Verborum mutatae significationes

Tracking Semantic Change in Latin with Distributional Semantic Models

Krzysztof Nowak

Lexicon Mediae et Infimae Latinitatis Polonorum
Institute of Polish Language
Polish Academy of Sciences

April 26 2017

Motivation and Outline

limiting data flood

	PHI	PL 1-60
tempus	12k	26k
cibus	1,7k	19k

overcoming biases

caro → *corpus* ? *fructus* ?

detecting patterns

- *distant reading* (Moretti)
- *macroanalysis* (Jockers)

side effects

- formalisation
- towards empirical definition of meaning

1. Methodology
   - Distributional Semantics
   - Corpora
2. Syntagmatic Perspective: tempus
   - Collocational stability
   - Clustering collocations
3. Paradigmatic Perspective: cibus

Method: Distributional Semantics

Distributional Hypothesis

"the degree of semantic similarity between two words (or other linguistic units) can be modelled ... as a function of the degree of overlap among their linguistic contexts" (Baroni and Lenci 2010)

DSMs in linguistics

- general: Turney and Pantel 2010
- accessible: Widdows 2004
- diachronic semantics: Heylen et co. 2008-
- Latin: Bamman and Crane 2011

	sacer	lego	ciuilis	gero	infero
liber	73	422	0	9	0
uolumen	234	158	0	12	0
bellum	38	16	481	831	331
pugna	0	0	8	37	17

Table: Similarity Matrix

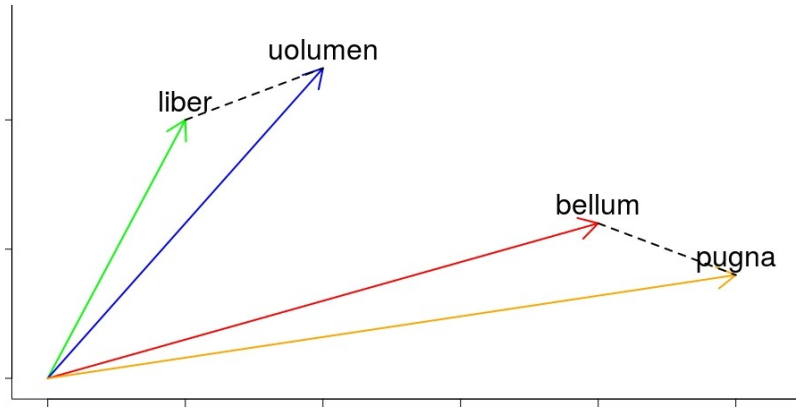


Figure: Measuring Distance

Method: Work Flow and Corpora

Workflow

1. frequency table

Term	Property	Freq
doceo	iudex	4
periculum	repello	3
visus	forma	4
addo	membrum	2

2.  `wordspace` package by Stefan Evert:
   - weighting raw counts
   - matrix reduction
3. word similarity functions
   - `pair.distances()`
   - `nearest.neighbours()`

Corpora

The PHI

- time: 2 BC - 6 AD
- ca. 8M tokens
- ca. 16k lemmas
- ca. 95k word forms

The PL 1-60

- time: 1 AD - 6 AD
- ca. 22M tokens
- ca. 16k lemmas
- ca. 115k word forms

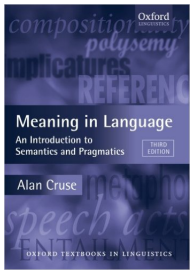
Syntagmatic Perspective: *tempus*

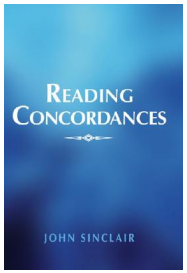
Collocation Extraction

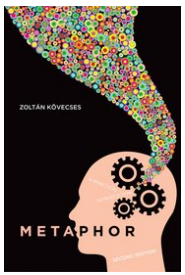
1. wide window
   vos tot milibus civium Romanorum^{5L} uno^{4L} nuntio^{3L} atque^{2L} uno^{1L} tempore necatis^{1R} quo^{2R} tandem^{3R} animo^{4R} esse^{5R} debetis
2. association strength
   - Dice: $\frac{2 \cdot f_{AB}}{f_A + f_B}$
   - minimum sensitivity: $min(\frac{f_{AB}}{f_A}, \frac{f_{AB}}{f_B})$
3. empirical collocations $\rightarrow$ clustering

