia entry: “Social Science” (elab. Taporware)

Co-occurrence

Co-occurrence produces as output the co-occurrences of two words within a context defined as the distance between them. In the example (fig. 7.5) , we can see the co-occurrences of “philosophy” and “science” in the context of 5 words and with reference to the Wikipedia entry *Social Science*.



8 co-occurrences found

Scientific method History of ***science*** **Philosophy** of ***science*** ***Science*** policy Humanities

Law Linguistics Media studies Methodology **Philosophy** Political ***science*** Psychology Public administration

disciplines such as psychology , **philosophy** , computer ***science*** , linguistics

however , with the positivist **philosophy** of ***science*** in the 19th

Scholarly method , Teleology , **Philosophy** of ***science*** , and Philosophy

Philosophy of ***science*** , and **Philosophy** of social ***science*** [edit

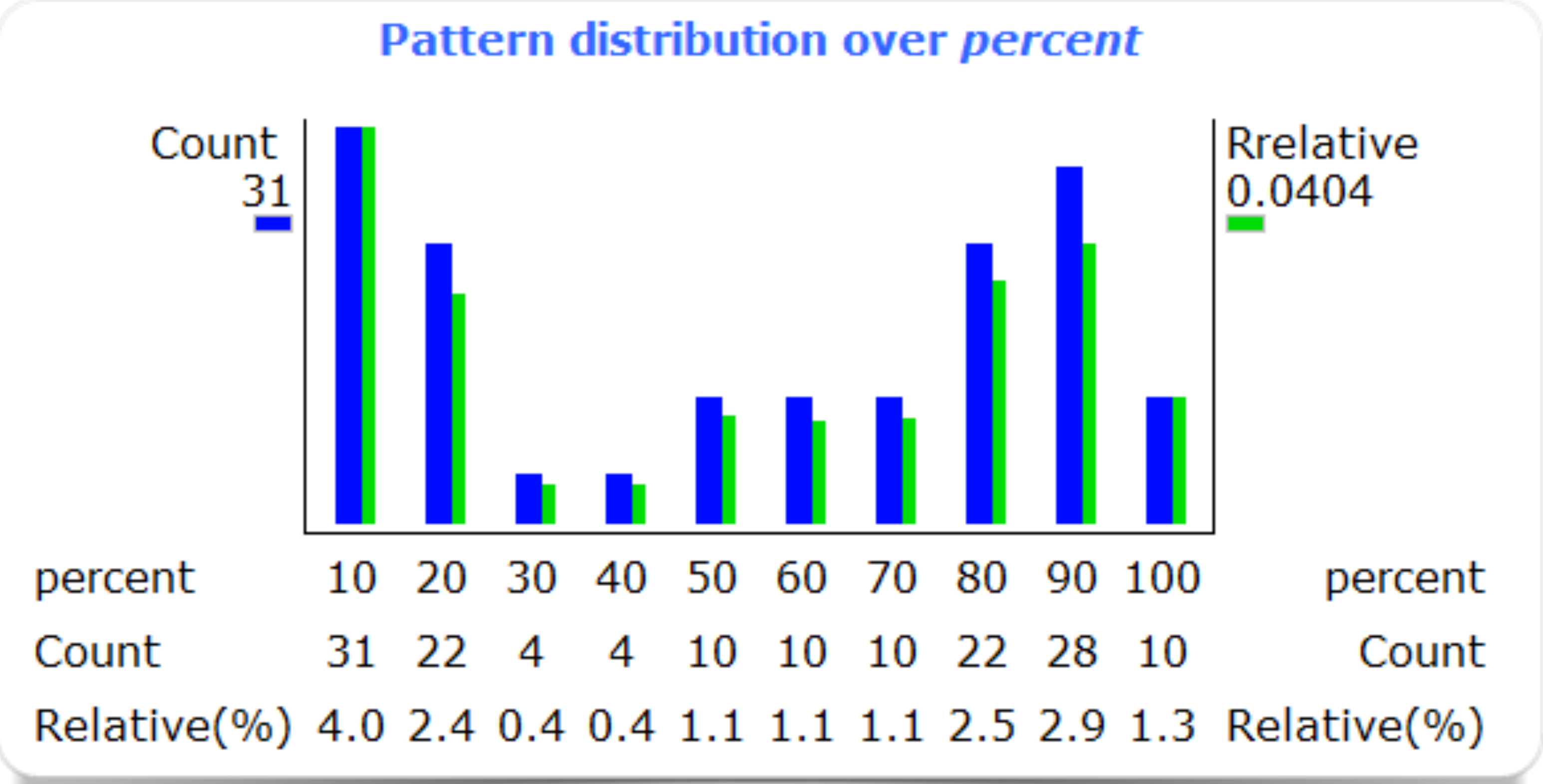
. Boston studies in the **philosophy** of ***science*** , 232 .

social ***science*** : a post-empiricist **philosophy** of social ***science*** . CUP

Fig. 7.5 Co-occurrences of the words *philosophy* and *science* in the Wikipedia entry “Social Science” (elab. Taporware)

Distribution

The tool Distribution allows to observe the distribution of the absolute and relative frequencies of the word selected in the corpus divided into 10 fragments (10% of the text for each fragment, or according to other fragmentation percentages). In the example (fig. 7.6), we can see the frequency distribution of the word *science* in the Wikipedia entry “Social Science” in 10% text blocks.

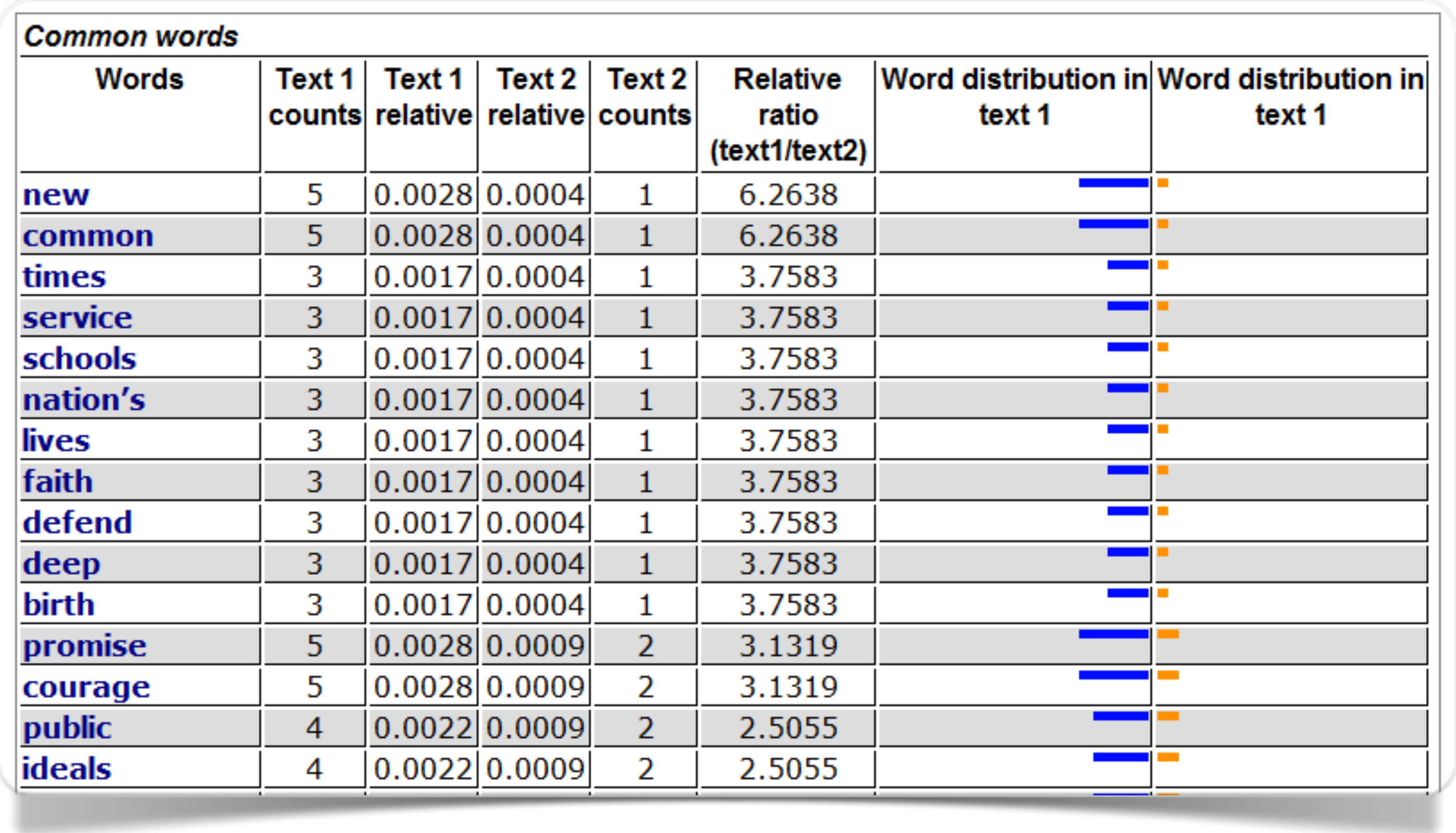


percent	10	20	30	40	50	60	70	80	90	100	percent
Count	31	22	4	4	10	10	10	22	28	10	Count
Relative(%)	4.0	2.4	0.4	0.4	1.1	1.1	1.1	2.5	2.9	1.3	Relative(%)

Fig. 7.6 Distribution of the absolute and relative frequency of the word *science* in the Wikipedia entry “Social Science” (elab. Taporware)

Comparator

Comparator compares two texts belonging to only one corpus in order to highlight the words in common and those exclusive of each text. In the example the corpus consists in the two swearing in speeches of the US president George W. Bush (taken from Bartleby.com): The first, of the 20th of January 2001; the second, of the 20th of January 2005. The two speeches take place in two very different contexts: The first has a more traditional tone

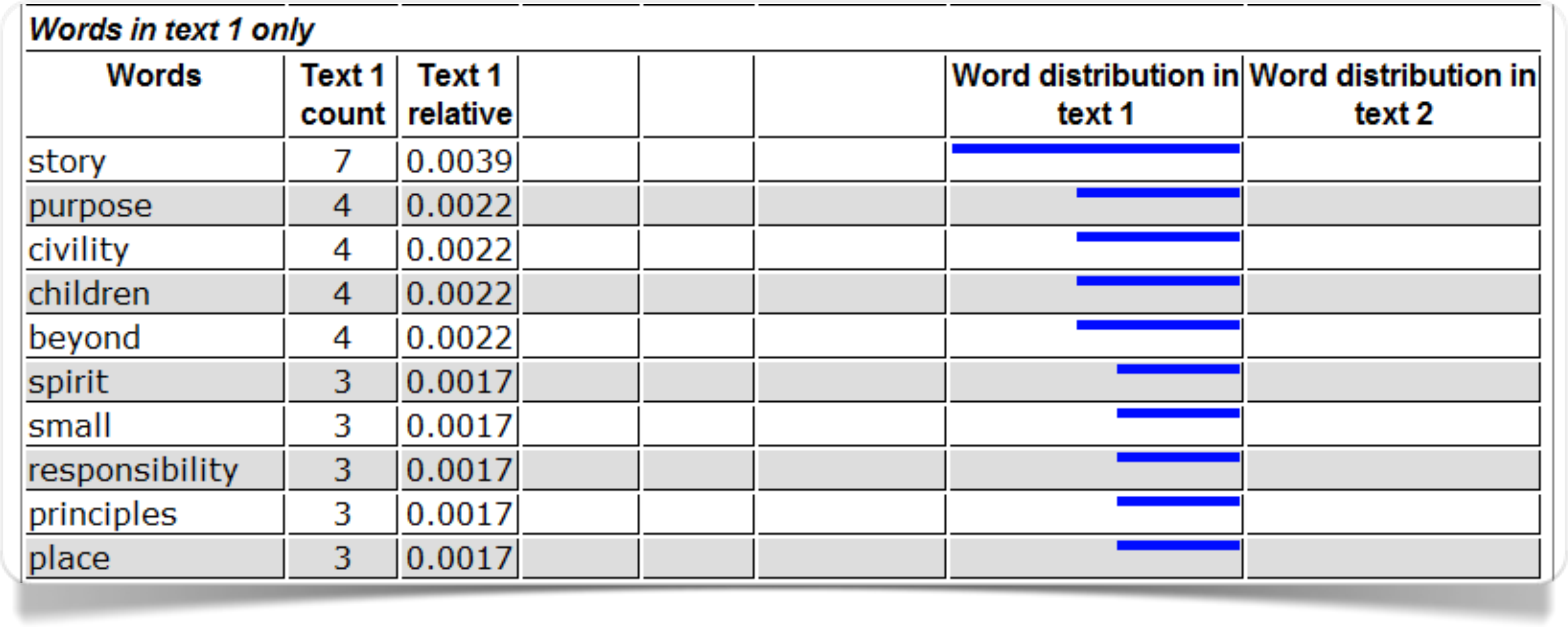


Common words

Words	Text 1 counts	Text 1 relative	Text 2 relative	Text 2 counts	Relative ratio (text1/text2)	Word distribution in text 1	Word distribution in text 1
new	5	0.0028	0.0004	1	6.2638	▬	▪
common	5	0.0028	0.0004	1	6.2638	▬	▪
times	3	0.0017	0.0004	1	3.7583	▬	▪
service	3	0.0017	0.0004	1	3.7583	▬	▪
schools	3	0.0017	0.0004	1	3.7583	▬	▪
nation’s	3	0.0017	0.0004	1	3.7583	▬	▪
lives	3	0.0017	0.0004	1	3.7583	▬	▪
faith	3	0.0017	0.0004	1	3.7583	▬	▪
defend	3	0.0017	0.0004	1	3.7583	▬	▪
deep	3	0.0017	0.0004	1	3.7583	▬	▪
birth	3	0.0017	0.0004	1	3.7583	▬	▪
promise	5	0.0028	0.0009	2	3.1319	▬	▬
courage	5	0.0028	0.0009	2	3.1319	▬	▬
public	4	0.0022	0.0009	2	2.5055	▬	▬
ideals	4	0.0022	0.0009	2	2.5055	▬	▬

Fig. 7.7 Comparison between the frequency of words in the swearing in speeches of G. W. Bush of the 20th of January 2001 (text 1) and of the 20th of January 2005 (text 2) (elab. Taporware)

and it calls on the elements of continuity of the American nation; The second is made in an international scenario strongly signed by the wars in the Middle-East, the wounds of the 11th of September 2001 and the fight against terrorism. In fig. 7.7 we can observe the first 15 common words ordered according to the decreasing values of the ratio between the occurrences of text 1 (2001) and the occurrences of text 2 (2005). As usual, the stop words are excluded. The column of relative values points out the words which – though common in both cases – are more representative of text1 than of text 2: *new*, *common*, *times*, *service*, *schools*, etc. In fig. 7.8 and fig. 7.9 we can see

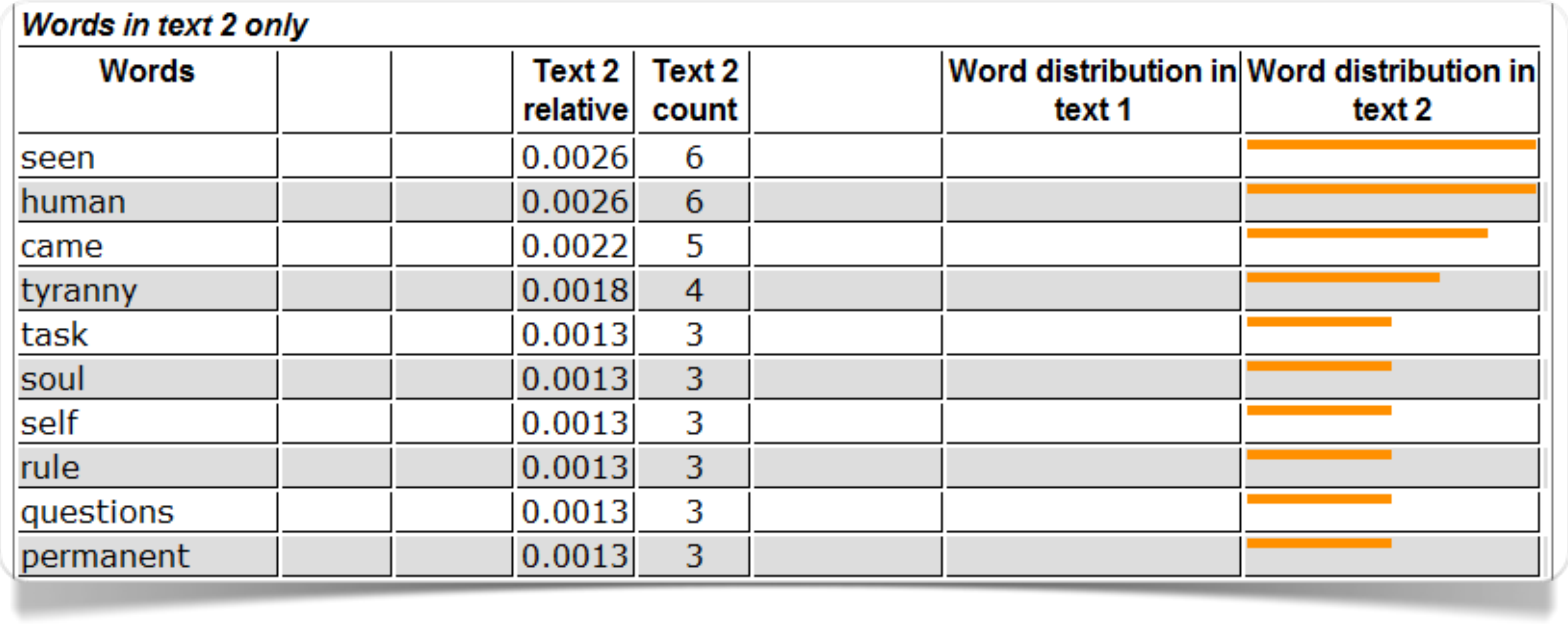


Words in text 1 only

Words	Text 1 count	Text 1 relative				Word distribution in text 1	Word distribution in text 2
story	7	0.0039					
purpose	4	0.0022					
civility	4	0.0022					
children	4	0.0022					
beyond	4	0.0022					
spirit	3	0.0017					
small	3	0.0017					
responsibility	3	0.0017					
principles	3	0.0017					
place	3	0.0017					

Fig. 7.8 Frequency of the words present only the swearing in speech of G. W. Bush,
the 20th of January 2001 (elab. Taporware)

the first 10 words present exclusively in one speech or in the other. The presence of references to the Twin Towers attack and to the war are obviously present in the second speech: *the history we have seen together*; *we have seen our vulnerability*; *there can be no human rights without human liberty*; *the ultimate goal of ending tyranny in our world*.



