



Tools for Historical corpus research, and a corpus of Latin

Barbara McGillivray
Oxford University Press

Adam Kilgarriff
Lexical Computing Ltd.



Outline

- Latin corpora
- Sketch Engine
- LatinISE: a Latin corpus for SkE
 - Collecting the texts
 - Metadata
 - Automatic annotation
 - Demo
- Conclusion



Latin corpora



Overview

- Index Thomisticus (1980) by R. Busa S. J.
 - First electronic corpus
 - 11 million words; lemmatized
- Digital editions
 - Perseus Digital Library (10 million words)
 - Corpus Grammaticorum Latinorum
 - Library of Latin Texts (50 million)
 - Musisque Deoque



Morphological annotation

- Manual
 - LASLA (1.5 million words)
- Automatic
 - Morpheus (Perseus)
 - CHLT-LEMLAT (ILC-CNR)
 - Words (W. Whitaker), Quick Latin



Treebanks

- Latin Dependency Treebank 53,000 tokens
 - Caesar, Cicero, Jerome, Ovid, Petronius, Propertius, Sallust, Vergil
- *Index Thomisticus* Treebank 100,000
 - Thomas Aquinas
- PROIEL Project 100,000
 - Translations of the New Testament in Latin, Greek, Old Church Slavonic, Armenian, Gothic



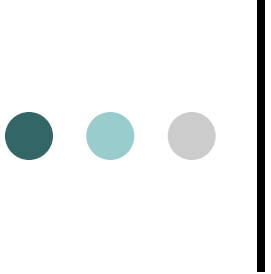
Motivation

- Latin is still a less-resourced language
- Features of our corpus
 - Size: 13 million words
 - Provided with metadata
 - Automatically annotated
 - Lemmatized
 - Part-of-speech tagged
 - Included in a clever corpus query system



Sketch Engine

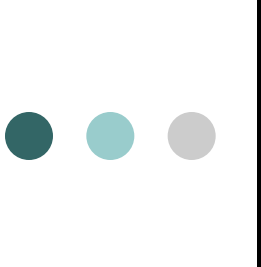
- 
- Corpus query tool, since 2003
 - Widely used by lexicographers
 - Commercial
 - OUP, CUP, Collins, Macmillan, Le Robert, Cornelsen, Shogakukan
 - National dictionary projects
 - Bulgaria, Czech Republic, Estonia, Netherlands, Slovakia, Slovenia
 - Universities
 - Linguistics, language research, NLP, language teaching



44 languages and counting

Large corpora ready-to-use for

*Arabic Bengali Bulgarian Chinese Czech
Croatian Danish Dutch English Estonian Finnish
French German Greek Gujarati Hebrew Hindi
Indonesian Irish Italian Japanese Korean Latin
Malay Malayalam Norwegian Persian Polish
Portuguese Romanian Russian Serbian
Setswana Slovak Slovene Spanish Swahili
Swedish Tamil Telugu Thai Turkish Urdu
Vietnamese*

- 
- Handles large corpora
 - Largest to date: 8 billion words
 - Fast
 - Web-based: no software to install
 - Build ‘instant corpora’ from the web
 - Load your own corpus
 - Quota of space on SkE server
 - ***Word sketches***
 - One-page, automatic accounts of a word’s grammatical and collocational behaviour
 - Free 30-day trial: sketchengine.co.uk

[Concordance](#)
[Word List](#)
[Word Sketch](#)
[Thesaurus](#)
[Sketch-Diff](#)
[Sketch-Eval](#)
[? Help on main menu](#)
[Save](#)
[Change options](#)
[Turn on](#)
[clustering](#)
[More data](#)
[Less data](#)
[Switch menu position](#)