Terminology

- **empirical** collocation $\approx$ co-occurrence: no lexicalization required (Evert 2009)
- lexical semantics, concordance analysis, CMT







Tempus: collocational stability (PHI)

Attested in 5 periods	Attested in 4 periods	Attested in 3 periods
vernus	annus	bellum
	brevis	medius
	certus	persona
	dies	venio
	differo	
	longus	
	praesens	
	spatium	

Tempus: collocations by century (PHI)

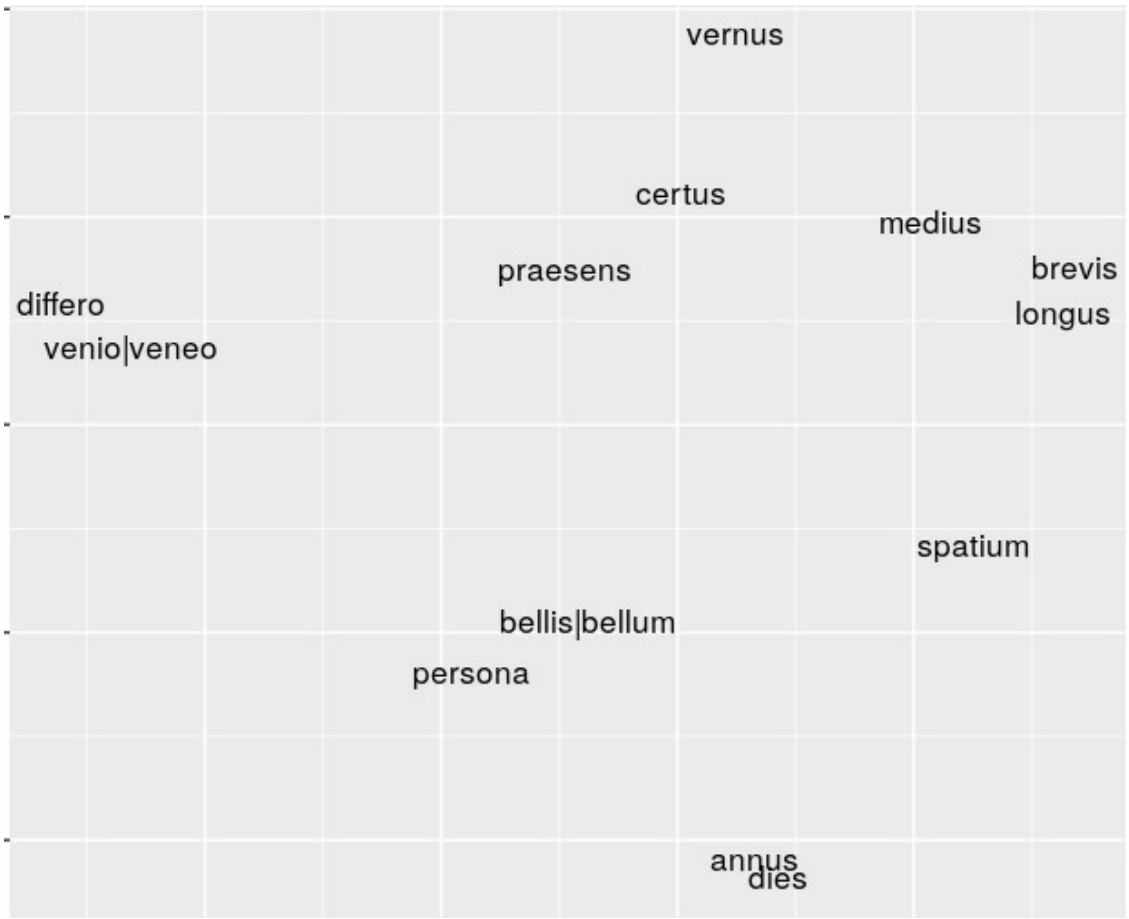
2 BC	1 BC	1 AD	2 AD	3 AD	4/5 AD	6 AD
occasio	**brevis**	**longus**	**spatium**	**vernus**	adverbium	mors
venio	aetas	**annus**	**brevis**	nox	**praesens**	testamentum
loquor	**certus**	nox	**longus**	medicus	futurus	**medius**
annus	publicus	**brevis**	momentum	hora	particeps	**certus**
quaero	superus	**certus**	diverto	verus	praeteritus	specto
terra	reliquus	tero	adverbium	orbis	perfectus	fructus
dies	**annus**	locus	syllaba	**annus**	**vernus**	possessio
video	Caesar	**dies**	aestivus	duco	significo	pubertas
	persona	nascor	vicis	urbs	autumnalis	condicio
	spatium	exiguus	disciplina	nomen	praetereo	manumitto

