Corpus collection and analysis for the linguistic layman: The Gromoteur

Kim Gerdes

LPP, Université Sorbonne Nouvelle & CNRS

Abstract

This paper presents a tool for corpus collection, handling, and statistical analysis: the Gromoteur. The paper explains the scientific needs that lead to the development of the tool and its main characteristics. Different usage schemes are explored in which the tool can contribute to simple NLP tasks that are often beyond the reach of the common linguist. In particular corpus collection from online resources or heterogeneous file structures becomes an easy task with the Gromoteur. But also tokenization, lemmatization, or part of speech annotation of various languages are a matter of a few clicks in this tool. The Gromoteur also allows for simple statistical analysis of the collected language data, including a section map, specificity computation for any set of sections, and collocation computation for n-grams.

Keywords: corpus collection, web crawler, online corpus, the Web as corpus, tagger, lemmatizer, statistical analysis, specificity computation, collocations

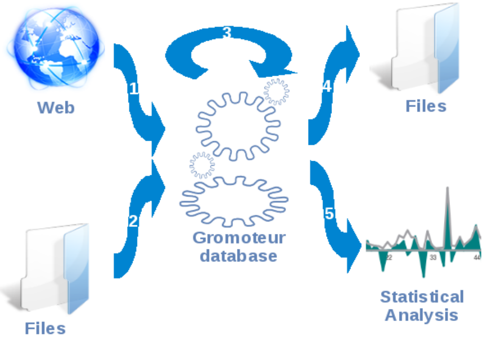
1. Introduction

The sudden abundance of linguistic data, written and spoken, through the World Wide Web, has brought linguistics from data scarcity and the obligation to rely on intuition in a short time to the possibility to verify and even quantify many of the formally purely theoretical postulates. In spite of this paradigm change, common linguists run up against a multitude of technical roadblocks to use “Big Data” for their research. If they want to use large corpora, they mainly end up either with simple results provided by all purpose search engines or with specific tools that have predefined corpora like Google’s Ngram viewer or the WebCorp Linguist's Search Engine. A simple task like downloading all available articles of an online newspaper or a blog, coding them uniformly, cleaning them and making simple measures of the distribution of words is beyond the reach of a linguist not trained on technical aspects of encoding and scripting.

The Gromoteur attempts to remedy this need. It is a tool for linguists that gives easy access to textual corpora. It allows to get pages from the Web or from local files, treat them, analyze them, and output results. It is tailored for common needs of linguists that want to work on data they either have in various file formats on the hard drive, or that they have access to on the Web. All configurations and usage is controlled from a graphical interface, and in common usage scenarios no scripting or tweaking with configuration files is necessary.

2. Usage Schemes

The following figure illustrates the different paths linguistic data can take in the Gromoteur: The Gromoteur can take data from the internet (1) and from local files (2), extract parts of pages, sort, and clean the data (3), export the data to different text files and concordancer formats, and a simple integrated tool allows for simple statistical analyses.



2.1. The crawler

The central part of the Gromoteur is the crawler with a wide range of configuration options.

After creating a database (or using an existing one), the user can fill it with data from the internet. A dialog wizard guides the user through the configuration. The dialog distinguishes between *normal* and *expert* configuration options, the latter being accessible by clicking on the “expert” button of the dialog. The normal user is thus exposed only to a few essential options among the *expert* options presented below, the simple options being only the website to start crawling, the restriction on the websites to visit, and the number of pages to retrieve.

The first question is only relevant if the user has filled the database with the crawler before, and wants to complete the database now:

- new data is appended to the tables
- only equal URLs are overwritten
- the database is completely erased and started from zero
- the search starts with URLs that were collected during the last crawl and remained to be explored

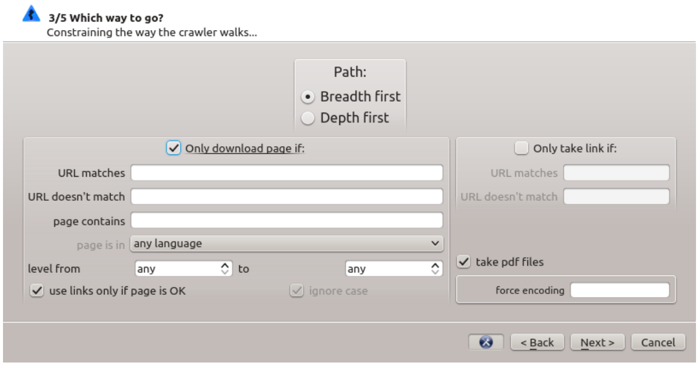
The user is then presented with the choice between the two basic ways of finding data on the internet:

1. by giving seed URLs as a starting point
2. by providing search terms for a search using the Bing search engine

The seed URLs can either be pasted directly into the in the configuration dialog or provided in a file containing a URL per line

The number of results for any Bing search are directly given in the interface and the result pages can also be opened and examined in a browser by a simple click. The Bing query syntax allows for all the common restrictions on language, URL, and file type (see http://onlinehelp.microsoft.com/ en-us/bing/ff808421.aspx for details). Moreover the Bing locale can be set automatically or manually. The user can decide to just download the search results or to use the results as seeds from which to keep crawling. To use the Bing search API, Gromoteur is preconfigured with a Bing user ID, but in case the users of Gromoteur want to make extensive use of this feature they are urged to get their own Bing user ID, in order not to attain the daily limit of searches per user ID.