Words in text 2 only

Words			Text 2 relative	Text 2 count		Word distribution in text 1	Word distribution in text 2
seen			0.0026	6			▬
human			0.0026	6			▬
came			0.0022	5			▬
tyranny			0.0018	4			▬
task			0.0013	3			▬
soul			0.0013	3			▬
self			0.0013	3			▬
rule			0.0013	3			▬
questions			0.0013	3			▬
permanent			0.0013	3			▬

Fig. 7.9 Frequency of the words present only in the swearing in speech of G. W. Bush, the 20th of January 2005 (elab. Taporware)

Chapter 8

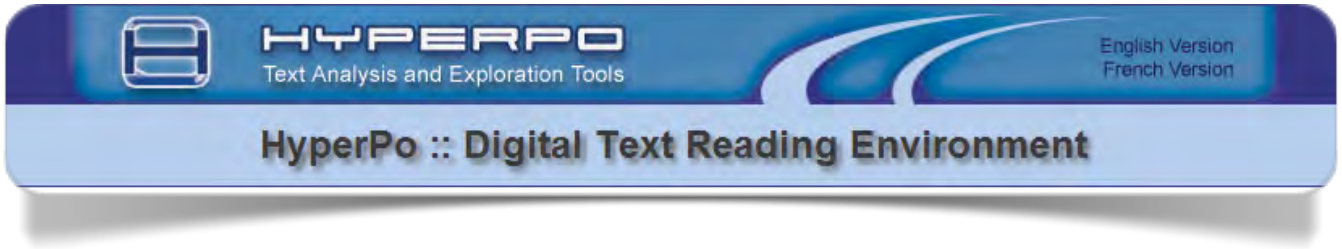
HyperPo Digital Text Reading Environment

HyperPo (now no longer available, but interesting to examine) was a complete text reading environment developed by Stéfan Sinclair. HyperPo performed linguistic analyses on texts in English, French, Spanish, German and Italian.

8

HyperPo

In *HyperPo* the text acquisition (fig. 8.2) could take place in three ways: via URL, “copy and paste” or uploading a file in plain text (Sinclair, 2003).



HYPERPO

Text Analysis and Exploration Tools

English Version

French Version

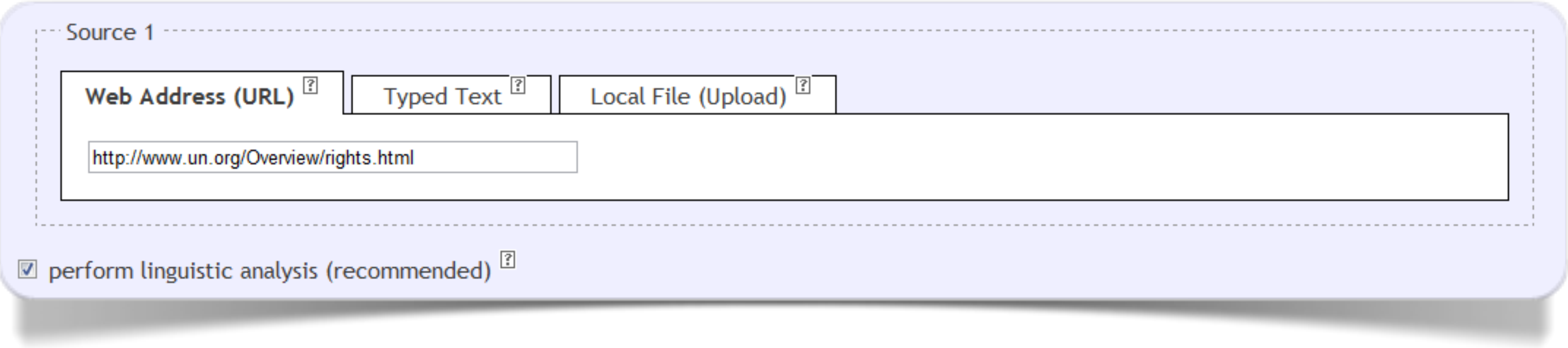
HyperPo :: Digital Text Reading Environment

Fig. 8.1 HyperPo Home Page (up to 2012)

Among the available menus, there were:

- *Words Tools*
- *Document Tools*
- *Visualization Tools* (currently usable only Mandala Browser).

In the following examples we used the text: *The Universal Declaration of Human Rights*.



Source 1

Web Address (URL) ? | Typed Text ? | Local File (Upload) ?

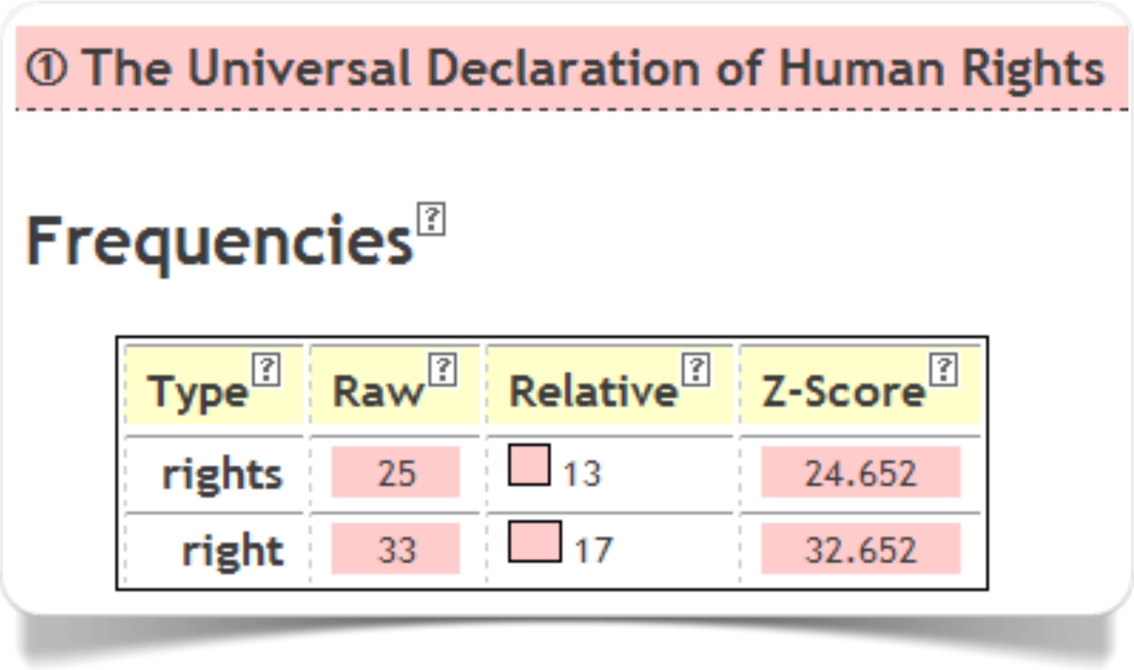
http://www.un.org/Overview/rights.html

☑ perform linguistic analysis (recommended) ?

Fig. 8.2 Hyperpo – Text acquisition modalities

Word Tools: Frequencies

The most interesting tool of this menu was *Study a word* which sums up all the information about a certain word. For example, in the output of the word *right* (with the option *match morphological variant*) we can see the inflected forms *right* and *rights* with absolute frequencies, relative frequencies per thousand and the *z-score*, that is a standard frequency measurement (fig. 8.3). In the same page there were other analysis and visualization modalities, such as *Word Series* (n-grams), *Word Collocates* (co-occurrences), *Distribution of Word* (a distribution chart of the word in the text divided in predefiined parts) and *Word in Context* (concordances of 5 words which come before and after the pivot word).



① The Universal Declaration of Human Rights

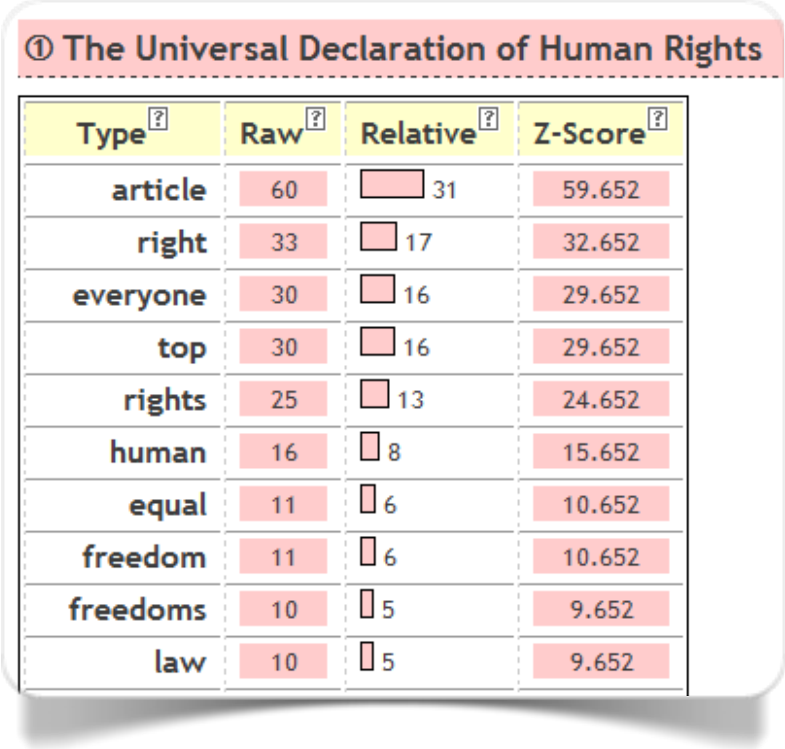
Frequencies

Type	Raw	Relative	Z-Score
rights	25	□ 13	24.652
right	33	□ 17	32.652

Fig. 8.3 Frequencies of the form right* (elab. HyperPo)

Document Tools

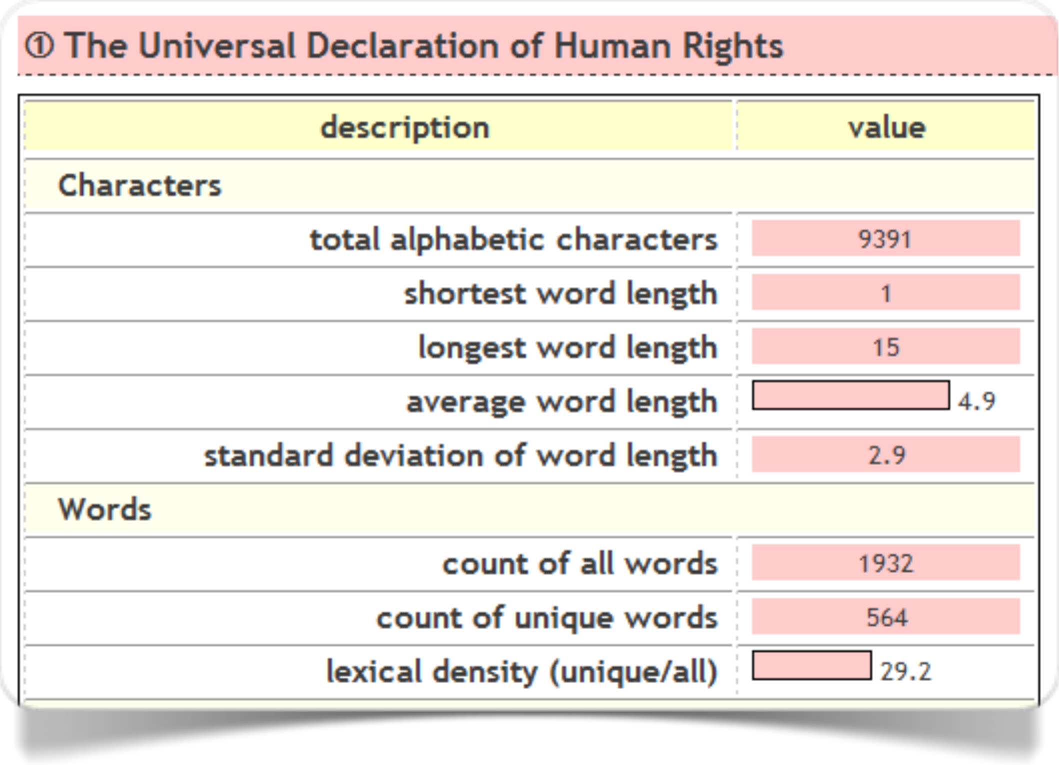
The tool *Frequencies* in the sub-menu *Document Context* listed the most frequent words in order of occurrences, of relative frequencies, of Z-scores or in alphabetical order. Among the options, it was possible to obtain the list of headwords, select the words which have a complete sense (*content words*) or the words which indicate a function (*function words*), such as articles and prepositions. In the example (fig. 8.4) are listed the 10 most frequent content words.



① The Universal Declaration of Human Rights

Type	Raw	Relative	Z-Score
article	60	31	59.652
right	33	17	32.652
everyone	30	16	29.652
top	30	16	29.652
rights	25	13	24.652
human	16	8	15.652
equal	11	6	10.652
freedom	11	6	10.652
freedoms	10	5	9.652
law	10	5	9.652

Fig. 8.4 List of the 10 most frequent content words (elab. HyperPo



① The Universal Declaration of Human Rights

description	value
Characters	
total alphabetic characters	9391
shortest word length	1
longest word length	15
average word length	4.9
standard deviation of word length	2.9
Words	
count of all words	1932
count of unique words	564
lexical density (unique/all)	29.2

Fig. 8.5 Basic lexicometric measurements (elab. HyperPo)

With the option *Statistical Summary* it was possible to obtain the lexicometric measurements of the text (fig. 8.5), among which the sum of the occurrences (*tokens*), the sum of distinctive words (*types*) and the *density* or lexicon extension ratio (*type/token ratio*).

Visualization Tools: Mandala Tokens

Mandala is an interface which enables us to surf among the words of a text to identify their relationship according to specific morphological or statistical properties. Stefan Sinclair and his collaborators chose the “mandala” metaphor (the Buddist cycle which describes the formation of the universe) because they thought it was suitable to express the sense of a trend that addresses the barycenter of the words according to the attraction poles which are added from time to time (Gainor et alii, 2009; Brown et alii, 2010). The objective is an experimental analysis, the identification of structures and relationships, the formulation of hypotheses.

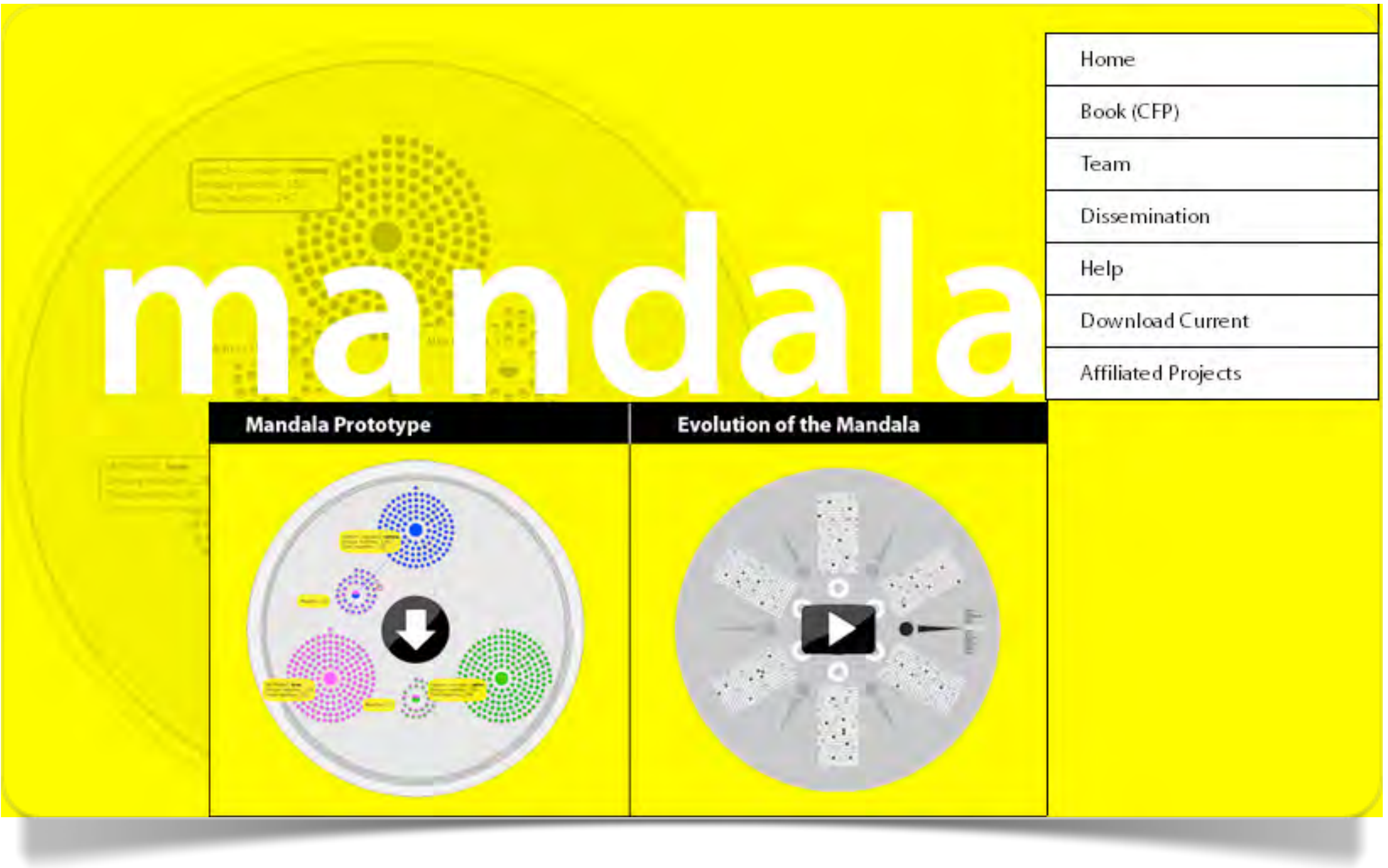


Fig. 8.6 Mandala Browser Home page

Mandala is no longer available in HyperPo, but the project goes on OS X operating system in Mandala Browser (fig. 8.6).

The model is applicable to various groupings of objects: in this case we are talking about words (fig. 8.7). The black points making up the circle are the occurrences (*word tokens*). The gray circle in the center is the trend of all the properties defining the words (frequencies, word types, stems, word length, etc.). The user interacts with the model defining the axes (or magnets) according to these properties, their measurements, modalities or typologies. A magnet works as an attraction for the words which correspond to the properties it is defined by. The result is a sort of dynamic Venn diagram that brings the words to gather into subsets defined by the properties and to converge according to the criteria chosen by the user in a display to the right of the interface.

