Extensive STSM report

Lexical semantic change detection in Latin: a case study on medical latin

Paola Marongiu (University of Neuchâtel)

Introduction

This is a supporting document to the report for the Short Term Scientific Mission (STSM) “Lexical semantic change detection in Latin: a use-case on medical latin”, started on 8/05/2023 and ended on 29/05/2023 and funded by the Cost Action **CA18209**. The use-case for this STSM was developed under the supervision of Dr. Barbara McGillivray.

This STSM should serve in particular Working Group 4, use-case 2.1 (Humanities), providing new evidence for the evaluation of word embeddings for historical languages (Latin).

In this document I will provide additional information to complement the main report with all the steps and main results of the application of word embeddings to the use-case on medical Latin.

Given that the difference in semantics of the terms analysed is related to the difference in the way these terms were used in different genres rather than diachrony, this case study will focus on semantic shifts, intended as a pair of meanings synchronically related by a genetic relation i.e., two meanings of a polysemous lexeme (Koptjevskaja-Tamm 2016: 1).

List of words

To compile the list of medical terms I referred to Langslow (2000). I selected words that from general meaning developed a specialised meaning in medical contexts.

1. *Anus*: ring/circle > anus > haemorrhoids
2. *Aranea* : spider > skin disease
3. *Cancer*: crab > cancer
4. *Caninus* : tooth of a dog > canine tooth
5. *Causa* : cause/legal cause > the cause of a disease/a disease/the case of a disease
6. *Colis* : stalk/stem > penis
7. *Fistula*: pipe/flute > fistula
8. *Folliculus* : seedsack > follicle
9. *Impetus*: attack > onset of a disease
10. *Lenticula*: lentil > freckle
11. *Lumen* : light > faculty of sight
12. *Malum*: a bad thing > a disease
13. *Menstrua* : monthly > menstrual cycle
14. *Mola*: millstone > molar teeth (plural form)

15. *Musculus* : mouse > muscle
16. *Patella*: small bowl > kneecap
17. *Pecten*: comb > pubic bone (comb-like thing)
18. *Plaga*: blow > wound, incision
19. *Potio* : drink > medicinal draught
20. *Pupilla* : young girl, ward > pupil
21. *Scrotum*: leather bag for arrows > scrotum
22. *Spina* : thorn > backbone
23. *Spiritus* : breath > intestinal gas
24. *Tibia* : flute > shin bone
25. *Vitium* : sin, fault > disease

Some words were excluded from the list, after a testing phase:

- *Testis, testes* 'witnesses' > 'testicles': these seem to be a case of homonymy. It might be more problematic to talk about semantic shift; moreover, this is a meaning that only emerges with the plural form of the lemma, which might be problematic for the algorithm to spot it. The same applies to *molae* 'molar teeth'.
- *Album*: 'white' > 'white of the eye'. Words that specialise too subtly e.g. that do not change the referent altogether but whose specialisation only depends on the context are hardly detected by the algorithm. E.g. *album*: still means 'white', but referred to the white of the eye. Moreover, the specialised meaning is linked to the specific word form *album* from nom. *albus*)

Gold Standard

Gold Standards for Latin lexical semantics already exist. One general benchmark is provided by Sprugnoli et al. (2019), and a specific Gold Standard for legal Latin is provided by Ribary and McGillivray (2020).
To build the Gold Standard for medical Latin I could not rely on any dictionary of synonyms as in Ribary and McGillivray (2020) for legal Latin, because we do not have one (Roelli 2021: 121). To look for synonyms I used the sources mentioned in Sprugnoli et al. (2020: 35). These are four dictionaries of Latin synonyms, all digitised and available online.

1. Hill (1804) https://github.com/latin-dict/HillJohn1804/blob/master/lexicon.json
2. Dumesnil (1819) https://archive.org/details/latinsynonymswit00garduoft/page/262/mode/2up
3. Von Doederlein and Taylor (1875)
4. Skrivan (1890) https://github.com/nikita-moor/latin-dictionary/blob/master/Skrivan1890/sources/lemmas.json

I could not find any online dictionary for 3, so I excluded it from the sample.
Medical words are rare in the Latin general lexicon, therefore it is not infrequent that they are not recorded in the dictionaries of synonyms. Whenever I could not find the word in any of the previously mentioned sources, I looked up the word in the Thesaurus Linguae Latinae

(1900–), a Latin dictionary in Latin, and used the Latin words used to describe the meaning of the target lemma in medical context.
The Gold Standard follows the same structure as the benchmark provided by LiLa, only it focuses on the list of 25 medical terms. It is a CSV file with four columns and it is structured as follows: 1) target lemma; 2) its synonym for the specialised meaning; 3) another two words that do not have any semantic similarity with the target word, randomly selected from Sprugnoli et al. (2020)

The corpus

I identified the LatinISE corpus (McGillivray and Kilgarriff, 2013) as a corpus of reference to perform the analysis. LatinISE is a corpus of Latin texts for a total of 13 million words, spanning from the 2nd century BC to the 21st century CE. The texts are annotated with lemmas and Parts of Speech, and come with a rich set of metadata, including author, genre, title, date and century.