goal (noun) ukWaC freq = 168345

object of	58924	3.0	and/or	16213	0.9	pp after-i	336	3.6	possessor	1934	3.4	predicate of	1778	3.3
score	8390	11.42	objective	858	7.01	minute	150	4.03	poacher	14	6.78	offside	22	6.87
achieve	9422	10.05	aspiration	159	6.73	break	11	1.55	opponent	39	4.44	cracker	8	5.45
concede	1421	9.43	ambition	151	6.42	goal	17	1.18	striker	15	4.29	elimination	9	4.7
accomplish	585	8.05	appearance	216	5.67	ball	9	0.85	defender	11	3.64	reward	10	2.9
reach	1924	7.84	penalty	102	5.32				visitor	67	3.42	finish	11	2.7
pursue	648	7.5	target	320	5.3				opposition	16	2.84	excellence	8	2.55
net	337	7.43	goal	315	5.3				government	169	2.78	goal	39	2.37
set	2413	7.42	dream	129	5.27				charity	38	2.72	destruction	9	2.27
attain	400	7.42	motivation	67	5.22				organization	21	2.66	fault	9	2.27
grab	406	7.39	aim	227	5.15				client	61	2.53	strike	10	2.27
pull	501	7.11	try	34	5.13				administration	18	2.43	creation	12	1.9
disallow	190	6.67	vision	154	5.1				learner	10	2.08	nothing	19	1.21
bag	186	6.64	ideal	52	5.1				organisation	71	1.72	bit	25	1.2
meet	1335	6.62	expectation	98	5.0				researcher	10	1.34	understanding	14	0.9
share	673	6.49	milestone	27	4.85				team	66	1.34	effort	17	0.87
notch	153	6.38	mission	97	4.73				author	16	1.17	one	10	0.73
realise	264	6.33	desire	73	4.7				nation	10	1.13	something	24	0.66
head	242	6.11	short-	13	4.61				club	26	1.12		8	0.47
desire	171	6.11	priority	128	4.58				company	84	1.11	production	11	0.07
define	371	5.99	purpose	223	4.49				game	40	1.01	growth	8	0.02
deserve	171	5.93	assist	11	4.36				side	31	0.81			
further	132	5.91	intention	61	4.36				user	29	0.76	pp inside-i	31	3.3
fulfil	131	5.69	striker	24	4.32				project	55	0.68	minute	21	1.2
fulfill	121	5.6	cap	34	4.3				life	63	0.59			
hit	227	5.56	goalkeeper	15	4.28				player	18	0.48			



Add your language/corpus?

- In your personal area
or maybe
- For all SkE users
 - Always interested in adding more resources
 - If it's a corpus that others may want:
quid pro quo: free use of tool
 - Contact: inquiries@sketchengine.co.uk



LatinISE: a Latin
corpus in the Sketch
Engine



Collecting the texts

- Three online digital libraries

- LacusCurtius

- <http://penelope.uchicago.edu/Thayer/I/Roman/home.html>

- IntraText

- <http://www.intratext.com>

- Musique Deoque

- <http://www.mqdq.it>

- From HTML to verticalised text



Metadata

- Author; title
- Genre (prose or poetry)
- Era; date; century
 - Oldest: Senatus consulta de Bacchanalibus (186 B. C.)
 - Most recent: Congregazione per la Dottrina della Fede, *Dominus Iesus* (2000)
- Metadata used to delete duplicated texts



Annotation

○ Natural Language Processing

- Lemmatization

- Proiel Project's morphological analyser (Dag Haug)
- Quick Latin

- Pos-tagging

- TreeTagger (H. Schmid, IMS, University of Stuttgart)

○ Advantages

- Not prone to human errors, fast, less costly



user: Barbara McGillivray used tokens: 56,706,751 / 100,000,000 days left: 630

Search

Corpora

[Create corpus](#)
[WebBootCaT](#)

Configuration templates

Sketch grammars

User groups

Corpus

[Add new file](#)
[Add web data \(BootCaT\)](#)
[Compile corpus](#)
[Open in SkE](#)
[Extract keywords](#)
[Configure corpus](#)
[Change sketch grammar](#)
[Expert mode](#)
[Download corpus](#)

LatinISE

[Add new file](#) / [Add data from web using WebBootCaT](#) / [Compile corpus](#) / [Open in SkE](#)

#	Original file	Plain text	Vertical	Tokens	Owner	
1	latin2_2.txt			13,583,877	Barbara McGillivray	



Subcorpora

○ Early (VII-II cent. B. C.)	401,557
○ Classical (I cent. B. C.)	2,275,030
○ Post-classical (I-VI cent. A. D.)	6,080,181
○ Medieval (VII-XIV cent. A. D.)	2,920,446
○ Modern (XV-XXI cent. A. D.)	2,034,940
○ Poetry	3,818,603
○ Prose	9,935,401



user: Barbara McGillivray corpus: Latin ISE

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Conc. menu](#)

[Save](#)

[View](#)

[concordance](#)

[Sample](#)

[Filter](#)

[Frequency](#)

[Node tags](#)

[Node forms](#)

[Doc IDs](#)

[Text Types](#)

[Collocations](#)

[ConcDesc](#)

[switch menu position](#)

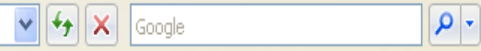
Frequency list

Frequency limit:

	<u>doc.era</u>	<u>Freq</u>	<u>Rel [%]</u>	
p/n	Romana, Postclassica	57777	76.8	
p/n	Mediaevalis	43766	190.1	
p/n	Romana, Classica	21192	108.8	
p/n	Nova	17803	111.0	
p/n	Romana, Antiqua	3544	34.4	
	<u>doc.genre</u>	<u>Freq</u>	<u>Rel [%]</u>	
p/n	prose	115405	223.3	
p/n	poetry	29118	31.4	



http://the.sketchengine.co.uk/auth/corpus/7259/ske/first_form?;lemma=;|pos=



File Modifica Visualizza Preferiti Strumenti ?

