ACORN
(the **A**ston **COR**pus **N**etwork**)**

Ramesh Krishnamurthy

ETAS 23rd AGM and Convention

January 2007

Solothurn, Switzerland

Corpora: first phase

- **A corpus is a large collection of texts stored in a computer, analysed using software tools, in order to show patterns of language use**
- **Corpus research began in the 1960s, and has developed alongside computer technology**
- **1960s-70s: 'small' corpora of English: word frequencies, concordances, mainly grammar studies**
- **Brown** (Francis & Kucera, US written, 1 million words)
- **LOB** (London-Oslo-Bergen, UK written, 1 million words)
- **OSTI** (Sinclair, UK spoken, 125,000 words)

Why bother with corpora?

- “Language users cannot accurately report language usage, even their own” (Sinclair, 1987)
- “There are many facts about language that cannot be discovered by just thinking about it, or even reading and listening very intently” (Sinclair, 1995)
- “Using a language is a skill that most people are not conscious of; they cannot examine it in detail, but simply use it to communicate” (Sinclair 1995)
- As language teachers and professionals, we often have strong intuitions about language use… *Corpus-based research, however, shows us that our intuitions are often completely wrong.* (Biber 2005)

Corpora: second phase

- **1980s-1990s: ‘larger’ corpora revolutionized language description and EFL dictionary compilation: Cobuild** (18 million words, written & spoken, UK & US), **Longman, Cambridge, Chambers-Harrap, Macmillan**
- Most frequent lexical words: ‘*time*’, ‘*people*’
- lexical verbs in ‘delexical’ senses (e.g. *take a bath, make a decision*)
- ‘*thing*’ and pronouns ‘this’ and ‘that’ refer to abstract entity (e.g. proposition/argument), not physical object: *A strange thing happened*… *Is that why you had a few days off?... This is why I'm opposed to the plan* (Willis 1990)
- ‘*see*’ means ‘understand’ (esp spoken: *I see, You see*)
- ‘*lamely*’ for excuses (not legs), ‘*crisply*’ with speech verbs

Corpora: third/current phase

- **1990s-2000s: ‘huge’ corpora: greater depth of analysis, variety of texts, statistical accuracy**
- **British National Corpus** (100 million words; UK)
- **Bank of English** (450 million words; UK, US, Aus, Can)
- **Oxford English Corpus** (over 1 **billion** words)

- **Corpora for many other languages (e.g. national dictionary projects)**

- **Specialist corpora: business, legal, translation (parallel), learner data, etc**

Changes in language pedagogy

- Aims: from intellectual, literary, cultural, academic study to **practical** use at work, in **business, media, tourism**
- Location: from universities to language schools
- Content: from invented examples, classroom discourse to **authentic, real world communication**
- Mode: from written to **spoken**, fiction to **non-fiction**
- Methodology: from homogeneous, passive/receptive, teacher-driven, grammar-translation to heterogeneous, active/transmissive, **learner-driven**, communicative
- Focus: from accuracy to fluency
- Age: from later start to earlier start (education policy, global population mobility)
- **Corpora are well-suited to many of these changes**

Pedagogic interest in corpora

- **CALL** (Computer-Assisted Language Learning) and **CorpusCALL**
- **Data-Driven Learning** (Tim Johns)
- **Lexical Syllabus** (Willis 1990)
- **TALC** (Teaching and Language Corpora) conferences (1994 – 2006)

  http://talc7.eila.jussieu.fr/previous_sites.en.shtml
- **CLLT** (Corpus Linguistics and Language Teaching) newsgroup

The ACORN project

- **initiated in the School of Languages and Social Sciences, Aston University, in 2005**
- **Stage 1: January 2006 – January 2007, funded by the Flexible Learning Development Centre, Aston University**
- **AIMS: Corpora in English, French, German and Spanish; translation corpora (original texts and translations); academic corpora (student data, textbooks, research articles); Software to process and analyse the data; Pedagogically-oriented outputs (exercises, testing); Training; Dissemination**

ACORN: Corpus Creation

- **Design** (staff questionnaires: current materials, topics)
- **Text Identification** (according to design)
- **Copyright Permission**
- **Text Acquisition** (download from Web; email attachments; copy from hard drives)
- **Data Conversion** (PDF, DOC, HTML, etc to plain TXT)
- **Indexing** (making texts available to software tools)

ACORN: data collected so far

- **English: 10 million words**
- **French: 12 million words**
- **German: 20 million words**
- **Spanish: 2 million words**
- **Parallel (translated) Texts: 6 million words**
- **TOTAL: 50 million words**

