



## INSTITUTIONEN FÖR SPRÅK OCH LITTERATURER

# CORPUS CHRISTI

En korpuslingvistisk studie av latinets semantiska utveckling i kristendomens spår.

**David Lafage**

---

Uppsats/Examensarbete:	Masteruppsats 30 hp
Program och/eller kurs:	LAT240
Nivå:	Avancerad nivå
Termin/år:	HT24
Handledare:	Sara Moding
Examinator:	Erik Bohlin
Rapport nr:	xx (ifylles ej av studenten/studenterna)

# Abstract

Uppsats/Examensarbete: Masteruppsats 30 hp  
Program och/eller kurs: LAT240  
Nivå: Avancerad nivå  
Termin/år: HT24  
Handledare: Sara Moding  
Examinator: Erik Bohlin  
xx (ifylles ej av studenten/studenterna)

Rapport nr:

Keywords: NLP, LatinBERT, XPLatinBERT, corpus linguistics, distributive semantics, Christian Latin, *Sondersprache*, word embeddings, MLM, machine-learning.

---

The objective of this master thesis is to measure the effect of Christianity on Latin semantics with a specially trained Large Language Model and departing from a corpus-driven approach. First, I am investigating if we can confirm that words selected in the literature about Christian Latin *de facto* have undergone a measurable semantic shift in the Christian age, and if we can enrich this list with previously unnoticed words. Next, I want to find out if the results differ significantly depending on how Christian Latin is defined.

The methodology is based on theories of distributional semantics and the Distributional Hypothesis, and follows other works in the field. First, an existing BERT model (LatinBERT) is trained on the *Patrologia Latina* corpus, under the assumption that this corpus is representative of Christian Latin. An algorithm is then selected from a metastudy to perform a Graded Change Detection and three different tests are performed in order to evaluate the model's performance. Finally, the results are computed and analyzed quantitatively and qualitatively, and inferential statistics are applied to the data.

The results show that the new model, XPLatinBERT (XPL), outperforms the SemEval2020 models for Latin on a benchmark based on a similar task. By and large all the words in the literature on Christian Latin are confirmed and other words are proposed by using a corpus-based (the third quantile) and a corpus-driven approach. Due to lemmatization issues in the corpora under investigation, some words are false-positives, which calls for a deeper qualitative investigation of the results.

Although a difference can be observed in the dataset as a whole, as well as on specific words, this difference is not strong enough to be statistically significant. It is therefore possible to consider Christian Latin as a register and to regard deviations as an effect of other factors such as Late and Medieval Latin, but more work has to be done.

XPL can now be found on Github (Lafage, 2025b).

# Förord

Till minne av Karin Hult, som handledde en av mina kandidatuppsatser och vars uppmuntran fick mig att fortsätta på den inslagna NLP-banan och läsa en masterutbildning i latin.

Denna uppsats har tillkommit tack vare så många andra människors bidrag, att min roll nog varit mer dirigentens än kompositörens: jag har egentligen bara pusslat ihop saker som andra mödosamt arbetat med. Därutöver har enstaka personers råd eller bidrag i kritiska lägen räddat projektet och dem vill jag rikta ett särskilt tack till:

Robert Adesam och Kaj Ailomaa, Göteborgs universitet.

Andreas Berglind, Göteborgs universitet.

Richard Johansson, Chalmers tekniska högskola.

Francesco Periti, KU Leuven.

Philipp Roelli, Universität Zürich.

Nina Tahmasebi, Göteborgs universitet.

Tack också till Johan Ahlberg och Rickard Roxvall på Motorit AB, som erbjudit mig anpassade anställningsformer för att frigöra tid till uppsatsen, och till examinatorn och handledaren för att ha erbjudit mig att läsa kursen på halvfart. Slutligen vill jag rikta ett särskilt tack till min make Adrián Fernández-Pello, som under de klosterlika former som uppsatsskrivandet (passande nog) i sitt slutskede tog, skötte bigården och andra sysslor så att jag till fullo kunde koncentrera mig på arbetet med kristet latin.

# Innehållsförteckning

1	Inledning.....	6
1.1	Bakgrund.....	6
1.2	Syfte och frågeställningar .....	7
1.3	Avgränsning.....	7
1.4	Motivering och ämnesrelevans .....	8
2	Teoretiskt ramverk .....	9
2.1	Kristet latin.....	9
2.1.1	Sondersprache.....	9
2.1.2	Tidsaspekten .....	10
2.1.3	Kategorisering .....	10
2.1.4	Lista över semantiskt utvidgade lemman .....	11
2.2	Distributiv semantik.....	12
2.3	Distributiv semantik och NLP .....	13
2.3.1	Lemma, type, token .....	13
2.3.2	Vektorer och word embeddings.....	13
2.3.2.1	Frekvensbaserade vektorer.....	13
2.3.2.2	Statiska embeddings.....	15
2.3.2.3	Kontextuella embeddings.....	15
2.3.3	Relevant forskning inom distributiv semantik.....	16
2.3.4	Kritik mot DH och word embeddings .....	17
3	Material och metod.....	18
3.1	Material.....	18
3.1.1	Patrologia Latina.....	18
3.1.2	Förkristet material.....	20
3.2	Metod.....	20
3.2.1	Förarbete.....	20
3.2.2	Träning och testning av XPLatinBERT.....	20
3.2.3	Graded Change Detection.....	22
3.2.3.1	Average Pairwise Distances (APD) .....	22
3.2.3.2	Word prototype (PRT) .....	23
3.2.4	Uträkning av APD och PRT per lemma .....	23
3.2.5	Avgränsning .....	23
3.2.5.1	Ordklasser och frekvenser.....	24
3.2.5.2	Tidsaspekten .....	24
4	Resultat.....	25

4.1	Framtagning av XPLatinBERT.....	25
4.1.1	Initiala överväganden .....	25
4.1.2	Deskriptiv statistik.....	26
4.1.3	Träning av XPLatinBERT .....	29
4.1.3.1	Inledande reflektioner .....	29
4.1.3.2	Resultat av träningen.....	31
4.1.4	Testning av XPLatinBERT.....	32
4.1.4.1	Test 1 – kvantitativ bedömning på testsetet.....	32
4.1.4.2	Test 2 – kvalitativ bedömning med MLM-task .....	33
4.1.4.3	Test 3 – kvantitativ benchmarkstudie .....	37
4.1.4.4	Sammanfattning av testresultaten .....	39
4.2	Extraktion av data i LA.....	40
4.3	Extraktion av embeddings.....	42
4.4	Diakron semantisk variation .....	43
4.4.1	APD-resultat för PL.....	44
4.4.1.1	Substantiv.....	44
4.4.1.2	Adjektiv .....	48
4.4.2	Sammanfattning av resultaten .....	52
4.4.3	Tröskelnivå för semantisk variation .....	53
4.4.4	APD-värde för lemman i litteraturen om kristet latin.....	54
4.4.5	Inferentiell statistik .....	57
4.5	Sammanfattande slutsatser och diskussion .....	63
4.5.1	$F_1$ .....	63
4.5.2	$F_2$ .....	63
4.5.3	Diskussion .....	63
4.5.4	Metodologisk diskussion .....	64
5	Förkortningar och termer.....	66
6	Referenslista .....	67
6.1	Tryckt litteratur .....	67
6.2	Webbsidor och blogginlägg .....	69

# 1 Inledning

Latin skulle kunna beskrivas som ett renodlat ”korpusspråk”. Liksom alla historiska språk saknar det nämligen en levande språkgemenskap som kan producera nytt (muntligt) material: allt som skrivits på latin kan anses bilda en färdig korpus, som i kraft av sin ovanligt stora historiska räckvidd lämpar sig för diakrona korpuslingvistiska studier. Detta flertusenåriga tidsspann som rymmer alla de texter som skrivits på latin möjliggör särskilt studier av specifika lexikala enheters semantiska utveckling. Som vi snart skall se, har utvecklingen av nya metoder de senaste åren lett till en betydande ökning av just den typen av studier. McGillivray beskriver hur nya datalingvistiska metoder håller på att förändra latinsk lingvistik och kallar detta nya fält ”latinsk datalingvistik” (2014:1–6). Materialet är fortfarande samma latinska texter; det är metoden som skiljer sig.

Avsnitt 1 ger en introduktion till ämnet, samt uppsatsens syfte och frågeställningar. Kapitlet följs av en genomgång av relevant forskning (avsnitt 2). Eftersom läsaren inte nödvändigtvis känner till hur LLM (Large Language Models) fungerar, ges en kortfattad genomgång, som även kan användas som introduktion till ämnet. Sedan fortsätter uppsatsen med material och metod (avsnitt 3) och resultaten i avsnitt 4. I avsnitt 5 hittar läsaren en översikt över förkortningar. Engelska nyckelord introduceras då den vetenskapliga litteraturen inom latinsk datalingvistik domineras av just det språket, men de ges i görligaste mån en svensk språkdräkt.

Notera att jag återger latinsk text med standardiserad stavning, enligt vilken bl.a. *v* skrivs *u* och *j* återges med *i*. Detta är standardprocedur i latinsk datalingvistik. Engelska ord stavar jag medvetet enligt brittiska stavningsregler.