Works per year in LatinISE until 1000 CE

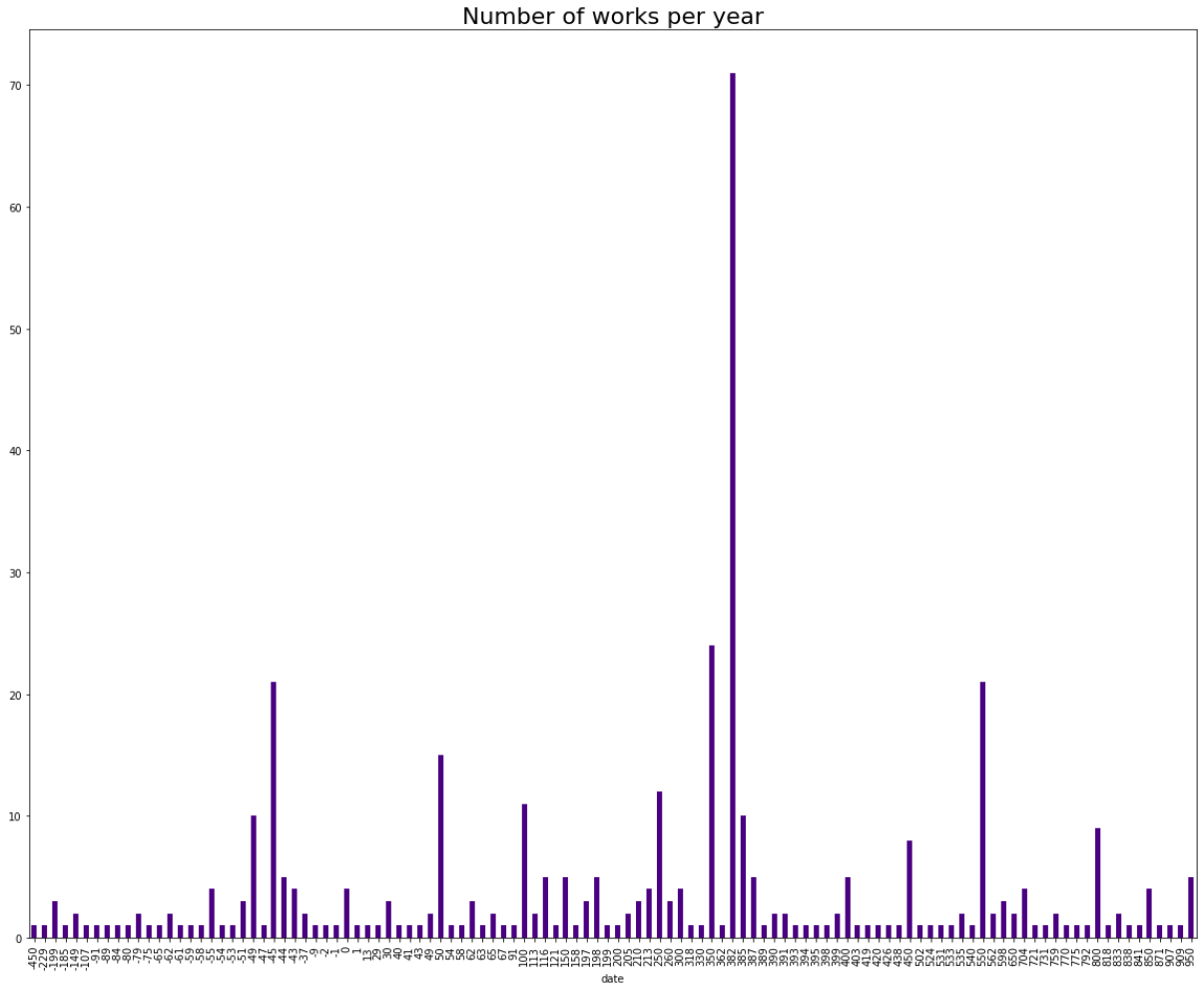

Figure 1. Works per year in LatinISE until 1000 CE.

Works for medical Latin in LatinISE

	id	title	creator	date	type	file	medical text
221	IT-LAT0382	De medicina	Celsus, Cornelius Aulus	30	prose	lat_0030_IT-LAT0382.txt	1
562	IT-LAT0895	Medicina ex oleribus et pomis	Gargilius Martialis, Quintus	260	prose	lat_0260_IT-LAT0895.txt	1
634	IT-LAT0987	De observatione ciborum	Anthimus	550	prose	lat_0550_IT-LAT0987.txt	1
837	ITMQDQ-247	medicamina faciei	Ovidius Naso, Publius	30	poetry	lat_+0030_MQDQ-247.txt	1
1221	ITMQDQ-440	liber medicinalis	unknown	350	poetry	lat_0350.0_MQDQ-440.txt	1

Table 1. Medical texts in LatinISE.

Distribution of texts between medical and non-medical sub-corpora (until 1000 CE)

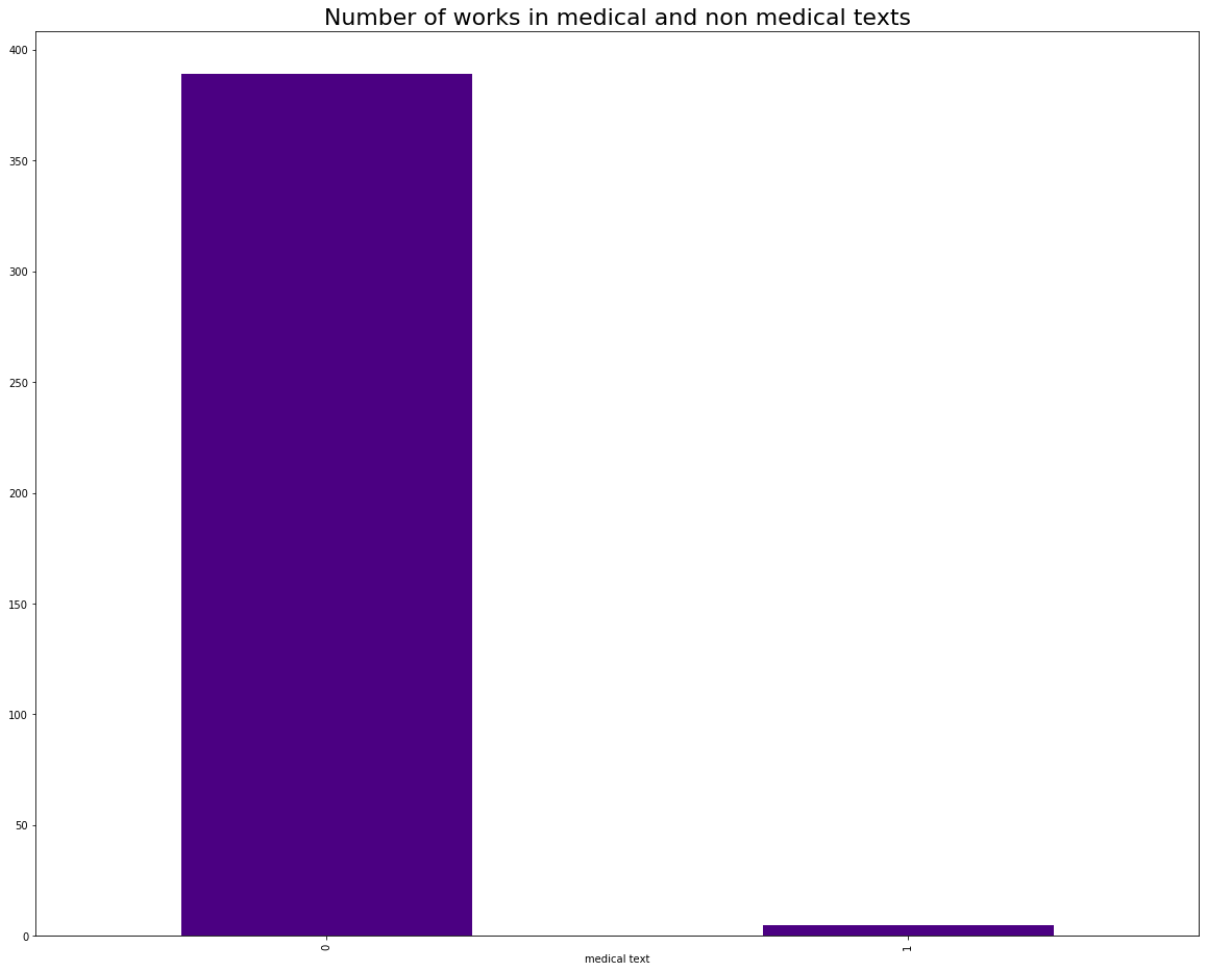

Figure 2. Number of works in the medical and non medical subcorpora.