ACORN corpus data: details

English: Business English, Academic Writing, Instruction Manuals, Political Speeches, Emails, EU legislation, Classic Literature (Shakespeare, Bronte, Darwin, Dickens, Poe, Shaw, Wilde), Nobel Speeches, University Job Advertisements, Junk Emails, Medical Abstracts, Fairy Tales
German: also Amnesty, Der Spiegel, Die Zeit, Book Reviews, Classic Literature (Goethe, Hesse, Kant, Lessing, Nietzsche, Schiller, Storm)
French: also Spoken Corpus, Classic Literature (Balzac, Daudet, Descartes, Maupassant, Verne, Zola)
Spanish: also Classic Literature (Cervantes, Zorilla)

ACORN Student Data

- **Obtaining material from students: consent forms; electronic submission (for plagiarism detection)**
- **Research will provide more information about the students:**
- general academic development; mastery of topics, themes, subjects
- development in academic writing style
- language development
- successful learning and teaching strategies which can be shared by staff and students
- possible problem areas, and the need to use alternative learning and teaching strategies
- the strengths and weaknesses of the current syllabus, and the need to adjust the focus, alter the sequence, and add or omit elements

Corpora vs Coursebooks

- Coursebooks often use made-up or heavily edited, unnatural text
- Glossaries and grammar explanations are severely restricted to specific context
- Often out-of-date
- Follow a grammatical syllabus
- Limited varieties of text

Corpora vs Dictionaries

- **Printed dictionaries**: limited by space; information is always partial, mediated, summarized, interpreted (sometimes wrongly!); internally inconsistent, or contradict each other; out of date; suffer from inertia, 'legacy' effect; under pressure from publishing deadlines, marketing, competition, etc;
- **Dictionary users**: receive no training and are impatient, so often miss the information, misinterpret it, or misuse it
- **Electronic dictionaries**: EFL ones are still in their infancy – copies of printed ones, with a few extra features; bilingual ones are often poor quality

Corpora vs Web and Search Engines		
	Web and Search Engines	**Corpora**
SIZE	**Vast**	**Manageable**
PROCESSING SPEED	**Slow**	**Fast**
ANALYSES	**Coarse-grained, General**	**Fine-grained, Detailed, Specific**
CONTENT RANGE	**No Overview; Diffuse, Uncategorized**	**Selected, Documented, Categorized**
CONTENT STABILITY	**Volatile/Dynamic: cannot replicate analyses**	**Stable: can replicate analyses**
CONTENT QUALITY	**uncontrolled**	**Controlled by selection**
SOFTWARE	**complex, 'black box'**	**simple, fully documented**

ACORN Software: current analytical functions and displays	
Frequencies: words and phrases (N-grams)	Is the word or phrase common or rare? (decide to pursue your query or not)
Distribution: i.e. which texts/authors use the word/phrase	Is it relevant to the text/context you are working in (reading/writing)
Concordances: examples of use	grammatical, phraseological and contextual behaviour of words/phrases
Collocation: 'word attraction'	*Less fixed aspects of phraseology*
Extended Contexts:	Examine discourse and textual features
Bibliographic information:	For quotation, referencing

The Future of ACORN

- Bidding to obtain more funding
- Stage 1 focussed on language learning; we want to extend to Social Sciences, and other Aston Schools: Business, Engineering, Life and Health Sciences
- More Displays, online Help, graded Access and customized Learning Paths, Exercise and Testing templates, integration with VLEs (Blackboard, WebCT), and e-learning software (Sanako)
- More Data

ACORN: DEMONSTRATION

- http://corpus.aston.ac.uk

- **We are hoping that the system will become available to all Aston staff and students during the next teaching period (i.e. early 2007)**