The next screen in the configuration wizard asks to put restrictions on the way to walk the web. As shown in the screenshot below, this includes the order in which the pages are taken, restrictions on the URL and the terms on the page, the depth of the crawl, and the file types. The Gromoteur can also check the language of each page, using the letter trigram based Language Detector (http://elizia.net/languageDetector/), thus staying on a unilingual search path. This is particularly interesting for the corpora creation of rare languages where the examples pages can be found by a keyword search, which is then controlled by the language testing.



3/5 Which way to go?
Constraining the way the crawler walks...

Path:
● Breadth first
○ Depth first

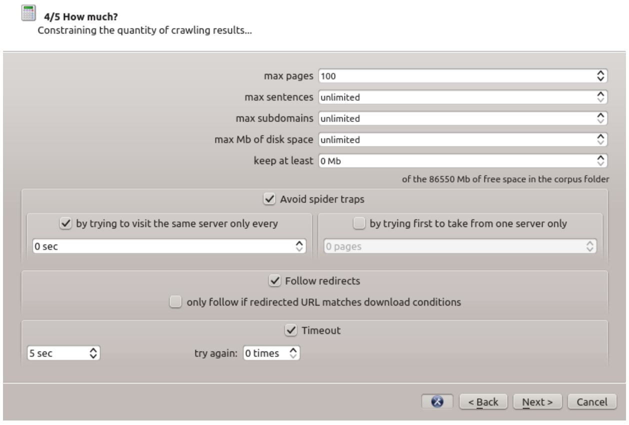
☑ Only download page if:
URL matches
URL doesn't match
page contains
page is in: any language
level from: any to: any
☑ use links only if page is OK
☑ ignore case

☐ Only take link if:
URL matches
URL doesn't match

☑ take pdf files
force encoding

The restrictions on the URLs and the web content can be a simple word or part of the URL, but they are in fact Perl/Python style regular expressions. Although regular expressions are a quite powerful tool allowing to go beyond simple keyword or URL matching, they turn out to be the biggest obstacle for the non-technical oriented users of the Gromoteur in the crawler configuration.

The following dialog page controls the quantity and speed of the crawl.



4/5 How much?
Constraining the quantity of crawling results...

max pages: 100
max sentences: unlimited
max subdomains: unlimited
max Mb of disk space: unlimited
keep at least: 0 Mb
of the 86550 Mb of free space in the corpus folder

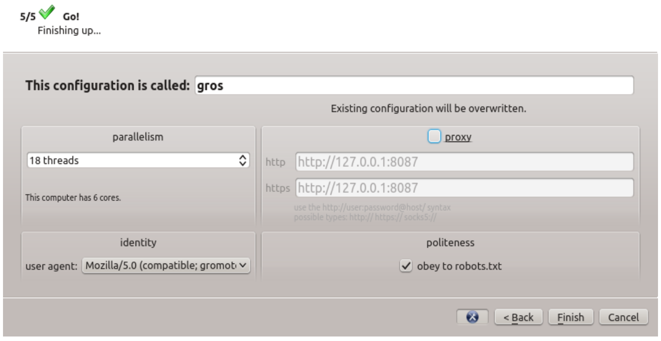
☑ Avoid spider traps
☑ by trying to visit the same server only every: 0 sec
☐ by trying first to take from one server only: 0 pages

☑ Follow redirects
☐ only follow if redirected URL matches download conditions

☑ Timeout
5 sec
try again: 0 times

The final configuration screen of the crawler controls the number of threads of the crawl, the usage of proxies, the identification of Gromoteur when crawling, and the handling of the robots.txt file.[1]

[1] The robots.txt file is a Web convention allowing web site owners to control the behavior of Web crawlers when visiting their websites.



5/5 ✓ Go!
Finishing up...

This configuration is called: gros

Existing configuration will be overwritten.

parallelism: 18 threads

This computer has 6 cores.