Concordance - First query form

Pagina iniziale



user: Barbara McGillivray corpus: LatinISE

Search

- [Concordance](#)
- [Word List](#)
- [? Help on main menu](#)

Query: cum

Make Concordance

Clear All

- [? Help on Expert Options](#)
- Expert options:
 - [Query Type](#)
 - [Context](#)
 - [Text Types](#)
- [Switch menu position](#)

[Concordance](#)

[Word List](#)

[Help on main menu](#)

[Help on Conc. menu](#)

[Save](#)

[View options](#)

[KWIC/Sentence](#)

[Sort](#)

[Left](#) | [Right](#)

[Node](#)

[References](#)

[Shuffle](#)

[Sample](#)

[Filter](#)

[Frequency](#)

[Node tags](#)

[Node forms](#)

[Doc IDs](#)

[Text Types](#)

Corpus: LatinISE

Hits: 87720

Page [xt](#) | [Last](#)

[Einsiedeln Eclogues](#)

[Einsiedeln Eclogues](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

talis Phoebus erat , **cum** laetus caede draconis

tempestas , moriens **cum** Roma supremas desperavit

trepidatione revocavit , **cum** fragilitas ingenii

auribus iudicaret . Nam **cum** esses in Campaniae

relevare fomentis . **Cum** itaque ad pristinum

fabulae prodiderunt , **cum** verae rationis explicatione

circumfusione concludat . Haec **cum** omnia mihi a te , Mavorti

desperatione deserui . Nam **cum** tibi totius Orientis

permodica ; quae omnia **cum** nos varia desperatione

admiratione dignum , **cum** sciamus inter ipsos

conantur evertere , **cum** alii deos non esse

hoc praesertim tempore **cum** aliud opus adgressi

vitia mala Mercurii **cum** Marte perfecit constellatio

Cum (conjunction)



user: Barbara McGillivray corpus: LatinISE

Search

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Expert Options](#)

Expert options:

[Query Type](#)

[Context](#)

[Text Types](#)

[Switch menu position](#)

Query Type:

Lemma



Lemma:

cum

conjunction



Make Concordance

Clear All

user: Barbara McGillivray corpus: LatinISE

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Conc. menu](#)

[Save](#)

[View options](#)

[KWIC/Sentence](#)

[Sort](#)

[Left](#) | [Right](#)

[Node](#)

[References](#)

[Shuffle](#)

[Sample](#)

[Filter](#)

[Frequency](#)

Corpus: **LatinISE**

Hits: **29476**

Page [Next](#) | [Last](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

auribus iudicaret . Nam **cum** esses in Cam
relevare fomentis . **Cum** itaque ad pri
fabulae prodiderunt , **cum** verae rationi
admiratione dignum , **cum** sciamus inter
paene inefficax sermo , **cum** deberemus ij
coruscatione terribilem , ecce **cum** ad Saturnum
ecce cum ad Saturnum , **cum** etiam ad eius
maius addiscimus , quod **cum** acciderit imp
opinor , aspiciat , **cum** in unum se lc
liberi fratres , et **cum** sit omnium n

Cum (preposition)



user: Barbara McGillivray corpus: LatinISE

Search

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Expert Options](#)

Expert options:

[Query Type](#)

[Context](#)

[Text Types](#)

[Switch menu position](#)

Query Type:

Lemma

Lemma:

cum

preposition

Make Concordance

Clear All

user: Barbara McGillivray corpus: LatinISE

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Conc. menu](#)

[Save](#)

[View options](#)

[KWIC/Sentence](#)

[Sort](#)

[Left](#) | [Right](#)

[Node](#)

[References](#)

[Shuffle](#)

[Sample](#)

[Filter](#)

[Frequency](#)

Corpus: **LatinISE**

Hits: **58073**

Page [next](#) | [Last](#)

[Einsiedeln Eclogues](#)

