Manual for the CRPC on the CQPweb interface

Manual 1.3

Version May, 2014

Amália Mendes, Michel Généreux, Iris Hendrickx

Centro de Linguística da Universidade de Lisboa
Complexo Interdisciplinar
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal

Table of Contents

Manual for the CRPC on the CQPweb interface 1
1. Corpus Queries 3
1.1 Concordances of word forms 3
1.2 Regular expressions 4
1.3 Part-of-speech tags 4
1.3.1 Auxiliary Verbs 6
1.3.2 Past Participles 7
1.3.3 Noun 7
1.3.4 Articles 8
1.4 Inflection tags (version CRPC POS fine-grained) 8
1.4.1 Annotation of Nouns and other categories 8
1.4.2 Annotation of Verbs 9
1.4.3 Queries for inflection tags 10
1.5 Lemmas 12
1.6 Word sequences 13
1.7 Contracted elements (*no*, *naquele*, *do*, etc…) 13
1.8 Sorting concordances 14
1.9 Sentences and Noun Phrases 14
1.10 Collocations 15
2. Main Left menu 15
2.1 Corpus Queries 15
2.2 User controls (registered version) 16

Preamble

The CRPC is a corpus of contemporary Portuguese which was automatically cleaned, part-of-speech tagged and lemmatized. In the current version of CRPC, version 2.0, 2010, the written part of the corpus that is available on CQPweb contains 309 million words. More information about the CRPC can be found here: http://www.clul.ul.pt/en/resources/183-reference-corpus-of-contemporary-portuguese-crpc

The corpus is available online at the following URL:
http://alfclul.clul.ul.pt/CQPweb/

This manual explains how to use the interface to query the CRPC. The query language (*Simple Query Syntax)* is almost the same as for the BNCweb which is described in detail in Chapter 6 of Hoffmann, Sebastian et al. (2008), *Corpus Linguistics with BNCweb - a Practical Guide*. Frankfurt/Main: Peter Lang.

1. Corpus Queries

1.1 Concordances of word forms

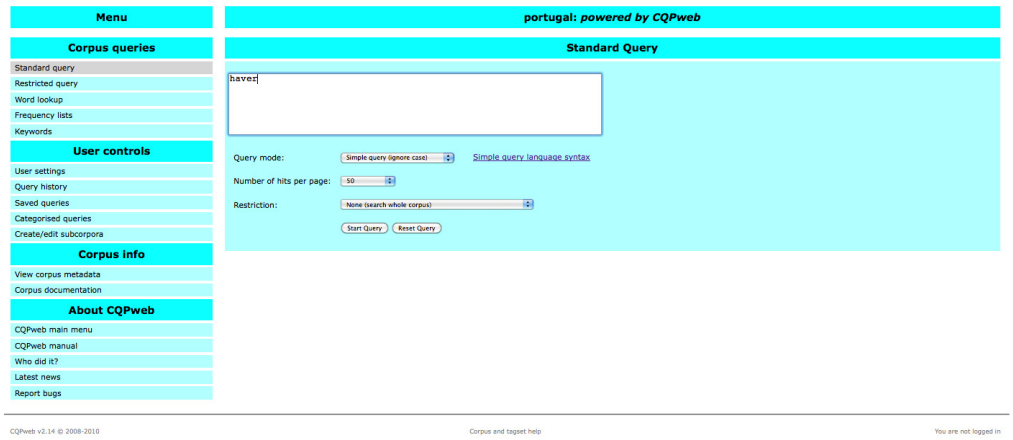
To conduct a simple query, go to the main page, without changing any option, insert a word or a sequence of words in the box and click on *Start Query*. At the top of the window with the results, there is information on the number of contexts, the number of texts in which the query occurs and information about the total corpus.

To make a new search, click GO (top right button).

To view information on the full text that matched the query for a particular concordance, click on the name in the left column "Filename". Any user can download its concordances by selecting “Download” on the drop-down list located at the top-right corner and click Go!.

To view a larger context of a particular concordance, click on the bold words on the intended line. You will see the words in a context of a few lines. In the top menu there is the possibility to enlarge the context, click on "More context". You can see the part-of-speech tags by clicking "Show tags".