ASTON UNIVERSITY

IN THE HEART OF BIRMINGHAM

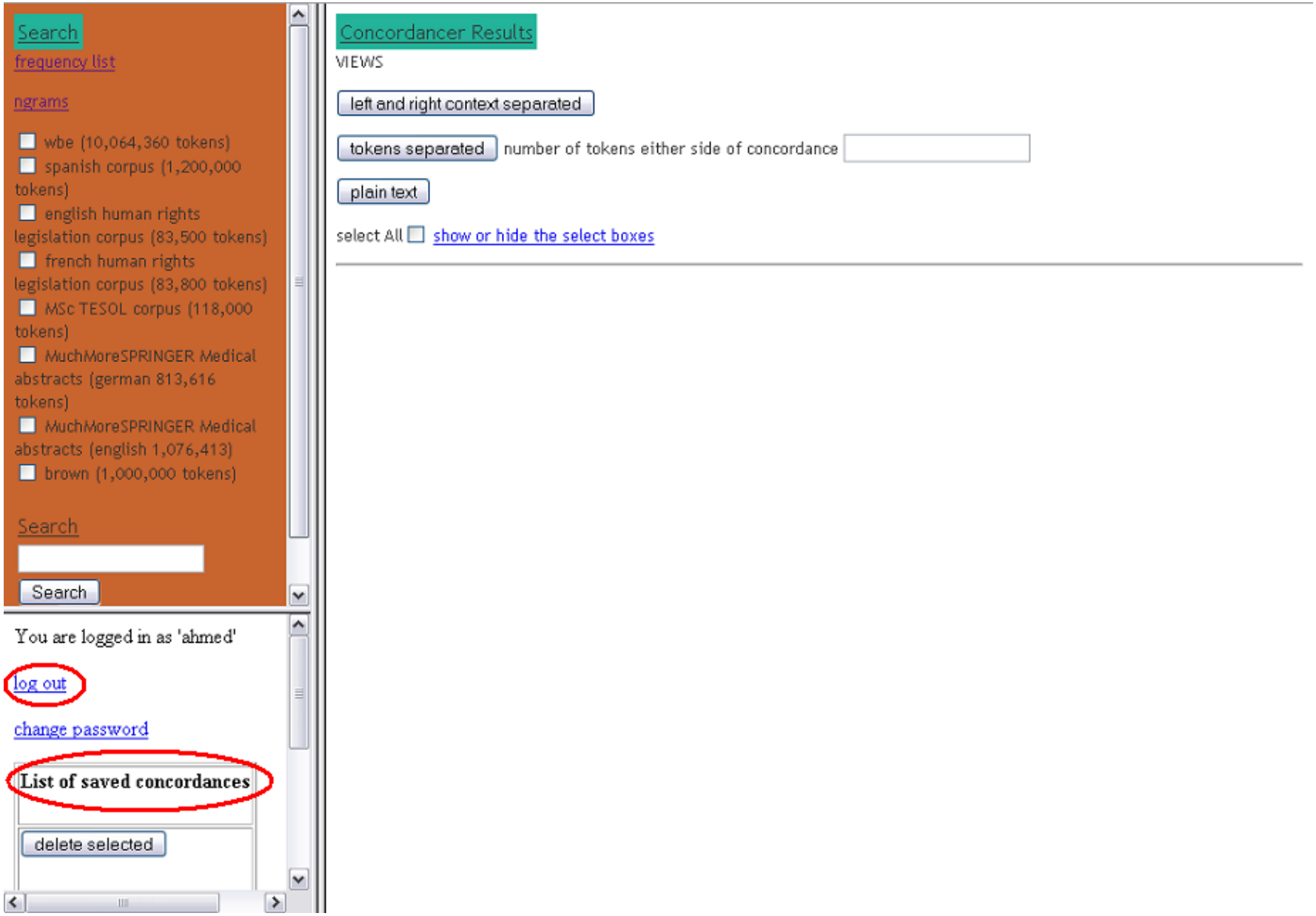
- Home
- Project Overview
- Log In
- ACORN Team
- Contributors
- Contact Us