□ proxy

http: http://127.0.0.1:8087

https: http://127.0.0.1:8087

use the http://user:password@host/ syntax
possible types: http:// https:// socks5://

identity

user agent: Mozilla/5.0 (compatible; gromot

politeness

☑ obey to robots.txt

The number of threads can significantly alter the speed with which the pages are taken from the Web. However, it is difficult to find the optimal configuration to enhance the speed. The main factors for the download speed are of course the user's internet connection, the distance to the webserver and the webserver's speed. So in theory, a higher number of parallel threads can only improve the speed, but in reality, a very high number of threads can slow down the hole process because of a slower performance of the crawler, but also because of the reaction of the server, in particular if the download is taken place from one unique server (and not from a variety of servers), as the multiple downloads can be interpreted as a denial of service attack by the webserver.

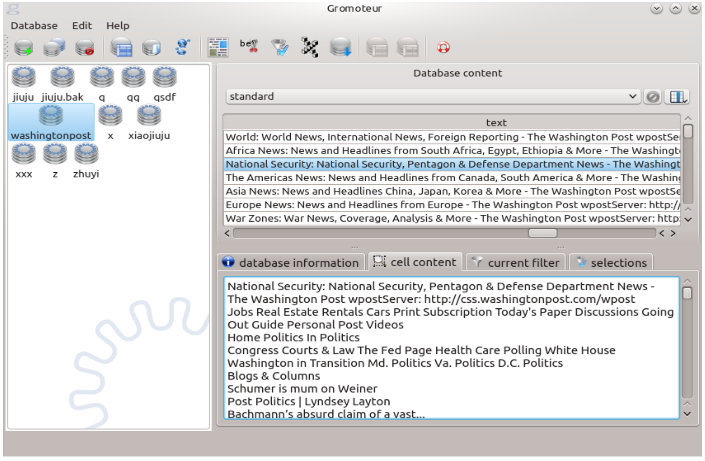
2.2. The file importer

Files can be imported by pointing to folder from which all text, html, or pdf files are imported. This can be very helpful for users that have their data in form of pdf ebooks are similar formats that are difficult to handle without some specialized tools. Alternatively textual data can be imported from a simple tab-separated (csv) format, as can easily be produced from a spreadsheet.

Both in the crawling and file importing process, a wide range of heuristics is applied for the cleaning of the import files and the transformation into uniform UTF-8 forms stored in a Sqlite data base. The database stores the source files as well as the text files, the text files are stored using the full text search module (FTS-4) that provides a full-text index allowing for very fast retrieval also in very large databases (of many gigabytes of text).

2.3. Cleaning, sorting, analyzing

Gromoteur's main window shows the different databases in the left panel and the the content of the selected database in the right panels. Each database has different columns whose complete content is shown when clicked upon. Different selection filters can be applied to any of the columns.

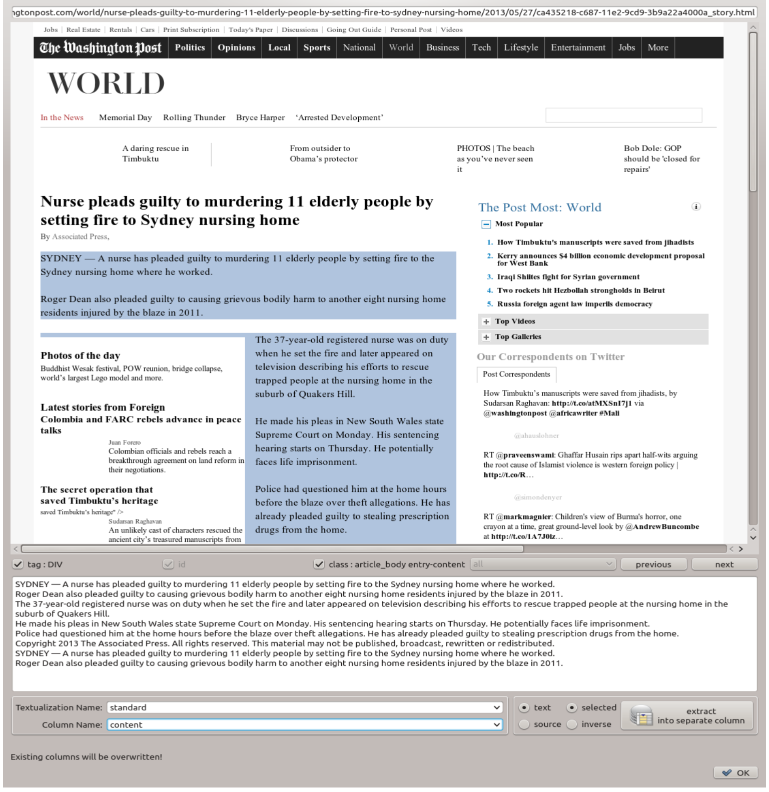


standard

text
World: World News, International News, Foreign Reporting - The Washington Post wpostSe
Africa News: News and Headlines from South Africa, Egypt, Ethiopia & More - The Washingt
National Security: National Security, Pentagon & Defense Department News - The Washingt
The Americas News: News and Headlines from Canada, South America & More - The Washin
Asia News: News and Headlines China, Japan, Korea & More - The Washington Post wpostSe
Europe News: News and Headlines from Europe - The Washington Post wpostServer: http://
War Zones: War News, Coverage, Analysis & More - The Washington Post wpostServer: http:

National Security: National Security, Pentagon & Defense Department News -
The Washington Post wpostServer: http://css.washingtonpost.com/wpost
Jobs Real Estate Rentals Cars Print Subscription Today's Paper Discussions Going
Out Guide Personal Post Videos
Home Politics In Politics
Congress Courts & Law The Fed Page Health Care Polling White House
Washington in Transition Md. Politics Va. Politics D.C. Politics
Blogs & Columns
Schumer is mum on Weiner
Post Politics | Lyndsey Layton
Bachmann's absurd claim of a vast...

A very important feature of the Gromoteur is the possibility to extract parts of webpages into separate columns. A graphical tool shows the downloaded webpage, and clicking on the parts of the page that are interesting for the linguistic analysis selects and highlights all similar sections of the page. The tool automatically show the specific HTML tags of the selected parts of the webpage. Those tags can also be adjusted manually. After testing on a number of pages, the linguist can decide to apply this extraction of segments on the whole corpus, resulting in a new column that only contains the central part of the page, for example the content of the article and not the surrounding information or advertisement.


ngtonpost.com/world/nurse-pleads-guilty-to-murdering-11-elderly-people-by-setting-fire-to-sydney-nursing-home/2013/05/27/ca435218-c687-11e2-9cd9-3b9a22a4000a_story.html
The Washington Post
Politics
Opinions
Local
Sports
National
World
Business
Tech
Lifestyle
Entertainment
Jobs
More
WORLD
Nurse pleads guilty to murdering 11 elderly people by setting fire to Sydney nursing home
The Post Most: World
Our Correspondents on Twitter
previous
next
extract into separate column
OK

☑ tag : DIV ☑ id ☑ class : article_body entry-content all

SYDNEY — A nurse has pleaded guilty to murdering 11 elderly people by setting fire to the Sydney nursing home where he worked.
Roger Dean also pleaded guilty to causing grievous bodily harm to another eight nursing home residents injured by the blaze in 2011.
The 37-year-old registered nurse was on duty when he set the fire and later appeared on television describing his efforts to rescue trapped people at the nursing home in the suburb of Quakers Hill.
He made his pleas in New South Wales state Supreme Court on Monday. His sentencing hearing starts on Thursday. He potentially faces life imprisonment.
Police had questioned him at the home hours before the blaze over theft allegations. He has already pleaded guilty to stealing prescription drugs from the home.
Copyright 2013 The Associated Press. All rights reserved. This material may not be published, broadcast, rewritten or redistributed.
SYDNEY — A nurse has pleaded guilty to murdering 11 elderly people by setting fire to the Sydney nursing home where he worked.
Roger Dean also pleaded guilty to causing grievous bodily harm to another eight nursing home residents injured by the blaze in 2011.

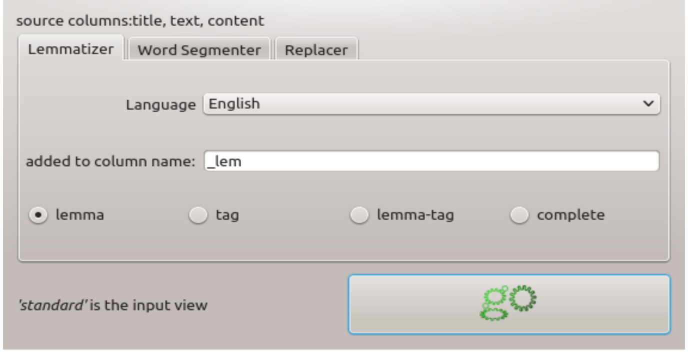
Textualization Name: standard
Column Name: content
● text ● selected
○ source ○ inverse

Existing columns will be overwritten!