[Einsiedeln Eclogues](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

[Matheseos libri VIII](#)

talis Phoebus erat , **cum** laetus caede draco

tempestas , moriens **cum** Roma supremas d

trepidatione revocavit , **cum** fragilitas ingenii

circumfusione concludat . Haec **cum** omnia mihi a te ,

desperatione deserui . Nam **cum** tibi totius Orienti:

permodica ; quae omnia **cum** nos varia despera

conantur evertere , **cum** alii deos non esse

hoc praesertim tempore **cum** aliud opus adgres:

vitia mala Mercurii **cum** Marte perfecit co

Libero tuo religiosa **cum** trepidatione , cui

Search a phrase



user: Barbara McGillivray corpus: LatinISE

Search

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Expert Options](#)

Expert options:

[Query Type](#)

[Context](#)

[Text Types](#)

[Switch menu position](#)

Query Type:

Phrase



Phrase:

magna pars

Make Concordance

Clear All

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Expert Options](#)

Expert options:

[Query Type](#)

[Context](#)

[Text Types](#)

[Switch menu position](#)

Corpus: **LatinISE**

Hits: **130**

Page [text](#) |

Saturnalia

feriae quas indulget

**magna
pars**

mensis Iano dicati

Historia de vita

quibus errorum tuorum

**magna
pars**

est , hæud dubie Caesarea

Omnia quae extant opera

omnes itemque senatus

**magna
pars**

sententiã eius laudant

Omnia quae extant opera

, praeterea senatus

**magna
pars**

gratia cepravata Adherbalis

Omnia quae extant opera

administrabat ; Gaetulorum

**magna
pars**

et Numidae usque ad

Omnia quae extant opera

hostium collocat . Eorum

**magna
pars**

superioribus locis

Omnia quae extant opera

profecto cuncti aut

**magna
pars**

Siccensium fidem mutavissent

magna

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Expert Options](#)

Expert options:

[Query Type](#)

[Context](#)

[Text Types](#)

[Switch menu position](#)

Corpus: **LatinISE**

Hits: **56**

Page [next](#) |

[Omnia quae extant opera](#)

Namque pauci libertatem ,

pars
magna

iustos dominos volunt

[Consolatio ad Marciam](#)

et spectator et ipse

pars
magna

conantium : disces

[Panegyricus Traiano](#)

fluminibus conferendus . Hinc

pars
magna

terrarum , mergi palanti

[Ab Urbe condita](#)

emeritis etiam stipendiis

pars
magna

voluntariorum ad nomina

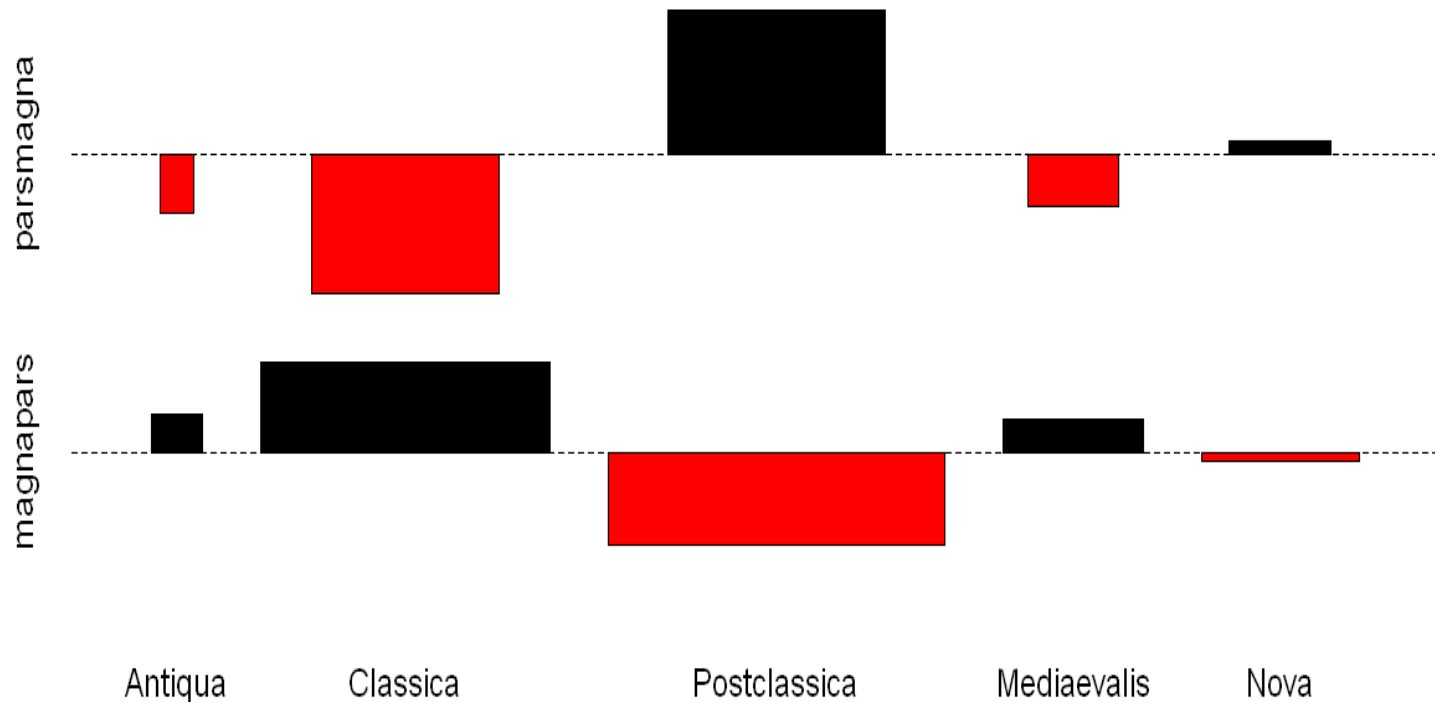
[Ab Urbe condita](#)