WARNING: very common grammatical words such as "que " "de", "o ", can be queried, but the lookup takes time due to very high frequency of these forms and the large size of the CRPC.



1.2 Regular expressions

A regular expression is a way of characterizing a string, you can view it as a pattern or a template in which you use wildcards to leave certain characters unspecified.

Wildcards	*example*	*matches with*
? a single arbitrary character	gat?	gato, gata
* zero or more characters	*mente	mente, absolutamente, provavelmente, etc.
+ one or more characters	+mente	absolutamente, provavelmente, etc. (but not: *mente*)

Simple Query Syntax uses a set of characters as meta-characters:

? * + [ ] ( ) { } , : @ / _ - < >

To query for the literal meaning of these characters, use a backslash in front.
E.g. to look for a question mark, type: *\?*

Query description	*example*	*matches with*
Alternatives: between square brackets	lind[o,a]	lindo, linda
Two alternatives followed by exactly 1 character	lind[o,a]?	lindos, lindas
Two alternatives followed by: 's' or nothing	lind[o,a][s,]	lindo, linda, lindos, lindas
Two alternatives followed by zero or more characters	lind[o,a]*	lindo, lindos, lindamente, lindinho, lindoso, lindano, etc.

1.3 Part-of-speech tags

You can search for ~~a~~ part-of-speech (POS) tag, or a combination of a word with a POS tag using '_'.

Notice that the tags have been assigned automatically and the corpus may contain errors.

Query description	*example*	*matches with*
the word form “desse” tagged as verb	desse_V	desse (verb)
words tagged as Indefinite	_IND	algo, tudo, nada, etc
Word string followed by zero or more character combined with POS	ante*_V	antecipar, antedatar, etc

The tag set follows the CINTIL tags (Barcelar et al, 2006) but with some modifications: multi-word units do not receive special POS tags as is the case in CINTIL (except a small list of latin expressions), and contracted forms (*pelo*, *do*) are kept and receive a double tag (pelo/PREP+DET), while in CINTIL these words are split into two separate tokens.

Tag	Category	Examples
ADJ	Adjectives	bom, brilhante, eficaz, …
ADV	Adverbs	hoje, já, sim, felizmente, …
CARD	Cardinals	zero, dez, cem, mil, …
CJ	Conjunctions	e, ou, tal como, …
CL	Clitics	o, lhe, se, …
CN	Common Nouns	computador, cidade, ideia, …
DA	Definite Articles	o, os, …
DEM	Demonstratives	este, esses, aquele, …
DFR	Denominators of Fractions	meio, terço, décimo, %, …
DGTR	Roman Numerals	VI, LX, MMIII, MCMXCIX, …
DGT	Digits	0, 1, 42, 12345, 67890, …
DM	Discourse Marker	olá, pois, então, pronto, …
EADR	Electronic Addresses	http://www.di.fc.ul.pt, …
EOE	End of Enumeration	etc
EXC	Exclamatives	que, quanto, ...
GER	Gerunds	sendo, afirmando, vivendo, …
GERAUX	Gerunds as auxiliary verbs	tendo, havendo
IA	Indefinite Articles	uns, umas, …
IND	Indefinites	tudo, alguém, ninguém …
INF	Infinitive	ser, afirmar, viver, …
INFAUX	Infinitive auxiliary verb	ter, havermos, ...
INT	Interrogatives	quem, como, quando, …
ITJ	Interjection	bolas, caramba, …
LTR	Letters	a, b, c, …
LADV1…LADVn	Latin Multi-Word Adverbs	a priori, a posteriori, per capita, sine die
MGT	Magnitude Classes	unidade, dezena, dúzia, resma, …
MTH	Months	Janeiro, Dezembro, …
ORD	Ordinals	primeiro, centésimo, penúltimo, …
PADR	Part of Address	Rua, av., rot., …
PNM	Part of Name	Lisboa, António, João…
PNT	Punctuation Marks	., ?, (, …
POSS	Possessives	meu, teu, seu, …
PPA	Past Participles not in compound tenses	sido, afirmados, vivida, …
PPT	Past Participle in compound tenses	sido, afirmado, vivido, …
PREP	Prepositions	de, para, em redor de, …