Moreover, a selection of tools allows to word-segment Chinese texts using Urheen[2] or Mmseg[3], lemmatize and tag English[4] and Chinese texts, and do global corrections using regular expressions. Gromoteur comes with the taggers of the Pattern module (De Smedt &

[2] http://www.openpr.org.cn/index.php/NLP-Toolkit-For-Natural-Language-Processing/68-Urheen-A-Chinese/English-Lexical-Analysis-Toolkit/View-details.html

Daelemans, 2012) which includes simple statistical taggers of varying quality for English, Spanish, German, French, Italian, and Dutch.



source columns:title, text, content

Language: English

added to column name: _lem

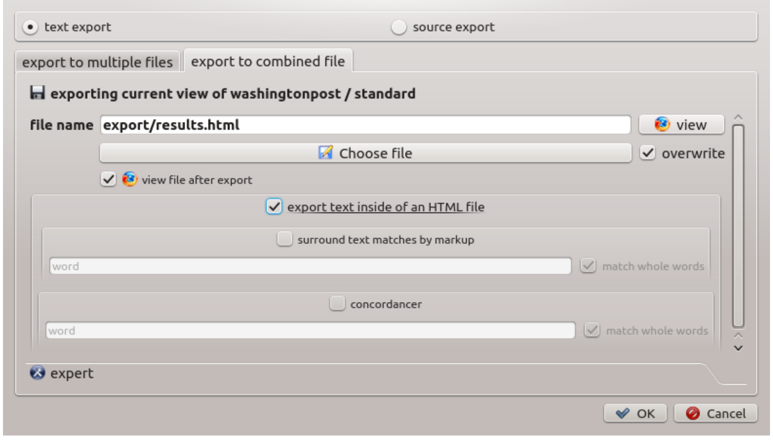
● lemma ○ tag ○ lemma-tag ○ complete

'standard' is the input view

A simple replacement option relying on regular expressions allows to make changes and corrections across the whole corpus.

2.4. Exporting the texts

Once the collection, cleaning, analysis, and correction process is finished and at least one of the column contains desirable linguistic data, there are different ways of exporting the column(s) for further inspection or use in other tools: After having selected the desired columns of the table in the main window, the export dialog offers two basic options: export to multiple files or to a unique combined file. The multiple file export provides the choice of including various separations between columns and naming options. The combined file export can just combine the data in a unique text file, but it can also produce html files with highlighted words as well as tables in a concordancer style.



● text export ○ source export

exporting current view of washingtonpost / standard

file name export/results.html

☑ overwrite

☑ view file after export

☑ export text inside of an HTML file

□ surround text matches by markup

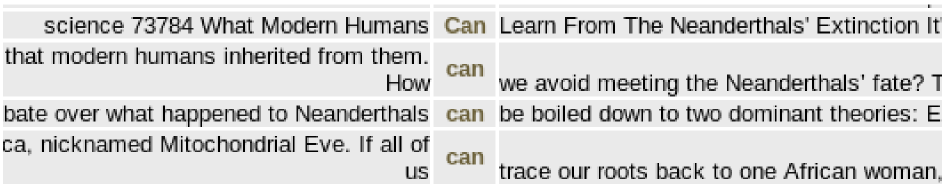
word ☑ match whole words

□ concordancer

word ☑ match whole words

[3] http://technology.chtsai.org/mmseg/

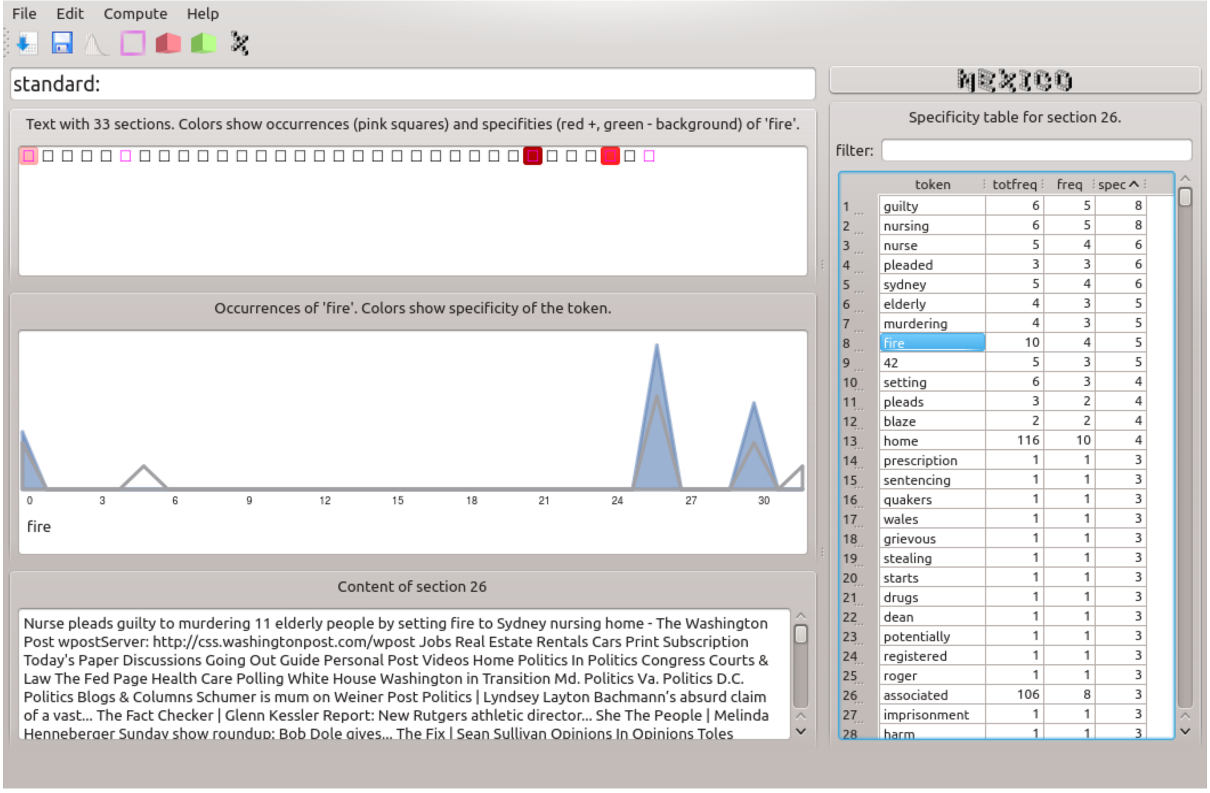
[4] Using Clips' Pattern tool: http://www.clips.ua.ac.be/pages/pattern



