



Tracing the semantic change of socio-political terms from Classical to early Medieval Latin with computational methods

Barbara McGillivray and Krzysztof Nowak

Motivation

- The central role of diachronic semantics
- Although existing academic dictionaries of Latin provide a rich description of a words meaning based on complete evidence, they do not offer ready access to the semantic trajectory of the word, nor explicit quantitative information on the sense distribution in different authors, genres, or periods.
- Challenges for large-scale analyses
- We aim investigate whether currently available corpora can form the basis for diachronic studies of this sort

Opportunities of corpora

- Growing number of large digital corpora for Latin (McGillivray 2014)
- Increasing number of computational methods for the detection of semantic change using distributional properties of words in corpora (Schlechtweg et al. 2020).
- This can enable new perspectives on lexical semantic change, particularly promising for under-resourced languages such as Latin

Contribution

- Preliminary study: new large-scale computational analysis of the sense evolution of a number of Latin lemmas from the Twelve Tables to the 9th century CE drawing on the LatinISE corpus (McGillivray and Kilgariff 2013)
- How well do these methods work?
- What results do they produce?
- How should they be improved for Latin linguistics?

Selected terms

- Polysemous socio-political terms (e.g. *ciuitas*, *populus*) referring to changing institutions of ancient and early medieval society
 - Amply attested in Latin texts through time
 - Their high frequency poses major challenges for manual analyses → perfect candidates for semi-automatic analyses
 - Includes person names and abstract nouns
 - Case study on *civitas*
- *civitas*
 - *consilium*
 - *senatus*
 - *hostis*
 - *imperator*
 - *natio*
 - *pontifex*
 - *potestas*

Previous studies

Gens, natio, populus, plebs	Lorenzo, J. (1976). Aportaciones al estudio léxico del latín de los cristianos. Emerita, 44(2), 357–371.
Civitas, oppidum, urbs, municipium, castrum, villa etc.; corpus, societas , collegium; ecclesia, capitulum, hospitale, pons, congregatio, conventus; communitas, communio, commune	Michaud-Quantin, P. (1970). Universitas: Expressions du mouvement communautaire dans le Moyen-Age latin. L'Eglise et l'Etat au moyen-âge.
populus, plebs, vulgus	Mouchová, Bohumila (1989). Kritik und Idealisierung des römischen Volkes. Acta Universitatis Carolinae. Graecolatina Pragensia, XII, 89-109.
res publica, ciuitas	Lyasse, Emmanuel (2007). Les rapports entre les notions de « res publica » et « ciuitas » dans la conception romaine de la cité et de l'Empire. Latomus, 66(3), 580-605.
Gens, imperium	Furtado, Rodrigo Correia (2008). From « gens » to « imperium »: a study of Isidore's political lexicon. In Wright, Roger (Ed.), Latin vulgaire – latin tardif. 8: Actes du VIIIème colloque international sur le latin vulgaire et tardif : Oxford, 6-9 septembre 2006 (pp. 408-414). Hildesheim ; Zürich: Olms-Weidmann.
Potentia, potestas, imperium; imperare, iubere; rex, gubernator, princeps	Thomas, Jean-François (2012). Sur le champ lexical du pouvoir en latin. Vita Latina, (185-186), 237-249.

The distributional hypothesis

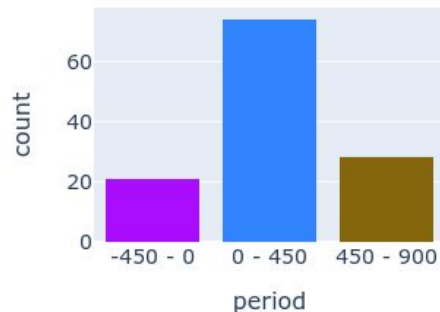
“You shall know a word by the company it keeps”

(J. R. Firth, Selected Papers, 1957)