I considered only IntraText texts because there are some repetitions with the other sources. I added 2 further medical texts that came from a different source (*Liber medicinalis* and *De observatione ciborum*)

The proportion between the two sub-corpora is highly unbalanced: the medical subcorpus includes only five texts, whereas the non medical subcorpus counts 389 texts.

Test for best configuration of parameters

To carry out this case study I adapted an existing code provided by Dr. McGillivray, from McGillivray and Nowak (2022) and available at https://github.com/BarbaraMcG/latinise/blob/master/lvlt22/Semantic_change.ipynb

The test was performed on the model trained on the whole selected corpus (i.e., all works from IntraText + 2 medical texts until 1000 CE)

The test was performed with *causa* (both a general term and specialised term in medical Latin)

Test 1: Minimal frequency = 5 (check for lower frequency words) and vector size = 100

Output closest neighbours for *causa* with this combination of parameters:

model.wv.similar_by_word('causa', 10)
[('causo', 0.8693389296531677),
 ('caussa', 0.8572415113449097),
 ('causatio', 0.828999936580658),
 ('causor', 0.8153892755508423),
 ('causalis', 0.8007003664970398),
 ('causarius', 0.7590205073356628),
 ('musa', 0.7423707842826843),
 ('causaque', 0.7420263886451721),
 ('Susa', 0.7409977912902832),
 ('causamque', 0.7246591448783875)]

The output is not correct: none of the neighbours are synonyms of *causa.* The algorithm focuses on orthographically similar words (*cause, caussa, causo* etc.). I try to correct this by raising the minimum frequency to 50 in Test 2.

Test 2: Adding max_n=0 and min_n=0 to avoid selection at the string level (turn off subwords), keep the other parameters steady

[('quaestio', 0.6342904567718506),
 ('accusatio', 0.6149478554725647),
 ('probatio', 0.6000396013259888),
 ('comparatio', 0.5965130925178528),
 ('lite', 0.5946058630943298),
 ('accusator', 0.5853508114814758),
 ('ratio', 0.5788862705230713),
 ('defensio', 0.5752539038658142),
 ('disceptatio', 0.5709359049797058),
 ('deliberatio', 0.5624971389770508)]

The output improves by turning off subwords: many of the words appearing among the closest neighbours are related to the legal sense of the word *causa* (*accusatio, probatio, lite, accusator, defensio, deliberatio*). Moreover the neighbour *ratio* is indicated as the direct synonym for *causa* in the Gold Standard provided by Ribary and McGillivray (2020).

Test 3: Turn off subwords, raise minimum frequency to 50, vector size 100

[('accusatio', 0.6261991858482361),
 ('lite', 0.6245417594909668),

('defensio', 0.6181176900863647),
('quaestio', 0.6128496527671814),
('probatio', 0.5928096771240234),
('litigator', 0.5784130096435547),
('accusator', 0.5771515965461731),
('culpa', 0.5738686323165894),
('lis', 0.565229594707489),
('actio', 0.552460253238678)]

The neighbours that point towards the legal meaning of *causa* are still there, but by raising the minimum frequency to 50 we lose the direct synonym *ratio*. **The combination no subwords + min frequency 5 seems to be the "best" one for this corpus.**

Try again with another word, *dies*:

Minimal frequency = 5 (subwords excluded)
[('hora', 0.6784750819206238),
('septimana', 0.6136382222175598),
('annus', 0.6078673005104065),
('mensis', 0.596966028213501),
('biduum', 0.5657556056976318),
('triduum', 0.5343977808952332),
('sabbatum', 0.527468204498291),
('vigilia', 0.525002121925354),
('triennium', 0.5221787691116333),
('nox', 0.5192839503288269)]

Minimal frequency = 50 (subwords excluded)
[('hora', 0.7011444568634033),
('septimana', 0.6784548163414001),
('feriae', 0.5813270211219788),
('mensis', 0.5806093811988831),
('triduum', 0.5749990940093994),
('pridie', 0.5530216097831726),
('biduum', 0.5481677055358887),
('annus', 0.5444790720939636),
('vigilia', 0.5228376984596252),
('quadriennium', 0.5126366019248962)]

With common words the minimum frequency threshold does not seem to noticeably impact results.

Another test with a word from the list: *potio*.

Minimum frequency = 5
[('potus', 0.7746923565864563),
('cibus', 0.7582693099975586),

('alimentum', 0.6682046055793762),
('extenuo', 0.6668452024459839),
('concoquo', 0.6660478711128235),
('medicamentum', 0.658385157585144),
('sitis', 0.6566506028175354),
('esca', 0.6549215912818909),
('febrem', 0.6542739272117615),
('pabulum', 0.6541865468025208)]

Minimum frequency = 50
[('cibus', 0.7652978301048279),
('esca', 0.713490903377533),
('potus', 0.7071375846862793),
('sitis', 0.6719212532043457),
('increscо', 0.666099488735199),
('alimentum', 0.6642388701438904),
('stomachus', 0.6602162718772888),
('sitio', 0.6561112999916077),
('alvus', 0.6513990163803101),
('frigidus', 0.6403871774673462)]

By raising the minimum frequency to 50, I lose the direct synonym *medicamentum* in the entire corpus.

Test on medical words: qualitative comparison of closest neighbours in the two models

In the following sections I list the closest neighbours for all the target words in the list under section “List of words in this document”. Words highlighted in green are the exact synonyms listed for each target word in the gold standard created for this case study. If none of the closest neighbours are synonyms, but some of them are close to the synonym or belong to that semantic frame and therefore suggest the semantic shift, I highlight them in yellow.
The problem with the test on medical words in LatinISE is two-folded: 1) LatinISE only contains 5 medical texts (listed in section ‘Works for medical Latin in LatinISE’). Medical words (or words with specialised meaning in medical language) are extremely rare in the corpus.