science 73784 What Modern Humans	Can	Learn From The Neanderthals' Extinction It
that modern humans inherited from them. How	can	we avoid meeting the Neanderthals' fate? T
bate over what happened to Neanderthals	can	be boiled down to two dominant theories: E
ca, nicknamed Mitochondrial Eve. If all of us	can	trace our roots back to one African woman,

2.5. *Analyzing the textual data*

The Gromoteur also includes a simple statistical tool named Nexico as it resembles the Lexico 3 tool (Fleury and Salem 2002) in its use of a section map, which allows for an easy overview of the collected data. For each section (i.e. each web page or local file), the specific words are computed, using the cumulative hypergeometric distribution of the words of the section compared to the complete corpus, i.e. an exact Fisher test is applied for each word in each section. As Pedersen 1996 remarks, the computation of the hypergeometric distribution is prohibitively heavy and, at the time, was only applicable to small corpora. He shows that the results (at least for the dependent bigrams he's working on) are more reliable than mutual information on sparse phenomena like collocations. The use of the hypergeometric distribution itself instead of some approximating function allows us to avoid the term "stop words" altogether: if a functional token (like a determiner or an auxiliary verb) appears in the results it is not due to its hight global frequency but rather because it really has a highly improbable frequency in the analyzed section if the words were distributed arbitrarily.



standard:

Text with 33 sections. Colors show occurrences (pink squares) and specifities (red +, green - background) of 'fire'.

Occurrences of 'fire'. Colors show specificity of the token.

Content of section 26

Nurse pleads guilty to murdering 11 elderly people by setting fire to Sydney nursing home - The Washington Post wpostServer: http://css.washingtonpost.com/wpost Jobs Real Estate Rentals Cars Print Subscription Today's Paper Discussions Going Out Guide Personal Post Videos Home Politics In Politics Congress Courts & Law The Fed Page Health Care Polling White House Washington in Transition Md. Politics Va. Politics D.C. Politics Blogs & Columns Schumer is mum on Weiner Post Politics | Lyndsey Layton Bachmann's absurd claim of a vast... The Fact Checker | Glenn Kessler Report: New Rutgers athletic director... She The People | Melinda Henneberger Sunday show roundup: Bob Dole gives... The Fix | Sean Sullivan Opinions In Opinions Toles