Table: 10 strongest collocations by period

period A	period B	shared	%
1	-1	18	36
2	1	11	22
2	-1	9	18
6	1	7	14
2	4/5	7	14

Table: Overlap of the collocations

Tempus: clustering recurring collocations (PHI)



Time is Space

longus, *brevis*, *medius*, *spatium*

movement

venio, *differo* (?)

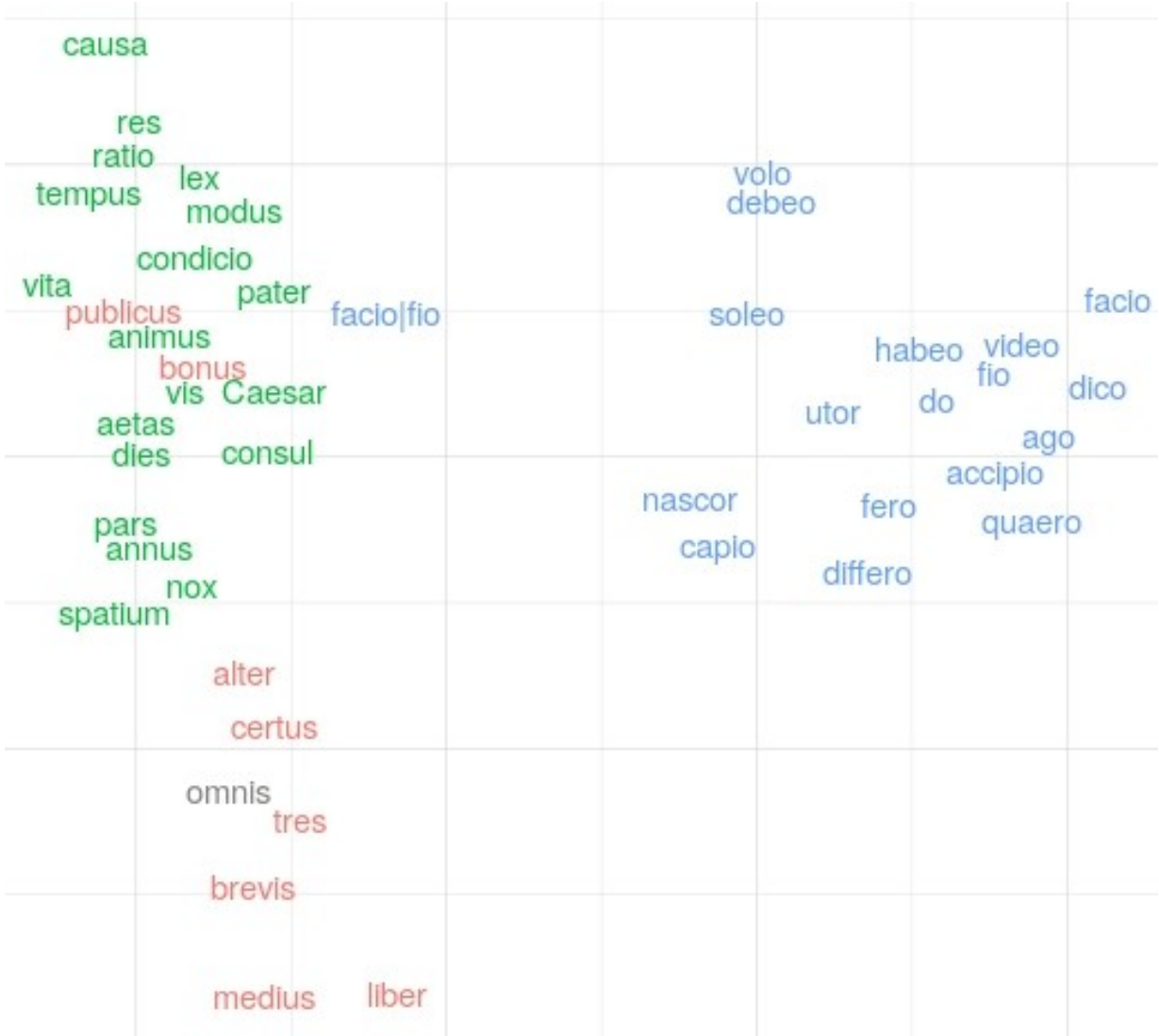
tempus + G

belli

hyponyms

annus, *dies*

Tempus: clustering all collocations (PHI)



Tempus: clustering nominal collocations (PHI)