Anus: ring/circle > anus > haemorrhoids

Neighbours in subcorpus 0 (non medical)
[('formosus', 0.7208097577095032),
('nutrico', 0.7165568470954895),
('delicatus', 0.6415420770645142),
('mamma', 0.6402621269226074),
('virguncula', 0.6257895827293396),

('nutrix', 0.6114704012870789),
('Sabina', 0.6049647331237793),
('lacto', 0.5958934426307678),
('puella', 0.5957138538360596),
('thalamus', 0.5906150341033936)]

The neighbours *formosus* 'beautiful', *delicatus* 'charming' suggest that in the non medical subcorpus the word *anus* is used with the meaning 'ring' rather than 'anus' or 'hamorrhoids'. Other words that might belong to the semantic frame of 'ring' and the word *thalamus* 'nuptial bed' and by extension 'marriage'.
Among the closest neighbours there are frequent references to women, and some specific references to breastfeeding *mamma* 'udder', *virguncula* 'little girl', *nutrix* 'nurse' but also 'breast', 'the one who feeds/raises', *puella* 'girl', *nutrico* 'to feed, raise', *lacto* 'to breastfeed'. This is probably due to the fact that *anus* is a case of homonymy with another entry with the meaning 'old woman'. The ThLL lists as synonyms *mater* 'mother', *vetula* 'old woman' and *puella, pupa* 'little girl' as antonyms. Both appear among the closest neighbours for *anus* 'circle, ring'. In this case the embeddings for *anus* in the non medical corpus might have also captured the distribution for the entry *anus* 'old woman'.

Neighbours in subcorpus 1 (medical)
[('maxilla', 0.9982499480247498),
('inguen', 0.9979749917984009),
('iuxta', 0.9974749088287354),
('ferramentum', 0.9971880912780762),
('scrotum', 0.9971383213996887),
('antequam', 0.9969991445541382),
('vinculum', 0.9969804883003235),
('glutino', 0.9967426061630249),
('sedeo', 0.9965815544128418),
('vulnero', 0.9965251684188843)]

Among the neighbours in the medical corpus the notion of 'ring' has not completely disappeared: the nieghbours *ferramentum* 'iron tool' and *vinculum* 'tie' might still be pointing towards this meaning. However, if we compare this set with the one obtained from the non medical subcorpus we obtain more references to body parts and anatomy: *maxilla* 'jaw', *inguen* 'groin', *scrotum* 'scrotum' (the latter located close to the anus).

Aranea: spider > skin disease

Neighbours in subcorpus 0 (non medical)
[('pharetra', 0.6134344339370728),
('vibro', 0.6050527095794678),
('transfigo', 0.6030136346817017),
('sulpur', 0.5997458696365356),
('hastile', 0.5971342325210571),
('texentium', 0.5743292570114136),
('stipula', 0.5697602033615112),
('ferrum', 0.5689342617988586),

('liciatorium', 0.5648847818374634),
('spiculus', 0.5626999139785767)]

Neighbours in subcorpus 1 (medical)
[('emortuus', 0.0),
('neve', 0.0),
('contusum', 0.0),
('hyoscyamus', 0.0),
('linamentis', 0.0),
('ater', 0.0),
('masculus', 0.0),
('caldus', 0.0),
('desido', 0.0),
('pilus#2', 0.0)]

The lemma *aranea* does not occur in the medical subcorpus.

Cancer: crab > cancer

Neighbours in subcorpus 0 (non medical)
[('aequinoctialis', 0.7350070476531982),
('Capricornus', 0.7211254239082336),
('lune', 0.7174912691116333),
('Chelis', 0.7116904854774475),
('exortus', 0.7026336193084717),
('Orion', 0.6997062563896179),
('Hydrae', 0.696772038936615),
('Cancri', 0.6959371566772461),
('brumalis', 0.676031231880188),
('exoriri', 0.6754188537597656)]

Many of the closest neighbours to the lemma *cancer* in the non medical subcorpus refer to the meaning ‘Cancer’ intended in the astronomical sense of the constellation of Cancer.

Neighbours in subcorpus 1 (medical)
[('perniciosus', 0.9970979690551758),
('muto', 0.9970343112945557),
('abscido', 0.9969481825828552),
('animal', 0.9968684315681458),
('vetustas', 0.9966973066329956),
('afficio', 0.9966148734092712),
('occupo', 0.9965019226074219),
('desum', 0.9964045882225037),
('vinculum', 0.9962295889854431),
('caelum', 0.9959255456924438)]

The closest neighbours of *cancer* in the medical subcorpus are not as straighforward as in the case of *anus*. The adjective *perniciosus* might be interpreted as belonging to the medical

sense of ‘cancer’ as a type of disease, but among the neighbours we find *animal* ‘animal’, which refers to the meaning of *cancer* ‘crab’, *caelum* ‘sky’ which seems to refer back to the meaning of Cancer as the constellation.

Caninus: tooth of a dog > canine tooth

Neighbours in subcorpus 0 (non medical)
[('palo', 0.7080064415931702),
 ('palustrem', 0.7017350196838379),
 ('unctus', 0.6942863464355469),
 ('sucidus', 0.6911542415618896),
 ('pampinus', 0.6911504864692688),
 ('myrteum', 0.687336504459381 1),
 ('caedito', 0.686789333820343),
 ('refoveo', 0.686775803565979),
 ('Arsa', 0.6860144138336182),
 ('balsamum', 0.6847855448722839)]

Neighbours in subcorpus 1 (medical)
[('emortuus', 0.0),
 ('neve', 0.0),
 ('contusum', 0.0),
 ('hyoscyamus', 0.0),
 ('linamentis', 0.0),
 ('ater', 0.0),
 ('masculus', 0.0),
 ('caldus', 0.0),
 ('desido', 0.0),
 ('pilus#2', 0.0)]

The lemma *caninus* does not occur in the medical subcorpus.

Causa: cause/legal cause > the cause of a disease/a disease/the case of a disease

Neighbours in subcorpus 0 (non medical)
[('ratio', 0.6195141077041626),
 ('accusatio', 0.6100706458091736),
 ('quaestio', 0.5647055506706238),
 ('deliberatio', 0.5530708432197571),
 ('lite', 0.5506374835968018),
 ('crimen', 0.5385777354240417),
 ('defensio', 0.5352333784103394),
 ('disceptatio', 0.5333311557769775),
 ('lis', 0.5312601327896118),

('petitor', 0.5270451903343201)]

Medical notion is completely absent from this list. Many neighbours point towards the use of *causa* in its legal sense. The first one (*ratio*) is the direct synonym in the benchmark by Sprugnoli et al. (2020).