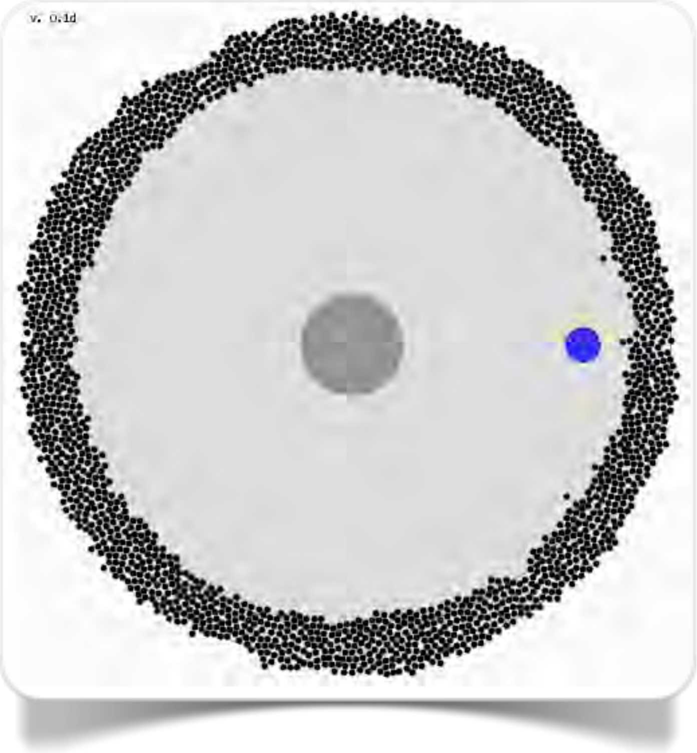


Fig. 8.7 Mandala Tokens of the *Declaration of Human Rights* (elab. HyperPo)

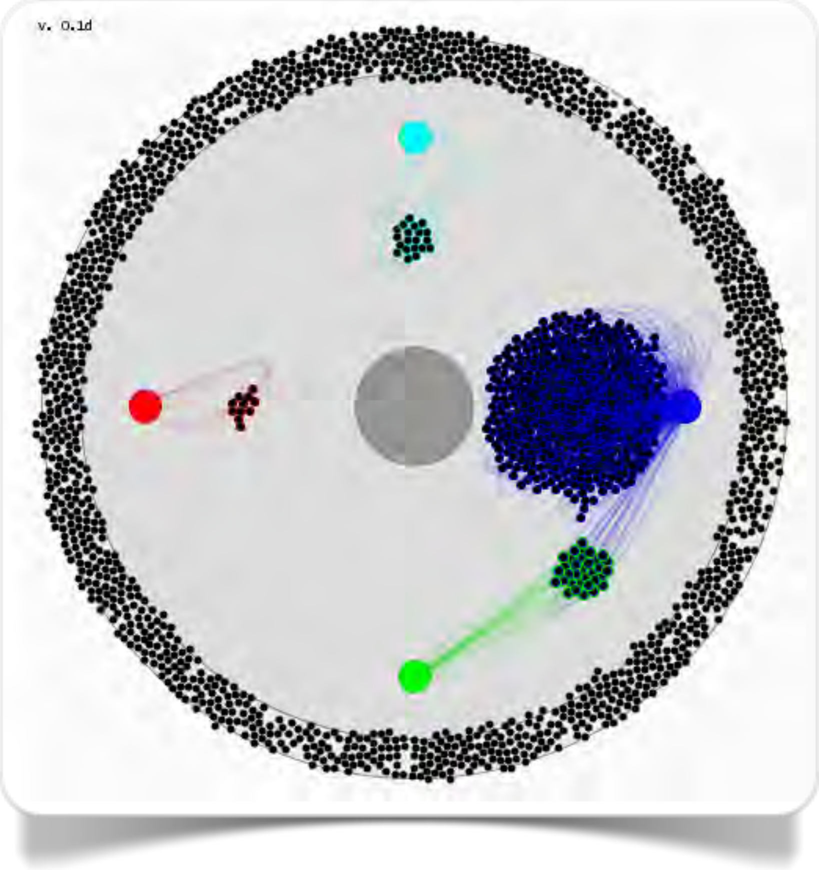
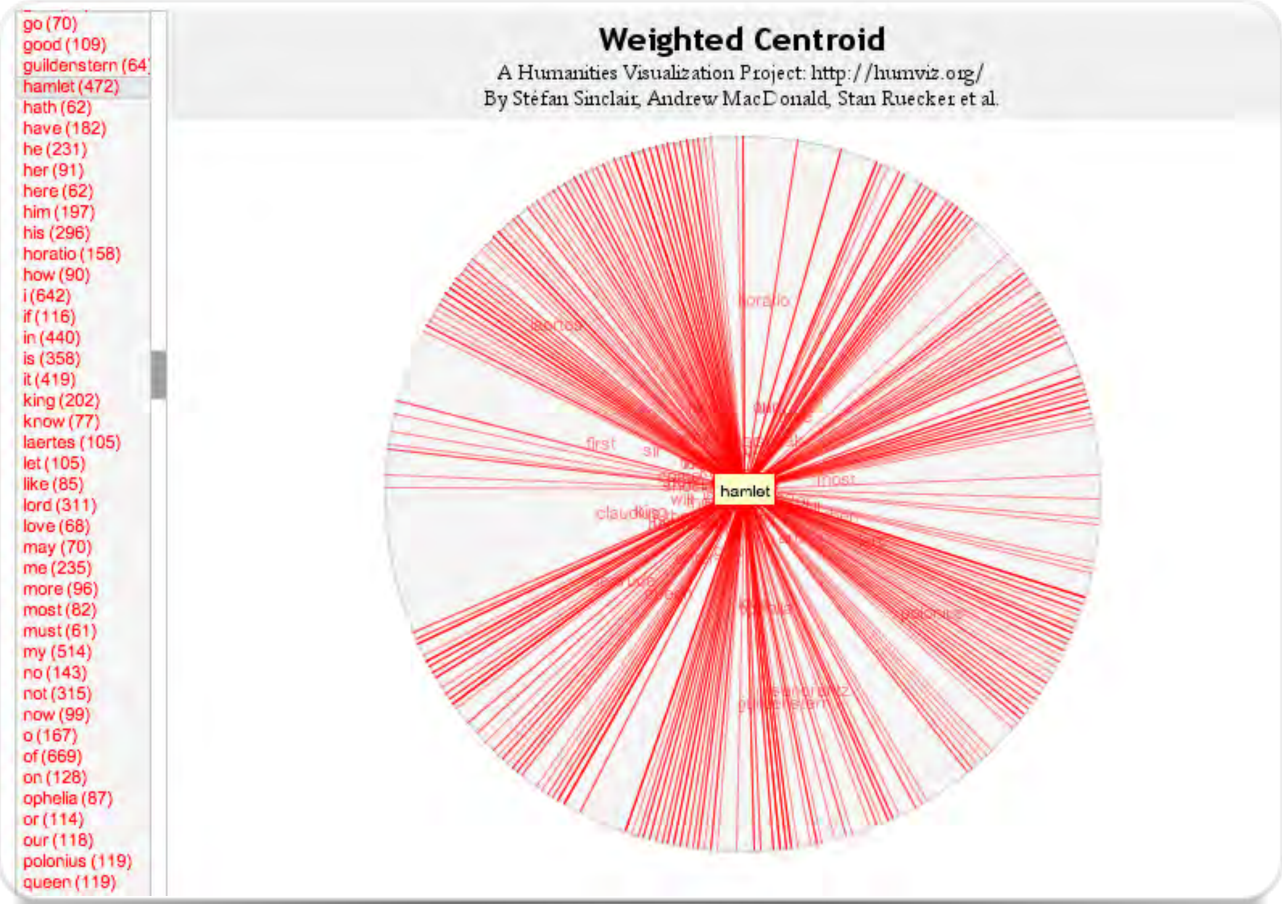


Fig. 8.8 Mandala Tokens of the *Declaration of Human Rights* : Words and occurrences of *human, right e rights* (elab. HyperPo)

In the example shown in fig. 8.8 the blue magnet is defined by the words with over 30 occurrences; the green magnet is defined by the occurrences (n=33) of the word *right*; the light blue magnet is defined by the occurrences (n=25) of the word *rights*; the red magnet, by the occurrences (n=16) of the word *human*. We can observe that only the group of green points, which represent the occurrences of the word *right*, tend to move close to and to show connections with the group of blue tokens, which represent the words with over 30 occurrences.

Frequencies Centroid



go (70)
good (109)
guildenstern (64
hamlet (472)
hath (62)
have (182)
he (231)
her (91)
here (62)
him (197)
his (296)
horatio (158)
how (90)
i (642)
if (116)
in (440)
is (358)
it (419)
king (202)
know (77)
laertes (105)
let (105)
like (85)
lord (311)
love (68)
may (70)
me (235)
more (96)
most (82)
must (61)
my (514)
no (143)
not (315)
now (99)
o (167)
of (669)
on (128)
ophelia (87)
or (114)
our (118)
polonius (119)
queen (119)

Weighted Centroid

A Humanities Visualization Project: http://humviz.org/
By Stéfan Sinclair, Andrew MacDonald, Stan Ruecker et al.

Fig. 8.9 Frequencies Centroid of *Hamlet* (Elab. HyperPo

Frequencies Centroid was an experimental text tool which worked in a similar way to *Mandala*. The principle is quite simple: The entire text is positioned clockwise along the circumference of a circle with start and end at 12:00 o'clock. Inside the circle the words are positioned in a way that is determined by the average weight of the word itself related to the distribution of its occurrences in the chapters or in the various parts the text is divided into. The more a word is distant from the circumference, the more it is used in different parts of the text. A word used frequently and in a balanced way all over the text, will be collocated at the center of the circle. In the example shown in 8.9, in which is represented *The Tragedy of Hamlet, Prince of Denmark* by William Shakespeare, *Hamlet*, with 472 occurrences, is obviously present in all the play, but with greater frequency at the hours four, six, eight and ten in which he is more active. At eleven there is the mortal duel with *Laertes*.

The character *Polonius* (119 occurrences) is present between the II and the III Act till his death, caused by Hamlet, at the end of the III Act (fig. 8.10). *Laertes* (105 occurrences) is sporadically present till the fifth scene of the IV Act, in which he talks with King Claudius and with Ophelia, and then expecially in the V Act in which the play ends (fig. 8.11).

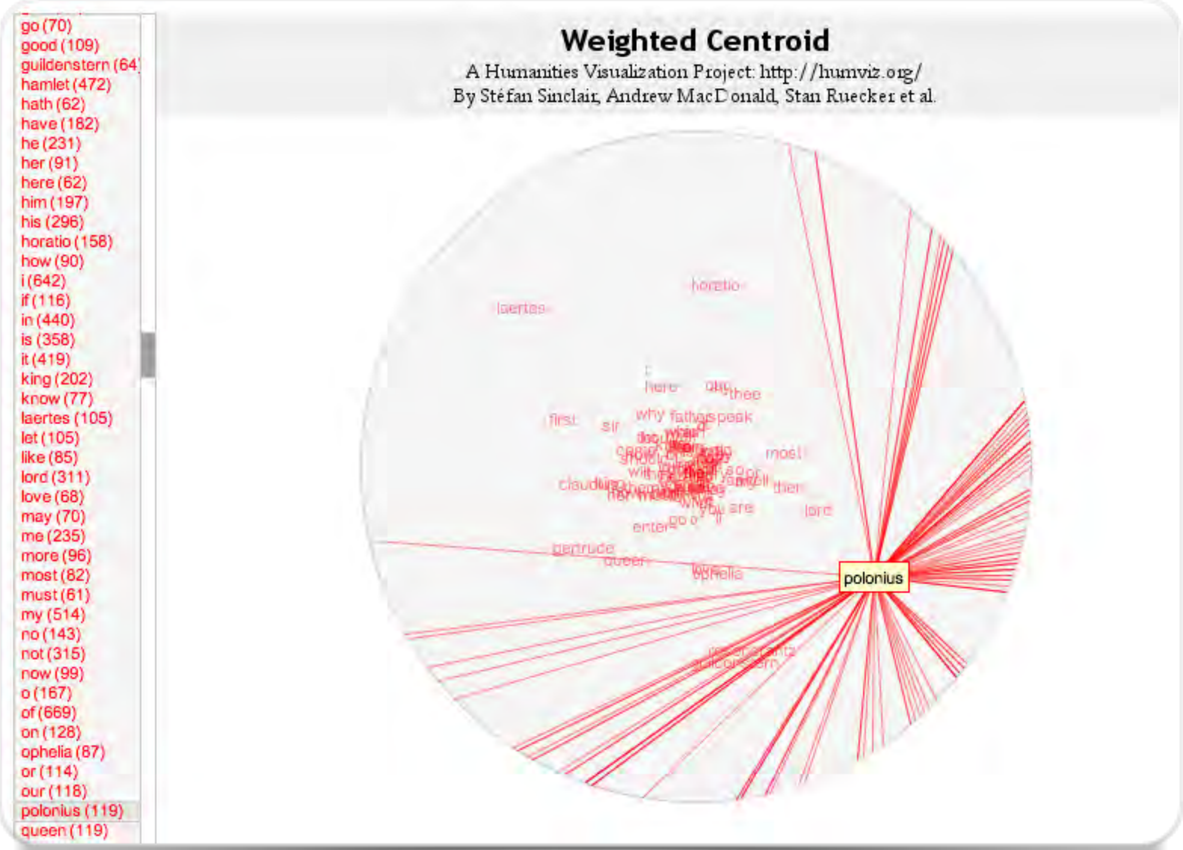


Fig. 8.10 Frequencies Centroid of *Polonius* (Elab. HyperPo)

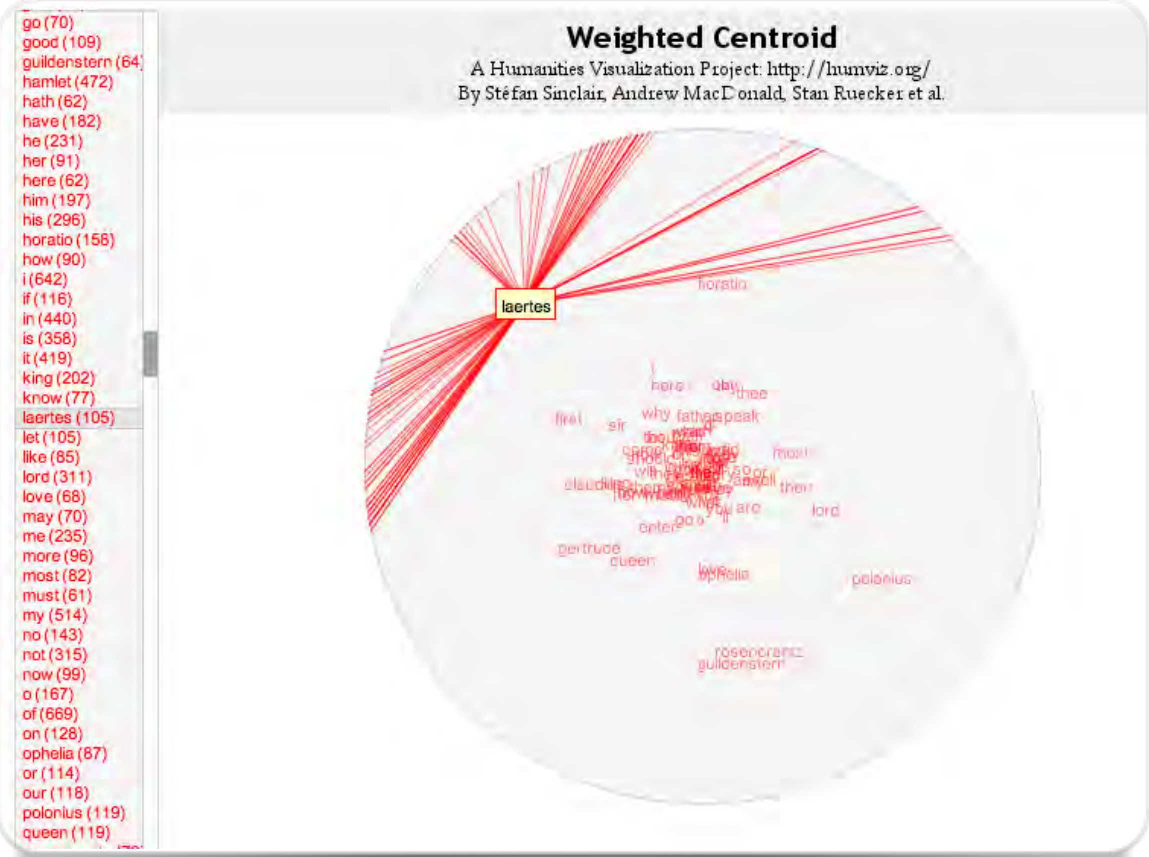


Fig. 8.11 Frequencies Centroid of *Laertes* (Elab. HyperPo)

Chapter 9

Text Arc

A very similar experiment (and which inspired Stéfan Sinclair) is TextArc, developed by W. Bradford Paley, a very original computer artist, who from the early '70s has worked on the representation of hidden information; the information that is not immediately perceived.

9

Text Arc

Paley considers himself an *interaction designer* and his formation is the combination of various disciplines: cognitive psychology, literature, computer science, experimental cinema (Paley, 2002). The Text Arc application provides an example on a text of Wonderland, by Lewis Carroll, and it is interfaced with the Gutemberg project for what concerns the text digitalization. There are therefore thousands of texts which can be analysed. As in *Frequencies Centroid*, the words are positioned along the circumference (in this case of an ellipsis). The most frequent words are highlighted in a more intense colour. Placing the cursor on top of a word, it is possible to observe the connection between the word and the point in which it appears in the text distribution.

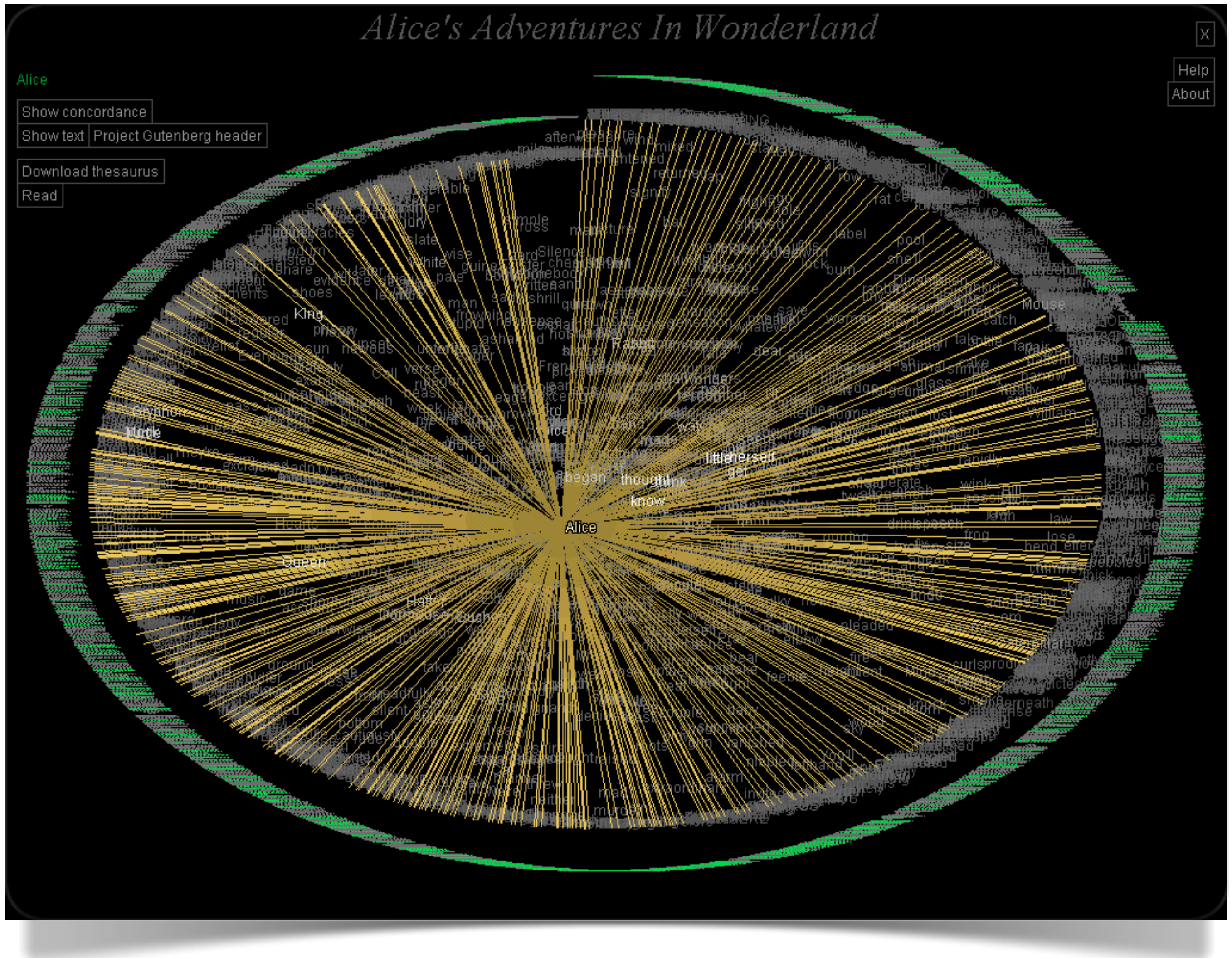
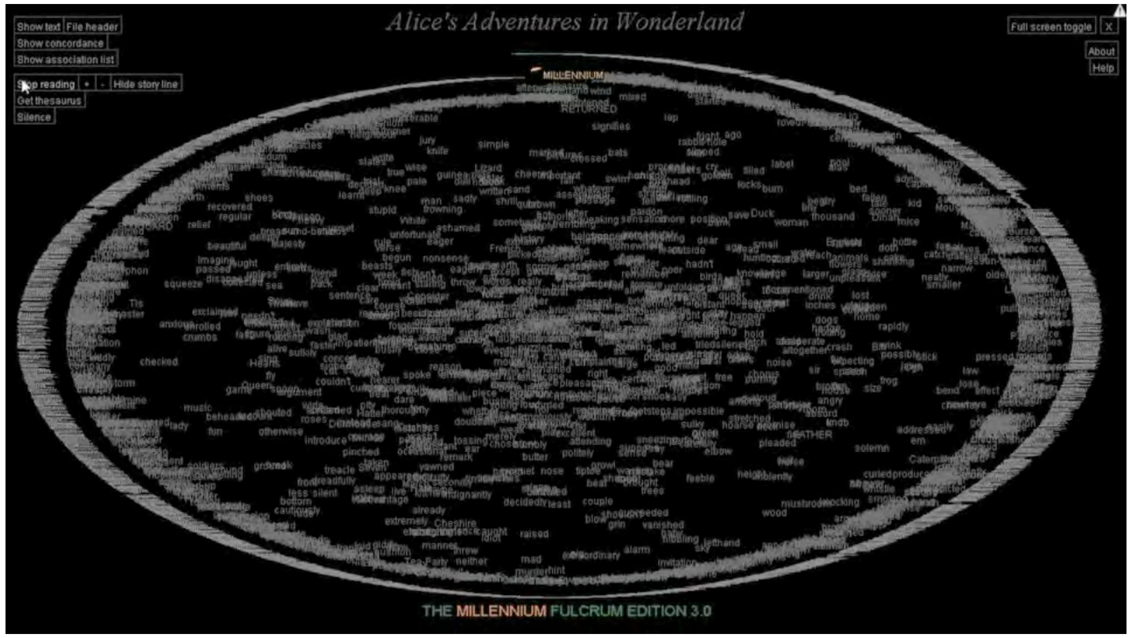