## 1.1 Bakgrund

Forskare har sedan 1800-talet uppmärksammat kristendomens effekt på det latinska språkets lexikala och semantiska utveckling. Olika strategier används för att beteckna de koncept som uppstår med den nya religionen: nya ord bildas, etablerade ord får en ny betydelse, som antingen lever parallellt med den ursprungliga eller ersätter den helt. 1800-talets kategoriseringsvurm ledde till samlingstermen ”kristet latin” för att beteckna summan av ovannämnda lexikala och semantiska säregenheter, men termen är svårdefinierad, både i tid och rum (López-Silva, 2003:117). Den s.k. Nijmegen-skolan, har mellan 1930- och 1970-talet gått så långt som att anfäktat idén om kristet latin som en särskild språkvarietet (*Sondersprache*), eller det vi idag kanske snarare skulle kalla ”sociolekt”, inom det latinska språket. Teorin, som inte gick att bevisa med annat än sporadiska nedslag i litteraturen, föll i glömska efter svåra fejder som involverade 1900-talets stora filologer (Denecker, 2018).

Inom det växande fältet som ovan kallats latinsk datalingvistik har de senaste decennierna stora korpora bearbetats med språkteknologier under paraplytermen NLP (Natural Language Processing), som står för utvecklingen av algoritmer och verktyg för automatisk eller semiautomatisk analys av språklig data (McGillivray, 2022b:2). Sedan genombrottet med språkmodeller som BERT (Bidirectional Encoder Representation from Transformers) omkring 2018, har fältet varit särskilt produktivt. Äldre teorier som inte gick att bevisa empiriskt har dammats av och utgör nu utgångspunkten för nya rön. Just företeelsen ”kristet latin” och andra specialiserade varieteter som ”rättslatin” har alldeles nyligen varit föremål för den typen av studier, som vi skall se i avsnitt 2.

Trots fältets växande popularitet bland filologer och de stora fördelar de kan hämta därur, trots en stor sammansvetsad online community som publicerar material för att underlätta ”inkörningen”<sup>1</sup> och stödet som AI-verktyg numera erbjuder, finns det i Sverige idag få kurser i NLP som riktar sig specifikt

---

<sup>1</sup> Jag vill tipsa om en bok av Patrick J. Burns som snart borde komma ut, och som förväntas få stor betydelse som introduktion till programmering och NLP för klassiska filologer. Jfr. litteraturlista.

till humanister. Det var en förmån för mig att komma in på en av de få kurserna i Sverige i programmering specifikt för humanister utan andra förkunskapskrav än en kandidatexamen<sup>2</sup>, samt i korpuslingvistik<sup>3</sup>, båda kurser på Linnéuniversitetet.

## 1.2 Syfte och frågeställningar

Syftet med denna uppsats är att mäta effekten av kristendomen på latinets semantik. Som vi kommer att se, använder sig Burton av ett kvalitetskriterium för att avgränsa det svårfångade begreppet ”kristet latin” diakront, som består i att endast inkludera texter författade mellan 200 e.Kr. och ca 600 e.Kr (2011:486). Frågeställningen bryts ner i två moment.

F<sub>1</sub>) Kan vi med hjälp av språkteknologiska verktyg och metoder bekräfta en mätbar semantisk förändring under kristendomens inflytande i substantiv och adjektiv som i litteraturen upptas som exempel på det som kallats ”kristet latin”, och kan vi i sådant fall berika denna lista med ytterligare ord?

F<sub>2</sub>) Har Burtons avgränsning av ”kristet latin” i en specifik tidsperiod en statistiskt säkerställd effekt på den observerade variationen jämfört med att betrakta det som en tidlös företeelse, och kan vi i sådant fall peka ut ord som mest påverkas av det?

## 1.3 Avgränsning

I denna uppsats studerar jag endast renodlade fall av semantisk variation (och bortser alltså från lexikal variation). Eftersom uppsatsen vidare utgår från specifika antaganden kring polysemi och homonymi definieras dessa termer här.

Polysemi är i det här fältet ett så grundläggande begrepp att det sällan definieras. Enligt Nationalencyklopedin är polysemt ett ord som har flera olika betydelser, som *lätt* i betydelserna ”som väger litet” och ”som inte är svårt” (NE, u.å., *polysemi*). Två ord kan råka ha samma form men två helt olika etymologier, i vilket fall vi talar om *homonymi*. T.e.x delas *ius*, *iuris*, *n* i L&S upp i två uppslag: det ena har etymologin ”kindred to Sanscr. *yūsh*, the same” med hänvisning till Gr. *ζωμός* och betydelserna ”broth, soup, sauce” samt ”juice, mixture”<sup>4</sup>, medan det andra har mycket vanligare betydelser kopplade till ”right, law, justice” och etymologin ”kindred with Sanscr. *yu*, to join” med hänvisning till *ζέβυνοι*, *iungo* samt *lex* from *ligo*<sup>5</sup> (McGillivray et al., 2022a:89–90).

Märk att vissa semantiker, som Tahmasebi et al., av praktiska skäl inte skiljer operativt mellan polysemi och homonymi (2021:6), eftersom detta inte alltid låter sig operationaliseras med NLP-verktyg. Av samma anledningar följer jag detta utvidgade synsätt, även om jag i denna uppsats använder mig av metadata (lemma- och ordklassannotering), vilket i teorin borde reducera effekten av polysemi och homonymi något. Jag gör vidare inga anspråk på WSD (Word Sense Disambiguation, McGillivray et al., 2022a:58), även kallat WSI (Word Sense Induction, Tahmasebi et al., 2021:41), vilka är tekniker som används inom NLP för att skilja mellan polysemers olika betydelser. Detta innebär att jag operationaliserar semantisk variation som förändringen av ett eventuellt polysemt lemmas *alla* betydelser, vilket är en tekniskt sett tacksam förenkling som tas upp i Tahmasebi et al. (2021:63).

---

<sup>2</sup> Programming for Digital Humanities, 15 ects. Se litteraturlistan.

<sup>3</sup> Corpus Methods in Practice, 15 ects. Se litteraturlista.

<sup>4</sup> L&S s.v. *jus*, *juris*, *n* 1.

<sup>5</sup> L&S s.v. *jus*, *juris*, *n* 2.

## 1.4 Motivering och ämnesrelevans

Kvalitativa studier om kristet latin har hittills kommit fram till teorier med begränsad förankring i empirin, i vilka några få ord tas upp som exempel för en hel population, utan hänvisning till hur frekventa de är. Sådana studier är mycket användbara som utgångspunkt i det att de bygger på framstående filologers ackumulerade intuition om det latinska språket. Problemet är dock att författarna omöjligen kan veta hur representativa deras i enstaka nedslag handplockade exempelord är i förhållande till andra. Själva begreppet ”kristet latin” såsom det definierats i den typen av studier är idag omtvistat, och detta beror till stor del på att det för samma aspekter som modern lingvistik gärna håller isär. Det är t.ex. oklart hur termen exakt förhåller sig till närbesläktade begrepp som ”de kristnas latin”, ”bibliskt latin”, ”Altchristliches Latein”, ”patristisches Latein” eller ”Kirchenlatein”. Sist men inte minst är en tidsram för kristet latin ofta inte angiven, och man undrar om det betraktas som en synkron eller diakron företeelse. Till skillnad från t.ex. Mohrmann (1977) gör Burton ett försök till temporal avgränsning av kristet latin mellan 200 e.Kr. och ca 600 e.Kr, med argumentet att man kan anse det material som producerats under den perioden komma från modersmålstalare (2011:486).

Därför har temat på senare år dragit korpuslingvisters uppmärksamhet. Hittills har datalingvistiska studier över kristet latin främst utgått från de ordlistor som de kvalitativa studierna kommit fram till (korpusbaserad forskning, på engelska *corpus-based*), främst med det metodologiska syftet att förbättra tekniken eller utgöra en s.k. goldstandard (manuellt annoterat dataset som används för att träna eller utvärdera maskininlärningsmodeller). Sprugnoli et al. (2020) har visserligen gjort ett försök till korpusdriven forskning (*corpus-driven*), där de inte utgår från några fördefinierade ord utan ”låter korpusen tala”, men tyvärr definierar de inte vad de menar med ”kristet/medeltida latin” och de väljer också ut en korpus som inte kan anses balanserad eller representativ för varken kristet eller medeltida latin (*Corpus Thomisticum*). Resultaten kan därför inte generaliseras på någon annan population än just *Corpus Thomisticum*. Det är således motiverat att försöka undersöka semantisk variation under kristendomens inflytande utan fördefinierade sökkriterier och med en korpus som kan anses representativ för kristet latin.

Universitetet i Zürich har också nyligen (2024) publicerat en open source-version av *Patrologia Latina*, som jag senare kommer att argumentera är representativ för just kristet latin. Till skillnad från den tidigare versionen som såg dagens ljus i samband med det s.k. OpenGreekAndLatin-projektet (Github, u.å., OpenGreekAndLatin project), har OCR-fel i denna version i stort sett rättats, vilket tidigare varit ett stort hinder (jfr. Bamman&Burns 2020:2). Materialet är också annoterat avseende lemma och ordklass (PoS för *Part of Speech*. Jfr. Rao & McMahan, 2019:34).