Neighbours in subcorpus 1 (medical)

[('ideoque', 0.9894729256629944),
('casus', 0.9861354827880859),
('propono', 0.9856301546096802),
('vinco', 0.9856215715408325),
('sanesco', 0.9854629635810852),
('solo', 0.9854286313056946),
('nosco', 0.9850139021873474),
('siquidem', 0.9836058020591736),
('evenio', 0.9828482270240784),
('praecipuus', 0.9825442433357239)]

Results are suboptimal. With *causa* the experiment does not work for its specialised meaning “disease” (nor for the supposedly more general meaning “cause of disease”). However, some neighbours point towards the use of *causa* as ‘disease’: see *sanesco* ‘to heal’.

Colis/Caulis: stalk/stem > penis

This lemma does not occur in the whole corpus

Fistula: pipe/flute > fistula

Neighbours in subcorpus 0 (non medical)
[('vicenum', 0.7571058869361877),
('centenum', 0.727862536907196),
('modulus', 0.7275359630584717),
('diametrum', 0.7077943682670593),
('plumbea', 0.7031548023223877),
('vicenarius', 0.6905785799026489),
('semuncia', 0.6882086992263794),
('pertundito', 0.6747350692749023),
('crocus', 0.6727489829063416),
('quadratus', 0.6727017164230347)]

Neighbours in subcorpus 1 (medical)
[('abscido', 0.9949777722358704),
('patefacio', 0.9944912195205688),
('membrana', 0.994250476360321),

('vinculum', 0.9940910935401917),
('cancer', 0.9940453171730042),
('rumpo', 0.9939383268356323),
('naturalis', 0.9937660694122314),
('quicumque', 0.9937065243721008),
('talus', 0.9936233758926392),
('ala', 0.9935315847396851)]

Folliculus: seedsack > follicle

Neighbours in subcorpus 0 (non medical)
[('pampinus', 0.8198416829109192),
('pineum', 0.8055068850517273),
('furfur', 0.8008548617362976),
('citrum', 0.796357274055481),
('callosus', 0.7912360429763794),
('lupinus', 0.7903788089752197),
('ficos', 0.787680983543396),
('sucidus', 0.7866107821464539),
('ulmeus', 0.7819167375564575),
('damascena', 0.78098064661026)]

Neighbours in subcorpus 1 (medical)
[('emortuus', 0.0),
('neve', 0.0),
('contusum', 0.0),
('hyoscyamus', 0.0),
('linamentis', 0.0),
('ater', 0.0),
('masculus', 0.0),
('caldus', 0.0),
('desido', 0.0),
('pilus#2', 0.0)]

The lemma *folliculus* does not occur in the medical subcorpus.

Impetus: attack > onset of a disease

Neighbours in subcorpus 0 (non medical)
[('eruptio', 0.718285858631134),
('incursus', 0.6353676319122314),
('concursus', 0.6190795302391052),
('acriter', 0.5763891339302063),
('concito', 0.5652066469192505),
('uiribus', 0.560851514339447),

('pavor', 0.5604428648948669),
('perrumpo', 0.559894323348999),
('inpetu', 0.5539030432701111),
('rabies', 0.5501555800437927)]

Neighbours in subcorpus 1 (medical)

[('scroto', 0.997705340385437),
('iacio', 0.9975391626358032),
('inbecillitas', 0.9973631501197815),
('mature', 0.9973325133323669),
('mano', 0.9972262382507324),
('numquam', 0.9970070719718933),
('aliquando', 0.9968353509902954),
('vaco', 0.996812105178833),
('muto', 0.9967483878135681),
('totus', 0.9967460632324219)]

Lenticula: lentil > freckle

Neighbours in subcorpus 0 (non medical)

[('perunges', 0.8608763813972473),
('asperso', 0.8605087995529175),
('cyperum', 0.8578701019287109),
('defritum', 0.8502394556999207),
('porrus', 0.8457130789756775),
('anethum', 0.8446499109268188),
('puleium', 0.8427358269691467),
('asperges', 0.841949999332428),
('silfi', 0.8419413566589355),
('pisum', 0.8400180339813232)]

The closest neighbours for the non medical texts point to the meaning of *lenticula* as a plant of lentils, or the lentil itself.

Neighbours in subcorpus 1 (medical)

[('lac', 0.9950354695320129),
('ruta', 0.9947835803031921),
('olea', 0.99403315782547),
('marrubium', 0.9940274357795715),
('rubus', 0.9916974902153015),
('adjicio', 0.9916856288909912),
('sucidus', 0.9906005263328552),
('malicorium', 0.9898480772972107),
('farina', 0.9896025061607361),

('furfur', 0.9895826578140259)]

In this case, from the analysis on the sub corpus of medical texts the meaning ‘freckle’ does not emerge. Instead among the closest neighbours there is a strong presence of herbs that could be used for medicinal purposes.

Lumen: light > faculty of sight

Neighbours in subcorpus 0 (non medical)

[('fulgor', 0.7066628932952881),
('lux', 0.6773566603660583),
('lucidus', 0.6475052237510681),
('radius', 0.6295573115348816),
('candor', 0.6255735754966736),
('caligo', 0.6172627210617065),
('luceo', 0.592840313911438),
('inlumino', 0.5927273035049438),
('splendor', 0.5899431109428406),
('caliginem', 0.5879647135734558)]

Neighbours in subcorpus 1 (medical)

[('expeditus', 0.9964432120323181),
('plerumque', 0.9961612224578857),
('paene', 0.9961373805999756),
('restituo', 0.9958831667900085),
('curiosus', 0.9958614706993103),
('infans', 0.9958545565605164),
('capillus', 0.9958263635635376),
('adversus', 0.9957664012908936),
('prolabor', 0.9956214427947998),
('planta', 0.995497465133667)]

Malum: a bad thing > a disease

Neighbours in subcorpus 0 (non medical)

[('malus#3', 0.7445393800735474),
('deterior', 0.6224241256713867),
('publicatio', 0.550349771976471),
('conscientia', 0.5132147073745728),
('fraudis', 0.503659725189209),
('invidus', 0.5022456645965576),
('malignitas', 0.48834073543548584),
('nequitia', 0.48028719425201416),
('malitia', 0.47844555974006653),
('pessimi', 0.4778210520744324)]

The medical sense for *malum* is absent from the non medical corpus.