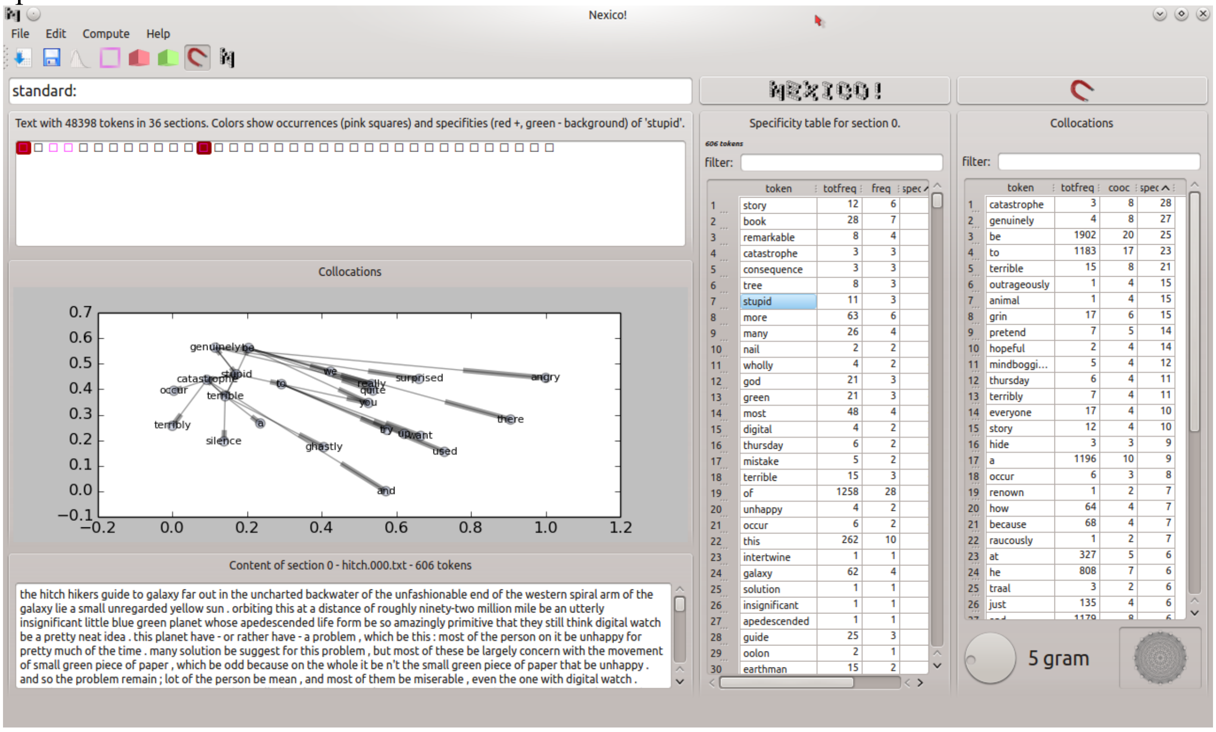
Specificity table for section 26.

filter:

	token	totfreq	freq	spec
1	guilty	6	5	8
2	nursing	6	5	8
3	nurse	5	4	6
4	pleaded	3	3	6
5	sydney	5	4	6
6	elderly	4	3	5
7	murdering	4	3	5
8	fire	10	4	5
9	42	5	3	5
10	setting	6	3	4
11	pleads	3	2	4
12	blaze	2	2	4
13	home	116	10	4
14	prescription	1	1	3
15	sentencing	1	1	3
16	quakers	1	1	3
17	wales	1	1	3
18	grievous	1	1	3
19	stealing	1	1	3
20	starts	1	1	3
21	drugs	1	1	3
22	dean	1	1	3
23	potentially	1	1	3
24	registered	1	1	3
25	roger	1	1	3
26	associated	106	8	3
27	imprisonment	1	1	3
28	harm	1	1	3

For our implementation, we developed a very effective computation of the cumulative hypergeometric distribution, with configurable thresholds. Contrary to the rest of Gromoteur, this function had to be written in C for performance reasons.

The specificity table shown on the right of the screenshot above shows the specificity of all the words in the selected sections compared to the rest of the corpus. Specificities are the logarithmic order of the p-value of the exact Fisher test. The positive specificity of “8” for “guilty”, for example, indicates that the surplus of the form “guilty” has a probability of approximately 10^{-8}. Inversely, negative specificities show an under-representation of a form in the selected parts. Moreover, the statistical window shows a graphical representation of the occurrence and specificity of forms over the whole corpus. All tables and graphics can easily be exported in order to be used in word processing software or for further analysis in a spreadsheet.



standard:

Text with 48398 tokens in 36 sections. Colors show occurrences (pink squares) and specifities (red +, green - background) of 'stupid'.

Content of section 0 - hitch.000.txt - 606 tokens

the hitch hikers guide to galaxy far out in the uncharted backwater of the unfashionable end of the western spiral arm of the galaxy lie a small unregarded yellow sun . orbiting this at a distance of roughly ninety-two million mile be an utterly insignificant little blue green planet whose apedescended life form be so amazingly primitive that they still think digital watch be a pretty neat idea . this planet have - or rather have - a problem , which be this : most of the person on it be unhappy for pretty much of the time . many solution be suggest for this problem , but most of these be largely concern with the movement of small green piece of paper , which be odd because on the whole it be n't the small green piece of paper that be unhappy . and so the problem remain ; lot of the person be mean , and most of them be miserable , even the one with digital watch .

Specificity table for section 0.