Som McGillivray understryker är forskning inom det hon kallar ”nya paradigmet” inom latinsk datalingvistik i högsta grad kollaborativ (2014:10). Forskare utvecklar språkmodeller (LLM), som anpassas av andra forskare och publiceras om i ny tappning för att uppfylla nya syften. Modellen som den här uppsatsen resulterar i bygger i högsta grad på modellen LatinBERT som i sin tur bygger på data som annoterats av olika aktörer och publicerats *open source* (t.ex. LASLAs under många år mödosamt manuellt lemmatiserade korpus, jfr. Longree & Fantoli, 2023). Som ett led i denna filosofi publicerar jag också min modell på Github, och hoppas att den eller att mina resultat skall kunna återanvändas av andra som intresserar sig för kristet latin (Lafage, 2025b).

## 2 Teoretiskt ramverk

Uppsatsen förankras först teoretiskt i forskning om kristet latin. Sedan kommer jag att redovisa nödvändiga begrepp och teoretiska utgångspunkter inom distributiv semantik. Avsnittet avslutas med en genomgång av hur dessa teorier tillämpas språkteknologiskt för att mäta semantisk variation, samt med en översikt av liknande arbeten.

### 2.1 Kristet latin

Forskning om kristet latin inleds redan på 1800-talet under det växande intresset för vulgärlatin och medeltidslatin, och på den tiden använde man termen ”kristet latin” för att beteckna de kristna författarnas latin (López Silva, 2003:117).

#### 2.1.1 Sondersprache

Istället för att beskriva de första kristnas latin (och medeltidslatin) som ett ”degenererat” språk, en vanlig tanke vid den tiden, valde Mohrmann i Schrijnens efterföljd att betrakta fenomenet som en form av språkvarietet: en *Sondersprache* (Schrijnen, 1932; Mohrmann 1977). Med *Sondersprache* menar Schrijnen en språkintern varietet som utvecklats till följd av en förändrad psykologisk verklighet och som manifesterar sig i de lexikalsemantiska, morfologiska, syntaktiska och fonologiska fälten. I det hänseendet har Schrijnen faktiskt beskrivits som en sociolingvistikens föregångare (Denecker, 2018:340–341).

Den nya religionen kräver ett nytt ordförråd för att beskriva nya koncept och eftersom kristendomen kommer till det latinska språkområdet från det grekiska, handlar det i mångt och mycket om ett försök att klä koncept som i grunden verbaliserats på grekiska (och till viss del hebreiska) i en latinsk språkdräkt (Mohrmann, 1977:147; López Silva, 2003:120). Schrijnen döper de uppståndna språkliga (hos honom mestadels lexikala) säregenheterna till *kristianismer* (1932:377) och skiljer mellan *direkta* (*unmittelbare*) och *indirekta* (*mittelbare*) *kristianismer*: det förra avser ord som uppkommit för att beskriva en ”kristen” verklighet såsom *incarnatio*, medan han med det senare menar ord som kommit till stånd under kristet inflytande men som inte nödvändigtvis beskriver en kristen verklighet, såsom *habitaculum* (1932:380–382).

Dessa kristianismer är inte enbart av lexikal art: de är i högsta grad även semantiska fenomen. Mohrmann skriver exempelvis om *direkta semantiska kristianismer* som ”neuer Wortinhalt, neue Bedeutung in bereits vorhandenen Wortformen” (1977:126). Semantiska kristianismer anses alltså vara ord som redan existerade i det latinska språket och som fick en utvidgad, kristen betydelse, såsom *pax*, *confessio* eller *conversio* (1977:14–16).

López Silva benämner detta *Sondersprache* med Mohrmanns terminologi ”gruppspråk” (*Gruppensprache*, *lengua de grupo*) eller ”specialspråk” (*lengua especial*) vilket han definierar som en samling kulturellt betingade diskurser som en specifik språkgemenskap använder (2003:117), och föredrar att använda termen ”de kristnas latin” (*el latín de los cristianos*), som han skiljer från andra grupp- eller fackspråk såsom medicinskt eller juridiskt latin i dess snabba expansion till breda folklager (2003:118–119).

Idag verkar en konsensus vara att de kristna i sina texter använde den varietet av latin som de faktiskt talade snarare än en specifik *Sondersprache* (Versteegh, 2017:72). Vidare anses de element som brukar lyftas fram som ”kristna” härröra från översättningar av bibliskt material, medan anhängare av Nijmegen-skolan påstod att de kristnas *Sondersprache* utvecklades först innan det i sin tur färgade de bibliska översättningarna (Burton, 2011:487–488). López Silva noterar att det till skillnad från andra

”fackspråk” som medicinskt latin endast finns några enstaka passager i samtida källor (hos Augustinus) som föreslår varianter på ett fåtal ord *ore christiano*, och att det dessutom är mycket svårt om ens möjligt att verifiera morfosyntaktiska särdrag som Nijmegen-skolan vill göra gällande (2003:120–121).

Denecker beskriver emellertid *Sondersprache*-hypotesens gradvisa nedgång under sociologins framfart som anmärkningsvärt med tanke på dess sociolingvistiska ansats. Skälen anges vara bl.a. det misslyckade förhållandet till senlatinets kontinuitet och komplexitet, samt att man förlitat sig på ett intuitivt tillvägagångssätt, när vad teorin i själva verket krävde var kvantitativa metoder (2018:351–352).

### 2.1.2 Tidsaspekten

Något som tycks skilja olika personer åt som skrivit om kristet latin som koncept är tidsaspekten. Liksom Nijmegen-skolan sätter exempelvis Ortuño Arregui inga tydliga tidsramar för företeelsen i fråga (2016), även om termer som *altchristliches Latein* indirekt skriver in kristet latin i en tidig period. Å andra sidan anger Burton perioden 200 e.Kr. – 600 e.Kr, under vilken skribenterna i högre grad än i senare epoker kan förväntas ha haft latin som sitt modersmål (2011:486). Burtons kapitel (Burton, 2011) i Clacksons antologi (Clackson, 2011) återfinns emellertid inte under den kronologiska indelningen utan under huvudrubriken ”register”. Språkhistoriskt överlappar kristet latin enligt Burtons avgränsning både det som kallats senlatin och medeltidslatin. I sociolingvistiskt hänseende hittar vi i kristet latin element som i andra sammanhang anses vara diastratiskt markerade såsom tillhörande en ”lägre” språkvarietet som ibland kallats ”vulgärlatin” (2011:486–487). Samtidigt innehåller samma korpus välpolerad ”klassisk” prosa (2011:488). Burton verkar definiera kristet latin i fråga om Augustinus språk som ”non-standard usages that might be heard among Christians” (2011:486).

Även om han inte skriver in företeelsen i någon avgränsad tidsperiod citerar López Silva författare från ungefär samma tidshorisont som Burton (2003:116). Han avfärdar idén om kristet latin som *Sondersprache* som grundlös och föredrar i stället konceptet ”kristendomens inflytande inom ramen för senlatinets”. Vidare anser han att bevisen för särskilda syntaktiska eller fonologiska kristianismer inte är övertygande och väljer att betrakta kristet latin endast som ett semantiskt fenomen (2003:121).

I mer recent forskning har korpuslingvister betraktat en korpus bestående av texter skrivna mellan Kristifödelse och modern tid som representativa för kristet latin (McGillivray et al., 2022a; Perrone et al. 2021, Shlechtweg et al. 2020), detta kanske som en respons på denna brist på tydliga tidsramar, men också förmodligen för att helt enkelt undvika att begränsa ett i förhållande till korpora för andra språk redan skralt material.

### 2.1.3 Kategorisering

López Silva listar tre typer av ”kristianismer” (2018:121–122):

- i. direkta låneord från främst grekiska såsom *apostolus* eller *baptismus*.
- ii. neologismer baserade på direktöversättningar av grekiska ord<sup>6</sup>: *glorificari*, *sanctificari*.
- iii. semantiska utvidgningar: ord som används i nya sammanhang såsom *fides*, eller *virtus*.

Burton skiljer mellan tre typer av semantiska utvidgningar (2011:490):

- i. semantiskt fokus: när delar av ett ords semantiska domän framhävs, såsom *gentes*, som får betydelsen ”icke-judar” utan att förlora sin grundbetydelse (”folk”).
- ii. semantisk innovation: då ett ord får en ny betydelse som inte är direkt relaterad till dess ursprungliga betydelse. T.ex. *sacramentum*, som i klassiskt latin oftast betyder ”faned”.

---

<sup>6</sup> Burton kallar detta *calquing* eller *loan translation*. Jfr Burton (2011:489).

- iii. kanoniserade metaforer: när ord används i första hand i stelnade metaforer, såsom *flagellum* ("piska") som får betydelsen "[Guds] gissel".

I denna uppsats kommer vi endast att behandla fall av semantisk utvidgning. Detta kan innebära olika saker, exempelvis att den ena av ett polysemt ords olika betydelser blir dominant, eller att en ny betydelse uppstår. Här presenteras en lista över substantiv och adjektiv som upptas i litteraturen som exempel på semantisk utvidgning under kristendomens inflytande. Listan kommer att komma till användning senare i uppsatsen.