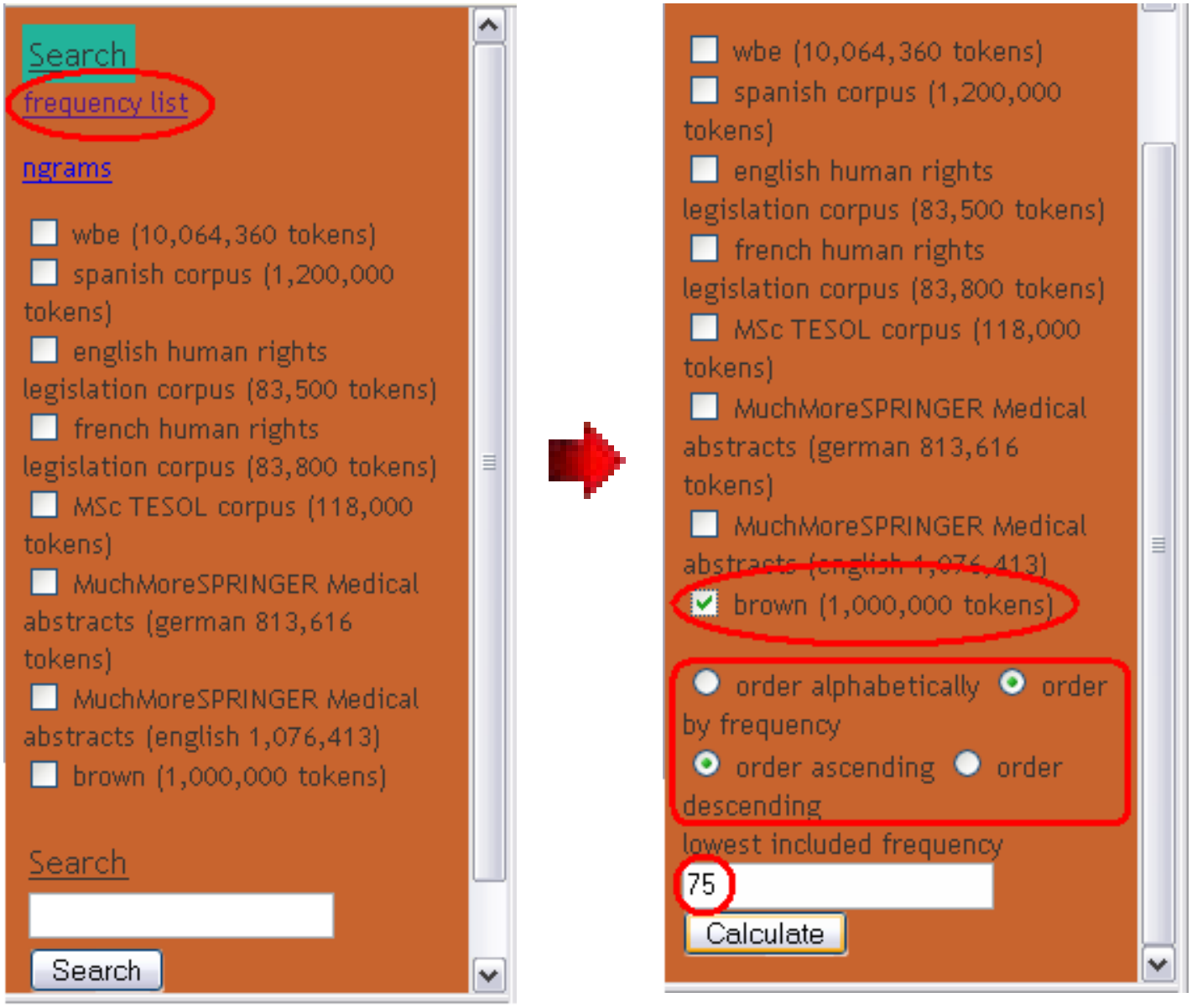
Log in to ACORN Software

Username: ahmed

Password: ••••••••••

Login

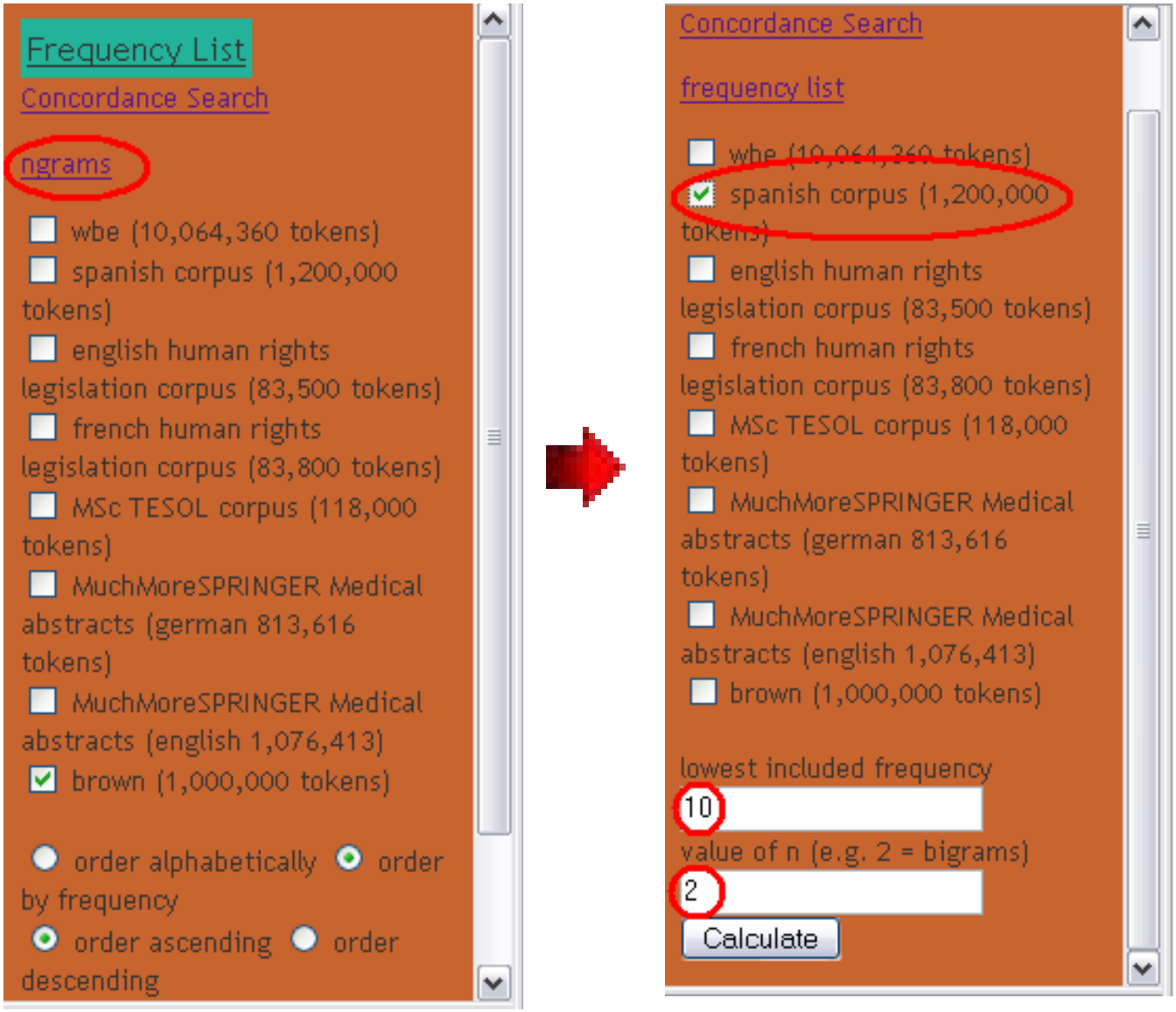

Search
frequency list
ngrams
wbe (10,064,360 tokens)
spanish corpus (1,200,000 tokens)
english human rights legislation corpus (83,500 tokens)
french human rights legislation corpus (83,800 tokens)
MSc TESOL corpus (118,000 tokens)
MuchMoreSPRINGER Medical abstracts (german 813,616 tokens)
MuchMoreSPRINGER Medical abstracts (english 1,076,413)
brown (1,000,000 tokens)
Search
Search
You are logged in as 'ahmed'
log out
change password
List of saved concordances
delete selected
Concordancer Results
VIEWS
left and right context separated
tokens separated
number of tokens either side of concordance
plain text
select All
show or hide the select boxes


Search
frequency list
ngrams
wbe (10,064,360 tokens)
spanish corpus (1,200,000 tokens)
english human rights legislation corpus (83,500 tokens)
french human rights legislation corpus (83,800 tokens)
MSc TESOL corpus (118,000 tokens)
MuchMoreSPRINGER Medical abstracts (german 813,616 tokens)
MuchMoreSPRINGER Medical abstracts (english 1,076,413)
brown (1,000,000 tokens)
Search
Search
wbe (10,064,360 tokens)
spanish corpus (1,200,000 tokens)
english human rights legislation corpus (83,500 tokens)
french human rights legislation corpus (83,800 tokens)
MSc TESOL corpus (118,000 tokens)
MuchMoreSPRINGER Medical abstracts (german 813,616 tokens)
MuchMoreSPRINGER Medical abstracts (english 1,076,413)
brown (1,000,000 tokens)
order alphabetically
order by frequency
order ascending
order descending
lowest included frequency
75
Calculate