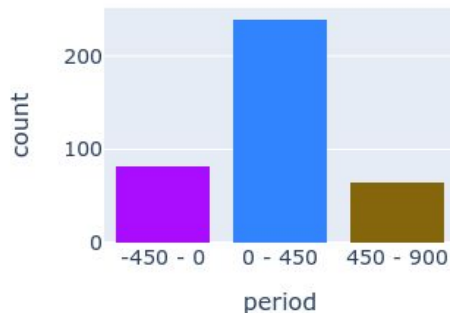
- At least certain aspects of the meaning of lexical expressions depend on their distributional properties in the linguistic contexts
- Usage-based semantics, in line with traditional philology

Corpus and Periodization

Number of unique authors



Number of texts

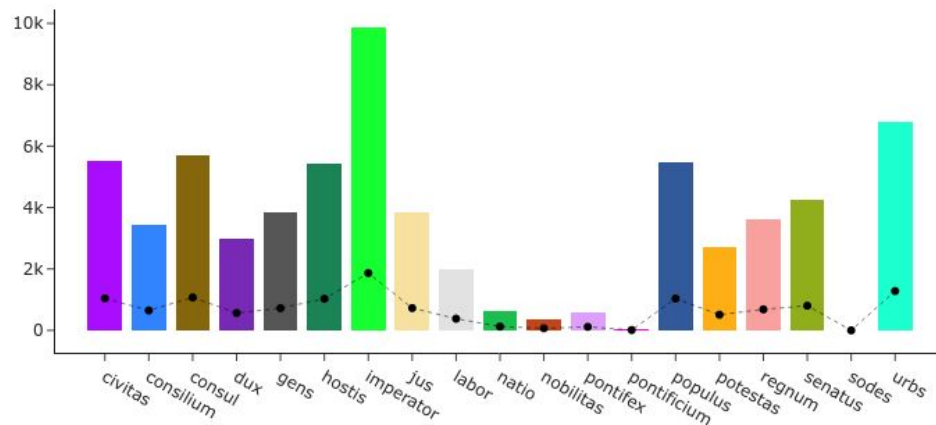


Period	Unique lemmas	Tokens	Unique tokens
-450 - 0	44861	1395858	103432
0 - 450	97396	2799762	195764
450 - 900	50265	1105116	97905
all	142051	5300736	262883

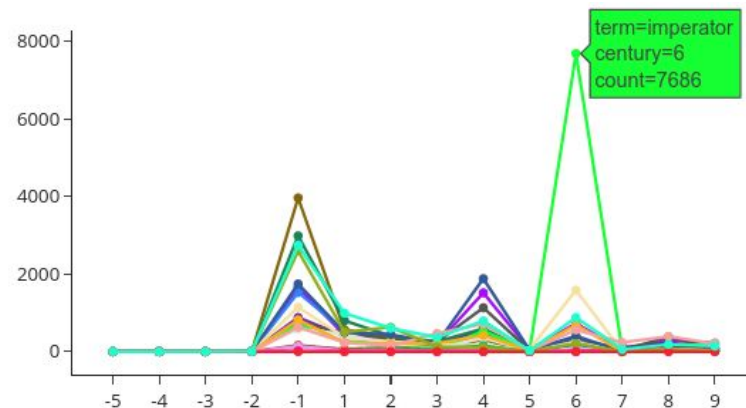
3 diachronic subcorpora

- 450 years
- time is a numeric variable
- size of subcorpora varies significantly
- alternative approaches to corpus slicing, e.g. Hilpert 2014

