

Documentation of CONTAWORDS¹



With ContaWords you can quickly and easily get a frequency analysis of your texts (pdf, html, txt). The results can be downloading and easily handled.

How it works?

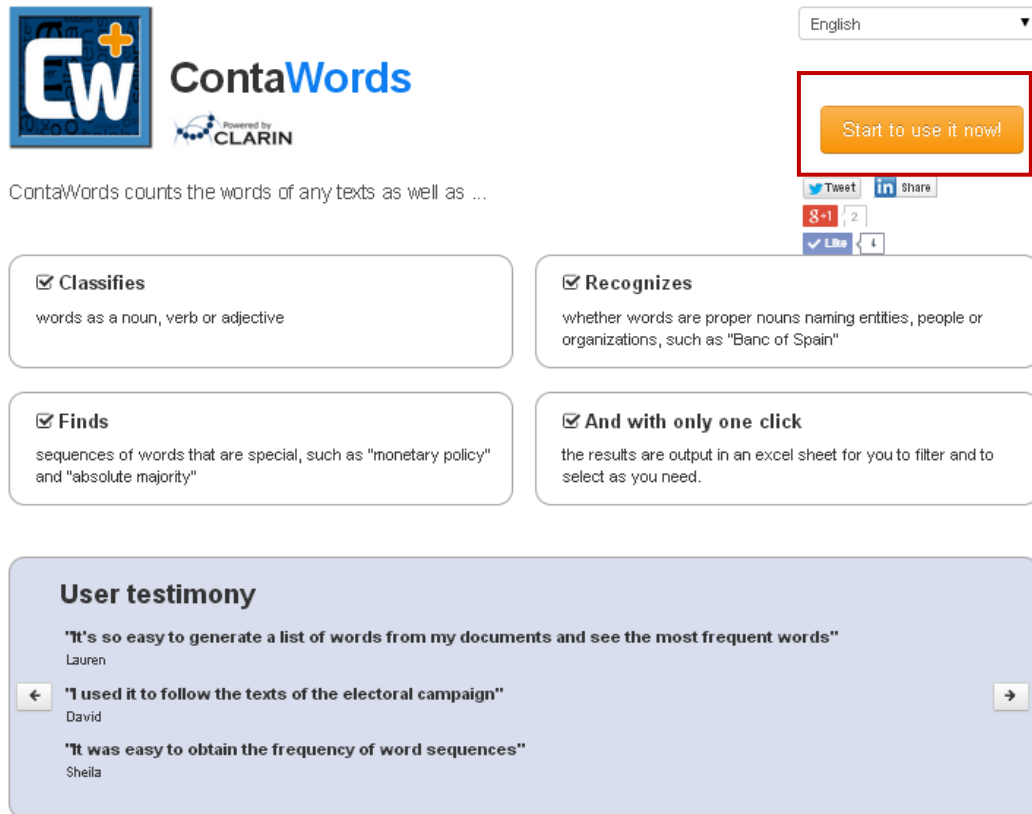
When you give ContaWords a task, it first reads the words of a text file and decides what part of speech to assign to each word (credit-Noun-credit but credit-Verb-to_credit). It then begins to count how many times a word appears in the text in every possible way (credits, credit, credited... etc).

ContaWords can also take into account that a sequence of words can, for example, be the name of an organization, a person or an entity (Named Entity Recognition).

Besides all this, ContaWords can find sequences of two words that seem to appear together more often than would be expected, such as "monetary policy " or "fiscal balance ". Maybe it can be useful for your study, too!

¹ <http://contawords.iula.upf.edu/>

Using Contawords



The image shows the ContaWords web application interface. At the top left is the ContaWords logo, which includes a stylized 'CW' with a plus sign and the text 'ContaWords'. Below the logo is the text 'Powered by CLARIN'. To the right of the logo is a language dropdown menu set to 'English'. Below the language menu is a red-bordered box containing an orange button that says 'Start to use it now!'. Below this button are social media sharing icons for Twitter, LinkedIn, Google+, and Facebook. Below the social media icons is a section with four features, each in a rounded rectangle:

- Classifies**: words as a noun, verb or adjective
- Recognizes**: whether words are proper nouns naming entities, people or organizations, such as "Banc of Spain"
- Finds**: sequences of words that are special, such as "monetary policy" and "absolute majority"
- And with only one click**: the results are output in an excel sheet for you to filter and to select as you need.