Frequency Results

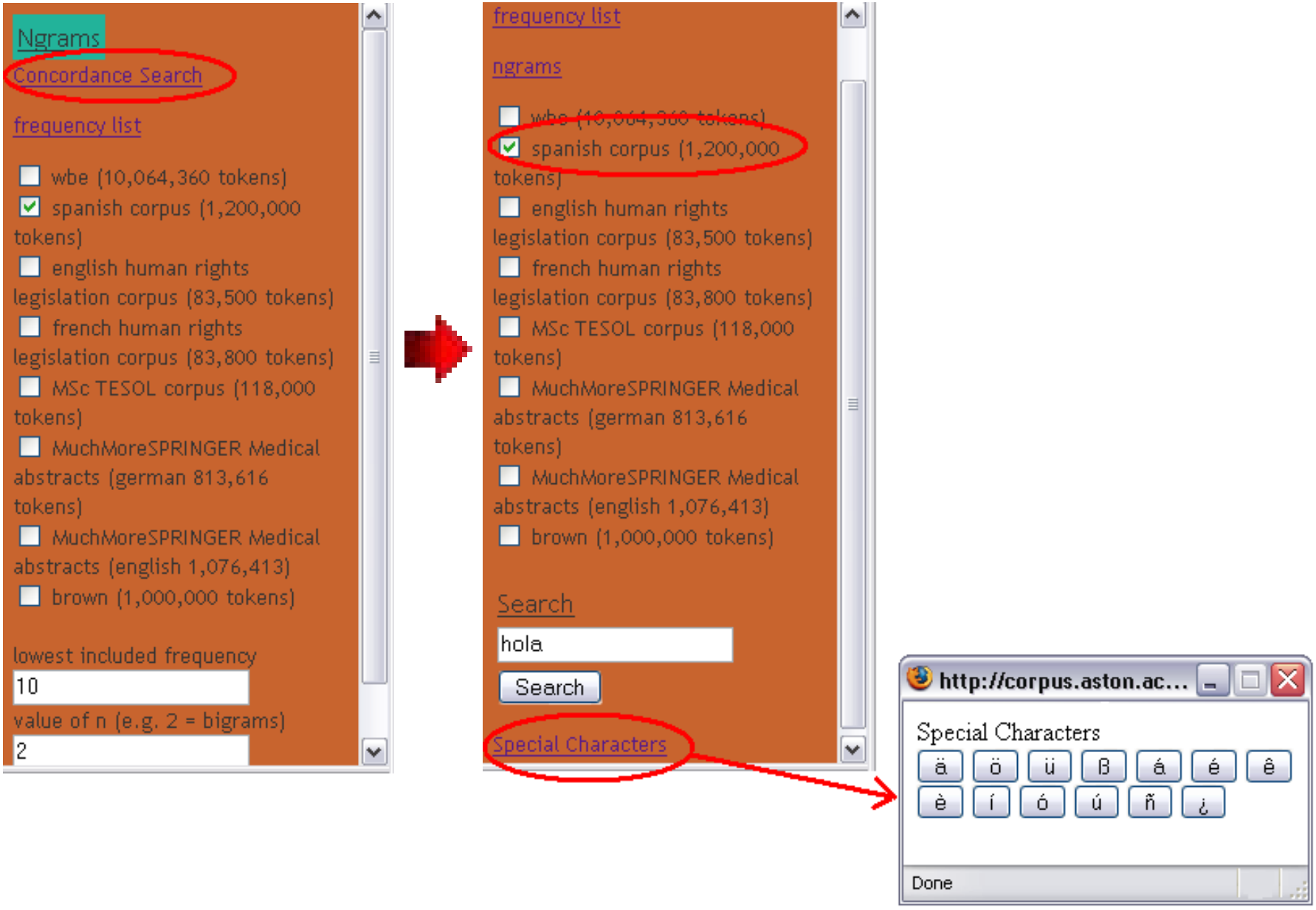
read from file

frequency of brown (1,000,000 tokens)			
total number of types = 44416			
total number of tokens = 1022053			
rank	**WORD**	**FREQUENCY**	**rate per 10,000**
1	the	69906	683.98
2	of	36430	356.44
3	and	28880	282.57
4	to	26210	256.44
5	a	23433	229.27
6	in	21382	209.21
7	that	10583	103.55
8	is	10102	98.84
9	was	9801	95.90
10	he	9536	93.30
11	for	9489	92.84
12	it	8765	85.76
13	with	7274	71.17
14	as	7242	70.86
15	his	6982	68.31
16	on	6755	66.09
17	be	6374	62.36


Frequency List
Concordance Search
ngrams
wbe (10,064,360 tokens)
spanish corpus (1,200,000 tokens)
english human rights legislation corpus (83,500 tokens)
french human rights legislation corpus (83,800 tokens)
MSc TESOL corpus (118,000 tokens)
MuchMoreSPRINGER Medical abstracts (german 813,616 tokens)
MuchMoreSPRINGER Medical abstracts (english 1,076,413)
brown (1,000,000 tokens)
order alphabetically
order by frequency
order ascending
order descending
Concordance Search
frequency list
wbe (10,064,360 tokens)
spanish corpus (1,200,000 tokens)
english human rights legislation corpus (83,500 tokens)
french human rights legislation corpus (83,800 tokens)
MSc TESOL corpus (118,000 tokens)
MuchMoreSPRINGER Medical abstracts (german 813,616 tokens)
MuchMoreSPRINGER Medical abstracts (english 1,076,413)
brown (1,000,000 tokens)
lowest included frequency
10
value of n (e.g. 2 = bigrams)
2
Calculate