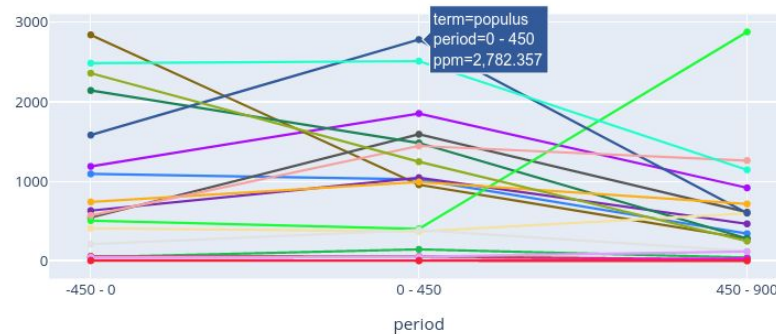
Distribution in the corpus



Absolute frequency in the entire corpus



**Absolute frequency
(centuries)**



**Relative frequency per 1M words
(periods)**

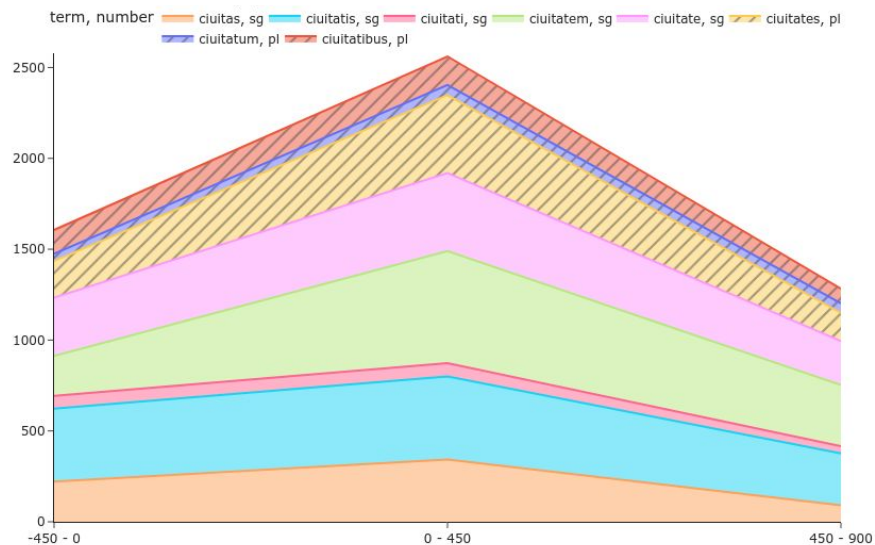
Focus on *civitas*: lexicographic account

uo, 197). *compus*: *conceivis*.
***civitas**, -ātis *f.* a *civis* (cf. VARRO ling. 10, 39 GRAMM. suppl. 68,14 LSID.orig.15,2,1 l.77). *scribitur*: *ceivitas*: LEX repetund. (Corp. I 198) 12 *al.* LEX Tarent. (Ephem. epigr. IX p. 18qq.) 39 CORP. V 7231,2 (*aet. Aug.*) *al.* IX 5834. FRONTIN. aq. 1² (*celintatis cod.*; *sed cf.* 89. 94 *al.*). *civitas*: LEX de flam. Narbon. (Corp. XII 6038) 17 *al.* CORP. V 4954. XII 1855. III 388 *et saepius*. *cibitas*: CORP. VI 34676. XIV 2080. IX 648 (*bis*). X 478. 6226. III 218 (*Cypr.*, a. 198 p. *Chr.*), 11 EDICT. imp.

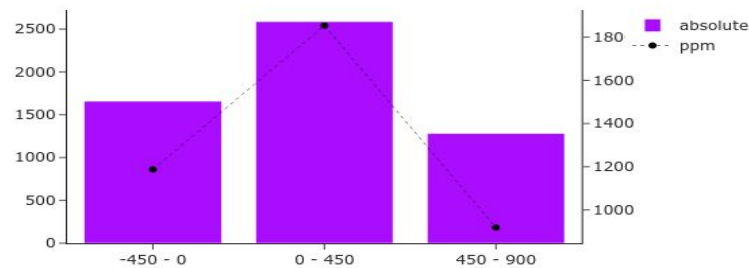
ThLL, vol. III, s.v. *civitas*

		BCE			CE					
		3	2	1	1	2	3	4	5	6
corporaliter	universitas civium		Plaut.							Vita C. Arel.
	urbs, oppidum			Rhet. Her.						Vita C. Arel.
	Platonis civitas			Cic.					Aug.	*
	tropologiae Christianae						Tert.	Apring.		
	singularia	Plaut.								Dionys. Exig.
de statu singulorum civium				Rhet. Her.						Cod. iust.