#### 2.1.4 Lista över semantiskt utvidgade lemman

Det finns ytterst få exempel från Nijmegen-skolan, trots omfattande teoretiska texter om kristet latin. Dessutom verkar senare författare i mångt och mycket ha upprepat varandras ord och sällan genererat nya exempel. Källor till listan är Burton (Bu 2011), Clackson & Horrocks (C&H 2007:265–304, främst 284–292), McGillivray et al. (Mc 2022a), López Silva (L 2003), Mohrmann (M 1977) och Otruño Aregui (O 2019). Jag har medvetet undantagit Sprugnoli et al. (2020) eftersom deras ord är hämtade endast på ett kvantitativt sätt och enbart bygger på texter av Thomas av Aquino. I Mc hittar vi därutöver en lista över lemman som manuellt annoterats som semantiskt oförändrade. Det förekommer dock en diskrepans: listan i Mc Stämmer inte riktigt överens med listan över samma lemman i goldstandarden i McGillivray et al. (2020). Detta gäller ord som *ius*, *uoluntas* och *templum*, som annoteras som utvidgade respektive outvidgade i de olika listorna. Jag väljer därför att exkludera listan över semantiskt oförändrade lemman och hämtar bidrag från McGillivray et al. endast i Mc, samt justerar för *salus*. Den resulterande sammanställningen redovisas i tabell 1 och kommer att användas längre fram i uppsatsen.

	Bu	Mc	C&H	L	M	O
substantiv	ciuitas	ciuitas	passio	fides	pax	lauacrum
	misericordia	cohors	missa	lauacrum	confessio	peccatum
	uirtus	consul	uersus	caro	conuersio	
	gentes	dolus	lectio	spiritus	salus	
	nationes	dux		confessio	caritas	
	sacramentum	humanitas		pax	uerbum	
	collyrium	imperator		paganus	lauacrum	
	flagellum	pontifex		gentes	gloria	
	plaga	potestas		sacramentum	caro	
		sacramentum		perpetratio		
		scriptura		ploratio		
		uirtus				
	adjektiv		beatus		arrepticus	
		fidelis		ceruicatus		
		sanctus				

Tabell 1. Ordlista över lemman som i litteraturen anges ha genomgått en semantisk utvidgning under kristendomens inflytande.

Vissa ord, som *imperator* eller *dux*, verkar spegla medeltida snarare än kristet latin, och detta beror kanske på att Mc har ett bredare fokus. Andra som *arrepticus* och *ceruicatus* framstår som ovanligt sällsynta, men jag inkluderar dem för att vara konsekvent.

I följande avsnitt presenteras distributiv semantik, som utgör den teoretiska grunden för den datalingvistiska metodologin i denna uppsats.

## 2.2 Distributiv semantik

Traditionellt har utgångspunkten inom både strukturalism och generativ grammatik varit det De Saussure kallade *langue*, dvs språk som ett abstrakt, regelstyrt system som föregår *parole*, vilket är konkreta realiseringar av detta abstrakta system. Förenklat kopplas enligt detta synsätt ett ord (*signifiant*) till ett koncept (*signifié*), ungefär som i en ordboksdefinition (Saussure et al., 1971:33–34). I sin nyutgivna bok förklarar gruppen Quantitative Lexicology and Variational Linguistics (QLVL) i stället semantisk variation från de s.k. referenterna, dvs de konkreta objekt/företeelser som finns i världen och som man refererar till med hjälp av språk (Geeraerts et al., 2024:8). Gränsen mellan ord, koncept och referent kan ibland vara diffus, vilket McGillivray problematiserar gällande det latinska ordet *uirtus*, som beroende på sammanhanget enligt Lewis & Short refererar till olika dock snarlika och ofta sammanvävda koncept som ”manhood”, ”power”, ”courage” eller ”value” (McGillivray et al., 2022a:54–55).

Med hjälp av prototyp teorin, enligt vilken referenters tillhörighet till en semantisk kategori anses vara mer eller mindre stark beroende på hur central en referent är för kategorin i fråga (en s.k. *prototyp*), beskriver QLVL relationen mellan ord inom samma kategori och mellan olika betydelser av samma ord (s.k. *polysemer*) således vara distributiv: vissa referenter anses ha mer framtoning (*saliency*) för ett visst koncept än andra (2024:9–12). Konsekvenserna av ett sådant synsätt är dels att förhållandet mellan centrum och periferin i detta semantiska rum anses kunna kvantifieras, dels att frekvenser (antal förekomster av något) spelar en roll vid bedömningen av denna framtoning.

En distributiv ansats bygger på den s.k. *Distributiva Hypotesen* (DH), enligt vilken ord som är distributivt lika också måste vara semantiskt lika (Geeraerts et al., 2024:67).<sup>7</sup> Om vi tillämpar ett distributivt perspektiv på polysemi kan vi säga att betydelsen av ordet *caput*, *-itis*, *n* utgörs av alla kontextuella instanser av ordet. Med ”kontext” menas här lingvistisk kontext, till skillnad från icke-lingvistiska kontexter som den kommunikativa situationen (Lenci, 2018:154). För att avgöra om betydelsen i de enskilda kontexterna är t.ex. ”mänskligt huvud”, ”huvudstad” eller ”kapitel”<sup>8</sup> måste sammanhanget observeras: om ordet *caput* förekommer i samma mening som *crinis*, *facies* eller *cerebrum* är sannolikheten att vi har med betydelsen ”mänskligt huvud” att göra större än om ord som *mundus*, *civitas* eller *munus* istället står där, eller med Firths i sammanhanget synnerligen väl citerade aforism ”You shall know a word by the company it keeps” (1962).

Olika förekomster av *caput* i en korpus kan enligt detta synsätt teoretiskt placeras i ett idealiserat tvådimensionellt rum, där semantisk närhet visualiseras genom avstånd mellan förekomsterna i rummet. Detta avstånd uträknas med hjälp av de ord som förekommer runt om varje förekomst av *caput* i korpusen (Geeraerts et al., 2024:15). Nästa avsnitt redogör för hur vektorer byggs i praktiken för att placera orden i detta rum.

Studieobjektet är nu inte längre *langue* utan *parole*: fokus ligger alltså på faktiska yttringar i verkligheten, vilkas kopplingar till koncept kan studeras kvantitativt i stället för kvalitativt (Geeraerts et al., 2024:12–13). Geeraerts et al. skiljer mellan *systemic meaning* (kontextoberoende ordboksdefinitioner) och *utterance meaning* (kontextuella betydelser, med varierande referenter beroende på situationen) och skriver: ”[...] even if you are primarily interested in systemic meaning, *utterance meaning is the primary observational basis of semantics.*” (2024:53).

---

<sup>7</sup> För en koncis och pedagogisk framställning av distributiv semantik, se McGillivray & Tóth (2020:63).

<sup>8</sup> Jfr Ahlberg et al. sv *caput*, *-itis*, *n*.

## 2.3 Distributiv semantik och NLP

Detta avsnitt introducerar några koncept inom NLP, som har relevans för distributiv semantik och som kommer att användas i uppsatsen. Avsnittet utgör en naturlig övergång till metodavsnittet och avslutas med en kort redogörelse av studier inom fältet som gjorts för latin och grekiska.

### 2.3.1 Lemma, type, token

Inom NLP kallas *lemma* ett ords grundform (Rao & McMahan, 2019:33–34) eller ordboksform (McGillivray & Tóth, 2020:23–24), medan lemmanas olika böjda former kallas antingen *type* om man buntar ihop alla likadana ordformer, eller *token* om man menar alla enskilda ordförekomster (McGillivray & Tóth, 2020:22).<sup>9</sup> Nedan följer några exempelord från texten med titeln *Canones* av författaren Abbo Floriacensis i XML-format efter TEI-standarden<sup>10</sup> från samlingen *Patrologia Latina*:

```
<cc:s n="1" parent_pid="9741:3;2">
<cc:w lemma="dico" pos="V:IND">dicit</cc:w>
<cc:s n="4" parent_pid="9741:16;2">
<cc:w lemma="dico" pos="V:IND">diximus</cc:w>
<cc:s n="3" parent_pid="9741:17;4">
<cc:w lemma="dico" pos="V:IND">diximus</cc:w>
<cc:s n="5" parent_pid="9741:45;6">
<cc:w lemma="disco" pos="V:IND">discimus</cc:w>
```

Av exemplet framgår fyra ordförekomster med respektive metadata (information om orden som exempelvis ordklass eller källa): två olika lemman (*dico* och *disco*), tre olika types (*dicit*, *diximus* och *discimus*) och fyra olika tokens: *dicit*, *diximus* (i mening 4, stycke 16;2), *diximus* (i mening 3, stycke 17;4), samt *discimus*. En korpus sägs vara *lemmatiserad* när varje ordförekomst har kopplats till ett lemma, som i exemplet ovan. Inom lexikografi används även begreppet *lexem*, vilket är en abstrakt meningsenhet (Tahmasebi et al., 2021:6). Exempelvis utgör *adversus* som preposition och substantiv två olika lexem (två olika betydelser) och likaså två olika homonyma lemman (de dyker upp i en ordbok som två skilda uppslagsord, även om de delar samma form). I praktiken brukar man inom NLP alltid använda termen ”lemma”.