Fig. 9.1 Alice in Wonderland of Lewis Carroll (elab. TextArc)

In the case of *Alice*, the word *mouse* appears 41 times in the second and third chapter, once in the first and once in the last. It is therefore collocated along the border of the ellipsis at 2:00; *King* appears 64 times only from the seventh chapter onwards and the word is collocated at 10:00 in a specular position; *Alice*, which is obviously present everywhere, is collocated at the center of the ellipsis. Starting the reading of the text, by pressing the button *Read* on the left of the monitor, the words flow along the story; selecting *Show story line* the connections between one word and another are visualized in sequence with curve lines which are accompanied by music (clicking on *Sound*). The music stresses the rhythm of the story according to the importance of a certain word in the set. From the same menu (*Show tex*t) it is possible to visualize the text and observe the scrolling, or to visualize the occurrences (*Show concordance*) of the words with reference to their position in the text.



Movie 9.1 Alice in Wonderland of Lewis Carroll
(elab. TextArc)

Chapter 10

Voyant
See through your text

Voyant is an Open Source collaborative project in a text analysis environment, developed by Stéfan Sinclair and Geoffrey Rockwell with different interactive tools.

10

Voyant Workbench Screen

Voyant is the access to where you can really manage the texts (fig. 10.1). The acquisition can take place in different ways: by inserting the URL; copying and pasting; uploading files from your own computer. It is also possible to analyse a corpus made up of more texts (or fragments). Voyant accepts documents in plain text, Word and Pdf format.

In the example we analyse the entire corpus of Shakespeare's works already present in the option *Open a pre-defined corpus*. Voyant's work platform shows the corpus in various ways (fig. 10.2).

The exploration takes place using six windows which can be hidden or called back according to necessity. During the first phase the three main windows are activated. The window high on the left visualizes a traditional word cloud (*Cirrus*). The text is positioned on the right (*Corpus Reader*) and down on the left there is a summary of the most important measurements of the corpus and of the most significant words of each play.

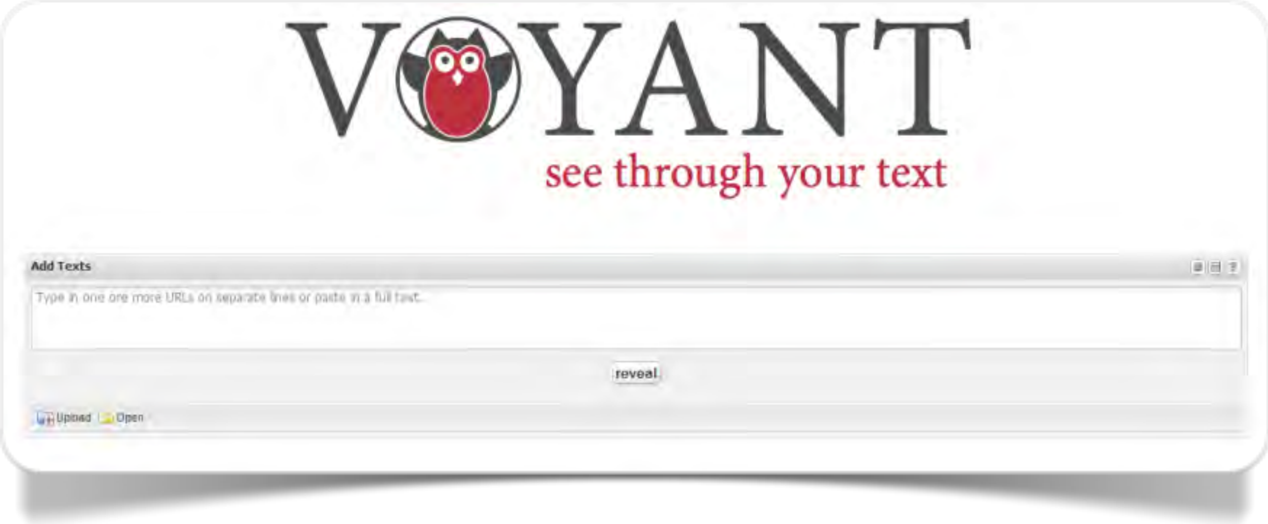
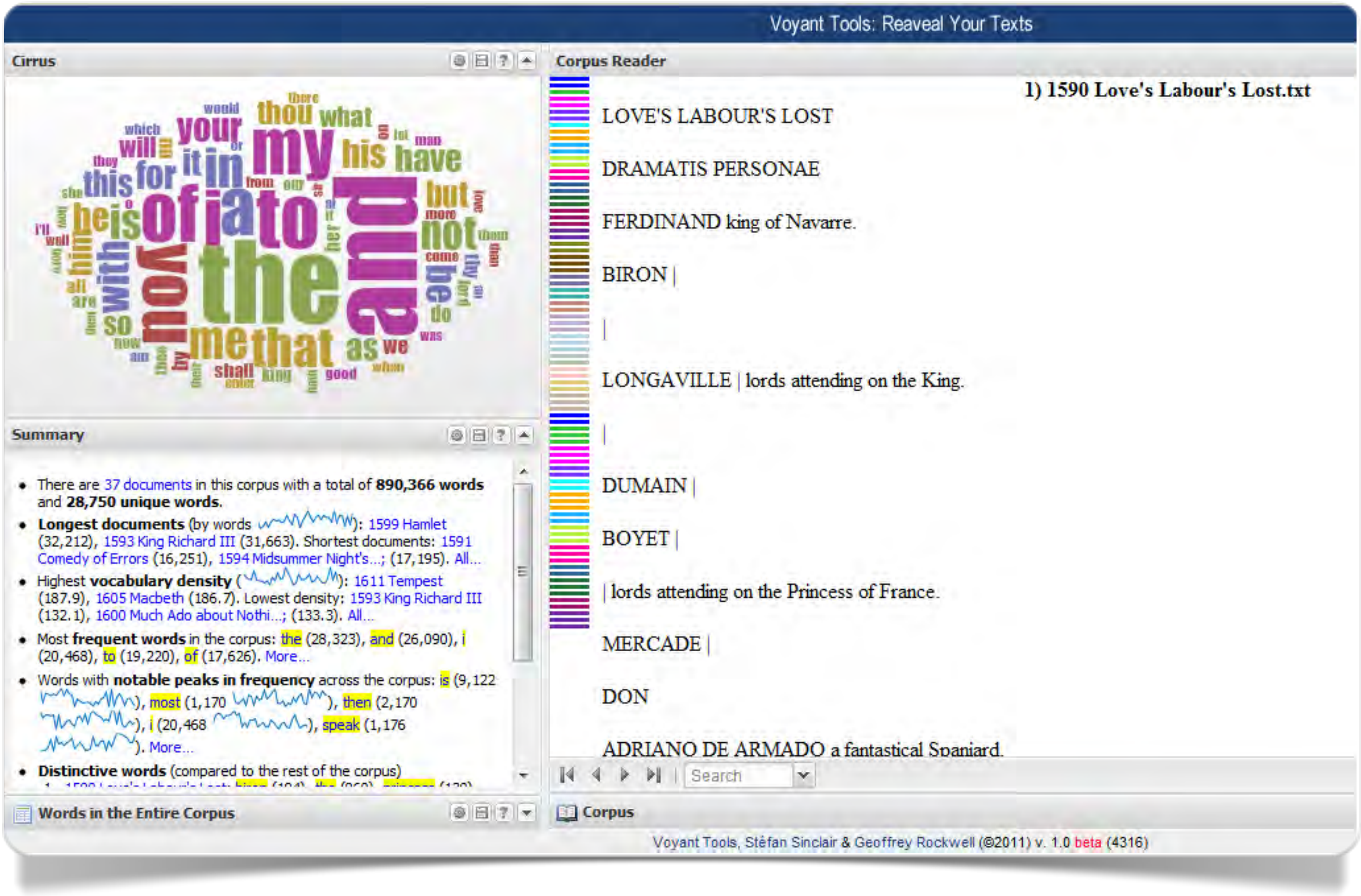

VOYANT
see through your text
Add Texts
Type in one ore more URLs on separate lines or paste in a full text.
reveal
Upload
Open

Fig. 10.1 Voyant: Work platform



Voyant Tools: Reaveal Your Texts
Cirrus
and
the
you
i
to
of
my
is
be
not
me
that
with
in
it
this
for
your
thou
his
have
but
Summary
Corpus Reader
Search
Words in the Entire Corpus
Corpus
Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (©2011) v. 1.0 beta (4316)

- There are 37 documents in this corpus with a total of **890,366 words** and **28,750 unique words.**
- **Longest documents** (by words): 1599 Hamlet (32,212), 1593 King Richard III (31,663). Shortest documents: 1591 Comedy of Errors (16,251), 1594 Midsummer Night's...; (17,195). All...
- Highest **vocabulary density** ( ): 1611 Tempest (187.9), 1605 Macbeth (186.7). Lowest density: 1593 King Richard III (132.1), 1600 Much Ado about Nothi...; (133.3). All...
- Most **frequent words** in the corpus: the (28,323), and (26,090), i (20,468), to (19,220), of (17,626). More...
- Words with **notable peaks in frequency** across the corpus: is (9,122), most (1,170), then (2,170), i (20,468), speak (1,176). More...
- **Distinctive words** (compared to the rest of the corpus)

1) 1590 Love's Labour's Lost.txt

LOVE'S LABOUR'S LOST

DRAMATIS PERSONAE

FERDINAND king of Navarre.

BIRON |

|

LONGAVILLE | lords attending on the King.

|

DUMAIN |

BOYET |

| lords attending on the Princess of France.

MERCADE |

DON

ADRIANO DE ARMADO a fantastical Spaniard.

Fig. 10.2 Voyant: Workbench screen

Cirrus

During the acquisition phase Cirrus (Click *Open*; select *Shakesperare’s play* and *OK*) visualizes the word cloud of the most frequent words in all the corpus. This default option is not very useful because the most frequent words are always empty or banal words. Clicking, however, on the options positioned on the top right of Cirrus's window, the menu allows to select the list of stop words in different languages (English, French, German, Spanish, Hungarian, Italian, Norwegian).

Selecting the list of words in English, the Cirrus cloud changes considerably (fig. 10.3), even though some empty words remain active because they are not present in the list. For example, archaic pronouns such as: *thou*, *thee*, *thy*.

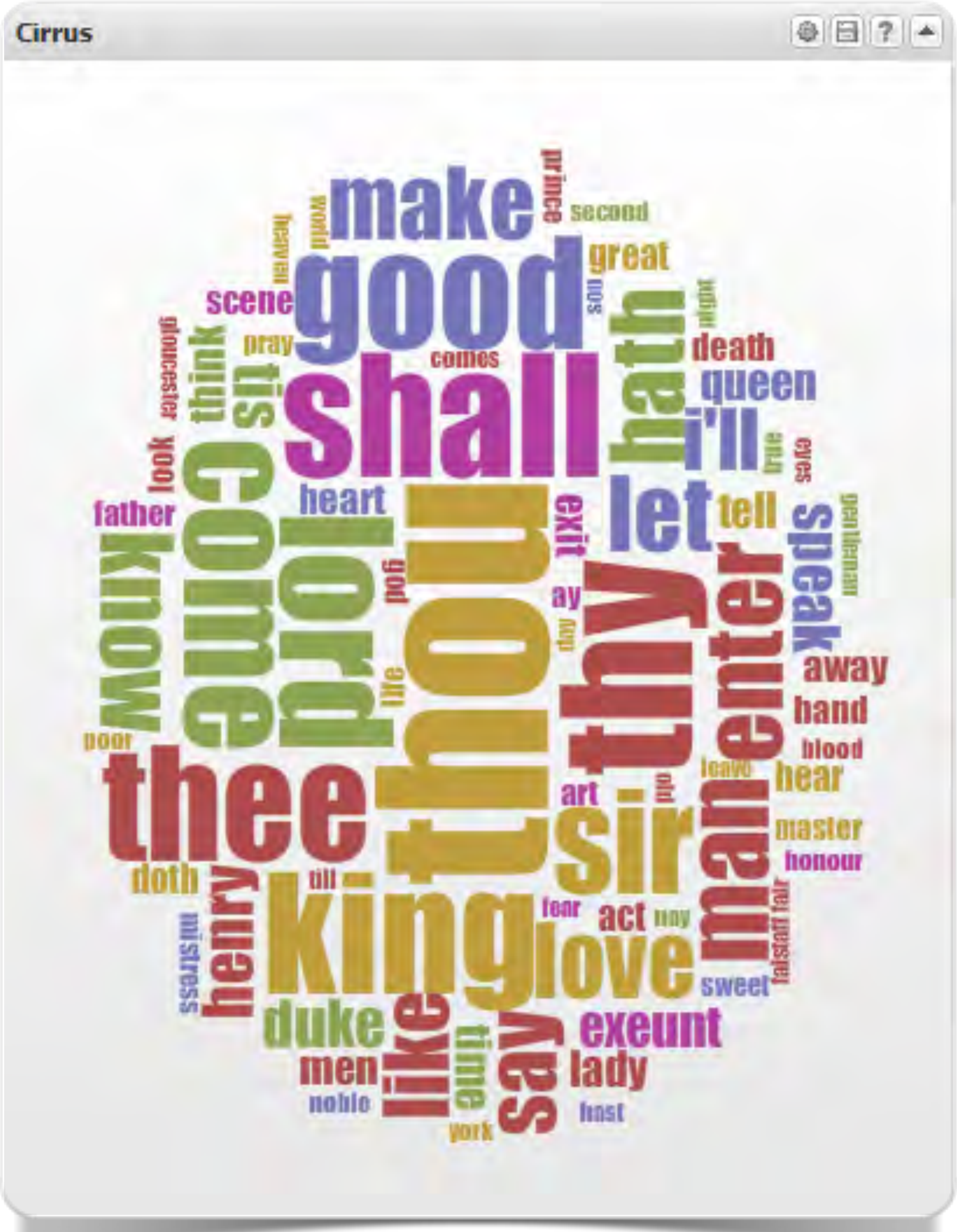


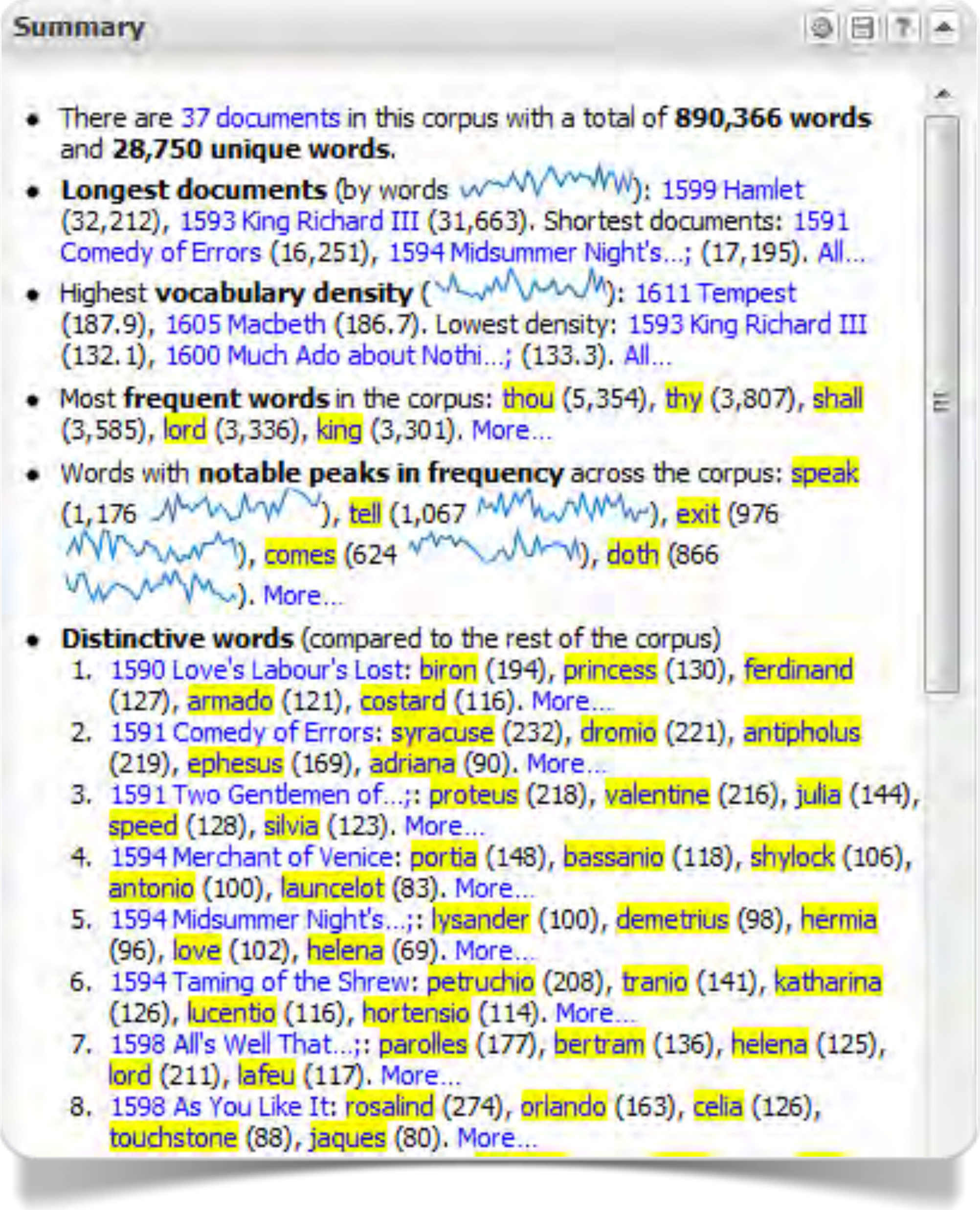
Fig. 10.3 Word cloud of the most frequent words in the corpus Shakespeare (Elab. Voyant)

Summary

In the window Summary (Click *Open*; select *Shakesperare's play* and *OK;* fig. 10.4) we find the traditional lexicometric measurements of the corpus: 37 documents (all Shakespeare's theatre plays) for a total of 890 366 occurrences (tokens) and 28 750 distinctive words (types). The document with the greatest number of occurrences is *Hamlet* (32 212).