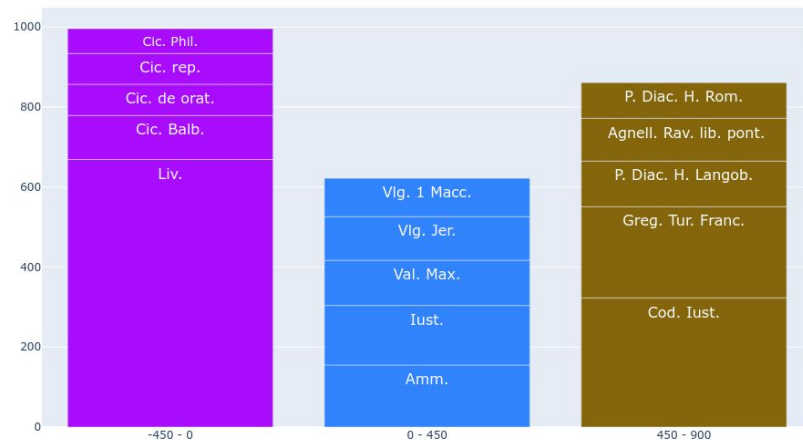
Focus on *ciuitas*: distribution in the corpus



Wordform distribution



Absolute and relative frequency



5 texts with largest number of occurrences

Collocations

	Lemma	Cooccurrences [?]	Candidates [?]	T-score	MI	↓	LogDice
1	episcopus	2,356	145,512	46.94	4.92		8.59 ...
2	Jerusalem	828	16,278	28.47	6.57		8.58 ...
3	porta	735	12,806	26.86	6.75		8.48 ...
4	extra	714	13,467	26.45	6.63		8.42 ...

Cooccurrence as an epiphenomenon (Evert)

- syntagmatic dimension
 - idioms
 - lexical collocates...
- paradigmatic dimension
 - synonymy, hyperonymy, antonymy ...
 - domain relations...

... uere glorie et **ciuitatis** eterne occiderunt ...