PRS	Personals	eu, tu, ele, …
QNT	Quantifiers	todos, muitos, nenhum, …
REL	Relatives	que, cujo, tal que, …
STT	Social Titles	Presidente, dr., prof., …
SYB	Symbols	@, #, &, …
TERMN	Optional Terminations	(s), (as), …
UM	"um" or "uma"	um, uma
UNIT	Measurement units in abbreviated form	Kg, h, seg, Hz, Mbytes,...
VAUX	Finite "ter" or "haver" in compound tenses	temos, haviam, …
V	Verbs (other than PPA, PPT, INF or GER)	falou, falaria, …
WD	Week Days	segunda, terça-feira, sábado, …
Contracted forms	**Combinations of :**	
CL+CL	Two clitics	-lha, lhos, -ma, ma, -tas, ...
PREP+ADV	Preposition and Adverb	dali, daì, daqui, ...
PREP+DA	Preposition and Definite Articles	aos, na, nos, da, dos
PREP+DEM	Preposition and Demonstratives	desse,deste, naquela
PREP+IND	Preposition and Indefinite	noutra, noutros, doutra, ...
PREP+INT	Preposition and Interrogative	aonde
PREP+PRS	Preposition and Personal pronoun	comigo, conosco, dela, nele, ...
PREP+QNT	Preposition and Quantifier	nalguns, noutro,noutras, ...
PREP+REL	Preposition and Relative	donde, aonde
PREP+UM	Preposition and "um" or "uma"	dum, duma

1.3.1 Auxiliary Verbs

Only occurrences of the verbs “ter” and “haver” in compound tenses are tagged as auxiliary verbs.

They may receive three tags:

_VAUX if the auxiliary verb is in a finite tense (*tinha* feito)
_INFAUX if the auxiliary verb is in the infinitive form (*ter* feito)
_GERAUX if the auxiliary verb is in the gerund form (*tendo* feito)

Examples:

O que não *tem* existido é a oferta de soluções alternativas (…).
(…) a capa negra de estudante que o bisavô e o próprio avô *haviam* usado nos seus tempos de estudante.

1.3.2 Past Participles

The past participle verb forms can receive one of two tags: PPT or PPA.

PPT

The PPT tag identifies exclusively past participle verb forms in compound tenses, with auxiliary verbs "ter" and "haver".

Examples

O que não tem *existido* é a oferta de soluções alternativas (…).
(…) a capa negra de estudante que o bisavô e o próprio avô haviam *usado* nos seus tempos de estudante.
Vendo que o Rafa já tinha *cumprido* o seu dever, felicitou-o efusivamente (…).

PPA

In all the remaining contexts (i.e., besides compound tenses) in which a past participle can occur the tag PPA is applied.
Some contexts are ambiguous between the PPA tag and the ADJ (adjective) tag.
We consider that the tag PPA applies when: (i) the past participle form occurs in a passive sentence; (ii) it is possible to establish a relation with a transitive verbal construction; and (iii) the past participle form occurs with a semi-auxiliary verb or a negation element.
The tag ADJ applies when it is not possible to establish a relation with a transitive verbal construction or, being it possible, the verb in a transitive construction has a different meaning than the past participle.
It should also be noted that it is possible to have coordinated structures with an adjective and a past participle (ex. Um copo bonito [ADJ] e partido [PPA]).

Examples

Past participles (PPA)
isto não se pode ser *resolvido* assim
(…) gostam da cidade como ela é e não queriam que ela fosse muito *adulterada* (…)
(…) e assim meio rústicas mas depois *pintadas* por nós (…)
(…) vai à esquadra entregar a tal nota *escrita* pelo punho do pai (…)
Estou *deprimida*.
Cheguei *cansado*.

Adjectives (ADJ)
As cidades do alentejo são muito *fechadas* (…)
(…) uns videozinhos mas em boneco *animado* (…)
(…) e é muito *complicado* para nós (…)
(…) é um biólogo e muito *interessado* nos problemas da vida (…)
(…) é a zona que fica mais bem *colocada*.

1.3.3 Noun

Proper names

The category "Proper names" includes anthroponyms, toponyms, titles of artistic works (literary works, songs, paintings, etc.), institutions, addresses, acronyms, and siglas.

Note that in cases of multiword proper names, only the words from open classes are tagged with PNM. Other words, such as prepositions, conjunctions, etc., are tagged according to the category that they belong to.