### 2.3.2 Vektorer och word embeddings

Ett vanligt sätt för att studera semantisk variation enligt den distributiva hypotesen (DH) är att räkna ut hur många gånger vissa ord i en text förekommer nära varandra och kondensera denna information i en vektor. I den här uppsatsen används s.k. *kontextuella* vektorer. För att illustrera detta förklaras härnäst hur de byggs upp, med jämförelse mellan s.k. *frekvensbaserade* och *statiska* vektorer.

#### 2.3.2.1 Frekvensbaserade vektorer

Vi betecknar en specifik odefinierad ordform  $w$  (t.ex. *caput*, *militibus* eller *amabat*). Enligt DH skall ett  $w$ :s samtliga kontextord i en specifik korpus, dvs de ord som förekommer i samma kontext som  $w$ , utgöra en representation av det specifika  $w$ :s betydelse(r). Därigenom kan en matris konstrueras över alla ord i en korpus och deras inbördes kombinationer. Summan av dessa kombinationer anses således utgöra varje ords semantiska innehåll och kan vektoriseras. I detta fall består vektorn av lika många

<sup>9</sup> Märk att skiljetecken också behandlas som token inom NLP.

<sup>10</sup> Text Encoding Initiative. Jfr. McGillivray & Tóth (2020:11).

dimensioner som det finns ordkombinationer i korpusen och man talar därför om en s.k. *explicit vektor* (Tahmasebi & Dubossarsky, 2023:8; McGillivray, 2022b:14). En sådan matris skulle kunna se ut som i tabell 2.

	miles	gallus	dux	agricola	bos	campus	aratum	rusticus	sagitta	pugna	castra
miles		3	126	24	12	35	2	0	98	103	186
gallus	3		0	58	67	91	24	76	0	4	0
dux	126	0		13	20	15	6	0	5	29	24
agricola	24	58	13		98	78	106	167	0	3	0
bos	12	67	20	98		56	129	59	1	5	6
campus	35	91	15	78	56		34	87	16	52	24
aratum	2	24	6	106	129	34		67	0	8	0
rusticus	0	76	0	167	59	87	67		1	3	0
sagitta	98	0	5	0	1	16	0	1		91	56
pugna	103	4	29	3	5	52	8	3	91		75
castra	186	0	24	0	6	24	0	0	56	75	

Tabell 2. Matris över frekvenser av samförekomst av lemman i en fiktiv korpus.

Det kan observeras att lemman *aratum* och *rusticus* i matrisen uppvisar höga frekvenser av samförekomst med liknande ord (*agricola*, *bos*, *campus*, osv) och att de skyr samma lemman (*miles*, *dux*, *sagittae*, etc). Enligt DH är detta ett tecken på att *aratum* och *rusticus* tillhör samma semantiska fält. På samma vis verkar *pugna* och *castra* dyka upp i liknande sammanhang. *Campus* är litet ambivalent, förmodligen då det förekommer med både militära ord när det betyder "(slag)fält" och med jordbruksord när det betyder "åker". För att filtrera bort funktionsord (*non-content words*), dvs frekventa ord som *et*, *cum* eller *hic*, vilka förekommer med alla typer av ord och därmed inte tillför användbar information, använder man sig av algoritmer som PMI (Pointwise Mutual Information), med vilka man räknar ut sannolikheten av att två ord samförekommer genom att jämföra med hur frekventa dessa ord är totalt i korpusen (Tahmasebi & Dubossarsky, 2023:4).

Om vi buntar ihop alla samförekomster av ett lemma med övriga lemman i korpusen kan vi producera en vektor. Vektorn för *miles* i vår fiktiva korpus skulle alltså kunna representeras på följande vis:

(3 126 24 12 35 2 0 98 103 186)

Det skulle i detta format kunna jämföras med hjälp av en enkel algoritm med de övriga lemmanas vektorer för att hitta dem som är mest semantiskt lika. I det här exemplet består vektorn av 10 värden då vi har 11 olika ord som jämförs med varandra, vilket resulterar i en vektor med tio dimensioner. I en korpus i verkligheten handskas man dock ofta med miljontals ord som skall jämföras. Vektorer i flera miljoners dimensioner är dock praktiskt taget ohanterliga och därför brukar man bygga s.k. *implicita vektorer* genom dimensionsreduktion till ett fixerat antal dimensioner (Lenci, 2018:156–157). Denna vektor utgör en kondenserad representation av ett ords semantiska kontext i en specifik korpus och benämns ofta en *tät vektor* (*dense vector*).

Under de senaste 10–15 åren har maskininlärningsalgoritmer använts för att utveckla s.k. *predictive models* (Lenci, 2018:160). Syftet är att med hjälp av neurala nätverksalgoritmer träna en modell till att producera implicita vektorer för alla ord i en korpus, som kallas *word embeddings* (McGillivray, 2022b:14). Det finns två typer, som båda inkluderas i termen *neural embeddings*: statistiska och kontextuella.

### 2.3.2.2 Statiska embeddings

Ett exempel på statiska word embeddings är Word2vec, vilket för latinets del genereras i exempelvis pipelinen Classical Language ToolKit eller CLTK (Github, CLTK, *CLTK*; McGillivray & Tóth, 2020:21).

För att träna statiska Word2Vec embeddings lär sig modellen först s.k. *vikter* i en korpus där experter manuellt annoterat morfosyntaktisk och ibland även semantisk information, och tillämpas sedan på en mycket stor oannoterad korpus, där den försöker gissa varje  $w$  med hjälp av vikterna och probabilistiska metoder som CBOW (Continuous Bag Of Words) och Skip-Gram (SG). Den förra syftar att gissa  $w$  utifrån ett fördefinierat kontextfönster runt om det, och den andra att gissa kontexten givet samma ord (McGillivray, 2022b:15).

Till skillnad från frekvensbaserade modeller representerar talen som vektorn består av inte ordfrekvenser utan s.k. *parametrar*. När modellen möter  $w$  första gången projiceras det på ett slumpmässigt valt ställe i det vektoriella rummet, och varje gång modellen möter  $w$  i en ny kontext uppdateras vektorns parametrar (Tahmasebi & Dubossarsky, 2023:5). Hur långt ordet flyttas i rummet (och därmed i vilken grad parametrarna justeras) beror på hur svårt det är för algoritmen att gissa  $w$  varje gång det stöter på det. En s.k. *cross-entropy loss function* räknas ut för att mäta skillnaden mellan gissningen och det riktiga utfallet och därmed göra en bedömning av hur väl modellen gissar ord (Rao & McMahan, 2019:3), och genom *backpropagation* beräknas hur mycket vikterna bidrog till det uppmätta felet (Eisenstein 2019:52–56). Därefter justeras vikterna för att minimera loss nästa gång modellen stöter på ordet i fråga tills modellen anses ha konvergerat, dvs alla ord har hittat sin plats i rummet (Dahl 2023:kapitel 10–11).

Dessa vektorer kallas även *type embeddings*, eftersom de producerar en vektor per  $w$ .

### 2.3.2.3 Kontextuella embeddings

Ett problem med statiska embeddings är bl.a. att polysemers olika betydelser går förlorade genom att det skapas endast en vektor per  $w$ , oavsett om det kan dölja sig olika betydelser bakom den (McGillivray, 2022b:17–19). Därför har modeller för s.k. *kontextuella word embeddings*<sup>11</sup> utvecklats, vilka producerar en vektor per *token* med hjälp av kontexten, dvs varje enskild ordförekomst i en korpus representeras i det vektoriella rummet. Med kontextuella modeller kan man teoretiskt skilja mellan olika homonymer genom att olika betydelser av ett ord placeras på olika ställen i rummet (Tahmasebi & Dubossarsky, 2023:6).

För att anpassa träningen av statiska och kontextuella modeller justeras s.k. hyperparametrar, såsom antal ord före och efter varje  $w$ , antal *epoker* eller ”varv” igenom träningsmaterialet, dvs hur många gånger modellen möter samma material, eller *learning-rate* (LR), dvs hur stora ”steg” i det vektoriella rummet modellen uppdaterar vektorerna med varje gång den gissar ett maskerat ord (Tahmasebi & Dubossarsky, 2023:6–7). Ett högt LR medför snabbare träning, samtidigt som modellen riskerar att bli mindre precis, medan ett lågt LR gör att modellen tar orimligt lång tid att tränas och kan bli överdrivet känslig, trots hög precision. En träning med lågt LR kräver också fler epoker, eftersom det krävs fler förekomster av varje ord för att alla ord skall hitta sin slutdestination i rummet (Tahmasebi & Dubossarsky, 2023:7).