Below the features section is a 'User testimony' section with a light blue background. It contains three testimonials, each with a left arrow, a quote, a name, and a right arrow:

- "It's so easy to generate a list of words from my documents and see the most frequent words"** by Lauren
- "I used it to follow the texts of the electoral campaign"** by David
- "It was easy to obtain the frequency of word sequences"** by Sheila

At the bottom of the page are four logos: 'upf. Universitat Pompeu Fabra Barcelona', 'gruptr', 'Unión Europea', and 'Generalitat de Catalunya Departament d'Economia i Coneixement'.

You can run multiple ContaWords simultaneously: when ContaWords starts to count, the status "In Progress" will appear in "Results" and when the task is complete the status will change to "Finished".



ContaWords

Powered by
CLARIN

English



Count!

More info.

FAQ

Credits

File

input_extract_chapter3_CharlesDarwin_SouthAmericanGeology.txt

input_extract_chapter1_CharlesDarwin_SouthAmericanGeology.txt

You can copy and select more
than one URL

ContaWords will count the
different files as one single text.

Upload more files

Workspace

External files

Tell us the documents language...:)

Run now!

Results

Started	Ended	Status	Inputs	Results
25/06/2014 12:54		Counting...	English 1. input_extract_chapter1_CharlesDarwin_SouthAmericanGeology.txt? 1403700889	
25/06/2014 12:53	25/06/2014 12:54	Done	English 1. input_extract_chapter3_CharlesDarwin_SouthAmericanGeology.txt? 1403700691	1. Download

When the status is "In Progress", you can already copy the URLs of new texts and click to run ContaWords again.

For each task that you give to ContaWords, it puts all of your results into a single spreadsheet file (Excel file; .xls) that you can download when all of the tasks are finished.

Results

Started	Ended	Status	Inputs	Results
25/06/2014 12:54	25/06/2014 12:55	Done	English 1. input_extract_chapter1_CharlesDarwin_SouthAmericanGeology.txt? 1403700889	1. Download

In each tab of the Excel file, you will find the following information:

- A- A general overview;
- B- Frequency counts for each individual part-of-speech: nouns, adjectives and verbs;
- C- Entities;
- D- Absolute frequency count for combinations of 2 words that ContaWords has deemed “significant”;
- E - Absolute frequency count for different lemmas (all parts of speech)
- F- Absolute frequency count for words by form (all parts of speech)
- G- Absolute frequency count for two word sequences by lemma

The screenshot shows an Excel spreadsheet with the following content:

	A1	Statistics
1	STATISTICS	
2		
3	word	1153 tokens, 416 types
4	lemma	1153 tokens, 371 types
5	pos	1153 tokens, 40 types
6		
7	CONTENT INDEX	
8		
9	TABS:	
10		
11	A	INDEX, statistics
12		
13	B	Nouns, adjectives and verbs frequencies
14		
15	C	Named Entities
16		
17	D	Absolute lemmas frequencies (all grammatical categories included)
18		
19	E	Absolute word frequencies by form (all grammatical categories included)
20		
21	F	Absolute frequency of two words sequences (bigrams)
22		
23	G	Absolute frequency of specially selected two word sequences
24		
25		
26		
27	PART OF SPEECH TAGS:	
28		
29	A	Adjective
30	B	Adverb

At the bottom of the spreadsheet, there is a tabbed interface with the following tabs: A-summary, B-adjectives, B-nouns, B-verbs, C-named_entities, D-freq-lemma, E-freq-word, F-lemma-bigrams, and G-freq-lemma. The 'A-summary' tab is currently selected and highlighted with a red border.

The above is all done with ContaWords. Now it's your turn! You can sort the words to see the most frequent nouns, or look to see if there are many negative adjectives, or ... you can discover new uses for the information provided by ContaWords.



This document is licensed under a Creative Commons Attribution 3.0 Spain License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/es/>.

Please send feedback and questions on this document to: iulatri@upf.edu

TRL Group (Tecnologies dels Recursos Lingüístics), Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (IULA-UPF)