Examples

Diário_PNM
De_PREP
Notícias_PNM

MINISTÉRIO_PNM
de_PREP
a_DA
CULTURA_PNM

1.3.4 Articles

Three tags are used in the annotation of articles: one for definite articles (DA), one for plural indefinite articles (IA), and one exclusively for the singular forms of the indefinite articles (UM), i.e., "um" and "uma".

Tag	Category	Examples
DA	Definite Articles	o, os, a, as
IA	Indefinite Articles	uns, umas
UM	Indefinite Articles – singular form	um, uma

1.4 Inflection tags (version CRPC POS fine-grained)

1.4.1 Annotation of Nouns and other categories

Category	Value	Tag	Examples
gender	masculine	m	gatos_CN#mp
	feminine	f	cadeira_CN#fs
	not applicable	g	context doesn't provide information on gender: apoio e segurança a banhistas_CN#gp
number	singular	s	mesa_CN#fs
	plural	p	livros_CN#mp
person	first person	1	eu_PRS#ms1, eu_PRS#fs1
	second person	2	tu_PRS#ms2, tu_PRS#fs2
	third person	3	ela_PRS#fs3, eles_PRS#mp3
diminutive		-dim	mesinha/MESA/CN#fs-dim
augmentative		-sup	facão/FACA/CN#ms-sup
superlative		-sup	normalíssimo/NORMAL/ADJ#ms-sup o maior/GRANDE/ADJ#ms-sup
comparative		-comp	eles são maiores_ADJ#mp-comp do que

Notice that the tags have been assigned automatically and the corpus may contain errors.

- The gender of invariable nouns and adjectives is determined by the context. If an adjective has only one form for the masculine and feminine (ex: grande), the gender value is marked according to the gender of the entity that the adjective modifies: in the context "casa grande", the adjective will be marked as feminine, while in the context "prédio grande" the adjective will be marked as masculine. The same applies to pronouns that do not show gender marks: "tu" will be either feminine or masculine according to the context. If it is not possible to determine the gender of a noun, adjective or pronoun, the tag "g" should be applied to indicate unknown gender (estudantes_CN#gp).

- The clitic "se" doesn't have features of gender. It is annotated as "se_CL#gs3".

¨ Denominators of Fractions have gender and number information, except symbol "*%*".

1.4.2 Annotation of Verbs

The verb forms may encode features of tense, mood, person, number and, in the case of the PPA tag, gender.

Tags for tense and mood:

Tense/Mood	Tag
Present – Indicative	pi
"Pretérito Perfeito" - Indicative	ppi
"Pretérito Imperfeito" – Indicative	ii
"Pretérito Mais que Perfeito" - Indicative	mpi
Future – Indicative	fi
Conditional	c
Present – Subjunctive	pc
"Pretérito Imperfeito" – Subjunctive	ic
Future – Subjunctive	fc
Affirmative imperatives[1]	imp impaf
Non inflected infinitives	ninf
Undetermined infinitives (context does not provide enough information to decide whether it is inflected or non inflected)	nef

Infinitives:

Non-inflected infinitives are tagged as _INF#ninf.
However, they are annotated as _INF with no subtag in some contexts of *partir*, *seguir*, *ser*, *calhar*, *ver*, *pôr*, *actualizar*, *deixar*, *ficar* and *passar*. This will be normalized in a future version.

Inflected infinitives are tagged with person and number tags (ex: _INF#1p). They are in some

[1] We recommend using both tags when querying for affirmative imperatives to make sure to retrieve all cases. Negative imperatives are tagged as a form of the subjunctive.

cases tagged with the subtag f- preceding person and number (ex: _INF#f-1p). This will be corrected in a future version.

Tags for person, number and gender:

Category	Value	Tag	Example
person	first person	1	quero_V#pi-1s mostraram_V#mpi-3p vires_INF#2s experimenta_V#imp-2s pensariam_V#c-3p
	second person	2	
	third person	3	
number	singular	s	
	plural	p	
gender	masculine	m	interessado_PPA#ms
	feminine	f	abandonadas_PPA#fp
	not applicable	g	entregue_PPA#gs

1.4.3 Queries for inflection tags

Inflection tags are encoded after the main category tag, separated by #.