Neighbours in subcorpus 1 (medical)

[('difficultas', 0.9630550742149353),
 ('articulus', 0.9567130208015442),
 ('consuesco', 0.9554041624069214),
 ('male', 0.9552204012870789),
 ('urina', 0.9537398219108582),
 ('voluntas', 0.9534554481506348),
 ('spes', 0.9531292915344238),
 ('sensus', 0.9526328444480896),
 ('insania', 0.9525879621505737),
 ('vesica', 0.9525068998336792)]

The synonym for *malum* (*morbus*) ‘disease’ does not appear among the closest neighbours. However, some of the lemmas in the list seem to point towards a use of the word related to the body (*urina* ‘urine’ and *vescica* ‘bladder’). The neighbour *insania* ‘madness’ gets closer to the notion of disease.

Menstrua: monthly > menstrual cycle

Neighbours in subcorpus 0 (non medical)

[('quintadecima', 0.6784973740577698),
 ('coagulo', 0.6650052070617676),
 ('effluxerit', 0.6644333600997925),
 ('caedito', 0.6587182283401489),
 ('Octobrium', 0.6507802605628967),
 ('seruentur', 0.6498989462852478),
 ('putrescit', 0.6493684649467468),
 ('decembri', 0.6474031805992126),
 ('autumnalis', 0.6436105966567993),
 ('uespera', 0.6390299797058105)]

Neighbours in subcorpus 1 (medical)

[('emortuus', 0.0),
 ('neve', 0.0),
 ('contusum', 0.0),
 ('hyoscyamus', 0.0),
 ('linamentis', 0.0),
 ('ater', 0.0),
 ('masculus', 0.0),
 ('caldus', 0.0),

('desido', 0.0),
('pilus#2', 0.0)]

The lemma *menstruus* does not occur in the medical subcorpus.

Mola: millstone > molar teeth (plural form)

Neighbours in subcorpus 0 (non medical)
[('plumbum', 0.746660053730011),
('aggero', 0.7344804406166077),
('cratis', 0.7252821922302246),
('bitumen', 0.703269362449646),
('crates', 0.7016198635101318),
('asinarius', 0.7005409002304077),
('funis', 0.6971569061279297),
('lamminis', 0.6944992542266846),
('condensus', 0.6931260228157043),
('coagmento', 0.6929770112037659)]

Neighbours in subcorpus 1 (medical)

[('emortuus', 0.0),
('neve', 0.0),
('contusum', 0.0),
('hyoscyamus', 0.0),
('linamentis', 0.0),
('ater', 0.0),
('masculus', 0.0),
('caldus', 0.0),
('desido', 0.0),
('pilus#2', 0.0)]

The lemma *mola* does not occur in the medical subcorpus.

Musculus: mouse > muscle

Neighbours in subcorpus 0 (non medical)
[('cratis', 0.8192490935325623),
('pluteum', 0.7833704352378845),
('tabulatus', 0.7824996113777161),
('ballista', 0.7676939368247986),
('vimen', 0.7657342553138733),
('trabes', 0.7561563849449158),
('stipites', 0.7425310015678406),
('sudibus', 0.7417828440666199),
('tignum', 0.7387229800224304),
('corium', 0.7376952171325684)]

Neighbours in subcorpus 1 (medical)
[('rumpo', 0.9939658641815186),
 ('incisum', 0.9939491152763367),
 ('exeo', 0.9936187267303467),
 ('subsum', 0.9931867122650146),
 ('decido', 0.993084192276001),
 ('mors', 0.992587685585022),
 ('vitio', 0.9925201535224915),
 ('cucurbitula', 0.9923760294914246),
 ('appareo', 0.9922506809234619),
 ('tumeo', 0.9920444488525391)]

Patella: small bowl > kneecap

Neighbours in subcorpus 0 (non medical)
[('patina', 0.8819414973258972),
 ('caccabus', 0.8786182999610901),
 ('oenogaro', 0.8742801547050476),
 ('mortarium', 0.8705757260322571),
 ('frico', 0.8621558547019958),
 ('cocta', 0.8574983477592468),
 ('perunges', 0.8552541732788086),
 ('liquamine', 0.8513573408126831),
 ('surclas', 0.8509286642074585),
 ('furnus', 0.8497797250747681)]

Neighbours in subcorpus 1 (medical)
[('emortuus', 0.0),
 ('neve', 0.0),
 ('contusum', 0.0),
 ('hyoscyamus', 0.0),
 ('linamentis', 0.0),
 ('ater', 0.0),
 ('masculus', 0.0),
 ('caldus', 0.0),
 ('desido', 0.0),
 ('pilus#2', 0.0)]

The lemma *patella* does not occur in the medical subcorpus

Pecten: comb > pubic bone (comb-like thing)

The lemma *pecten* does not occur in the whole corpus.

Plaga: blow > wound, incision

Neighbours in subcorpus 0 (non medical)
[('contritio', 0.49932730197906494),
 ('septentrionalis', 0.47354528307914734),
 ('aquilo', 0.4706868827342987),
 ('pessima', 0.4643608033657074),
 ('septentrionem', 0.4558558166027069),
 ('meridies', 0.4414684474468231),
 ('desolo', 0.43876615166664124),
 ('serpens', 0.43028056621551514),
 ('languor', 0.43025338649749756),
 ('sibilo', 0.4296463131904602)]

Neighbours in subcorpus 1 (medical)

[('crus', 0.9959562420845032),
 ('ibi', 0.994205892086029),
 ('scalpellum', 0.9938284158706665),
 ('extendo', 0.9937617182731628),
 ('descendo', 0.9934535622596741),
 ('excidium', 0.993284285068512),
 ('ergo', 0.9932499527931213),
 ('proximus', 0.9931142330169678),
 ('integer', 0.9930745363235474),
 ('iacio', 0.9929271936416626)]

The only neighbour that points to the meaning of *plaga* as 'wound' is *crus*, which means 'leg, shin'.

Potio: drink > medicinal draught

Neighbours in subcorpus 0 (non medical)
[('potus', 0.596651554107666),
 ('sopio', 0.5308095216751099),
 ('quies', 0.5303115844726562),
 ('medicamentum', 0.5270382165908813),
 ('cibus', 0.5206639766693115),
 ('satio', 0.515001654624939),
 ('potionis', 0.5114859342575073),
 ('somnoque', 0.5112212896347046),
 ('haurio', 0.5046415328979492),
 ('adlici', 0.4984215795993805)]

Surprisingly, the medical meaning associated with *potio* 'remedy, medicine' appears in the closest neighbours for the non-specialised subcorpus, and not for the specialised one. The use of *potio* in the sense of medical remedy might have been more frequent in the common language than its general meaning 'drink'.