window: left 2 .. right 2

continuous

glorie et **ciuitatis**

non-continuous

glorie et **ciuitatis**

unity: lemma

gloria et **ciuitas** aeterna

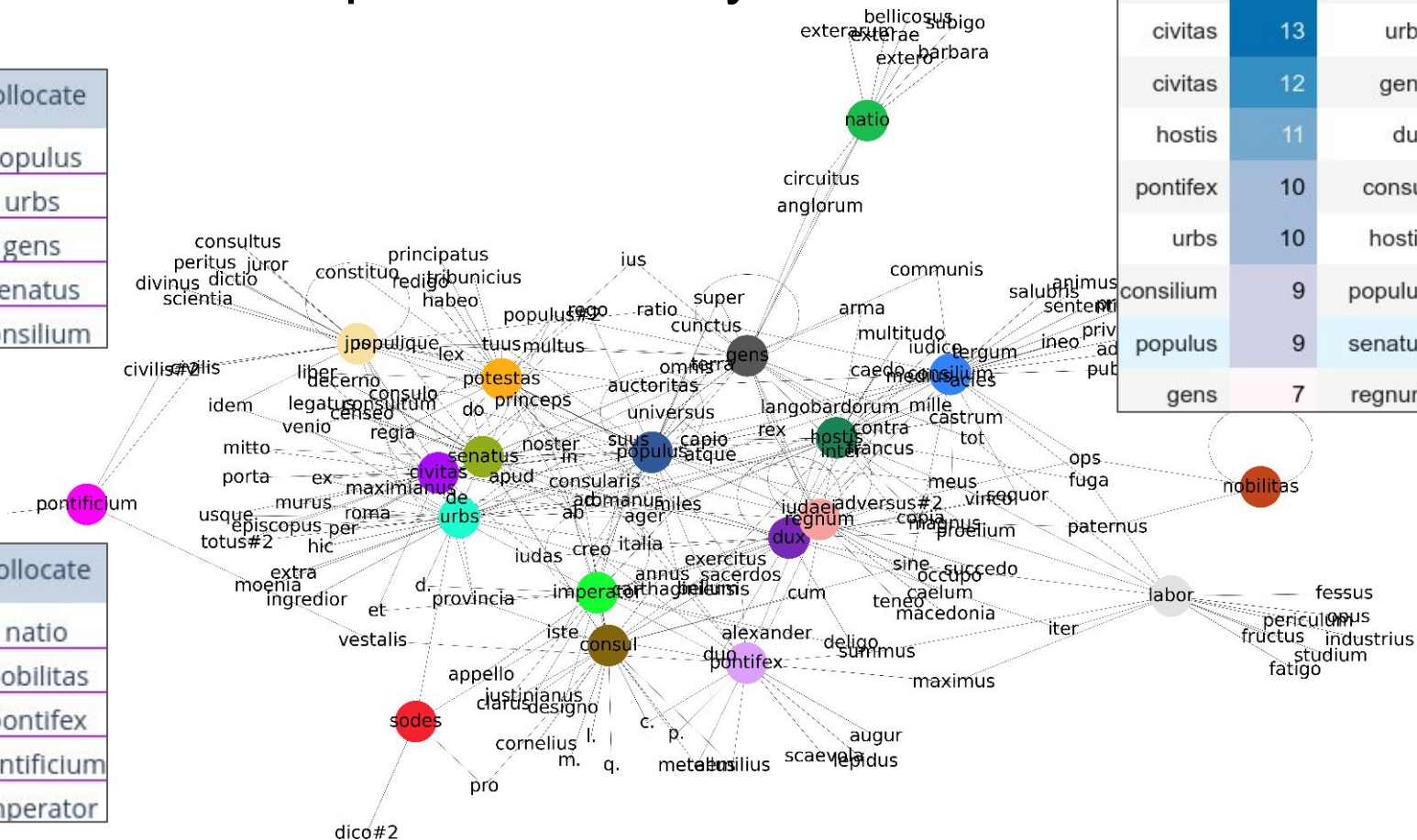
unity: word form

gloriae et **ciuitatis** aeternae

Collocations: overlap and similarity

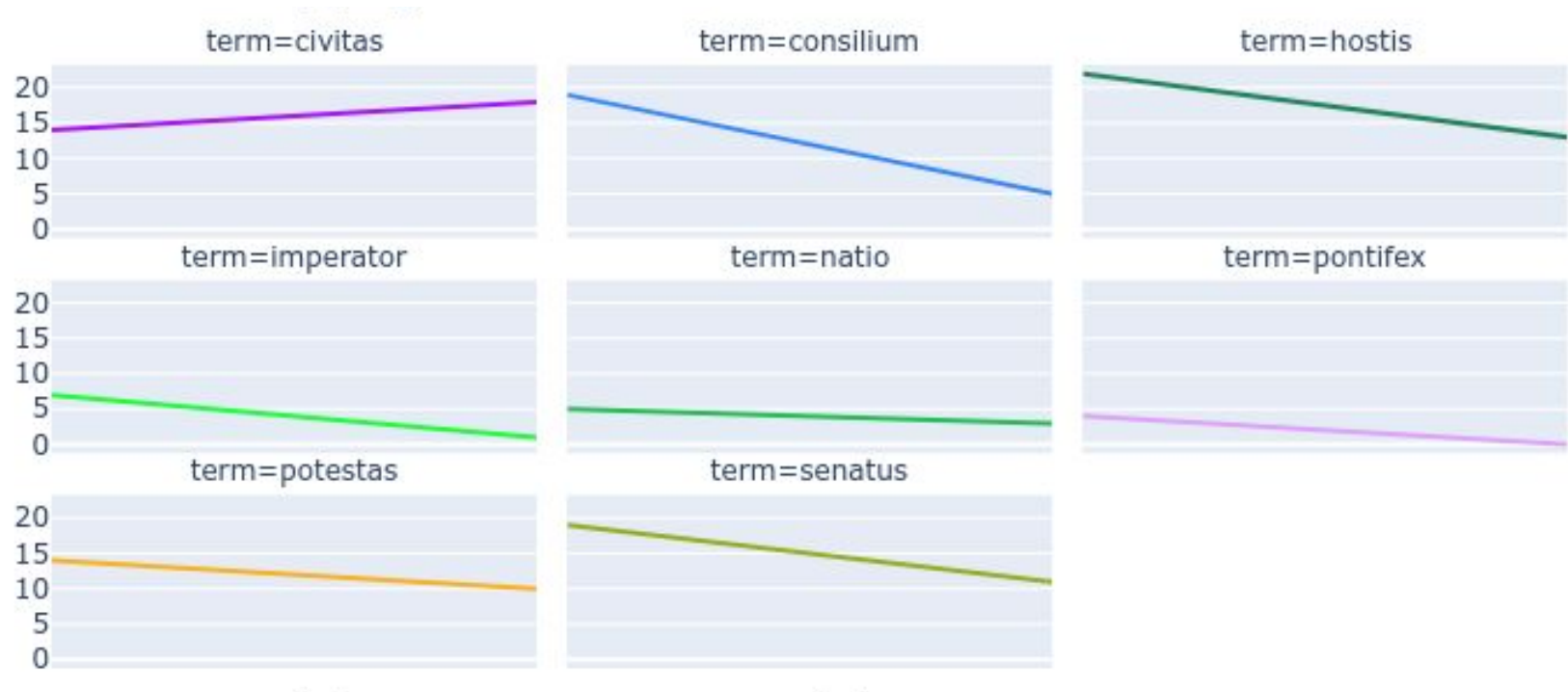
count	collocate
13	populus
13	urbs
12	gens
7	senatus
6	consilium