The diminutive, augmentative, superlative and comparative tags are preceded by a hyphen.

The tense and mood tags are followed by a hyphen (if more tags occur).

Examples of tag order with Nouns and other categories:

word	main tag	gender	number	diminutive augmentative superlative comparative --------- person	Example
gatinho	_CN#	m	s	-dim	gatinho_CN#ms-dim
cadeirinhas	_CN#	f	p	-dim	cadeirinhas_CN#fp-dim
facão	_CN#	m	s	-sup	facão_CN#ms-sup
normalíssima	_CN#	f	s	-sup	normalíssima_ADJ#fs-sup
maiores	_ADJ#	m	p	-comp	eles são maiores_ADJ#mp-comp do que
ela	_PRS#	f	s	3	ela_PRS#fs3

Examples of tag order with Verbs and other categories:

word	**main tag**	**tense/mood**	**person** **1, 2, 3**	**number** **m s**	**Example**
cantaste	_V#	pi- …	2	s	cantaste_V#pii-2s
tem	_VAUX#	pi- …	3	s	tem_VAUX#pi-3s feito
cantar	_INF#	ninf			cantar_INF#ninf
fecharmos	_INF#		1	s	fecharmos_INF#1p
ter	_INFAUX#		3	s	ter_INFAUX#3s bebido
sorrindo	_GER				sorrindo_GER
tendo	_GERAUX				tendo_GERAUX sido
word	**main tag**	**gender**	**number**		
feito	_PPT				tinha feito_PPT
abandonadas	_PPA#	m f g			abandonadas_PPA#fp

Queries over POS and Inflection tags:

Query description	example	matches with
common nouns in the diminutive form	*_CN*dim	casinha, remoinho, caldinho, mulherzinha, …
plural common nouns in the diminutive form	*_CN#?p-dim	carrinhos, fitinhas, quadradinhos, …
masculine adjectives, in the superlative form	*_ADJ#m?-sup	maior, amicíssimo, aflitíssimo, belíssimo, …
personal pronouns, first person singular, followed by a verb	*_PRS#?s1 *_V*	eu gostava, eu considero. …
verbs in the present tense of the indicative	*_V#pi*	quero, traz, tem, vais, …
verbs in the present tense of the indicative, second person plural	*_V#pi-2p	vedes, estais, sois, pensais, …
auxiliary forms, not infinitive nor gerund	_*AUX*	teria, tinham, tenha, …
inflected infinitives	_INF#*[1,3]?	levarem, terem, para se distrair,
inflected infinitives followed by a personal pronoun	_INF#*[1,3]? *_PRS*	sermos nós, contratarem elas, defendermos nós, estar eu,

1.5 Lemmas

You can search for a lemma or root form of a word by using curly brackets. You can combine your search with POS tags.

Nominal lemmas: the lemma is the masculine singular form, if it exists. If not, it is the masculine plural form, or else the feminine singular form or else the form itself.

Verbal lemmas: the lemma is the infinitive form.

Query description	example	matches with
lemma	{poder}	poder, posso, podes, podia, etc
lemma with POS tag	{poder}_CN	poder, poderes
Word string followed by zero or more character combined with POS	ante*_V	anteceder, antecipar, antedatar, etc
lemma “ler” in the “pretérito pefeito” tense of the indicative	{ler}_V#ppi*	li, leste, leu, lemos, lestes, leram

Special cases:

- Inflected words from certain classes are annotated with a lemma in the masculine form and a lemma in the feminine form.
This applies to: Definite Articles (DA), Indefinite Articles (IA and UM), Personal Pronouns (PRS), Clitics (CL), Demonstratives (DEM), Possessives (POS), Quantifiers (QNT), Relatives (REL):

Query description	example	matches with
lemma “teu”	{teu}	teu, teus
lemma “tua”	{tua}	tua, tuas
lemma “a”, definite article	{a}_DA	a, as (definite articles)
lemma “a”, clitic pronoun	{a}_CL	a, as (clitics)
lemma “cujo”	{cujo}	cujo, cujos
lemma “cuja”	{cuja}	cuja, cujas

- The lemma of the past participle forms in compound tenses (PPT) is the infinitive. Past participles that do not form compound tenses (PPA) are lemmatized with both infinitive and past participle verb form. For instance, “apresentadas”, if not in compound tenses, is lemmatized as “apresentar,apresentado”.