Neighbours in subcorpus 1 (medical)
[('utor', 0.9829772114753723),
 ('abstineo', 0.9741743803024292),
 ('pridie', 0.9734511971473694),
 ('uti', 0.9712179899215698),
 ('egelidus', 0.9695520401000977),
 ('adsumo', 0.9688200354576111),
 ('acer', 0.9610457420349121),
 ('sumo', 0.9605630040168762),
 ('modicus', 0.9598569869995117),
 ('debeo', 0.9592777490615845)]

However, there are some interesting words among the closest neighbours for *potio* in the medical subcorpus. The verbs *adsumo* and *sumo* 'to receive, to take up' seem to refer to a substance that is imposed on the subject, rather than a drink that the subject willingly drinks. The verb of necessity *debeo* seems to point in that direction with a stronger nuance.

Pupilla: young girl, ward > pupil

Neighbours in subcorpus 0 (non medical)
[('pupillus', 0.6586187481880188),
 ('tutor', 0.6093322038650513),
 ('administrasse', 0.6037088632583618),
 ('tutela', 0.5639089941978455),
 ('obligatus', 0.5460118651390076),
 ('curator', 0.5336513519287109),
 ('patrona', 0.5176621675491333),
 ('vidua', 0.5000547766685486),
 ('pupillaris', 0.4934292137622833),
 ('pignus', 0.48506319522857666)]

Neighbours in subcorpus 1 (medical)
[('commode', 0.9960851073265076),
 ('protraho', 0.995667040348053),
 ('exaspero', 0.995612621307373),
 ('coicienda', 0.9955204129219055),
 ('specillo', 0.9954385161399841),
 ('inde', 0.995216429233551),
 ('labrum', 0.9949774146080017),
 ('mamma', 0.9946674108505249),
 ('conpremo', 0.9946157932281494),

('testa', 0.994404137134552)]

Scrotum: leather bag for arrows > scrotum

The lemma *scrotum* does not occur in the whole corpus.

Spina: thorn > backbone

Neighbours in subcorpus 0 (non medical)
[('folium', 0.7782173752784729),
 ('vepres', 0.7522402405738831),
 ('stramentum', 0.738955020904541),
 ('locusta', 0.7326549291610718),
 ('oliva', 0.7314444184303284),
 ('testa', 0.7297854423522949),
 ('arista', 0.721219003200531),
 ('spica', 0.7201421856880188),
 ('granum', 0.719623863697052),
 ('virgultum', 0.7191399931907654)]

Neighbours in subcorpus 1 (medical)
[('vertebra', 0.9980313777923584),
 ('retundo', 0.9976984858512878),
 ('uncus', 0.9976956248283386),
 ('sinuo', 0.997674822807312),
 ('apprehendo', 0.9971036911010742),
 ('cerebrum', 0.9969908595085144),
 ('inhaeresco', 0.9969809055328369),
 ('costa', 0.996886730194091 8),
 ('rima', 0.9968589544296265),
 ('transversum', 0.9968309998512268)]

Not the exact synonym, but very close to the meaning of ‘back’, *dorsum.*

Spiritus: breath, air > intestinal gas

Neighbours in subcorpus 0 (non medical)
[('trinitas', 0.5778629779815674),
 ('angelus', 0.5743215680122375),
 ('Trinitatis', 0.5146111845970154),
 ('diabolus', 0.5006036758422852),
 ('invocatio', 0.4707029163837433),
 ('inspiratio', 0.46047791838645935),
 ('corruptio', 0.4583539068698883),
 ('caro', 0.4577744901180267 3),
 ('angelicus', 0.4566404223442077 6),

('anima', 0.44951263070106506)]

Neighbours in subcorpus 1 (medical)
[('nam', 0.9927971959114075),
('intus', 0.9913724064826965),
('pus', 0.9886324405670166),
('frango', 0.9884505271911621),
('tumor', 0.9881983995437622),
('relinquo', 0.988068163394928),
('motus', 0.9880082011222839),
('febre', 0.9873912930488586),
('nonnumquam', 0.9869715571403503),
('intendo', 0.9869352579116821)]

Tibia: flute > shin bone

Neighbours in subcorpus 0 (non medical)
[('fides#2', 0.7541792988777161),
('cantus', 0.7499963641166687),
('cymbalum', 0.7445164918899536),
('lyra', 0.7413572072982788),
('cithara', 0.7400211095809937),
('tympanum', 0.7399448752403259),
('tinnitus', 0.7395069003105164),
('organum', 0.7257954478263855),
('plectrum', 0.7043542861938477),
('chorda', 0.6989166736602783)]

Neighbours in subcorpus 1 (medical)
[('emortuus', 0.0),
('neve', 0.0),
('contusum', 0.0),
('hyoscyamus', 0.0),
('linamentis', 0.0),
('ater', 0.0),
('masculus', 0.0),
('caldus', 0.0),
('desido', 0.0),
('pilus#2', 0.0)]

The lemma *tibia* does not occur in the medical corpus.

Vitium: sin, fault > disease

Neighbours in subcorpus 0 (non medical)

[('avaritia', 0.6334289312362671),
('uitiis', 0.62870192527771),
('inertia', 0.6192713379859924),
('vitiosus', 0.6179677248001099),
('culpa', 0.6062299609184265),
('stultitia', 0.6019758582115173),
('malitia', 0.5940094590187073),
('insolentia', 0.5920464992523193),
('neglegentia', 0.5816465616226196),
('nequitia', 0.5792486667633057)]

The closest neighbours for the non medical subcorpus suggest the meaning 'sin, fault', by referring to various subtypes of fault: *avaritia* 'greed'*, culpa* 'fault'*, stultitia* 'stupidity'*, malitia* 'ill will, malice'*, insolentia* 'haughtiness'*, neglegentia* 'carelessness'*, nequitia* 'wickedness'*.*

Neighbours in subcorpus 1 (medical)
[('video', 0.9915851354598999),
('ferus', 0.9847535490989685),
('medicina', 0.9810498952865601),
('casus', 0.9806272387504578),
('sub', 0.9806148409843445),
('autem', 0.9795443415641785),
('incido#2', 0.978618860244751),
('auxilium', 0.9781866073608398),
('orior', 0.9775609374046326),
('causa', 0.9765673875808716)]