The *lexical density* or vocabulary extension (*type/token ratio*) is equal to 32 per thousand and it is always higher in each single document (186 per thousand in Macbeth) than in the corpus. From the statistical point of view, the vocabulary is considered "representative" of the language (in this case texts written by Shakespeare) if it is less than 200 per thousand.

The *distinctive words* are the words more present in a document than in the overall corpus. In this case the words highlighted are those identifying the characters of each play.



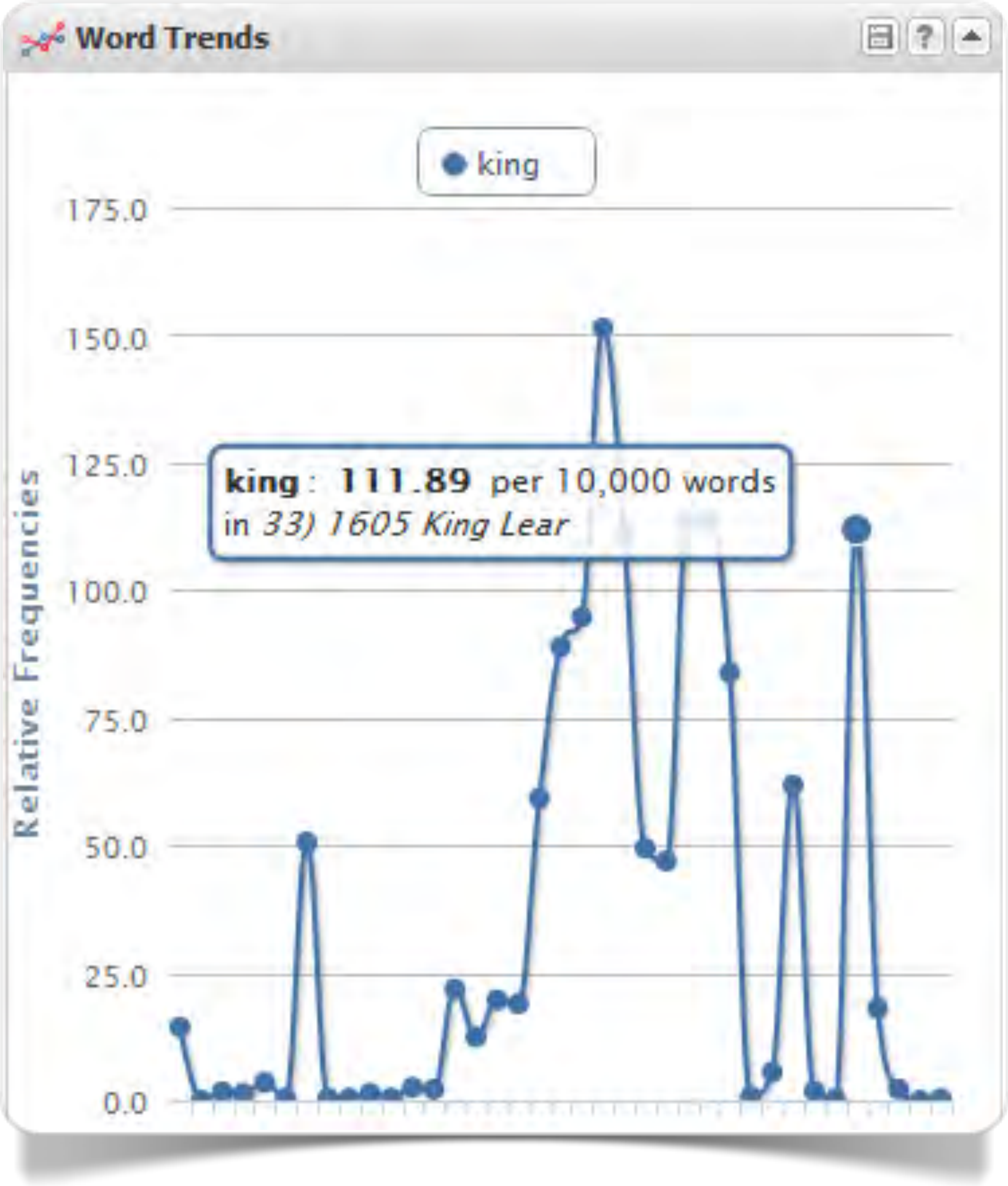
Summary

- There are 37 documents in this corpus with a total of **890,366 words** and **28,750 unique words.**
- **Longest documents** (by words): 1599 Hamlet (32,212), 1593 King Richard III (31,663). Shortest documents: 1591 Comedy of Errors (16,251), 1594 Midsummer Night's...; (17,195). All...
- Highest **vocabulary density** ( ): 1611 Tempest (187.9), 1605 Macbeth (186.7). Lowest density: 1593 King Richard III (132.1), 1600 Much Ado about Nothi...; (133.3). All...
- Most **frequent words** in the corpus: thou (5,354), thy (3,807), shall (3,585), lord (3,336), king (3,301). More...
- Words with **notable peaks in frequency** across the corpus: speak (1,176), tell (1,067), exit (976), comes (624), doth (866). More...
- **Distinctive words** (compared to the rest of the corpus)
  1. 1590 Love's Labour's Lost: biron (194), princess (130), ferdinand (127), armado (121), costard (116). More...
  2. 1591 Comedy of Errors: syracuse (232), dromio (221), antipholus (219), ephesus (169), adriana (90). More...
  3. 1591 Two Gentlemen of...;: proteus (218), valentine (216), julia (144), speed (128), silvia (123). More...
  4. 1594 Merchant of Venice: portia (148), bassanio (118), shylock (106), antonio (100), launcelot (83). More...
  5. 1594 Midsummer Night's...;: lysander (100), demetrius (98), hermia (96), love (102), helena (69). More...
  6. 1594 Taming of the Shrew: petruchio (208), tranio (141), katharina (126), lucentio (116), hortensio (114). More...
  7. 1598 All's Well That...;: parolles (177), bertram (136), helena (125), lord (211), lafeu (117). More...
  8. 1598 As You Like It: rosalind (274), orlando (163), celia (126), touchstone (88), jaques (80). More...

Fig. 10.4 Summary : Lexicometric measurements in the corpus Shakespeare (Elab. Voyant)

Word Trends

Clicking on the words of the Cirrus cloud or on the words highlighted in *Summary* the window *Word Trends* is activated on the top right of the screen. This visualizes the trend of the relative frequencies (per 10 000 words) of the selected word in the corpus of 37 documents. Fig. 10.5 shows the trend of the word *king*. On the MIT website, from which Voyant takes the corpus, Shakespeare's works are organized in sequence in the following order: *Comedy* (17); *History* (10); *Tragedy* (10); *Poetry* (5, but not considered in the corpus taken as example). For each category the higher relative frequencies can be found in document n. 7 (*All's Well That Ends Well*, 1590), in document n. 21 (*King Henry VI*, *part III*, 1595) e in document n. 33 (*King Lear*, 1605).

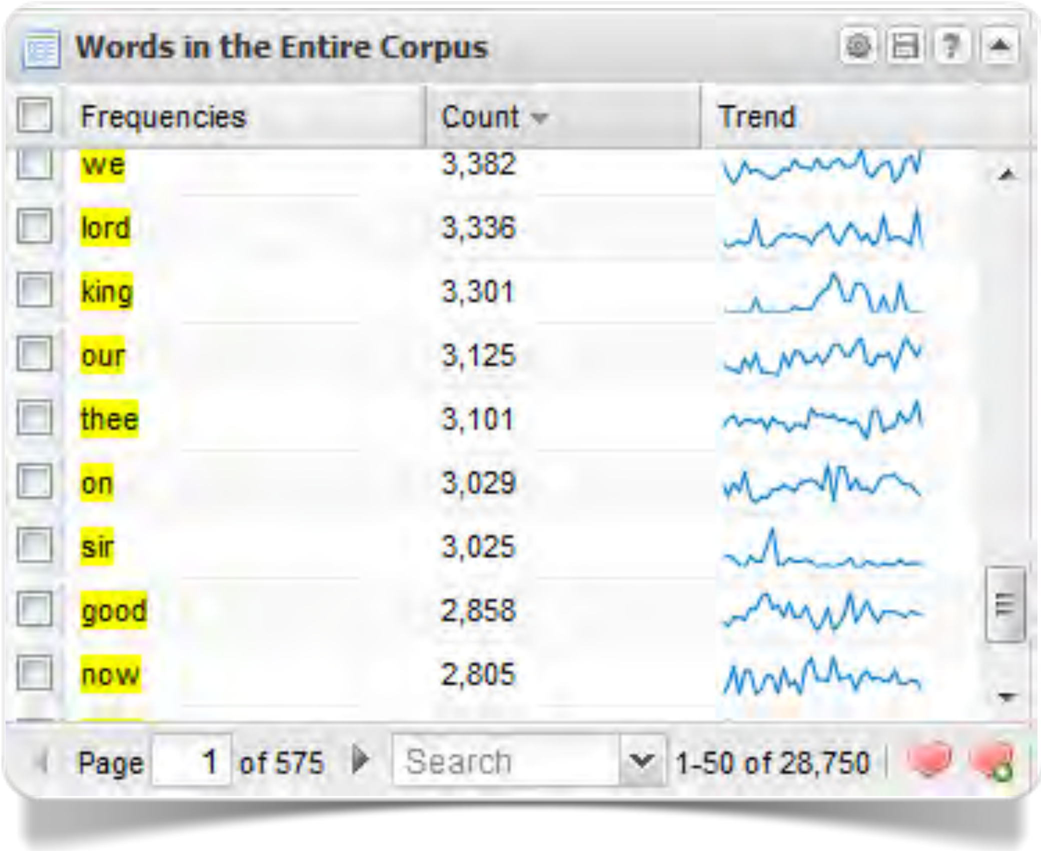


king : 111.89 per 10,000 words
in 33) 1605 King Lear

Fig. 10.5 Word Trends: relative frequencies of *king* in the corpus Shakespeare elab. Voyant)

On the bottom left side of the window is visualized in relative frequency graphs the distribution of the occurrences of all the words in the corpus (fig. 10.6).

The bottom center window visualizes a summary of the lexicometric measurements of the documents belonging to the corpus (fig. 10.7).

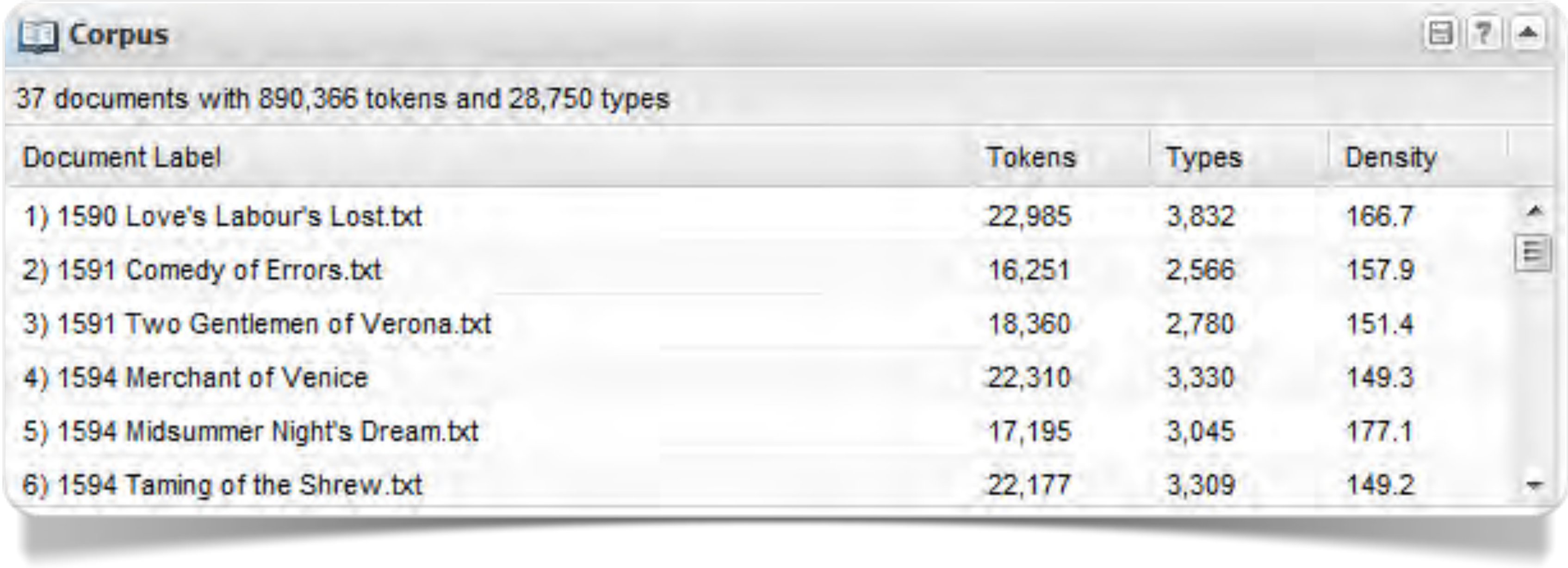


Words in the Entire Corpus

□ Frequencies	Count	Trend
□ we	3,382	
□ lord	3,336	
□ king	3,301	
□ our	3,125	
□ thee	3,101	
□ on	3,029	
□ sir	3,025	
□ good	2,858	
□ now	2,805	

Page 1 of 575 Search 1-50 of 28,750

Fig. 10.6 Distribution of the occurrences in the corpus Shakespeare (elab. Voyant)



Corpus

37 documents with 890,366 tokens and 28,750 types

Document Label	Tokens	Types	Density
1) 1590 Love's Labour's Lost.txt	22,985	3,832	166.7
2) 1591 Comedy of Errors.txt	16,251	2,566	157.9
3) 1591 Two Gentlemen of Verona.txt	18,360	2,780	151.4
4) 1594 Merchant of Venice	22,310	3,330	149.3
5) 1594 Midsummer Night's Dream.txt	17,195	3,045	177.1
6) 1594 Taming of the Shrew.txt	22,177	3,309	149.2

Fig. 10.7 Lexicometric measurement of corpus Shakespeare (elab. Voyant)

Bubblelines

Voyant offers many other text analysis tools. Some are integrated in the interface (Cirrus, Corpus Grid, Corpus Summary, Reader, ecc.); others are accessible one by one. Among the most interesting, we can suggest *Bubblelines*, *Document Term Frequencies* and *Lava*.

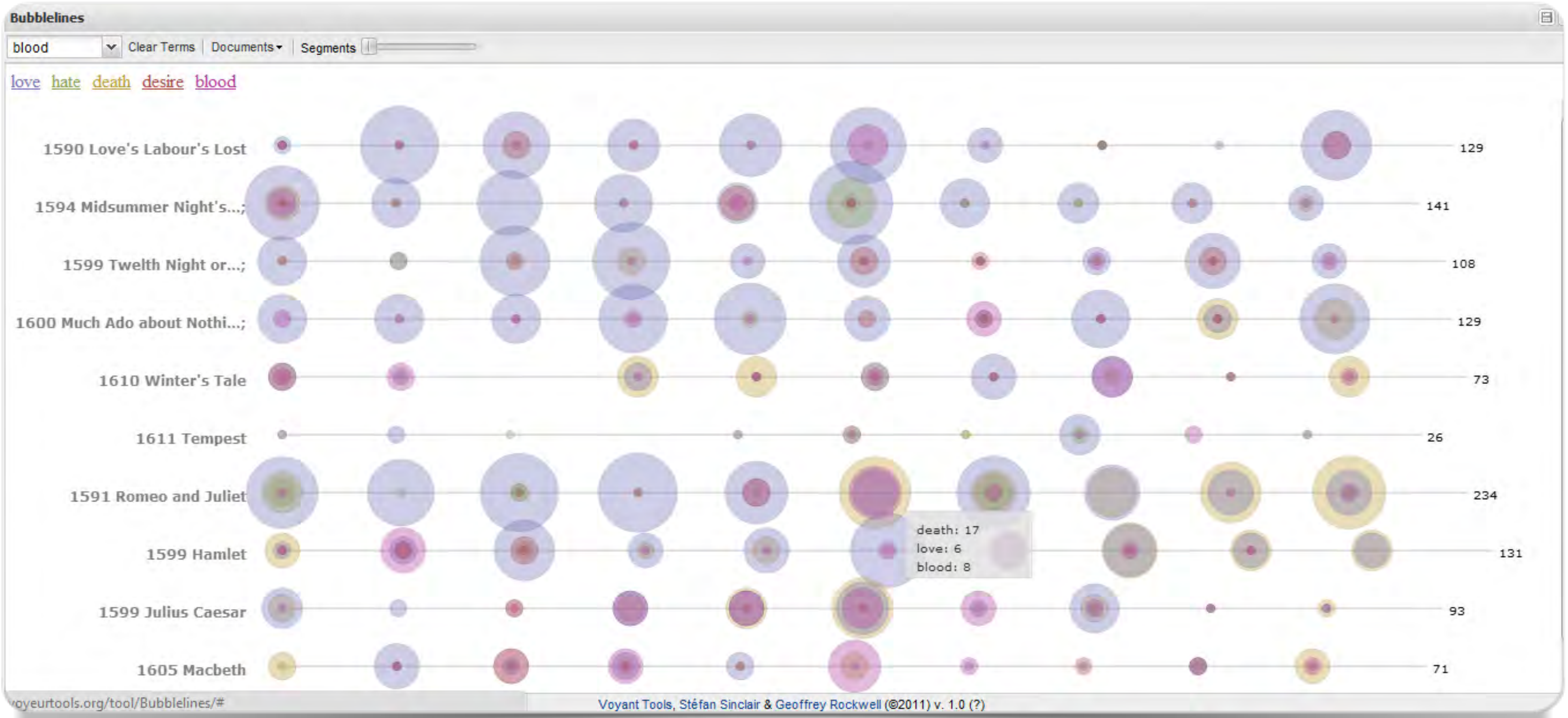

Bubblelines
blood
Clear Terms
Documents ▾
Segments
love hate death desire blood
1590 Love's Labour's Lost 129
1594 Midsummer Night's...; 141
1599 Twelth Night or...; 108
1600 Much Ado about Nothi...; 129
1610 Winter's Tale 73
1611 Tempest 26
1591 Romeo and Juliet 234
1599 Hamlet 131
1599 Julius Caesar 93
1605 Macbeth 71
death: 17
love: 6
blood: 8
voyeurtools.org/tool/Bubblelines/#
Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (©2011) v. 1.0 (?)