Googles BERT är ett exempel på en kontextuell modell (Devlin et al., 2018). Det bör noteras att BERT använder sig av en s.k. Wordpiece tokenizer, vilket är mycket viktigt i ett formrikt språk som latin. Först segmenteras texten i meningar och ord med hjälp av en menings- respektive ordtokenizer (Rao & McMahan, 2019:30). Senare kan ord brytas ner i rot och ändelse av Wordpiece tokenizer, och en vektor för ordet beräknas som medelvärde av de olika vektoriella representationerna för ordets olika

---

<sup>11</sup> Dessa kallas även bl.a. *dynamic embeddings*, jfr. Tahmesbi et al. (2021:27).

delar (Wang & Choi, 2023:3). BERT-modeller, såsom LatinBERT, brukar ha 768 parametrer per token (Bamman & Burns, 2020). Eftersom träning av en BERT-modell kräver mycket stor datakapacitet, brukar man i praktiken träna om en redan existerande modell på en specifik uppgift, när man väl vill använda den (en s.k. *fine-tuning*), även om detta kan variera med uppgiftens art. En fine-tuning kan t.ex. innebära att man exponerar modellen för många texter som tillhör en specifik domän eller att man tränar den på annoterat material, för att den skall lära sig automatisera en specifik uppgift (Tahmasebi & Dubossarsky, 2023:6).

Efter introduktionen av kontextuella word embeddings har antalet studier om diakron semantisk variation ökat, då metoden lämpar sig väl till det (ibid).

### 2.3.3 Relevant forskning inom distributiv semantik

Även om forskningsfältet fortfarande är i sin linda, har olika modeller, metoder för insamling av data och bedömning av resultaten börjat etableras (Tahmasebi & Dubossarsky, 2023:2).

Rodda et al. skrev 2017 om en studie över kristendomens effekt på det grekiska språket genom att beräkna vektorer med hjälp av PMI, med ett fönster av 11 ord runtom  $w$  och SVD (Singular Value Decomposition) som dimensionsreduktionsmetod, för alla ord i en automatiskt lemmatiserad korpus, som de delar upp i två tidsbaserade subkorpora med Kristi födelse som skiljelinje. Alla lemmans vektorer jämförs med övriga lemmans vektorer i respektive korpus och den erhållna datan jämförs sedan mellan subkorpora, bl.a. genom att hämta *nearest neighbours*. Med denna metod hittar de ord som visade semantisk variation mellan subkorpora, vilket tolkades som ett tecken på kristendomens inflytande, även om författarna inser effekten av genre och av den förenklade tidsbaserade uppdelningen (Rodda et al., 2017:22–23). Samma metod (*k-nearest neighbours*) från word embeddings användes av Ribary & McGillivray för att skilja ut specialiserad från ”konkret” betydelse av ord i Iustinianus *Corpus iuris civilis* på latin (Ribary & McGillivray, 2020).

Barbara McGillivray har varit drivande i distributiv semantik för latin och grekiska, och hon har skrivit handböcker i NLP för historiska språk. I sin studie 2022a använder sig McGillivray et al. av ett distributivt ramverk för att studera latinsk semantisk variation i samband med kristendomens utbredning, med ett lexikografiskt fokus enligt vilket olika förekomster av polysema ord manuellt kopplas till specifika ordboksdefinitioner av mänskliga annoterare. Forskarna utgår från en uppsättning latinska ord som i litteraturen anges ha genomgått en semantisk förändring under kristendomens inflytande (korpusbaserad forskning) och mäter deras grad av semantisk variation.

Sprungnoli et al. har meriten av att ha gjort en explorativ studie över ord som förändrat sin betydelse från klassiskt latin till (sic) ”medeltida/kristet latin” (2020). De använder sig av type embeddings för att mäta diakron semantisk variation mellan en förkristen korpus och Thomas av Aquinos texter från den stora korpusen *Index Thomisticus*. I senare delen av sin studie analyserar de varje ords s.k. *top-K neighbours*, dvs de ord vars vektorer mest liknar varandra, något som används som proxy för semantisk närhet, i vardera korpusen. Därefter tolkar de förändringar i dessa top-K neighbours mellan de två korpora som ett tecken på semantisk variation, och presenterar en lista över 20 ord som mest skall ha ändrat sin betydelse. Notera dock att de inte definierar varken medeltida eller kristet latin, och att man kan ifrågasätta om *Index Thomisticus* verkligen kan betraktas som representativ för endera. Ändå är deras bidrag intressant i så måtto att de till skillnad från Burton (2011) och McGillivray et al. (2022a) inte utgår från fördefinierade ord.

Alldeles nyligen publicerade Caffagni et al. en studie i vilken de tränar om olika LLM för latin för att automatiskt identifiera olika typer av bibliska referenser i patristiskt material (2025). Genom att använda sig av en delvis manuellt annoterad benchmark kan de mäta effekten av träningen på resultaten, och bevisar att de omtränade modellerna (fine-tuning) överträffar förtränade modeller i samma uppgift.

Liu et al. drog slutsatsen i oktober 2024 att kontextuella word embeddings överträffar statiska i att fånga upp semantiska förändringar i en specialiserad rättslatinsk korpus. De visar också att BERT-modeller direkt tränade på målkorpusen överträffar förtränade BERT-modeller såsom LatinBERT.<sup>12</sup>

Effekten av genre har studerats av bl.a. Perrone et al. (2021). Enligt dem är genre i småkorpusspråk som latin och klassisk grekiska en bidragande faktor i WSD. McGillivray et al. använder samma beräkningsmodell för att avgöra om förändrade frekvenser av grekiska ord som *mus*, *harmonia* och *kosmos* kan anses bero på genre eller tid (2019). Rodda et al. lyfter även fram effekten av genre på distributiv semantik (Rodda et al., 2017:20). En annan faktor som kan påverka resultaten är vad texterna handlar om (*topic*) (Tahmasebi & Dubossarsky, 2023:9).

### 2.3.4 Kritik mot DH och word embeddings

En distributiv ansats kombinerad med word embeddings innebär en tämligen snäv definition av mening: icke-lingvistisk kontext såsom talsituationen förbises av praktiska skäl helt (McGillivray, 2022b:5). Vidare bäddar sådana metoder inte för en analys av mening som övergår ordgränserna, som i idiomatiska uttryck (McGillivray & Tóth, 2020:64). Ett annat problem är att distributiva modeller fångar liknelser i betydelse mycket bättre i vissa ordklasser som substantiv jämfört med t.ex. verb, och att de inte skiljer mellan olika typer av relation mellan ord som hypernyymi, antonymi eller homonymi (Lenci, 2018:161). Kritiken riktar sig dock inte specifikt till BERT-modeller, utan snarare till statiska (Lenci, 2018, McGillivray, 2022b:16–17). Vid SemEval-2020 visade sig kontextuella embeddings visserligen prestera sämre än statiska för att lösa task 1, men detta skulle enligt Schlechtweg et al. kunna bero på att materialet var lemmatiserat (2020:10, Tahmasebi et al., 2021:47). Notera att skillnaderna i testresultaten avser specifika uppgifter i SemEval och att de är rätt små. Lenci anser överlag att distributiv semantik lämpar sig väl för att studera polysemi och semantisk variation (2018:165).

Word embeddings beskylls också ofta för att vara s.k. ”black-box modeller”, där förhållandet mellan parametrar och frekvenser gått förlorat (McGillivray, 2022b:16). I mina ögon ligger modellens validitet inte i reproducerbarhet genom transparens av data utan i att den faktiskt levererar resultat som stämmer överens med forskning gjord med andra metoder. Även om frekvensbaserade modeller speglar riktiga frekvenser, hur hade en granskare kunnat verifiera hela matrisen för en korpus av flera miljoner ord? När man väl kastat word embeddings med badvattnet måste man också kasta all generativ AI i ett svep, som faktiskt även det är en LLM som bygger på word embeddings. Med tanke på hur väl dessa modeller idag förmår översätta texter och därmed visa prov på semantisk akribi, vore detta enligt mig en högst problematisk ståndpunkt.

---

<sup>12</sup> Tahmasebi & Dubossarsky rapporterar samma observation (2023:9).

## 3 Material och metod

För att analysera inflytandet av kristet latin på det latinska språket behöver man bygga en måttstock som definierar det kristna latinets mest typiska semantiska och lexikala kännetecken, från vilken avvikelser kan mätas och särdrag identifieras. Härnäst kommer jag att beskriva planen för omträningen av en BERT-modell på en omfångsrik, representativ och balanserad specialiserad korpus av kristna texter, och dess tillämpning för att mäta semantisk utvidgning.

### 3.1 Material

Utgångspunkten för korpuslingvistik är att det är omöjligt att analysera en hel population: för att undersöka exempelvis svenskarnas politiska åsikter kan man i teorin kunna tillfråga alla svenska hushåll, men det skulle vara tidsödande. I stället gör man ett urval (*sample*). Förutsatt att urvalet är representativt för den undersökta populationen och balanserat, kan de slutsatser som dras för urvalet generaliseras för hela populationen. Statistiska tester brukar användas för att avgöra hur starka eller pålitliga resultaten är (McEnery et al., 2006:13).

Märk att representativitet för specialiserade korporas vidkommande, dvs korpora som till skillnad från allmänna korpora förenar texter från en specifik genre eller domän, ökar med korpusens s.k. *saturering*. För att mäta graden av saturering delas korpusen i lika stora segment: om varje inläsning av ett nytt segment tillför ungefär lika många nya lexikala enheter till korpusen som det föregående segmentet anses korpusen vara saturerad (McEnery et al., 2006:15–16). I en balanserad korpus är vidare ingen text, författare eller genre kraftigt överrepresenterad (McGillivray, 2014:11; McEnery, 2006:16–19).