The closest neighbours of the lemma *vitium* in the medical subcorpus suggest in some cases the meaning 'disease' e.g. in the case of *medicina* 'medicine', *causa* 'disease'

Concluding remarks

For none of the 25 words selected for the analysis were word embeddings trained on the medical subcorpus able to directly identify the synonym in the medical field. This might be due to the limited size of the medical subcorpus, and to the low frequency of some of the words in Latin e.g. *pecten.*
In only one case was such a synonym identified in the non-medical corpus: for *potio* with the meaning 'medicinal draught' the closest neighbours for the word embeddings trained on the non medical subcorpus include the lemma *medicamentum,* which was identified as the direct synonym for *potio* in the medical context.
However, it should be noted that, although the direct synonyms were not listed among the closest neighbours, for 11 words in the list at least one of the ten closest neighbours belongs to the semantic frame of the medical synonym.

	N° of words	Lemmas
The synonym appears as closest neighbour	1	*Potio*[1]
(Some of) the closest neighbours belong to the semantic frame of the medical sense	10	*Anus, cancer, causa, fistula, impetus, malum, plaga, pupilla, spina, vitium*
No synonyms nor neighbours	4	*Lenticula*[2]*, lumen, musculus, spiritus*
Does not occur in the medical subcorpus	7	*Aranea, caninus, folliculus, menstrua, mola, patella, tibia*
Does not occur in the whole corpus	3	*Colis, pecten, scrotum*

Table 2. Results of word embeddings trained on the medical subcorpus for the 25 lemmas in the list.

Appendix. Gold Standard

In this Appendix I show the Gold Standard created during this STSM (see section 'Gold Standard' in this document).

Target	**Synonym**	**Random1**	**Random2**	**Random3**	**Random4**
Anus	culus	incola	legatus	priuo	media
Aranea	morbus	plausus	nego	perscribo	luteus
Cancer	vulnus	arcesso	ordo	occulte	longaeuus
Caninus	dens	acrimonia	salus	sempiternus	mutabilis
Causa	morbus	contingit	moueo	lictor	lambo
Colis/caulis	penis	accliuis	inuenio	pirata	manumitto
Fistula	vulnus	decliuis	quasi	postumus	mauretania
Folliculus	utriculus	inclamo	signum	oraculum	lupa
Impetus	morbus	incendo	loquor	insolitus	iubar

[1] But the synonym appears among the closest neighbours in the non medical subcorpus.
[2] The closest neighbours of *lenticula* belong to the medical context, but they do not belong to the semantic frame of 'freckle', therefore I excluded them from '(Some of) the closest neighbours belong to the semantic frame of the medical sense'.

Lenticula	lentigo	propinquo	frater	inhibeo	intromitto
Lumen	visum	cumulate	quin	stabilis	niualis
Malum	morbus	euenio	ualeo	medicina	legatum
Menstrua	decurso	praecingo	regnum	moderor	leonnatus
Mola	dens	commode	aduersus	profugus	menapii
Musculus	lacertus	arripio	puer	murena	liuius
Patella	os genus	apto	occursatio	procurro	medus
Pecten	os pubis	appositus	sententia	rigidus	mollitia
Plaga	ulcus	accessio	sentio	latium	iulianus
potio	medicamentum	redundo	defendo	illo	iniuste
Pupilla	oculus	delator	hinc	scriptor	munificentia
Scrotum	follis	discumbo	crimen	rite	monitum
Spina	dorsum	crudus	finis	sosia	necopinans
Spiritus	flatus	incuso	cur	sacerdotium	munditia
Tibia	sura	inclino	fortis	petitio	manilius
vitium	morbus	uehemens	seruus	sella	munusculum

References

Dumesnil, J. B. G. (1819). *Latin synonyms: With their different significations: and examples taken from the best Latin authors*. GB Whittaker.

Hill, J. (1804). *The Synonymes in the Latin Language, Alphabetically Arranged; with Critical Dissertations Upon the Force of Its Prepositions, Both in a Simple and Compounded State: By John Hill, LL. D. Professor of Humanity in the University, and Fellow of the Royal Society of Edinburgh*. London: Longman and Rees.

Koptjevskaja-Tamm, Maria. (2016) The lexical typology of semantic shifts: An introduction. In P. Juvonen and M. Koptjevskaja-Tamm (Eds.), *The lexical typology of semantic shifts.* Berlin, Boston: Mouton de Gruyter.

Langslow, D. R. (2000). *Medical Latin in the Roman Empire*. Oxford: Oxford University Press.

McGillivray, B. & Kilgarriff, A. (2013). Tools for historical corpus research, and a corpus of Latin. In P. Bennett, M. Durrell, S. Scheible, R. J. Whitt (Eds.), *New methods in Historical Corpus Linguistics*. Tübingen: Narr.

McGillivray, B. & Nowak, K. (2022). Tracing the semantic change of socio-political terms from Classical to early Medieval Latin with computational methods. In *Latin vulgaire – latin tardif XIV. 14th International Colloquium on Late and Vulgar Latin. September 5-9, 2022, Ghent University.* Book of Abstracts. Ghent University. https://www.lvlt14.ugent.be/wp-content/uploads/2022/09/LVLT14-Book-of-abstracts.pdf (last accessed date: 31/01/2023).

Ribary, M, and McGillivray, B. (2020). A corpus approach to Roman law based on Justinian's digest. *Informatics*. Vol. 7. No. 4. MDPI.

Rodda, M, Probert, P. and McGillivray, B. (2019). "Vector space models of Ancient Greek word meaning, and a case study on Homer." *Traitement Automatique Des Langues* 60.3

Skřivan, A. (1890). *Latinská synonymika pro školu i dum*. V CHRUDIMI.

Sprugnoli, R., Passarotti, M. and Moretti, G. (2019). Vir is to Moderatus as Mulier is to Intemperans-Lemma Embeddings for Latin. *CLiC-it*.

Sprugnoli, R., Moretti, G. and Passarotti, M. (2020). Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas. *IJCoL. Italian Journal of Computational Linguistics* 6.6-1: 29-45.

ThLL = Thesaurusbüro München Internationale Thesaurus-Kommission (Ed.) (1900–). *Thesaurus Linguae Latinae*. Berlin: De Gruyter.

Von Doederlein, L. and Taylor, S. H. (1875). *Döderlein's hand-book of Latin synonymes*. WF Draper.