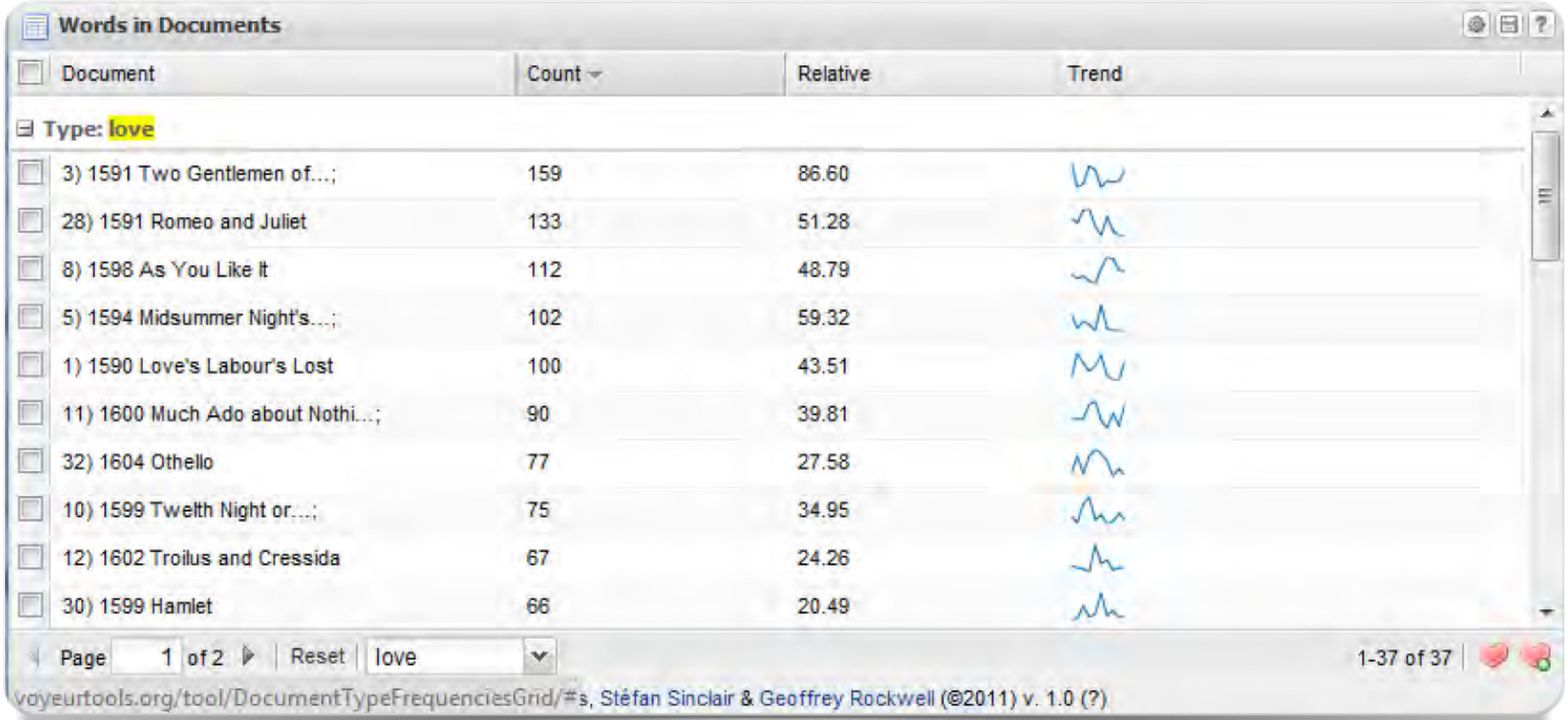
Fig. 10.8 Bubblelines: corpus Shakespeare (elab. Voyant)

Bubblelines visualizes the documents as straight lines positioned horizontally (Click *Open*; select *Shakesperare's play* and *OK*). Each line-document is divided in segments of the same length. The words are identified by a colour and are represented by bubbles, whose size is proportional to the frequency. The co-occurrent words in the same fraction of text overlap each other. The objective is to identify similarities and differences in the common repetition schemes.

In Fig. 10.8 are represented the words: *love*, *hate*, *death*, *desire*, *blood*. The horizontal lines, divided into ten segments, represent 10 plays: *A Midsummer Nigt's Dream*, *Twelfth Night*, *Much Ado about Nothing*, *Winter's Tale*, *The Tempest*, *Romeo and Juliet*, *Hamlet*, *Julius Caesar*, *Macbeth*. As one can see, considering the words selected, *The Tempest* has a completely different structure from the other plays. The first four have some similarities, in particular for what concerns the word *love* (in light blue). *Romeo and Juliet* is characterized by the persistent association of the words *love* and *death*.

Document Term Frequencies

The visualization of the word frequencies in Voyant's interface refers to all the occurrences of the corpus (fig. 10.6). *Document Term Frequencies* produces instead a table of the absolute and relative frequencies of a selected word for every document making up the corpus. In Fig.10.9 we can see the occurrences of the word *love* in the ten works with higher values in decreasing order.



Words in Documents

□ Document	Count	Relative	Trend
⊟ Type: love			
□ 3) 1591 Two Gentlemen of...;	159	86.60	
□ 28) 1591 Romeo and Juliet	133	51.28	
□ 8) 1598 As You Like It	112	48.79	
□ 5) 1594 Midsummer Night's...;	102	59.32	
□ 1) 1590 Love's Labour's Lost	100	43.51	
□ 11) 1600 Much Ado about Nothi...;	90	39.81	
□ 32) 1604 Othello	77	27.58	
□ 10) 1599 Twelth Night or...;	75	34.95	
□ 12) 1602 Troilus and Cressida	67	24.26	
□ 30) 1599 Hamlet	66	20.49	

voyeurtools.org/tool/DocumentTypeFrequenciesGrid/#s, Stéfan Sinclair & Geoffrey Rockwell (©2011) v. 1.0 (?)

Fig. 10.9 Frequencies of the word *love* in the corpus Shakespeare (elab. Voyant)

Lava

The visualization of the corpus in a tridimensional dimension is an interesting and unusual challenge. Lava (Click *Open*; select *Shakesperare's play* and *OK)*, in the first display, presents the corpus as a vertical cylinder in which each segment represents a document. Clicking on one of the cylinders, a document is selected and the application creates a ring, orientable in various directions. The words, in decreasing number of occurrences, are positioned along the circumference of the ring. In fig. 10.10 we can observe *Macbeth*'s ring. It is possible to visualize more rings and compare their properties.

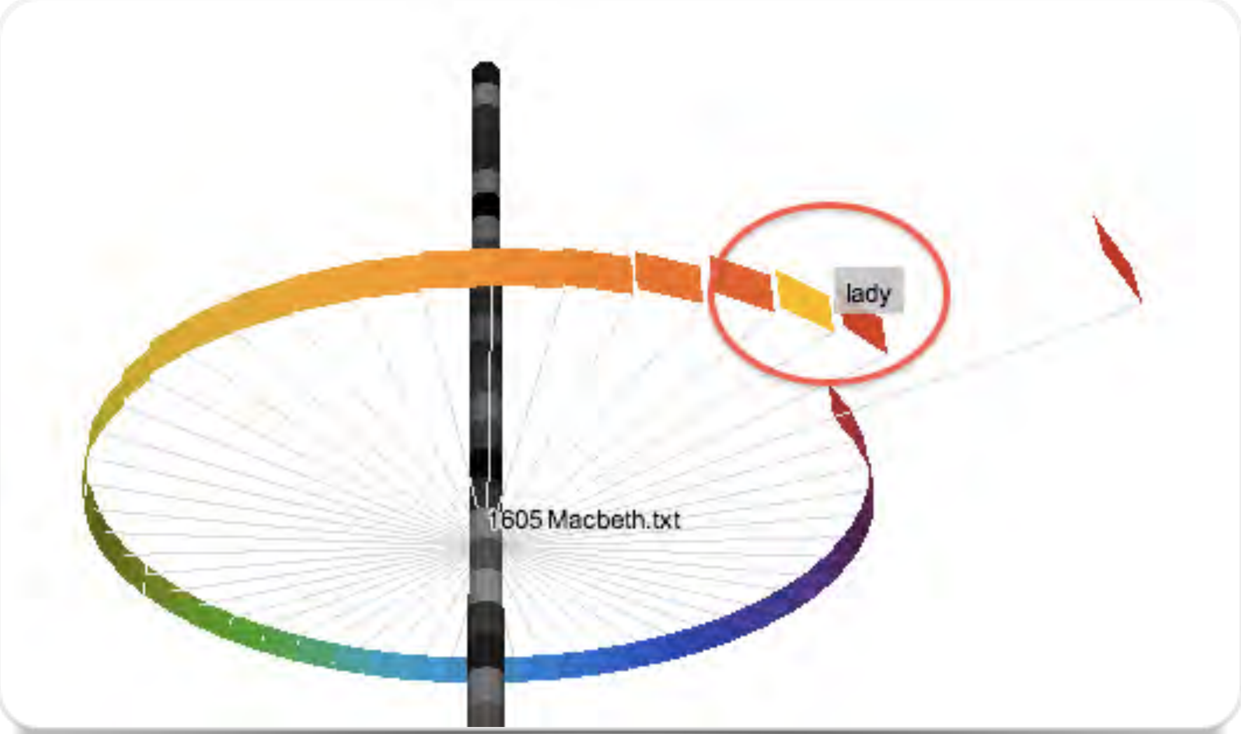


Fig. 10.10 Visualization of *Macbeth*'s ring (elab. Voyant)

Chapter 11

TerMine

The National Centre for Text Mining (NaCTeM) is operating at the University of Manchester in collaboration with the University of Tokyo and provides a set of tools and services for the academic community.

11

TerMine: C-Value

TerMine has been developped for the identification of specialistic teminology, but, more in general, it is tool for the extraction of keywords and of significant segments from the documents. The NaCTeM researchers' interests concern the medical field and the development potential of Text Mining in the context of hypothesis discovery and formulation.

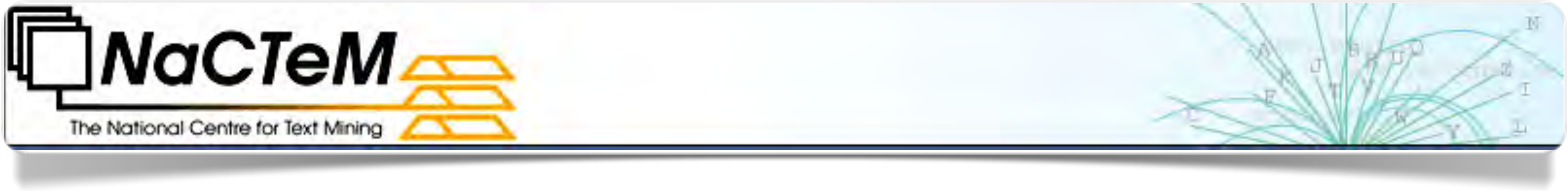
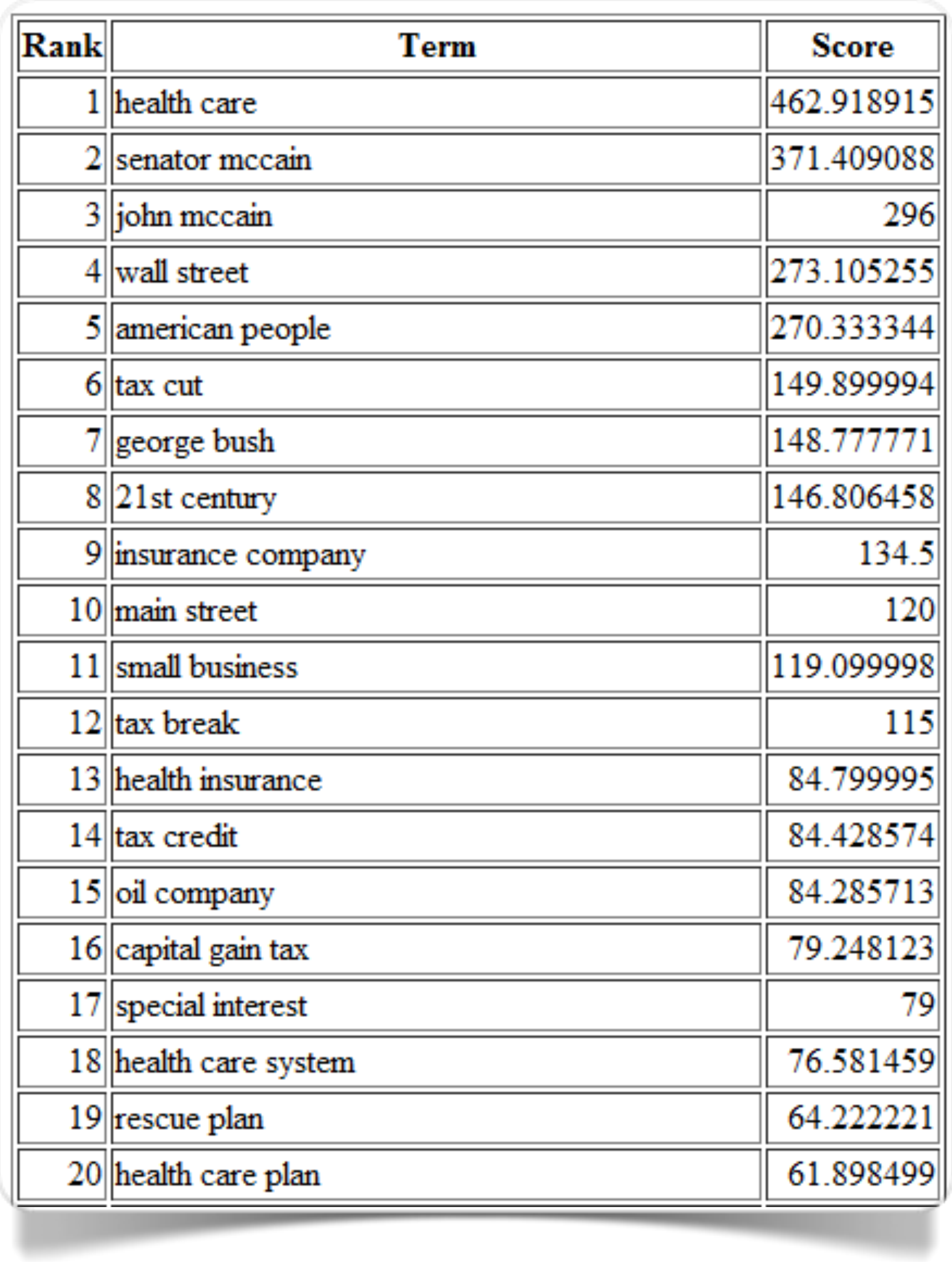


Fig. 11.1 NaCTeM – The National Centre of Text Mining – Logo

The text to be analysed can be inserted in an acquisition window, uploaded from an external source as a plain text file or indicating the web address of the document to be analysed (up to a maximum of 2 Mb). TerMine uses a method called C-value, a domain-independent method for automatic term recognition which combines linguistic and statistical analyses (Frantzi et alii, 2000).

The words extracted are ordered according to decreasing C-value values: The higher values correspond to more relevant words (for example, 2-grams and 3-grams have a greater relevance in conditions of equal occurrences with reference to single words). In fig. 11.2 we can observe the first twenty most relevant words in a corpus made up of 113 speeches delivered by Barak Obama during his presidential campaign from the 3rd of January to the 4th of November 2008 in different cities of the United States.



Rank	Term	Score
1	health care	462.918915
2	senator mccain	371.409088
3	john mccain	296
4	wall street	273.105255
5	american people	270.333344
6	tax cut	149.899994
7	george bush	148.777771
8	21st century	146.806458
9	insurance company	134.5
10	main street	120
11	small business	119.099998
12	tax break	115
13	health insurance	84.799995
14	tax credit	84.428574
15	oil company	84.285713
16	capital gain tax	79.248123
17	special interest	79
18	health care system	76.581459
19	rescue plan	64.222221
20	health care plan	61.898499

Fig. 11.2 C-values of the 20 most relevant words in the corpus of Barak Obama's speeches (Elab. TerMine)

In the promises of the new president there are clear references to the healthcare reform: *health care* (1); *insurance company* (9); *health insurance* (13); *health care system* (18); *health care plan* (20).

Chapter 12

LIWC Linguistic Inquiry and Word Count

The tools we have analysed in this book are mostly for the automatic analysis of texts from the point of view of statistical linguistics. LIWC is a real content analysis software.