angustiis portarum ;

pars

aggerem vallumque conscendi

Magna pars vs. pars magna



Context: *Dico/puto/credo quod*

Query Type:
Lemma: PoS:

Context

Lemma filter

Window: tokens.

Lemma(s): of these items.

PoS filter

Window: tokens.

PoS: of these items.

(use Ctrl+click for multiple selection)

- preposition
- pronoun
- punctuation
- verb

user: Barbara McGillivray corpus: LatinISE

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Expert Options](#)

Expert options:

[Query Type](#)

[Context](#)

[Text Types](#)

[Switch menu position](#)

Corpus: **LatinISE**

Hits: **56**

Page

Isidorus Hispalensis	clamem , quare <i>putas</i> quod taceam? plenus est
Ovidius Naso, Publius	adulter et <i>credi</i> quod non contigit esse ,
Thomas Aquinas: Sanctus	quia posset <i>credi</i> quod illud quod infirmum
Thomas Aquinas: Sanctus	posset aliquis <i>credere</i> quod non esset timendus
Thomas Aquinas: Sanctus	. Et ne <i>credatur</i> quod sit altus ne tu possis
Thomas Aquinas: Sanctus	est , quia <i>credunt</i> quod propter suam altitudinem
Augustinus, Aurelius	prodesse <i>creditur</i> quod delectat . Denique
Bonaventura: Santo	, ne forte <i>credat</i> quod sibi sufficiat lectio
Ioannes Paulus PP. II	quis forte <i>credat</i> quod sibi sufficiat lectio
Apuleius, Lucius	id esse <i>crederem</i> quod esset , sed omnia prorsus
Apuleius, Lucius	adseveranti posse <i>credere</i> quod tu quicquam in meam
Apuleius, Lucius	hercules <i>dicerem</i> quod sciebam , si loquendi
Cicero, Marcus Tullius	alienum a te <i>putabam</i> quod et in Africano fuisset

user: Barbara McGillivray corpus: LatinISE

[Concordance](#)

[Word List](#)

[? Help on main menu](#)

[? Help on Expert Options](#)

Expert options:

[Query Type](#)

[Context](#)

[Text Types](#)

[Switch menu position](#)

Corpus: **LatinISE**

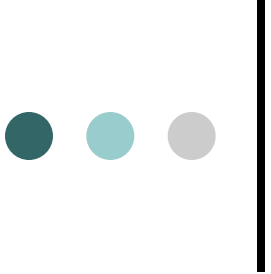
Hits: **56**

Page [text](#)

prose	clamem , quare putas	quod	taceam?' plenus est
poetry	adulter et credi	quod	non contigit esse ,
prose	quia posset credi	quod	illud quod infirmum
prose	posset aliquis credere	quod	non esset timendus
prose	. Et ne credatur	quod	sit altus ne tu possis
prose	est , quia credunt	quod	propter suam altitudinem
prose	prodesse creditur	quod	delectat . Denique
prose	, ne forte credat	quod	sibi sufficiat lectio
prose	quis forte credat	quod	sibi sufficiat lectio
prose	id esse credere	quod	esset , sed omnia prorsus
prose	adseveranti posse credere	quod	tu quicquam in meam
prose	hercules dicerem	quod	sciebam , si loquendi
prose	alienum a te putabam	quod	et in Africano fuisset



Conclusion

- 
- A new **large** resource for a **less-resourced** language
 - **NLP** tools on a **dead** language
 - Advanced corpus queries with **Sketch Engine**
 - <http://www.sketchengine.co.uk>
 - **Future**
 - Morphological tags (case, mood, voice, ...)
 - Syntactic tags (Word Sketches)