Korpuslingvistik har primärt utvecklats för moderna språk som engelska eller franska, för vilka ett rikt digitalt material finns. Till skillnad från moderna språk kan historiska språk inte producera nytt material. För att träna sin BERT-modell LatinBERT lyckades Bamman & Burns med stor möda skrapa ihop en imponerande korpus på 647,2 miljoner ord, förlitande sig i brist på bättre material tyvärr upp till 50% på delvis felaktigt OCR-inlästa texter (Bamman & Burns, 2020:2). Jämför detta med NOW-korpusen (News On the Web), som växer med ca 1,6 miljarder ord *varje år* och nu uppgår till över 20 miljarder ord (English Corpora, u.å., *NOW Corpus*). Dessutom har ett ”förurval” i praktiken skett med de latinska texter som traderats till oss, i det att många inte har motstått tidens tand medan andra egentligen mer perifera texter bevarats. När man bygger en latinsk korpus gör man alltså alltid ett urval ur ett urval och frågan är om materialet någonsin kan vara representativt nog för att man skall kunna uttala sig om någon form av latinitet eller litteratur överhuvudtaget. Är t.ex. en korpus bestående av Senecas tragedier representativ nog för att dra slutsatser om populationerna ”romersk tragedi” eller ”tragiskt latin”? Försiktighetsprincipen borde leda oss till att acceptera att de åtminstone är representativa för populationen ”den traderade romerska tragedin”, eftersom Senecas tragedier i stort sett är det vi har kvar av den tragiska genren.

Nedan beskrivs de två korpora som kommer att ingå i studien: *Patrologia Latina* och förkristet material.

#### 3.1.1 Patrologia Latina

I den här uppsatsen tränas BERT-modellen LatinBERT av David Bamman och Patrick J. Burns (Bamman & Burns, u.å.)<sup>13</sup> om på specifikt kristet material, och den resulterande modellen döps till

---

<sup>13</sup> Webbsida.

XPLatinBERT. Eftersom modellen tränas på oannoterat material behöver detta material vara rikt och balanserat. För låga ordfrekvenser kan nämligen resultera i att modellen inte konvergerar (jfr avsnitt 2.3.2). Denna korpus hittar vi i *Patrologia Latina* (Universitat Zurich, u. a., *Patrologia Latina*), en mycket omfattande korpus sammanstalld under 1800-talet av J.-P. Migne, vars onlineversion pa Zurichs universitets plattform *Corpus Corporum* motsvarar ca 84,5 miljoner ord (Universitat Zurich, u. a., *Corpus Corporum*).

*Patrologia Latina* ar resultatet av ett alldeles unikt redaktionellt och inte helt okontroversiellt arbete, lett av den excentriske abbe Jacques-Paul Migne, som pagick mellan 1844 och 1855 (Bloch, 1994:1). Ar 1854 hade foretaget *les Ateliers Catholiques* minst 596 anstallda (1994:13), vilket gjorde det till en av Frankrikes allra storsta arbetsgivare vid den tiden.<sup>14</sup> Det kolossala verket uppkom i en tid av okande religiositet efter Revolutionen (1994:10–11). Visserligen publicerade Migne nagra manuskript som dittills varit okanda (1994:58–60), men att han lyckades redigera ett sa omfangsrikt verk pa sa kort tid beror pa att han oftast nojde sig med att trycka om aldre (tryckta) editioner.<sup>15</sup> Han tog alltsa ofta inte hansyn till textvittnena enligt god filologisk sed utan polerade material som redan givits ut (1994:77), aven om han skriver i sin reklam att han valt ut de basta editionerna (1994:62). For att undvika upphovsrattsliga tvister skall han ibland aven ha valt en mindre tillganglig edition, trots att en mer etablerad sadan fanns tillganglig (1994:65). Hans medarbetare verkar ocksa ha varit praster som pa ett eller annat satt hamnat i onad och darmed nojde sig med en lag lon (1994:17–19), vilket kan ha paverkat slutresultatet negativt. Målet var alltsa helt enkelt att producera sa manga bocker som mojligt till ett sa attraktivt pris som mojligt, kvantitet fore kvalitet (1994:90). Korpusen ar saledes ingen bra kalla for ett kvalitativt filologiskt arbete, men dess omfattning gor den lamplig som specialiserad korpus for studier av kvantitativ art. I den har uppsatsen ar det alltsa irrelevant vilkendera lasarter som valts i en specifik passage, eller att vissa texter skulle ha retuscherats nagot under 1600-talet. Syftet ar namligen att hitta aterkommande semantiska monster, som LatinBERT kan tranas pa, och dar spelar det ingen roll om nagra ord inte motsvarar det som faktiskt stod i textvittnena: det enorma textomfanget gor att XPLatinBERT kommer att mota korrekta relationer tillrackligt manga ganger for att effekten av fel eller samre filologisk akribi skall vara forsumbar.

Korpusen har publicerats digitalt pa *Corpus Corporum*, en digital plattform som drivs av Zurichs universitet under ledning av Philipp Roelli, vars mal ar att samla open-source latinska texter i XML-format (enligt TEI-standard) for forskning (Universitat Zurich, u. a., *Corpus Corporum*).<sup>16</sup> Enligt Clerice ansags den digitala korpusen pa *Corpus Corporum* 2022 fortfarande lida brister i XML-formatet, i det att t.ex. ingen atskillnad gjordes mellan texterna och den kritiska apparaten (som alltsa hamnar i texten), aven om han noterar att en viss forbattring har skett (Clerice, 2022:4). En genomgang av texter ur korpusen visar att detta i stor utstrackning har korrigerats<sup>17</sup>, men det forekommer fortfarande vissa parsningsfel<sup>18</sup> och Philipp Roelli har i ett mejl till mig bekraftat att det aven foreligger lemmatiseringsfel. Gallande OCR-fel (inlasningsfel) noterar jag en vasentlig forbattring med den senaste versionen av korpusen pa *Corpus Corporum* jamfort med de XML-dokument som publicerats open source i samband med OpenGreekAndLatin-projektet (Github, u. a., OpenGreekAndLatin project, *Patrologia Latina*).

Mark att LatinBERT tranades pa 29,3 miljoner ord ur *Patrologia Latina* (Bamman & Burns, u.a.). Detta innebar att LatinBERT inte har exponerats for ca 65% av materialet. Da en korpus pa ca 30 miljoner ord anses kunna fanga upp semantiska relationer med hjalp av word embeddings (McGillivray

---

<sup>14</sup> Som jamforelse anger Bloch att de 125 000 storsta industriforetagen i Frankrike under den tiden anstallde i genomsnitt 10 personer.

<sup>15</sup> For en lista over nagra kallor som Migne tappade ur, se Bloch (1994:60–63).

<sup>16</sup> Ett stort tack till Philipp Roelli for att han tillgangliggjorde hela *Patrologia Latina* for mig, i bade annoterat och oannoterat format.

<sup>17</sup> Det skulle kunna bero pa uppdateringen som enligt *Corpus Corporum*-webbsidan skall ha gjorts 2024-03-15. Manga texter i korpusen verkar enligt XML-metadatan mycket riktigt ha uppdaterats i mars 2024.

<sup>18</sup> T.ex. har nagra meningar avgransats fel.

& Tóth, 2020:68) borde *Patrologia Latina* vara lagom stor. Dessutom borde den vara representativ, och med sina ca 5 700 texter i olika genrer någorlunda balanserad, vilket diskuteras vidare nedan.

### 3.1.2 Förkristet material

Referensmaterialet är texter som av kronologiska skäl inte kan förväntas förete några semantiska utvidgningar under kristendomens inflytande. Materialet hittas på *Corpus Corporum* och består av alla lemmatiserade texter valda efter Burtons tidskriterium (fram till 200 e.Kr.) ur följande korpora: *Latinitas Antiqua* från Perseus-projektet, *Auctores scientiarum varii*, och *Antiquitas Posterior* från digilibLT.<sup>19</sup> För enkelhetens skull kallas korpusen LA för *Latinitas Antiqua*, som är den med råge största källan till texterna.

## 3.2 Metod

Det är viktigt att poängtera att den till synes linjära beskrivningen av den i uppsatsen tillämpade metoden döljer faktumet att en sådan metodik i själva verket är cirkulär. Ett senare moment i pipelinen kan alltså ådagalägga ett initialt problem, vilket föranleder en omkörning av en ny reviderad kod. Alla korpus-specifika överväganden som gjorts till följd av uppstådda problem redovisas i avsnitt 4.

### 3.2.1 Förarbete

Alla texter i båda korpora bearbetas noggrant så att så få felaktigheter som möjligt smyger sig in. Ur varje text extraheras textens metadata (såsom författare och datum), alla meningar, ord, samt i görligaste mån ordens metadata (lemman och PoS).<sup>20</sup> Till varje mening associeras en unik identifikator, som i denna uppsats benämns ”sent\_id”, direkt från de tokeniserade XML-filerna. Här följer ett exempel:

40:3;2,9

Det första talet är vad som i XML-filerna kallas *cc\_idno*, dvs en unik identifierare för dokumentet i fråga, och i exemplet ovan står 40 för *Ad martyres* av Tertullianus. Talen efter kolonet identifierar meningen (i exemplet avsnitt 3, stycke 2, mening 9).<sup>21</sup> De extraherade texterna standardiseras (t.ex. ändras *j* till *i* och *v* till *u*). Beräkningar och stickprov görs för att sälla bort problematiska meningar eller texter, om dessa inte kan förbättras på annat sätt. Målet är att extrahera så många *fullständiga* meningar och så mycket metadata som möjligt.