12

LIWC

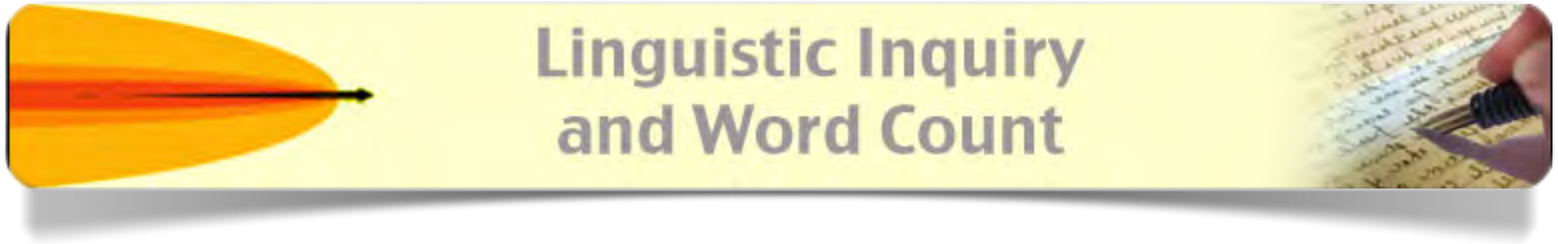


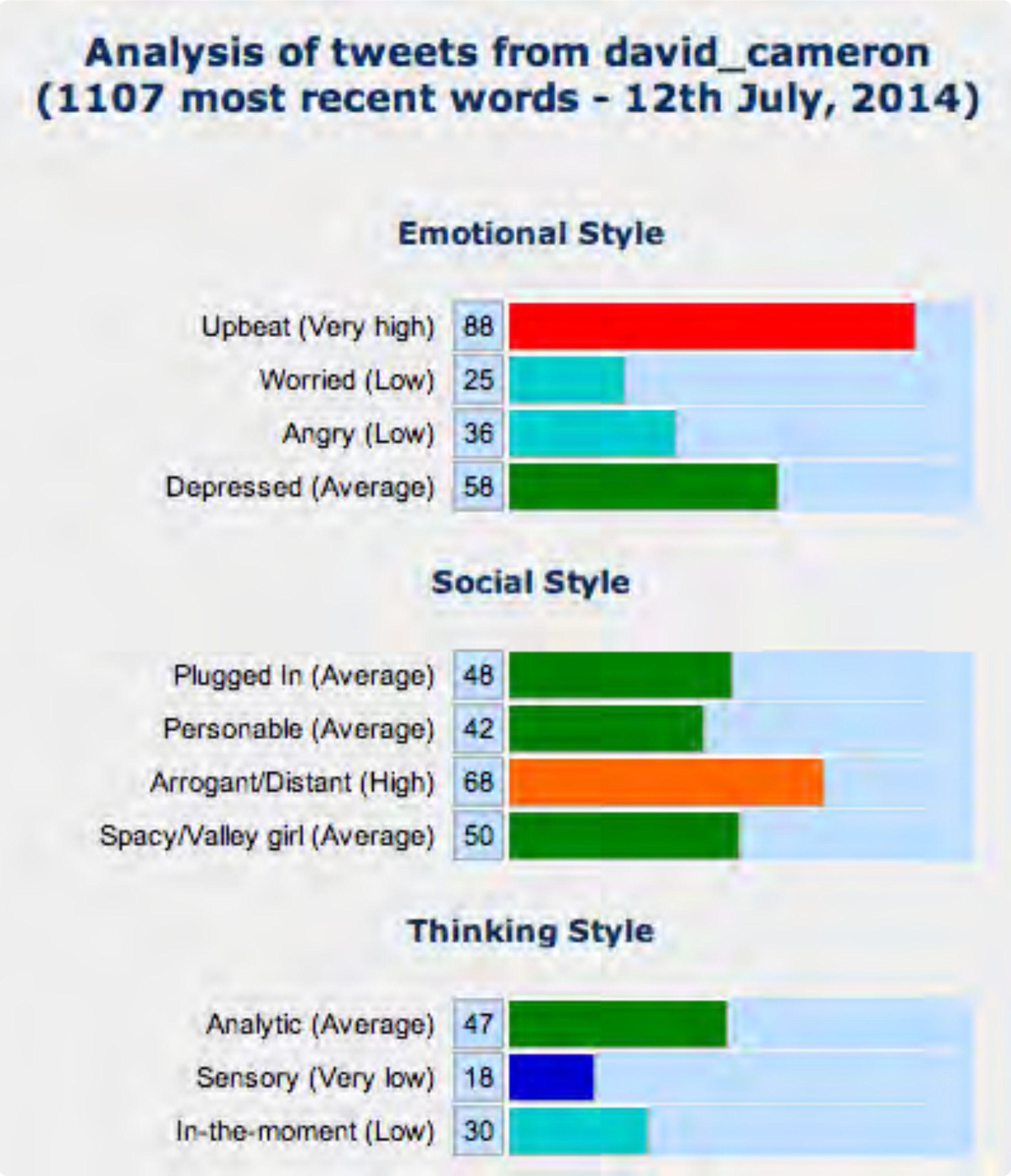
Fig. 12.1 LIWIC – Linguistic Inquiry and Word Count - Logo

LIWC (Linguistic Inquiry and Word Count) - developed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis – is a real content analysis software which classifies words according to certain categories (identified by dictionaries or specific lexicon) and compares the result with reference values taken from sample texts. The hypothesis of the authors is based on a psychometric approach according to which from the use of the words it is possible to obtain indications on some cognitive and emotional aspects of the personalities of the speakers.

The lexicon has been built using different sources for a total of 168 million words and 24 000 writers/speakers. The attribution of words to the categories took place in different phases (How it Works), also with the intervention of “independent judges”. In addition to the usual linguistic and grammatical categories, we find social, emotional, cognitive and behavioural processes.

For a complete use, the software needs to be installed on the computer and it would not be therefore totally in agreement with the other tools gathered in this volume. We have included it in the study because it offers access on line (Try Online) and this makes it possible to observe some of the most relevant results on a trial text inserted in the acquisition window. With Analyze Words we can reveal emotion, social and thinking style of the tweets (Gallery 12.1).

Gallery 12.1 Analyse Words (elab. LIWIC)



Chapter 13

Tree Cloud

Tree Cloud is an application which moves from the “word cloud” model to offer a visualization of the co-occurrences between words according to their closeness in a text.

13

Tree Cloud

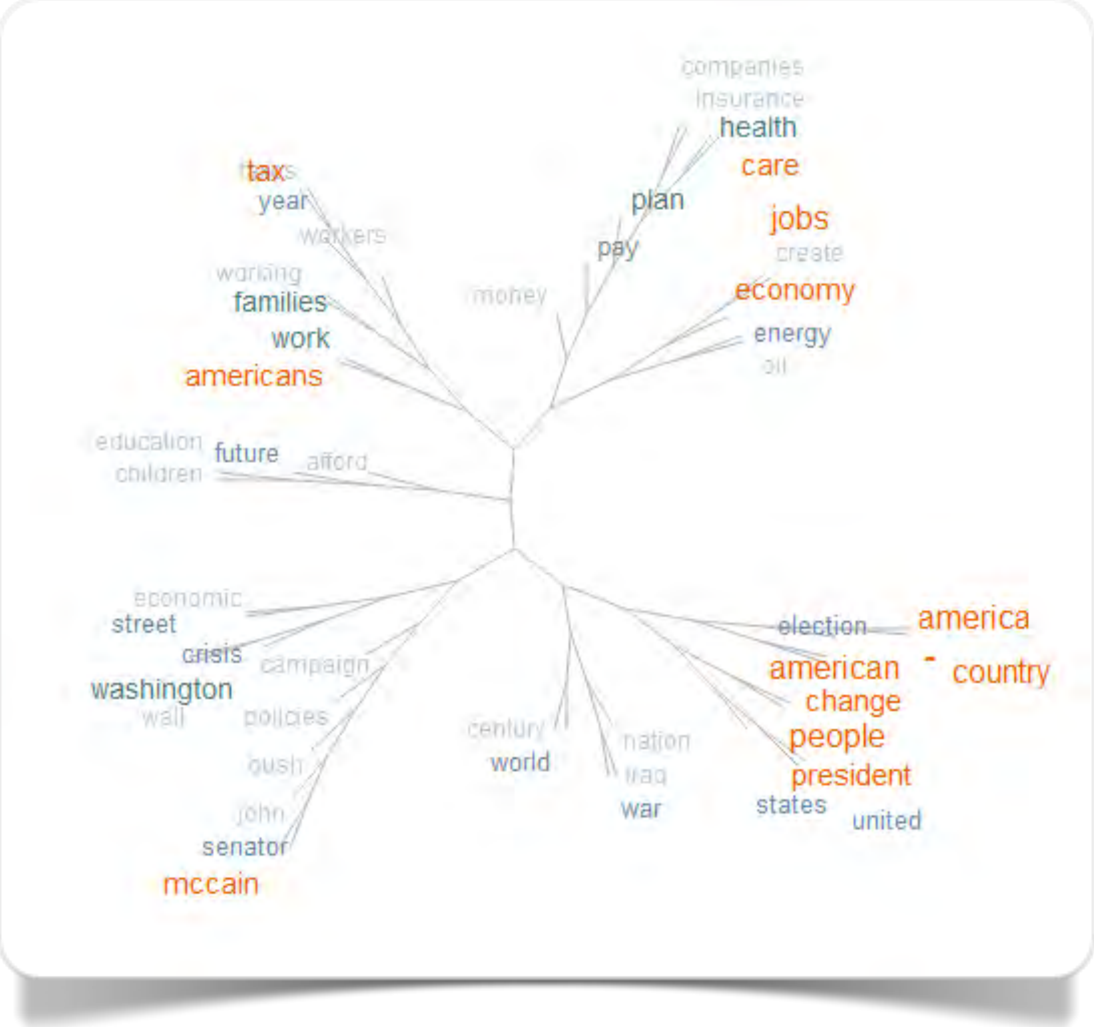


Fig. 13.1 The main topics in Barak Obama's presidential campaign speeches (elab. Tree Cloud)

The algorithm of Tree Cloud has been developed by Philippe Gambette and Jean Véronis for a software to be installed on computer, but it also permits the elaboration of texts from a web access page, copying and pasting in plain text. It is possible to gain access from Create! The words are grouped like a tree according to measures of vicinity and they produce “spontaneous” classifications according to their closeness. The result of a visualization can be therefore interpreted as a synthesis of the topics presented in the text (Gambette & Véronis, 2010; Amstuz & Gambette, 2010).

In fig. 13.1 we can see the visualization that summarizes the content of the corpus of the 113 speeches delivered during Barak Obama's presidential campaign. We can observe how the tree structure has developed in eight main branches. Moving from 12:00 we have: Health-care policies; energetic policies; presidential elections; international politics; the references to his republican opponent; conservative politics identified with Washington; the future of the young generations; American families and fiscal policies.

Chapter 14

Google Books Ngram Viewer

Google Books Ngram Viewer has been designed by a team of researchers of the University of Harvard directed by Jean-Baptiste Michel in collaboration with Google.

14

Google Books Ngram Viewer

Within Google Books' digitalization project (currently more than 30 million scanned books), the Harvard researchers and Jean-Baptiste Michel have built a corpora system in different languages (American English, British English, French, German, Spanish, Russian, Hebrew, Chinese and Italian). Ngram Viewer uses 4% of the books published from 1500 onwards for a total of 500 billion occurrences. (Michel et alii, 2010; Dekahaye & Gauvrit, 2013).

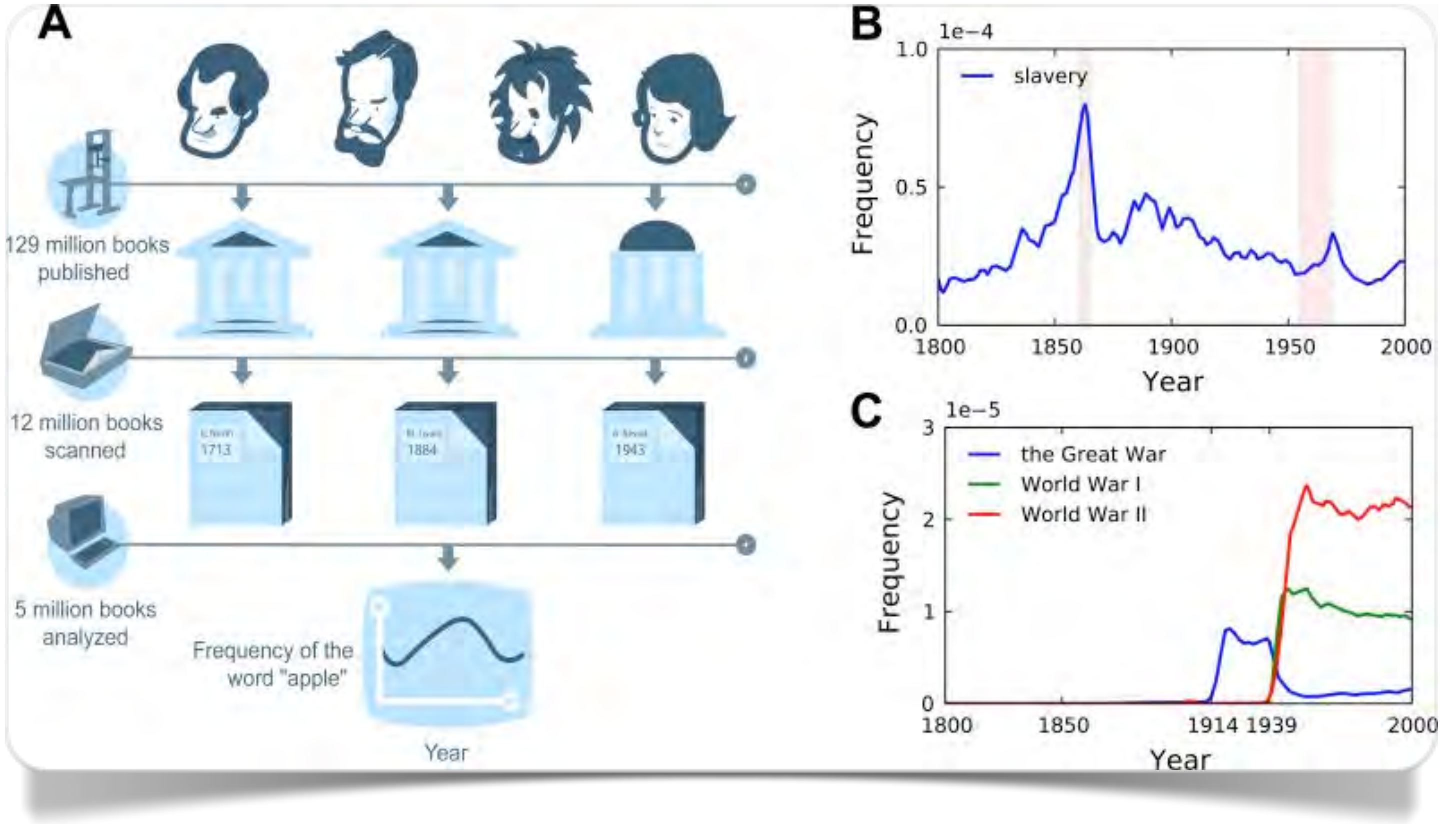
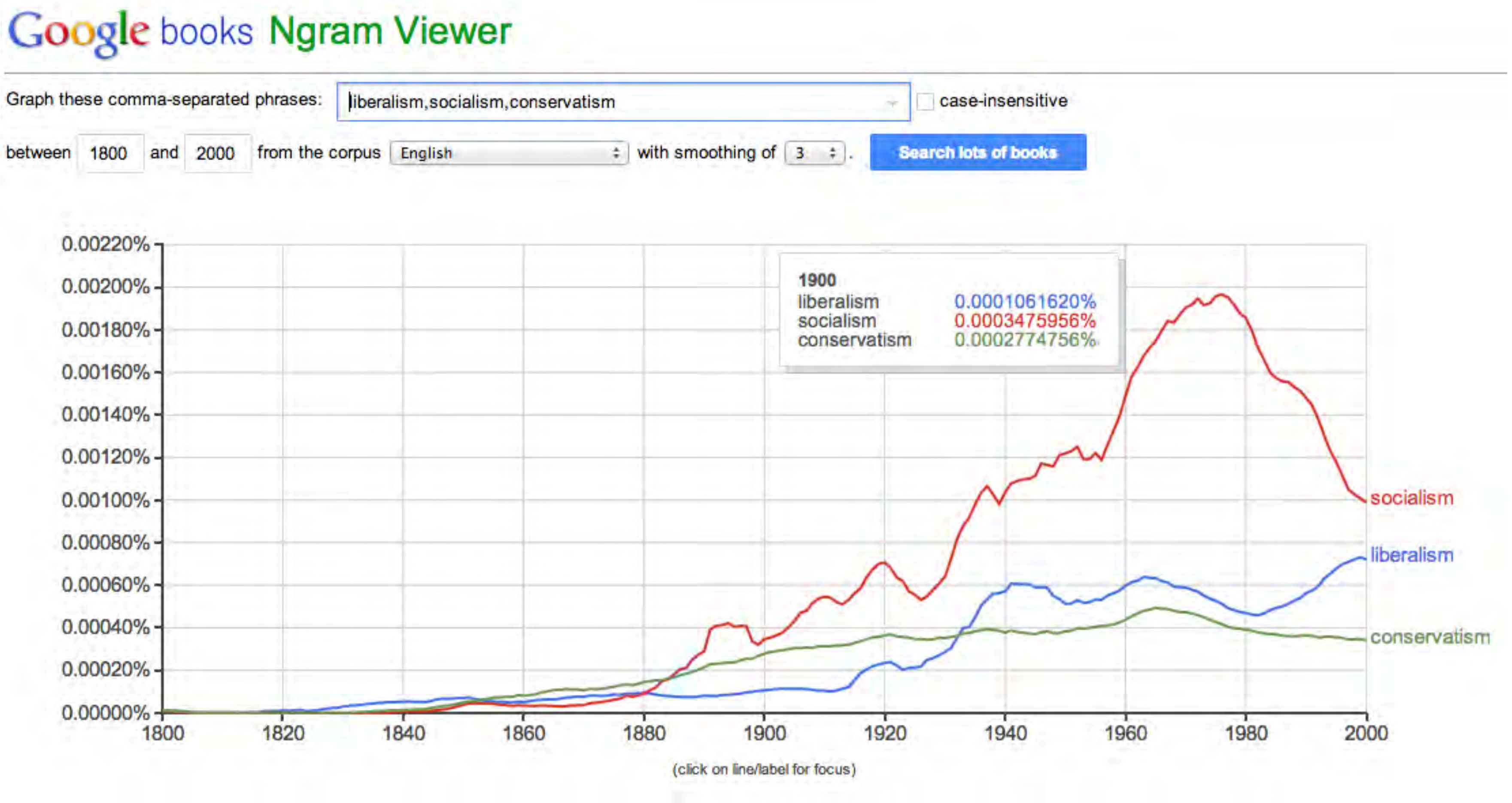


Fig. 14.1 Source: Jean-Baptiste Michel et alii, 2010

Also in this case, we are talking about an anomalous application in comparison to the others presented so far because it works on a pre-defined corpus. It is, however, a tool of absolute importance from the linguistic point of view when making comparisons and evaluations about the presence of a word over a chronological period of 500 years, although the authors suggest to use with caution the indications referring to the period before 1800. In Figure 14.2 we can see an example which compares *conservatism*, *liberalism* and *socialism*.



Graph these comma-separated phrases: liberalism,socialism,conservatism □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

1900
liberalism 0.0001061620%
socialism 0.0003475956%
conservatism 0.0002774756%

Fig. 14.2 Relative frequencies in percentage for three-year moving average of the words: *conservatism*, *liberalism*, *socialism* (elab. Google Books Ngram Viewer)

The ngrams are matched by case-sensitive spelling and plotted on the graph if found in 40 or more books during each requested year. Released in December 2010, Ngram Viewer currently supports wildcard search, inflection search, case insensitive search, part-of-speech tags and ngram compositions (About Ngram Viewer) The plot is viewed as a moving average. A smoothing of 3 means that the data shown for 1980 will be an average of the raw count for 1980 plus 3 value on either side: ("count for 1977" + "count for 1978" + "count for 1979"+ "count for 1980" + "count for 1981" + "count for 1982" + "count for 1983"), divided by 7. A smoothing of 0 means no smoothing at all: just raw data as in Figure 14.3.