count	collocate
0	natio
0	nobilitas
0	pontifex
0	pontificium
2	imperator



gens	15	populus
civitas	13	populus
civitas	13	urbs
civitas	12	gens
hostis	11	dux
pontifex	10	consul
urbs	10	hostis
consilium	9	populus
populus	9	senatus
gens	7	regnum

Collocations: diachronic overlap

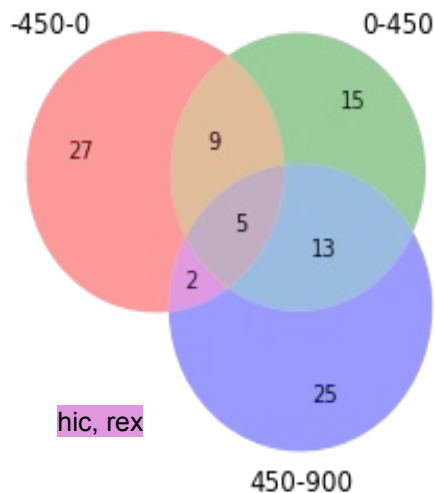


Focus on *ciuitas*: diachronic overlap

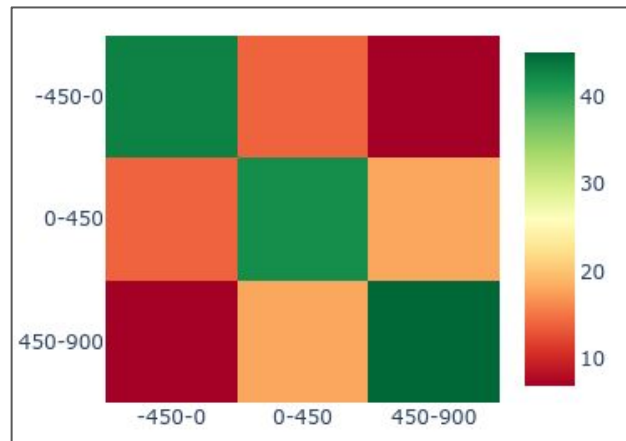
alius, asia, atque, ceterus, consilium, ex,
finitimus, homo, impero, jus, legatus, lex,
liber, libertas, mitto, mos, muto, non,
nullus, primoris, status, sum, summus,
teneo, tollo, unus, vir

de, in, quis#2, suus, totus#2

do, dono, et, graecia, is, noster,
omnis, princeps, universus



hic, rex



aedifico, civitas, david, extra, fugio, hierosolyma,
israel, iudas, magnus, munio, platea, regnum,
sanctus, urbs, villa

ad, capio, ingredior, intro, multus, murus, per, porta,
quidam, romanus, singulus, usque, venio

apud, classis, coepio, cum, curialis, defensor, dux,
egredior, episcopus, idem, igitur, infra, ipse, italia, iuxta,
proprius, provincia, quoque, ravenna, regia, reliquus, ubi,
vel, velovocorum, verus