606 tokens

filter:

	token	totfreq	freq	spec
1	story	12	6	
2	book	28	7	
3	remarkable	8	4	
4	catastrophe	3	3	
5	consequence	3	3	
6	tree	8	3	
7	stupid	11	3	
8	more	63	6	
9	many	26	4	
10	nail	2	2	
11	wholly	4	2	
12	god	21	3	
13	green	21	3	
14	most	48	4	
15	digital	4	2	
16	thursday	6	2	
17	mistake	5	2	
18	terrible	15	3	
19	of	1258	28	
20	unhappy	4	2	
21	occur	6	2	
22	this	262	10	
23	intertwine	1	1	
24	galaxy	62	4	
25	solution	1	1	
26	insignificant	1	1	
27	apedescended	1	1	
28	guide	25	3	
29	oolon	2	1	
30	earthman	15	2	

Collocations

filter:

	token	totfreq	cooc	spec
1	catastrophe	3	8	28
2	genuinely	4	8	27
3	be	1902	20	25
4	to	1183	17	23
5	terrible	15	8	21
6	outrageously	1	4	15
7	animal	1	4	15
8	grin	17	6	15
9	pretend	7	5	14
10	hopeful	2	4	14
11	mindboggi...	5	4	12
12	thursday	6	4	11
13	terribly	7	4	11
14	everyone	17	4	10
15	story	12	4	10
16	hide	3	3	9
17	a	1196	10	9
18	occur	6	3	8
19	renown	1	2	7
20	how	64	4	7
21	because	68	4	7
22	raucously	1	2	7
23	at	327	5	6
24	he	808	7	6
25	traal	3	2	6
26	just	135	4	6
27	and	1179	8	6

The window of collocation analysis by default is the whole text (webpage, import document, ..), but the window can be reduced to any kind of n-gram for the retrieval of collocations of a more syntactic nature. The same approach based on the hypergeometric distribution avoids again the usage of stopwords, contrarily to, for example, the computation of collocations based on mutual information. In the above screenshot the right most table shows the statistical collocates of the word “stupid”, which contains “be” and “to” because they genuinely are collocates of “stupid” also in a linguistic sense (the function verb and the governed preposition in construction such as “Is is stupid to ...” or “He was so stupid to ...”). The collocation graph shows the strength of the collocative relations and allows for an easy explorations of the collocative relations by “clicking” through the graph.

3. General aspects, comparison, limitations

The Gromoteur is programmed in python and QT using some modules written in C. All data is stored in Sqlite databases. This allows for easy scaling on tens of thousands of webpage but the Gromoteur is essentially geared towards small size corpora consisting of a few hundred pages. The tool has already been used in various projects even before its recent completion including various newspaper analyses and the construction of a German-Chinese corpus from a bilingual website.

Most other crawlers are either command line tools built around the *wget* command that need hand-crafted scripting and tweaking or are web-based tools that have to be installed on a server, like for example the Babouk crawler (de Groc 2011). To other knowledge, no other desktop tool exists for crawling mid-size corpora that is geared toward linguists and sufficiently easy to install and to use. Webcorp Live (Kehoe & Renouf 2002) is a Web-based tool that allows to use the web as a concordancer as the results taken from Google are presented in a more linguist-friendly manner. However, Webcorp is not a crawler as the whole web pages cannot be downloaded for further use. This also excludes statistical analysis of the results or further linguistic transformations like tagging or lemmatization. The Webcorp Linguist's Search Engine gives access to a set of preconfigured and preanalyzed English language webdata. In this sense, Gromoteur is a very different tool, as it collects the data locally and gives complete freedom to the user which data to collect and how to transform it.

Gromoteur is distributed under a GPL license in a Linux and a Windows version on http://gromoteur.ilpga.fr. The source code is available on https://launchpad.net/grosmoteur

However, the desktop architecture imposes certain limitations: The crawling cannot be distributed on different machines with various IP addresses, which makes it more vulnerable to server-side exclusions from downloading. Work in progress attempts to make the heuristic tools of Gromoteur accessible independently of the graphical interface and thus scriptable.

References

Fleury, S., and A. Salem (2002), *Lexico 3. Outil de statistique textuelle.* Paris: Université Sorbonne Nouvelle-Paris 3

Gerdes, K., & Samvelian, P. (2008), A Statistical Approach to Persian Light Verb Constructions., *Proceedings of the Colloque Lexique et grammaire*, L'Aquila.

de Groc, C. (2011), Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Lyon, France.

Kehoe, A. & A. Renouf (2002), “WebCorp: Applying the Web to Linguistics and Linguistics to the Web”. *WWW2002 Conference.*

Pedersen T. (1996), « Fishing for exactness », in *Proceedings of the South-Central SAS Users Group Conference*, Austin, TX. www.d.umn.edu/~tpederse/Pubs/scsug96.pdf

De Smedt, T. & Daelemans, W. (2012), “Pattern for Python”. *Journal of Machine Learning Research*, 13: 2031–2035.