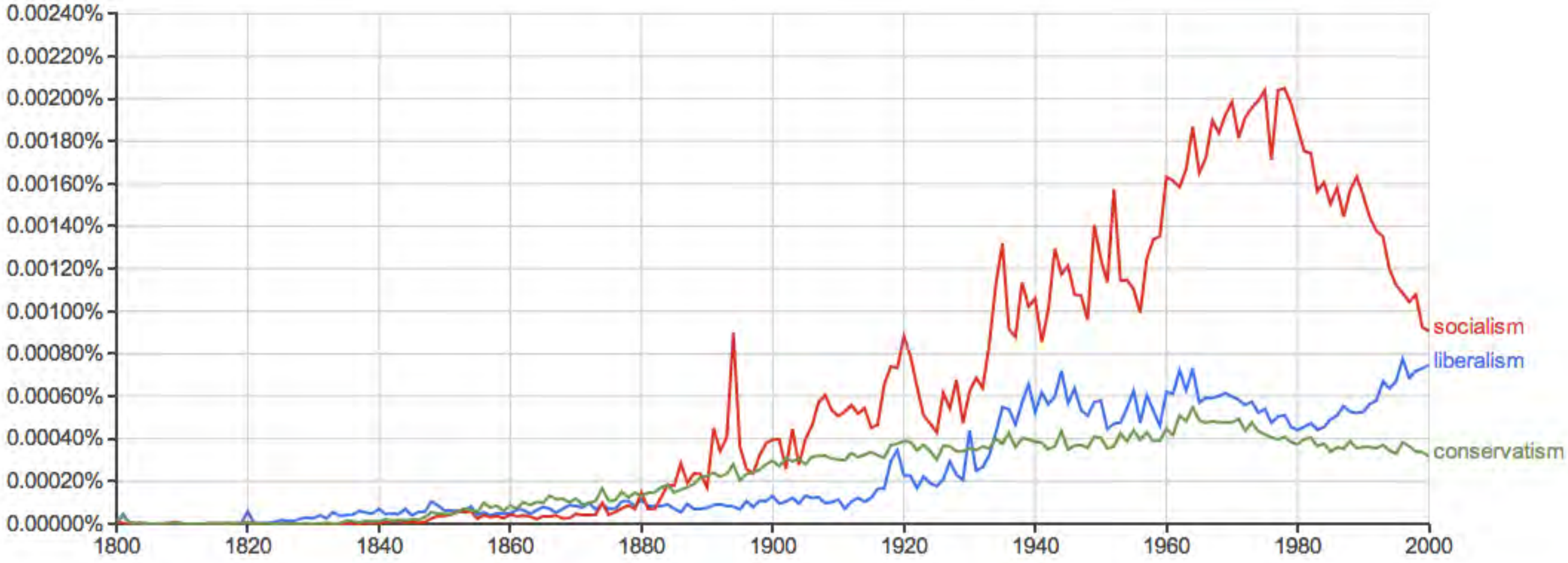
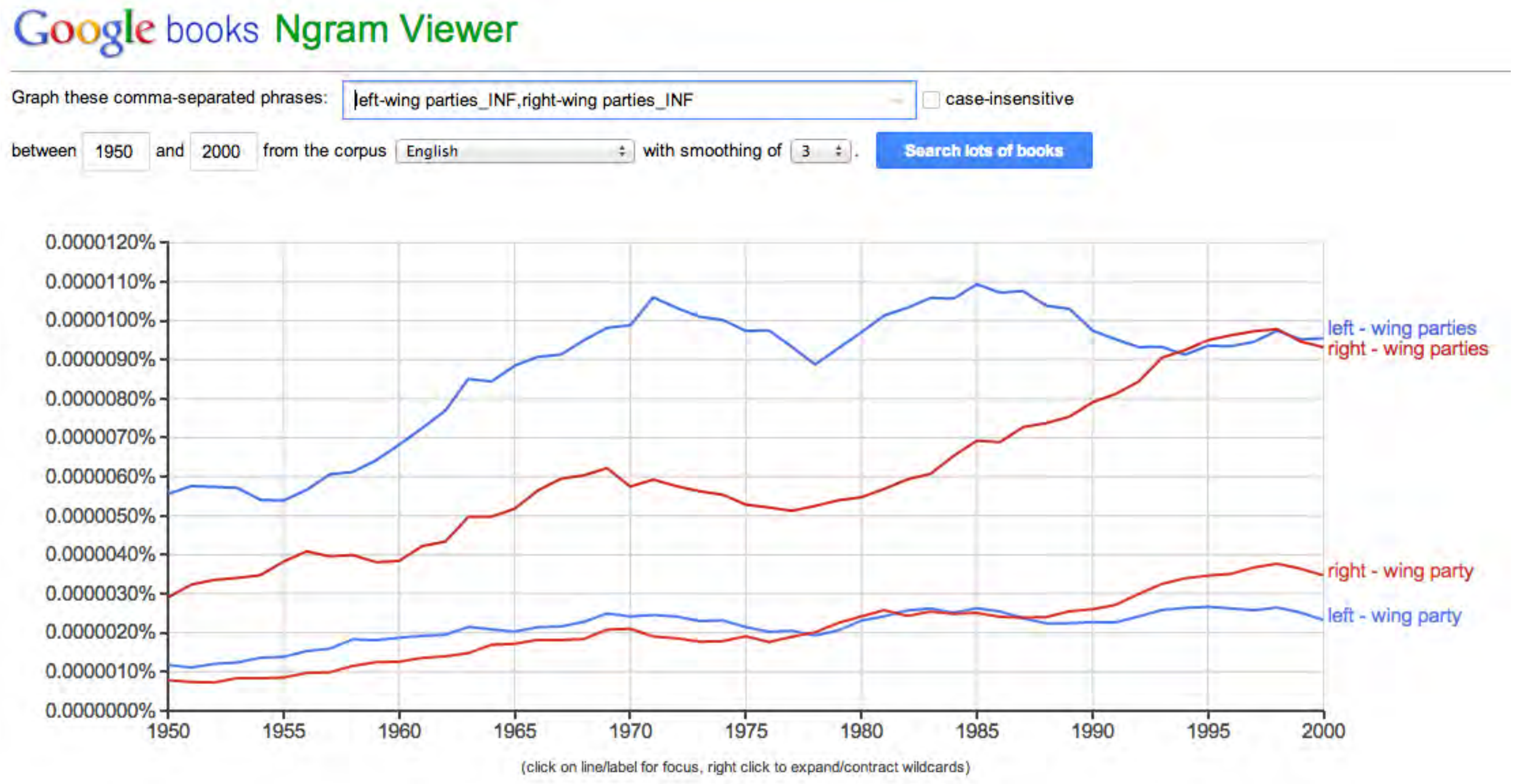


Fig. 14.3 Relative frequencies in percentage for year of the words: *conservatism*, *liberalism*, *socialism* (elab. Google Books Ngram Viewer)

A moving average is a technique to get an overall idea of the trends in a data set and it is commonly used with time series data to smooth out short-term fluctuations.

The Figure 14.4 shows trends in two 3-grams from 1950 to 2000: "left-wing parties” and “right-wing parties”. To get all the different inflections of the word “parties”, we append _INF to a word “parties”.



Google books Ngram Viewer

Graph these comma-separated phrases: left-wing parties_INF,right-wing parties_INF ☐ case-insensitive

between 1950 and 2000 from the corpus English with smoothing of 3.

Fig. 14.4 Relative frequencies in percentage for three-year moving average of the words: left-wing and right-wing parties (elab. Google Books Ngram Viewer)

Chapter 15

A “double movement” between quality and quantity

In general, all these tools have a potential that goes beyond the purpose for which they have been built.

15

Total recall

As often happens in computer science, the applications have to be explored and used with a fair dose of wisdom, but also of inventive and creativity. In the treatment of digital texts there is a fertile meeting between disciplines that study the uniqueness and the particularity of their subject and disciplines that try to generalize observations by selecting their properties and creating classes of objects.

This distinction brought in the past to the separation of human and natural sciences, of interpretation and explanation sciences. Now a synthesis is possible and it is up to the human and social sciences to accept the challenge and to move in the direction of eliminating the presumed contrast between quality and quantity.

With texts consisting in hundreds of thousands or millions of words, the researcher, not being able to read the text directly, is forced to make use of lexicometric and quantitative strategies to identify keywords and semantic units of some relevant interest or some particular element inside the text he/she is studying. Later on he/she will be able to assume an hermeneutical point of view selecting the text fragments of greater interest on the basis of the

Fig. 15.1 – *Total Recall* directed by Paul Verhoven (TriStar Pictures © 1990)

keywords previously identified. This “double movement” from quality to quantity and from quantity to quality becomes essential in the management of an e-memory, the “total memory” we will not be able to do without in a future not so distant from us (Bell e Gemmel, 2009).

The pioneers (or the prophets) of this revolution have ideally made reference to the tale by Philip Dick in *We Can Remember It For You Wholesale* (1966), made popular by the film *Total Recall* directed by Paul Verhoven (1990).

None of us would be ready to bet on these future anticipations because too often the future has followed paths which nobody had predicted. There are signs, however, indicating that this the direction we are taking.

Digital tracks

IWeb 2.0 with its introduction of concepts such as interactivity, performance, open work, subjective interpretation and cognitive space of the user, construction of sense, experiential pragmatics, computer science system exchange, has brought to all the essential elements at the basis of social interaction, communication and "total connectivity" which characterize knowledge, its economic valorisation and the improvement of decision strategies.

More and more numerous are the applications which favour the sharing of personal (photography, consumption behaviours, readings, GPS satellite positions, hotels and restaurants visited, wine lists, films suggested to friends) and institutional (e-government and open data) information. In this network of mass inter-connection will develop the digital tracks of e-life and of collective memory in the near future. There will no longer be borders between Real Life and Second Life (a distinction belonging to the era of "pre-Total Recall").

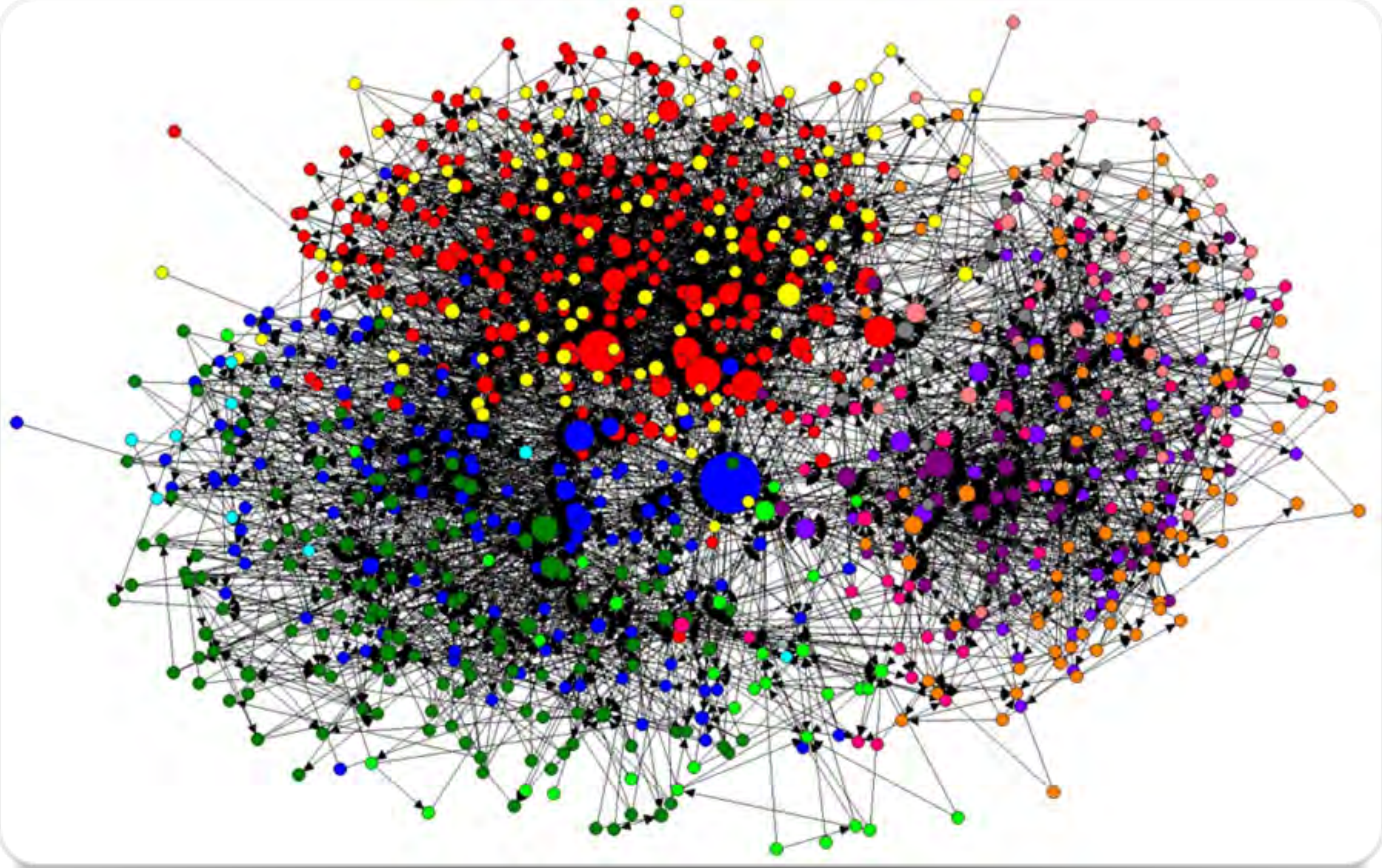


Fig. 15.2 *Social Network Analysis of the Orma* (Jean Ensminger © 2011)

Text-mining has a crucial role in the management of information and communication in the Network Society. Statistics and probability are indispensable to move with awareness and intelligence in this ocean of words, in which disorder seems to reign unchallenged and in which, instead, according to David Weinberger, is coming to life a new order where the protagonists and main organizers of contents and meanings are the users themselves (Weinberger, 2009).

Chapter 16

Bibliography and Web Resources

16

Bibliography

Amstuz D., Gambette P. (2010) Utilisation de la visualisation en nuage arboré pour l'analyse littéraire. In S. Bolasco, I. Chiari, L. Giuliano (Eds). *Statistical Analysis of Textual Data* (JADT2010: 10th International Conference on statistical analysis of textual data, Rome: 9-11 June). Milano: LED, pp. pp. 227-238.

Brown S., Ruecker S., Antoniuk J., Farnel S., Gooding M., Sinclair S., Patey M., Gabriele S. (2010) Reading Orlando with the Mandala Browser: A Case Study in Algorithmic Criticism via Experimental Visualization. *Digital Studies / Le champ numérique*, 2 (1) - http://www.digitalstudies.org/ojs/index.php/digital_studies/article/viewArticle/191

Bell G., Gemmel J. (2010) *Total Recall. How the E-memory Revolution Will Change Everything*. Penguin Group, 2009.

Bolasco S. (2013) *L'analisi automatica dei testi. Fare ricerca con il text mining*. Milano: Carocci.

Bush V. (1945) As We May Think. *Atlantic Montly*, 176 (July), pp. 101-108.

Chiari I. (2007) *Introduzione alla linguistica computazionale*. Bari: Laterza.

Delahaye J-P., Gauvrit N. (2013) *Culturomics. Le numérique et la culture*. Paris: Odile Jacob.

Feinberg J. (2010) Wordle, in J. Steele e L. Iliinski, *Beautiful Visualization*. Sabastopol, CA: O'Reilly Media, pp. 37-58.

Frantzi K., Ananiadou S., Mima H. (2000) Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3 (2) pp. 115-130.

Gainor R., Sinclair S., Ruecker S., Patey M., Gabriele S. (2009) A Mandala Browser User Study: Visualizing XML Versions of Shakespeare's Plays. *Visible Language*, 43 (1), pp. 60-85.

Gambette P., Véronis J. (2010) Visualising a text with a tree cloud. In Locarek-Junge H. and Weihs C. (Eds). *Classification as a Tool for Research. Studies in Classification, Data Analysis, and Knowledge Organization*. Part 3, SpringerLink, pp. 561-569.

Giuliano L., La Rocca G. (2008) *L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso*. Milano: LED.

Grbich C. (2012) *Qualitative Data Analysis: An Introduction*. London: SAGE.

Gruzd A. (2010) Exploring Virtual Communities with the Internet Community Text Analyzer (ICTA). In Danile Ben Kei (Ed.). *Handbook of Research on Methods and Techniques for Studying Virtual Communities. Paradigms and Phenomena*. Hershey, PA: IGI Global.

Haythornthwaite C., Gruzd A. (2007) A Noun Phrase Analysis Tool for Mining Online Community, in *Proceedings of the 3rd International Conference on Communities and Technologies*, Michigan State University.

Jurafsky D., Martin J.H. (2000) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Sadle River, NJ: Pearson Prentice Hall.

Krippendorf K. (2004) *Content Analysis: An introduction to its Methodology*. Thousand Oaks, CA: Sage Pub. (2nd ed.).

Lo Tzai-Wai R., He B., Ounis I. (2005) Automatically building a stopword list for an information retrieval. *Journal of Digital Information Management*, 3(1), 17-97.

Losito G. (2002) *L'analisi del contenuto nella ricerca sociale*. Milano: Franco Angeli (IV ed.).

Madge J. (1962) *The origins of scientific sociology.* New York: Free Press.

Mayer-Schönberger V., Cukier K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray.

Michel J-B., Yuan Kui Shen Y., Presser Aiden A., Veres A., Gray M.K., Pickett, J.P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M.A., Lieberman Aiden E. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books, www.sciencexpress.org /16 December 2010/Page 1/10.1126/science.1199644 (downloaded www.sciencemag.org - December 29, 2010).

Nelson T.H. (1981) *Literary Machine*. Swarthmore.

Paley W.B. (2002) TextArc: An alternate way to view a text (retrieved http://www.textarc.org/posters/TextArc_PosterNotes.pdf).

Pang L., Lee L. (2009) Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, Vol. 2, Nos. 1–2, pp. 1–135.

Pennebaker J.W. , Graybeal A. (2001) Patterns of Natural Language Use: Disclosure, Personality, and Social Integration. *Current Directions in Psychological Science*, 10(3), pp. 90-93.

Porter M.F. (1980) An algorithm for suffix stripping. *Program*, 14(3) pp 130–137.

Robertson S.E., van Rijsbergen C. J., Porter M.F. (1981). Probabilistic Models of Indexing and Searching, in R.N. Oddy R.N., S.E. Robertson, C.J. van Rijsbergen, P.W. Williams (Eds.). *Information Retrieval Research*, Proc. Joint ACM/BCS Symposium in Information Storage and Retrieval, Cambridge, June 1980, Butterworths, pp. 35-56.

Rockwell G. (2003) What is Text Analysis, Really? *Literary and Linguistic Computing*, 18(2), pp. 209-219.

Sinclair S. (2003) Computer Assisted Reading: Reconceiving Text Analysis. *Literary and Linguistic Computing*, 18(2), pp. 175-184.

Weinberger D. (2007) *Everything is miscellaneous: the power of the new digital disorder*. New York: Times Books.

Web Resources

Google Books Ngram Viewer (https://books.google.com/ngrams/)

LIWC – Linguistic Inquiry and Word Count (http://www.liwc.net/)

Many Eyes (http://www-958.ibm.com/software/data/cognos/manyeyes/)

NETLYTIC - Internet Community Text Analyzer (http://netlytic.org/)

Tagxedo (http://www.tagxedo.com/)

TAPoR – Text Analysis Portal of Research (http://portal.tapor.ca/portal/portal)

TAPoRwere – Prototype of Text Analysis Tools (http://taporware.ualberta.ca/~taporware/htmlTools/comparator.shtml?)

TerMine (http://www.nactem.ac.uk/software/termine/)

Textalyser (http://textalyser.net/)

Text Analyzer (http://www.online-utility.org/text/analyzer.jsp)

TextArc (http://www.textarc.org/)

TreeCloud (http://treecloud.univ-mlv.fr/)

Voyant – See through your text (http://voyeurtools.org/)

Wordle (http://www.wordle.net/)

SAPIENZA
UNIVERSITÀ DI ROMA

Department of Statistical Sciences

Data Science Series

1. L.Giuliano, *Il valore delle parole. L'analisi automatica dei testi in Web 2.0*.
2. D. Schiavon, Luca Giuliano. *Wizard grafico. Una guida alla visualizzazione dei dati numerici*.
3. L. Giuliano. *The Value of Words. Automatic Text Analysis Tools in Web 2.0.*
4. D. Frongia. *Introduzione alla Social Network Analysis* (in preparazione).

Luca Giuliano
The Value of Words. Automatic Text Analysis
Tools in Web 2.0.
Roma : Department of Statistical Sciences,
[2014] 112 p.

ISBN 978-88-908757-2-4
9 788890 875724