Query description	example	matches with
lemma	{apresentar}	verb forms of the verb *apresentar*, including past participles that occur in compound tenses (PPT), e.g: *tinham apresentado*
lemma	{apresentar\,apresentado}	occurrences of the past participle forms of *apresentar* that do not occur in compound tenses (PPA)

- Words with suffixes
Diminutives *-inho*. *-zinho*, *-ito*, and *-zito*; regular superlatives (ending in *-íssimo*), augmentatives: the lemma is the regular adjectival form, in the masculine singular.
Irregular comparatives: lemma is the form itself, in the masculine singular

Examples:
“lindíssima” -> lemma “lindo”
“grandalhão” -> lemma “grande”
“maior” -> lemma “maior”
“maiores” -> lemma “maior”

- “Irregular” feminine forms
The lemma is the form itself (e.g. actriz, etc)

- Foreign words
The lemma of foreign words is the occurring form itself.

Notice that the lemma tags have also been assigned automatically and the corpus may contain errors.

1.6 Word sequences

You can also search for multiple words. Notice that:

- punctuation marks are split from words and are separate tokens
- special characters need a backslash
- you can combine + and * to define a sequence of arbitrary words in your query. E.g. the pattern +** represents a sequence of one to three tokens.

Query description	example	matches with
Adjective followed by the lemma of the noun 'jantar'	*_ADJ {jantar}_CN	célebre jantar, breve jantar, grandes jantares, bom jantar, etc.
The word 'se' followed by an optional word and a comma	se * \,	se trata, se, se vê, se calhar, etc.
The lemma 'célebre' followed by the lemma of the noun 'jantar'	{célebre} {jantar}_CN	célebre jantar, célebres jantares
The preposition ‘de’ followed by the lemma of the noun 'jantar', separated by a minimum of one and a maximum of three words	{de} +** {jantar]_CN	de estar presente num jantar, de fazer um jantar, de nosso jantar, etc

1.7 Contracted elements (*no*, *naquele*, *do*, etc…)

Contractions of two words are annotated with double POS tags and lemmas. For example "no" has POS-tag "PREP+DA" and lemma "em+o". Below are some examples of how to search for these particular words, the '+' character is a meta- character, therefore you need to use a backslash.

Query description	example	matches with
To search for a contracted form, use '\+*'	{em\+*}	no,nas, naquele, etc.
Contracted forms followed by 'o'	{em\+o}	no, nos
To find both contracted and separate forms, the '|' means "or".	({em\+}|{em})	em, no, nos, na, naquele, etc.

1.8 Sorting concordances

After searching for a word or expression, you can sort the concordances obtained: open the window New Query, select Sort and click on Go!
By default, the concordances are sorted alphabetically by the first word on the right. You can change this option in “Position” and then click on "Update sort".

1.9 Sentences and Noun Phrases

Version 2.2 of the CRPC has been tagged with Noun Phrases (NPs). You can query those NPs provided you use the CQP syntax. Here are a few examples:

All NPs: (this will take a very long time!)
/region[np];
<np> []* </np>;

NPs with exactly 3 words:
<np> []{3} </np>;

V at the start of a sentence:
<s> [pos = "V"];

V at the start of a sentence:
[(pos = "V") & lbound(s)];

V at the end of a sentence:
[pos = "V"] [pos = "PNT"]? </s>;

NP with at least 3 adjectives:
<np> []* ([pos="ADJ.*"] []*){3,} </np>;

Sentences that start and end with a NP:
<s><np>[]*</np> []* <np>[]*</np></s>;

CN that is not contained in a noun phrase:
[(pos = "CN") & !np];

Sequence of two singular nouns within the same NP:
[pos="CN"] []* [pos="CN"] within np;

1.10 Collocations

You can obtain additional collocation information for a retrieved word or lemma pattern from a standard or restricted query. Instead of choosing 'new query', choose 'collocations' from the menu drop box and click 'go'. Click on "Create collocation database" to get the list of words that co-occur with the retrieved word pattern.