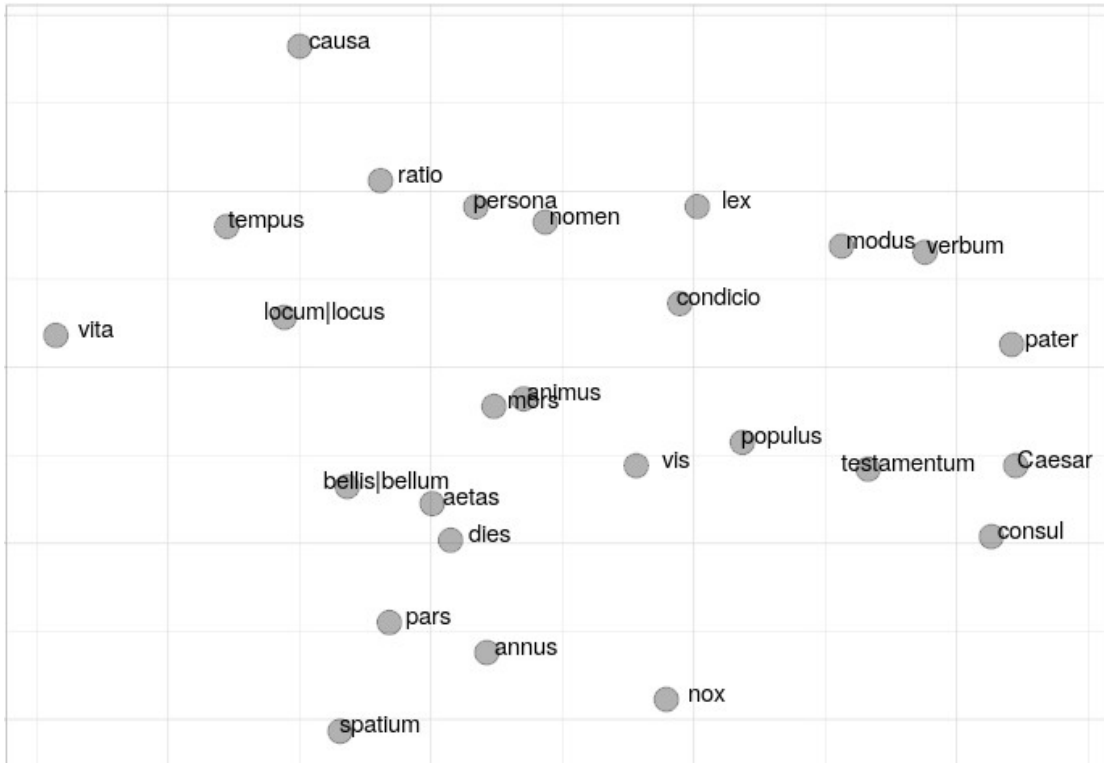


Figure: *Tempus*. Noun collocates

Time is Space

spatium

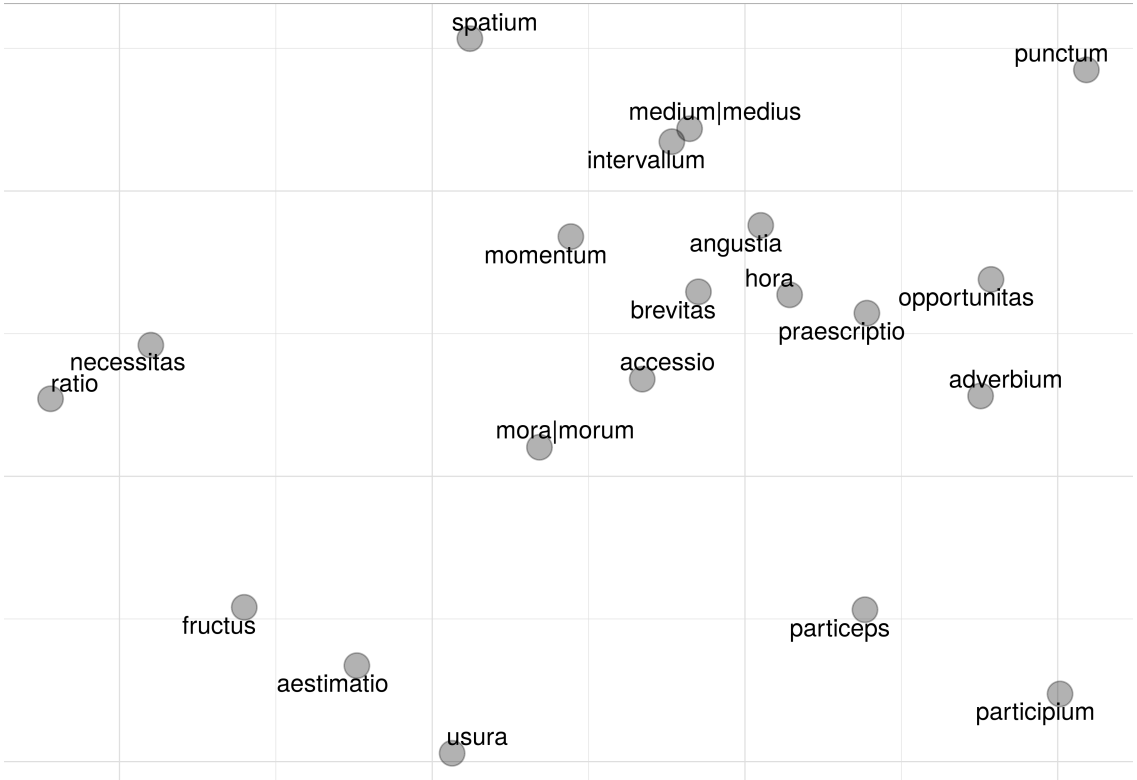
tempus + G

- hyponyms and near synonyms: *annus*, *dies*, *nox*, *aetas*
- various activities: *bellum*, *vita*

socio-political terms

- actors: *populus*, *Caesar*, *consul*
- law: *vis*, *lex*, *condicio*, *testamentum* (usually Iust. *Dig.*)

Temporis: clustering nominal collocations (PHI)