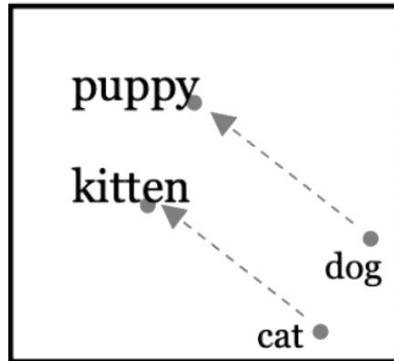
Word vectors

Target
word

The domestic **cat** is a small, typically furry carnivorous mammal.

Context
words

Context
words



cat

0.52 0.84 0.01

dog

0.40 0.90 0.10

Word vectors

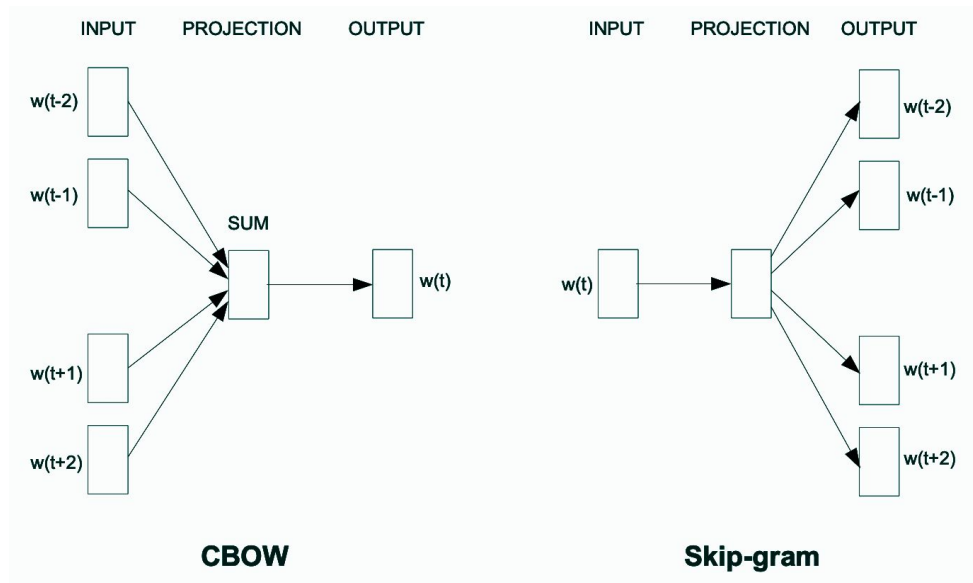
- Geometric distance is interpreted as semantic distance between words
- Nearest neighbours: closest words in the space to a certain word, occur in similar contexts
→ capture word relatedness