On top, "Collocation controls", you can:
- change the statistical measure used (statistics: Mutual information, t-score, etc.) and compare the results
- change the distance between words (collocation window)
In "submit changed parameters", press Go!

By clicking on the frequency, you get the concordances in which the word you searched co-occurred.

Below is a screenshot for collocations for the word 'janela' in a search window of 3 words to the left and right using Log-likelihood as distance measure, and a frequency threshold of 5.

Collocation controls

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

There are 7,462 different words in your collocation database for "[word="janela"%c]". (Your query "janela" returned 5,283 matches in 2421 different texts) [34.006 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood value
1	pela	459,379	47.001	606	460	1991.248
2	uma	1,971,066	201.666	1,077	799	1882.461
3	ABERTA	22,282	2.28	238	206	1744.769
4	da	4,537,364	464.233	1,480	764	1433.505
5	à	10,957,196	1121.067	2,476	1075	1274.909
6	quarto	18,750	1.918	126	107	807.36
7	Abrir	15,686	1.605	99	89	622.026
8	Indiscreta	96	0.01	39	26	586.634
9	abriu	8,028	0.821	70	53	484.527
10	parapeito	158	0.016	35	31	475.353
11	vidro	3,971	0.406	58	52	460.675
12	para	2,629,825	269.066	666	504	418.273

2. Main Left menu

2.1 Corpus Queries

- **Standard query** See section 1 above about standard searches.
- **Restricted queries** This enables you to search in a particular sub set of the corpus. A query can be restricted to searching in documents from a particular country (Portugal, Brazil, Angola, Mozambique, etc.) or on the text genre which offers the following choices: correspond (letters), direito (legal documents), folheto (flyers) , jornal (newspapers), livro (books), politica (politics), revista (magazines), varia.
- **Word lookup** Use this option to get frequency information about a particular word. You can also use regular expressions or only specify the beginning or end of a word. When you click on a word in the result page, you will get a concordance list.
- **Frequency lists** Gives a list of all word forms or lemmas from the corpus and their frequency.
- **Key Words** This rather advanced option allows you to compare a query in a restricted sub corpus against the full corpus.

2.2 User controls (registered version)

User controls are only available for registered users (the green version). This means essentially that unregistered users (the blue version) cannot *save* data (settings, queries and sub-corpora) on our server. However, they can *download* their results and benefit from exactly the same searching power available to registered users.

User settings Various user-oriented options.

Query History Shows all previously entered queries.

Saved queries When making a standard or restricted query, results can be saved. These saved queries are listed here. Registered users should keep the number of saved queries to a useful minimum by using the delete function.

Categorized queries The set of concordances obtained through a regular or restricted query can be organized using a set of labels applied to each individual context.

- Select the option "categorize" on the top right menu and click Go!

- Enter a name for the set of categories. For example, if you want to label each sense of a highly polysemous verb like "abater" (*move downwards / eliminate / negatively affect*) the set of values could be named "abater" or "verbpolysemy".

- Enter the names for each category. For example, considering the different senses of “abater”, the set could be: movement, movement_pronominal, psych, psych_pronominal, affect, affect_eliminate, subtract, etc.

- select the default value (for example, if the verb has a more frequent sense)

- click on Submit

The set of concordances will appear with a new column named 'Category' on the right, with the set of values to select. Two categories are automatically added to the set you have created: 'other' and 'unclear'.
After selecting a value for each context, select "save values and leave categorisation mode".

The set of categorised concordances remains available on the left menu. There are two interesting options under User Controls:
- add categories
- separate categories: this creates a separate list of concordances for each category, with information on the number of hits of each.

Create/edit sub corpora You can create separate sub corpora based on several criteria such as using the meta data from the corpus or using the matches from a query. For example, you can create a sub corpus containing only Portuguese news texts:
- In “Define new subcorpus via”, select ‘corpus meta data’
- click Go!
- enter a name for this new subcorpus
- choose the text-type restrictions 'Portugal' and 'jornal'
- click on ‘Create subcorpus from selected categories’.

Next you can compile a frequency list for this sub corpus by clicking ‘Compile’ under Frequency Lists on the left Menu. This frequency list can be further inspected using the option " Frequency lists" in the main menu. Registered users should keep the number of saved subcorpora to a useful minimum by using the delete function.