Ngram List results

frequency of 2-grams in spanish corpus (1,200,000 tokens)			
total number of tokens = 1157827			
rank	**2-gram**	**frequency**	**rate per 10,000**
1	de la	7167	61.900436
2	en el	3505	30.272224
3	lo que	3199	27.629343
4	en la	3143	27.145678
5	que no	3058	26.411545
6	de los	2930	25.306025
7	de su	2758	23.820484
8	que se	2570	22.196753
9	don Quijote	2077	17.938776
10	de las	1860	16.064575
11	que la	1623	14.017637
12	que le	1605	13.862174
13	que el	1498	12.938029
14	no se	1354	11.69432
15	que en	1301	11.236566
16	en su	1207	10.424701
17	de que	1165	10.061952
18	con la	1147	9.906488
19	de un	1114	9.621471


Ngrams
Concordance Search
frequency list
wbe (10,064,360 tokens)
spanish corpus (1,200,000 tokens)
english human rights legislation corpus (83,500 tokens)
french human rights legislation corpus (83,800 tokens)
MSc TESOL corpus (118,000 tokens)
MuchMoreSPRINGER Medical abstracts (german 813,616 tokens)
MuchMoreSPRINGER Medical abstracts (english 1,076,413)
brown (1,000,000 tokens)
lowest included frequency
10
value of n (e.g. 2 = bigrams)
2
frequency list
ngrams
wbe (10,064,360 tokens)
spanish corpus (1,200,000 tokens)
english human rights legislation corpus (83,500 tokens)
french human rights legislation corpus (83,800 tokens)
MSc TESOL corpus (118,000 tokens)
MuchMoreSPRINGER Medical abstracts (german 813,616 tokens)
MuchMoreSPRINGER Medical abstracts (english 1,076,413)
brown (1,000,000 tokens)
Search
hola
Search
Special Characters
http://corpus.aston.ac...
Special Characters
ä ö ü ß á é ê
è í ó ú ñ ¿
Done

Concordancer Results

VIEWS

left and right context separated

tokens separated number of tokens either side of concordance

plain text

select All show or hide the select boxes

Save selected concordances

search for "hola"

found 40 items (showing 40 items)

view *amó por lo bajo Pepa Frías después de darle la mano - . ¡ Qué afeminado es este Ramoncito ! - ¡* ***Hola*** *, barbián ! - dijo el joven tomando de la barba con gran familiaridad a Pinedo - . ¿ Qué te has h*

view *alada , entró , y la volvió á cerrar . No bien desapareció D . Fadrique , llegó la criada . - ¡* ***Hola*** *! - dijo el P . Jacinto . - ¿ Está Doña Blanca sola ? - Sí , padre . ¿ No entra su merced á verla*

view *ros que vm . ve ? El patron pidió diez mil duros , y Candido se los ofreció sin rebaxa . ¡ Hola ,* ***hola*** *! dixo entre sí el prudente Vanderdendur , ¿ con que esté extrangero da diez mil duros sin rega*

view *a maliciosa que mostraba que no sin razón la hermanita fiaba en sus profundos conocimientos . -* ***Hola*** *, Ramoncillo - dijo acercándose a Maldonado y dándole una palmada en la mejilla con familiaridad*

view *io . . . » . Sintió pasos sobre la arena , levantó la cabeza y vio a su lado a Frigilis . - ¡* ***Hola*** *! parece que se ha madrugado - dijo Crespo , que gustaba de ser siempre el primero . - Vamos ,*

view *nuto jardín . Al subir las pocas escaleras del piso bajo salió a la puerta una criada joven . -* ***Hola*** *, Petra : ¿ y tu ama ? - Duerme todavía , señor duque . - Pues ya son las doce - dijo sacando s*

view *con una catarata de gritos guturales con que significaba su inmensa alegría . - ¡ Hola , hola ,* ***hola*** *! . . . - y daba palmaditas en el hombro al otro . El Magistral no pudo saborear tranquilamente a*