Time is Space

spatium, *intervallum*, *angustia*, *brevitas*

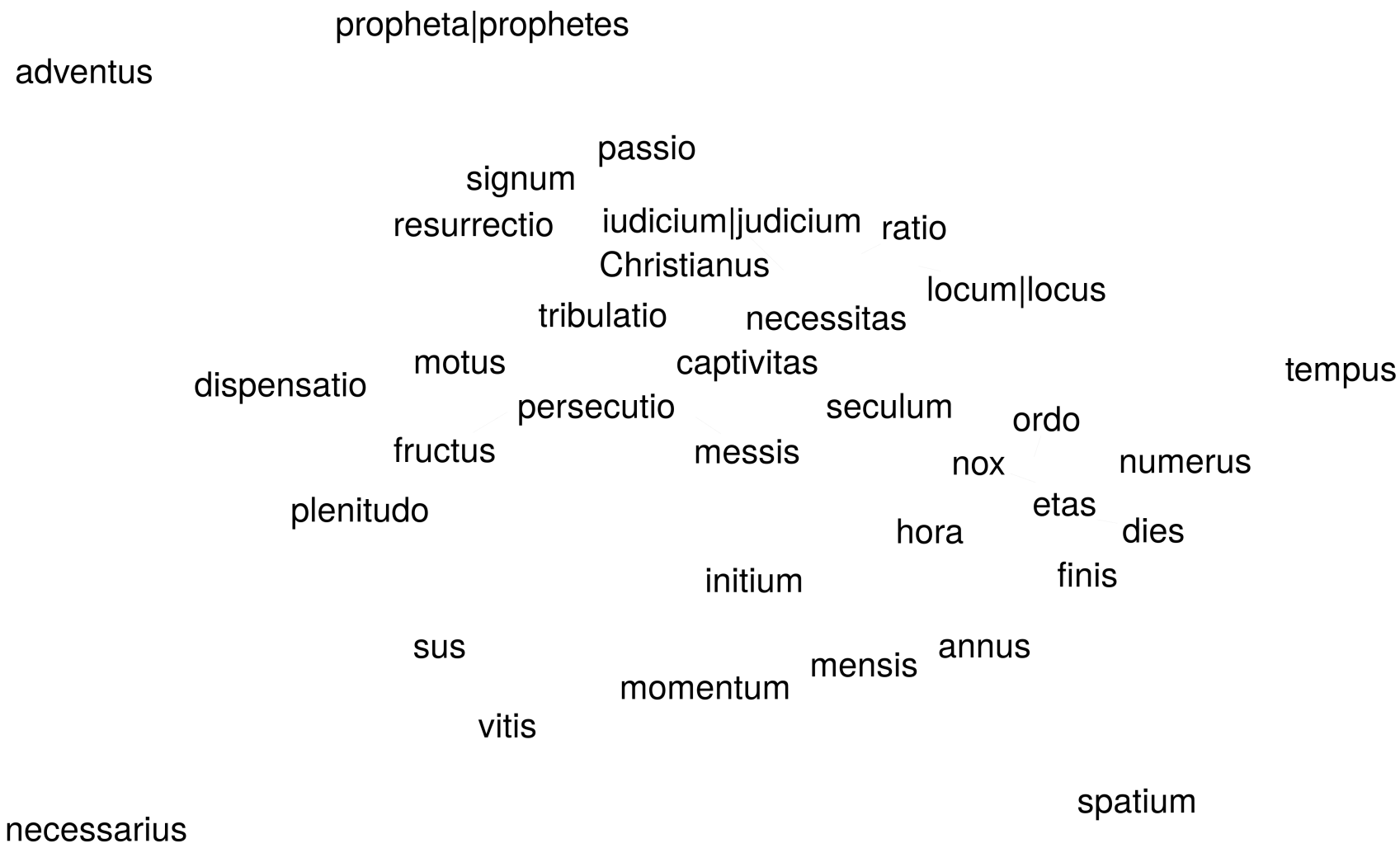
Time is Money

fructus, *aestimatio*, *usura*

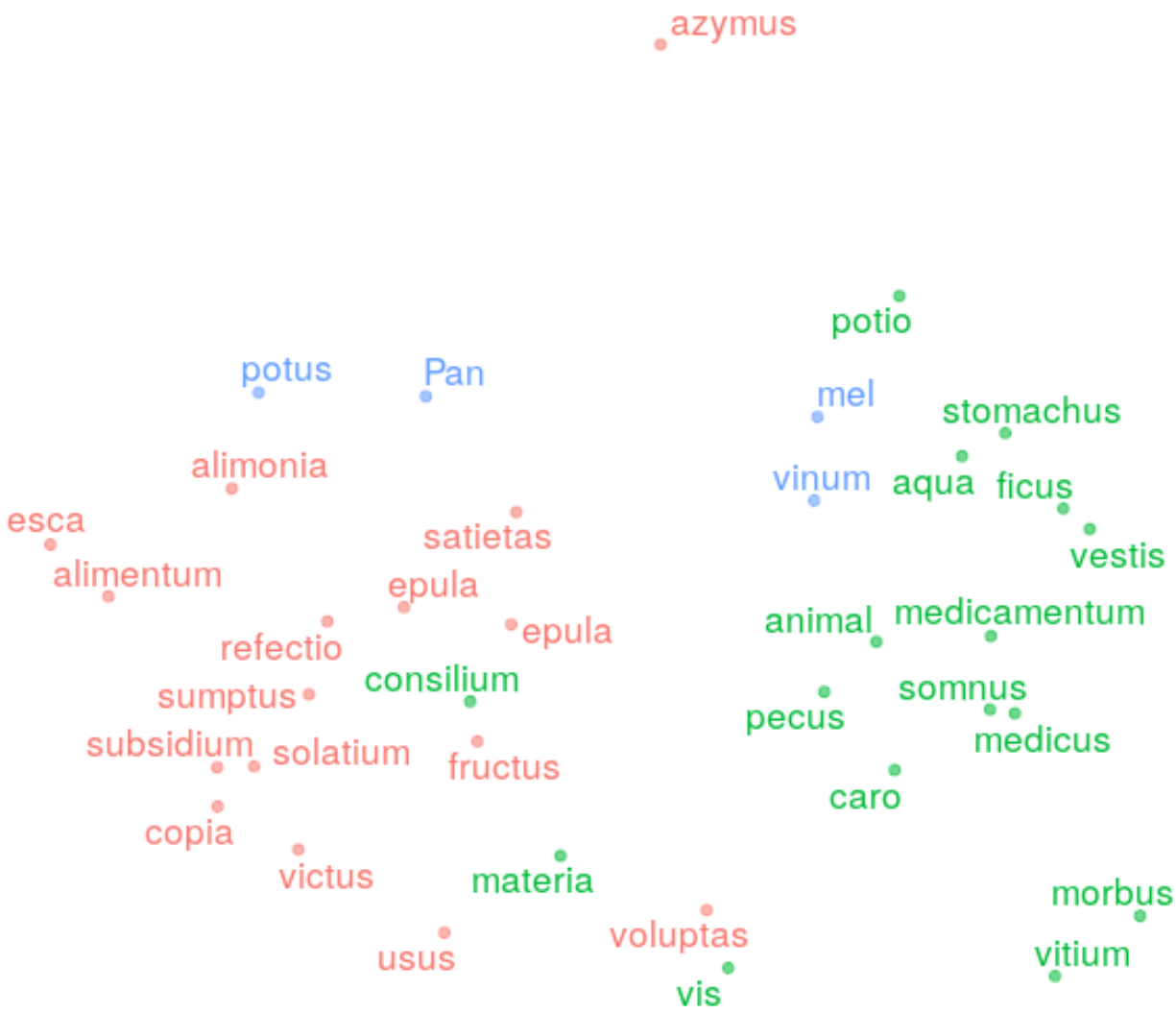
Figure: *Temporis*. 20 strongest nominal collocates

Tempus: clustering all collocations (PL)

Tempus: clustering nominal collocations (PL)



Paradigmatic perspective: *cibus* (PHI + PL)



Conclusions

- the DSMs are exploratory technique, do not provide out-of-the-box interpretation
- specific task (e.g. metaphor extraction) require targeted query and a finer setup
- room for improvement
  - cluster coherence
    - balancing the corpus
    - narrowing collocational window
    - refining properties (semantic and syntactic role)
  - better understanding of retrieved phenomena (arguments, semantic preference, semantic prosody, topical properties)

Thank you!

Questions?

krzysztof.nowak@ijp-pan.krakow.pl

Our Projects: eDictionary

http://scriptores.pl/elexicon/en

Our Projects: eCorpus

http://scriptores.pl/en/efontes/