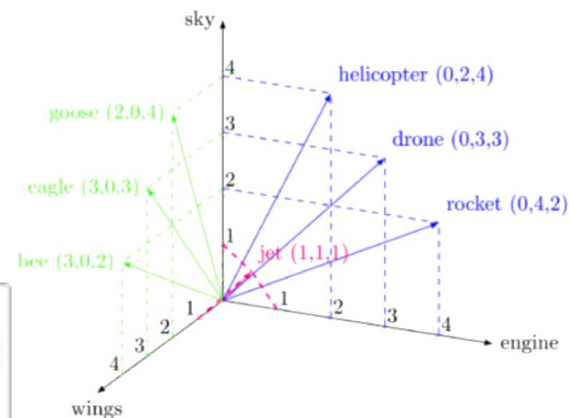
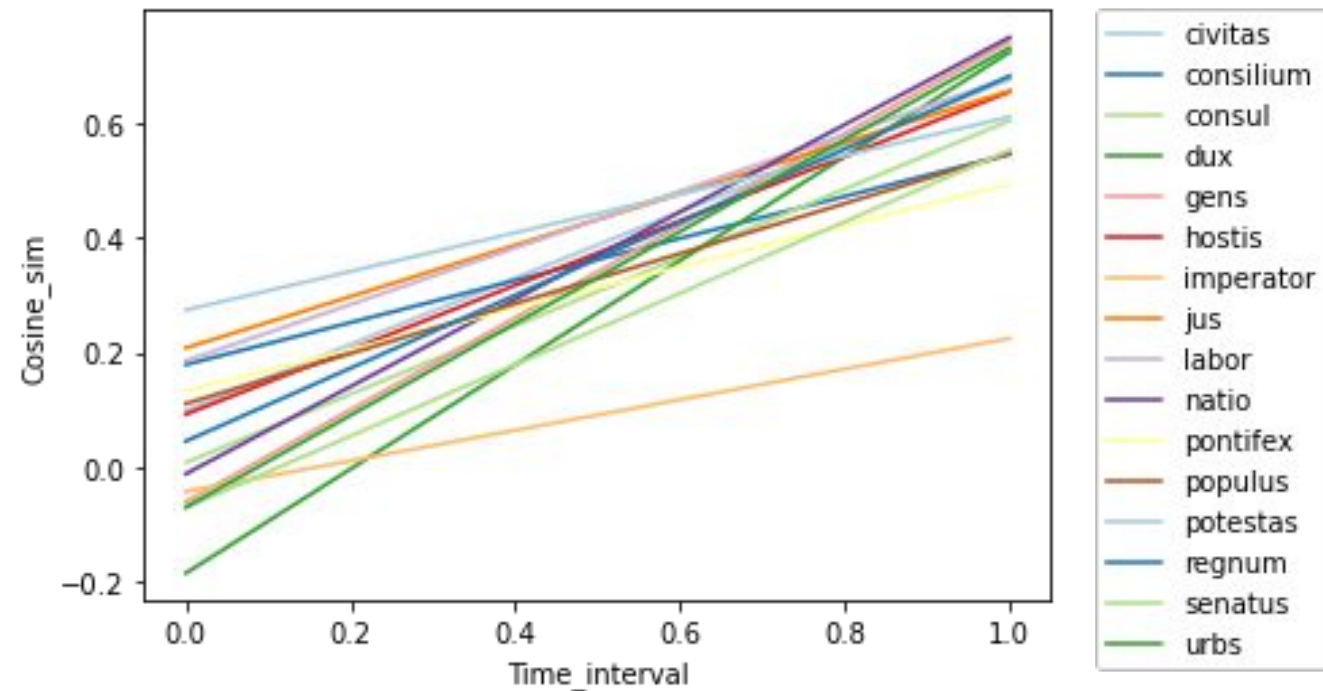
Word embeddings

- Word embedding models, e.g. Word2Vec (Mikolov et al. 2013)
- Dense vectors via Neural Networks.

- ★ Two architectures: **Skip-Gram** and **Continuous Bag-of-Words**
- ★ Important **parameters** (i.e. settings for training): context window, vector dimension, epochs.



Vectors' similarity



Civitas: from citizenship to city

450BCE-1BCE	1BCE-450CE	450CE-900CE
gens, libertas, legatio, servitus, societas, ditio ,potens, potentia, princeps, status	urbs, oppidum, murus, porta, provincia, regio, domus, castellum, vasto, Italia	urbs, insula, villa, oppidum, castrum, regio, castellum, vicus, Gallia, sedes

Method comparison

Collocation analysis	Embeddings
More interpretable	Less interpretable
Focus on co-occurrence (syntagmatic relations)	Focus on distributional similarity and relatedness (paradigmatic relations)
Does not need large corpora	Need large corpora

Conclusions

- Any results heavily depend on the corpus design
- Continuum between usage and semantic change
- We focus on gradual shifts in distributions rather than drastic changes

Future work

- Systematic definition of lexical fields
- Empirical approach to periodization
- Include different dimensions of variation, such as author, genre, domain, register, geographical provenance
- Extend chronological scope of the corpus

Thank you!

barbara.mcgillivray@kcl.ac.uk

krzysztof.nowak@ijp.pan.pl

Code: <https://github.com/BarbaraMcG/latinise>