Concordancer Results

VIEWS

left and right context separated

tokens separated number of tokens either side of concordance

plain text

select All show or hide the select boxes

Save selected concordances

search for "hola"

found 40 items (showing 40 items)

view	amó por lo bajo Pepa Frías después de darle la mano - . ¡ Qué afeminado es este Ramoncito ! - ¡	**Hola**	, barbián ! - dijo el joven tomando de la barba con gran familiaridad a Pinedo - . ¿ Qué te has h
view	alada , entró , y la volvió á cerrar . No bien desapareció D . Fadrique , llegó la criada . - ¡	**Hola**	! - dijo el P . Jacinto . - ¿ Está Doña Blanca sola ? - Sí , padre . ¿ No entra su merced á verla
view	ros que vm . ve ? El patron pidió diez mil duros , y Candido se los ofreció sin rebaxa . ¡ Hola ,	**hola**	! dixo entre sí el prudente Vanderdendur , ¿ con que esté extrangero da diez mil duros sin rega
view	a maliciosa que mostraba que no sin razón la hermanita fiaba en sus profundos conocimientos . -	**Hola**	, Ramoncillo - dijo acercándose a Maldonado y dándole una palmada en la mejilla con familiaridad
view	io . . . » . Sintió pasos sobre la arena , levantó la cabeza y vio a su lado a Frígilis . - ¡	**Hola**	! parece que se ha madrugado - dijo Crespo , que gustaba de ser siempre el primero . - Vamos ,
view	nuto jardín . Al subir las pocas escaleras del piso bajo salió a la puerta una criada joven . -	**Hola**	, Petra : ¿ y tu ama ? - Duerme todavía , señor duque . - Pues ya son las doce - dijo sacando s
view	con una catarata de gritos guturales con que significaba su inmensa alegría . - ¡ Hola , hola ,	**hola**	! . . . - y daba palmaditas en el hombro al otro . El Magistral no pudo saborear tranquilamente a
view	ba , he inferido que lo mismo sabia yo que él , y que para ser ignorante á nadie necesitaba . ¡	**Hola**	! ochenta tomos de la academia de ciencias ; algo bueno podrá haber en ellos , exclamó Martin . S
view	ones . Anduvieron algunos pasos en silencio . - ¿ Qué has visto tú . . . en ella ? - ¡ Hola ,	**hola**	! Parece que pica . - ¡ Ya lo creo ! ¿ Y dónde creerás que pica ? Vegallana se volvió para mira

tokens separated | number of tokens either side of concordance: 5

plain text

select All ☐ show or hide the select boxes

Save selected concordances

search for "estremo"

found 49 items (showing 49 items)

view ☐	del	cuerpo	como	en	el	**estremo**	del	estado	y	de	la
view ☐	nunca	entendí	que	llegaba	el	**estremo**	que	decís	'	'	.
view ☐	solo	en	la	cortesía	,	**estremo**	en	la	gentileza	,	fénix
view ☐	una	doncella	hermosísima	en	todo	**estremo**	,	y	de	muy	principales
view ☐	que	entonces	llegó	en	todo	**estremo**	aderezada	y	en	todo	estremo
view ☐	hombre	;	y	holgóse	en	**estremo**	de	haberle	encontrado	,	para
view ☐	manos	lo	sabía	hacer	por	**estremo**	.	Sucedió	,	pues	,
view ☐	y	el	serlo	también	en	**estremo**	el	bachiller	Sansón	Carrasco	.
view ☐	Espejos	y	su	escudero	En	**estremo**	contento	,	ufano	y	vanaglorioso
view ☐	necesidad	del	Agua	era	en	**estremo**	,	iendo	cerca	de	Costa
view ☐	peso	,	de	que	en	**estremo**	se	alegraron	.	Finalmente	,
view ☐	,	ni	llegar	tan	al	**estremo**	de	serlo	,	mientras	no
view ☐	declaró	el	último	punto	y	**estremo**	adonde	llegó	y	pudo	llegar
view ☐	los	repique	y	sacuda	por	**estremo**	;	de	zapateadores	no	digo
view ☐	esto	,	parece	bien	por	**estremo**	,	porque	tiene	la	boca
view ☐	y	la	hermosura	en	su	**estremo**	.	Admirado	quedó	el	oidor
view ☐	Y	tú	,	¡	oh	**estremo**	del	valor	que	puede	desearse
view ☐	.	-	Es	liberal	en	**estremo**	-	dijo	don	Quijote	-
view ☐	Costa	,	llegaron	à	tal	**estremo**	,	que	se	comieron	los