### 3.2.2 Träning och testning av XPLatinBERT

Gururangan et al. kommer i sin studie om förtränade BERT-modeller, vilket LatinBERT också är, fram till att de blir mer effektiva i olika typer av domän-specifika NLP-uppgifter om de tränas om på en oannoterad korpus som är representativ för domänen i fråga (2020:2). De döper denna typ av ”omträning” till *DAPT* eller *Domain-Adaptive PreTraining* (2020:2). I denna uppsats ämnar jag alltså utföra en *DAPT* på domänen ”kristet latin” med materialet från *Patrologia Latina*. Det bör dock noteras att modellen i detta fall inte tränas på annoterad data (det jag gör är alltså en s.k. *unsupervised training*, jfr. Schlechtweg et al., 2020:2).

---

<sup>19</sup> Se litteraturlistan för hyperlänkar.

<sup>20</sup> Som vi skall se i avsnitt 4 var detta en tämligen svår uppgift, inte minst på grund av den undermåliga parsningen i XML-filerna.

<sup>21</sup> Märk att detta system inte följer vanlig filologisk praxis med bok-, kapitel- och paragrafindelning, och att det därmed inte går att söka upp ett sent\_id direkt i en textedition.

*Patrologia Latina* delas automatiskt upp i ett träningsset motsvarande ca 95% av alla ord, ett valideringsset (ca 2.5%) och ett testset (ca 2.5%), med hjälp av en randomiseringsfunktion som tar hänsyn till antalet ord utan att kapa texter. Detta är en vanlig procedur i maskininlärning och MLM (Rao & McMahan, 2019:53–54). Träningssetet består av de texter som XPLatinBERT tränas på. Valideringssetet används under träningen för att testa modellen på okänd data och se om den överanpassar sig till träningssetet, samt för att justera hyperparametrarna baserat på hur väl modellen presterar. En överanpassad modell är en modell som har tränats för hårt på träningsmaterialet, vilket medför att den inte förmår leverera precisa resultat på okänt material (*overfitting*, Rao & McMahan, 2019:53–54). Testsetet används när modellen är färdigtränad för att bedöma hur väl den fungerar i jämförelse med andra modeller (i vårt fall med LatinBERT).

Märk att modellen medvetet tränas på olemmatiserat material. Skälen därtill är dels ovannämnda skepticism i Schlechtweg et al. 2020, dels ett samtal med Francesco Periti som bekräftade att det formrika latinska språket innebär att olika ordformer kan förväntas förekomma i olika semantiska kontexter. En latinist kan exempelvis förväntas ha en intuitiv förståelse för att *corpore* skulle kunna tendera att oftare förekomma efter *in* än *corpus*. Dessutom kunde av skäl som redovisas i avsnitt 4 en större andel lemman inte extraheras, och att träna modellen på den typen av data hade tveklöst inverkat negativt. Icke desto mindre avser jag att använda ordens metadata i nästa steg, när word embeddings hämtas. Anledningen är att vi genom att gruppera ordformer under ett lemma arbetar med högre frekvenser, vilket skulle kunna leda till mer precisa resultat. Även om ordformer som *corpore* och *corpus* skulle kunna tendera att förekomma i skilda omedelbara semantiska konstruktioner, kan vi åtminstone anta att de i meningar där betydelsen är ”armékår” förekommer tillsammans med andra ord än om de dyker upp i meningar där de betyder ”kropp” eller ”lik”.<sup>22</sup>

Jag har upptäckt ett fel i koden för tokeniseraren bakom CLTK, vilken används av LatinBERT (och därmed också av min omtränade modell). Det lilla felet gjorde att tokeniseraren inte kunde skilja enklitiska ord och betraktade ord som *Christusque* som en enda token istället för *Christus* + *-que*. Jag har följt ett blogginlägg på Github, där också LatinBERT publicerades, för att åtgärda felet men noterar att den lagade tokeniseraren misstolkar ablativändelser på *-ne*, i exempelvis *ratione*, som den enklitiska partikeln *-ne*, något jag i projektets inledande fas varken hade tid eller förmåga att åtgärda (Polycrates, 2022). Min bedömning var nämligen att enklitiska ord, särskilt *-que*, är så frekventa att det är bättre att acceptera det relativt sällsynta felet med *-ne*, som jag förresten först upptäckte när modellen hade tränats färdigt, än att modellen skulle lära sig vektoriella representationer för talrika ord som *Christusque*. Notera också att den felaktiga tokeniseraren publicerades tillsammans med LatinBERT på Github, vilket tyder på att LatinBERT skulle kunna ha tränats utan att enklitiska ord tokeniserats.

Jag har i övrigt i detta projekt inte hunnit fördjupa mig i hur väl Wordpiece tokenizer fungerar för latinets del, utan har accepterat den som följde med LatinBERT. Eftersom BERT kan hantera högst 512 wordpiece tokens åt gången filtrerar jag således bort alla meningar som överskrider detta antal, vilket gäller t.ex. särdeles långa ciceronska perioder. Alternativet hade varit att trunkera dessa meningar, men detta är inte förenligt med mitt kriterium om att träna modellen på endast *fullständiga* meningar. Risken är annars att den lär sig att meningar kan starta och sluta hur och var som helst. Notera att filtreringskriteriet även gäller när word embeddings i ett senare skede genereras.

Sist provas den framtagna modellen i följande tre tester. Först en MLM-uppgift (Masked Language Modeling) på testsetet. Sedan en kvalitativ uppgift i vilken olika ord maskeras i några utvalda eller egenhändigt författade meningar och modellerna producerar gissningar. Sist testas modellerna på ett s.k. evaluation dataset (även kallat *gold standard*) speciellt framtaget för latinsk semantik i samband med SemEval-tävlingen 2020 (Schlechtweg et al. 2020).

---

<sup>22</sup> För dessa olika betydelse, jfr. Ahlberg 1966, s.v. *corpus*, *-oris*, *n.* a), b) och c) β).

### 3.2.3 Graded Change Detection

Uppgiften att mäta diakron semantisk variation operationaliseras med utgångspunkt från Periti & Tahmasebi (2024) framgångsrika utvärdering av de olika metoder som de senaste tio åren tillämpats i s.k. *Graded Change Detection* (GCD). Målet är att använda XPLatinBERT för att räkna ut hur mycket ett ord har förändrats semantiskt mellan två diakront baserade korpora med hjälp av kontextuella embeddings. Periti & Tahmasebi skriver att de mest framgångsrika metoderna för GCD är APD (Average Pairwise Distances) och PRT (Prototype Embeddings, 2024:4266), vilket bekräftades senare i Periti & Montanelli (2024:11) samt i ett personligt samtal med Periti.

Till skillnad från meningsbaserade metoder, enligt vilka semantisk förändring utgår från betydelser (*word senses*) som härleds genom kluster och WSI, baseras s.k. formbaserade metoder som APD och PRT på ordformer  $w$  vars embeddings jämförs i två tidsbaserade korpora som vi kallar  $C_1$  och  $C_2$ , i den här uppsatsen en förkristen ( $C_1$ ) respektive *Patrologia Latina* ( $C_2$ ). För varje  $w$  extraheras respektive word embedding, och alla dessa embeddings samlas i ett set kallat  $\Phi_1$  respektive  $\Phi_2$  för  $C_1$  respektive  $C_2$ , enligt metoden beskriven hos Periti & Tahmasebi (2024:3). Således samlas exempelvis ett set  $\Phi_1$  av  $n^{23}$  förekomster av ordformen *corporis* i den förkristna korpusen och ett set  $\Phi_2$  av  $n$  förekomster av *corporis* i *Patrologia Latina*. I den här uppsatsen bygger jag även set av word embeddings för varje lemma  $l$  i respektive korpus.

Den initiala tanken var att följa Keidar et al. (2022) och använda mig av dimensionsreduktionstekniken PCA (Principal Component Analysis) och därmed reducera dimensionerna från 768 till 100 innan APD och PRT räknades ut. Den datakapacitet som dessa beräkningar förutsätter visade sig dock vara för stor, inte minst då fler studenter mot slutet av terminen tränade sina modeller på samma gpu-server, vilken i optimeringssyfte kastade ut min session när mitt script gjorde anspråk på för mycket CPU och RAM-minne och därmed sänkte hastigheten på mina medstudenters arbeten. Därför har jag inte utfört någon PCA. Jag har också anpassat algoritmen för den lemmavisa uträkningen av APD och PRT av samma anledningar.

#### 3.2.3.1 Average Pairwise Distances (APD)

I APD tillämpas enligt Periti & Tahmasebi (2024:4265) en enkel s.k. *distance metric* för att räkna ut hur mycket varje  $w$  i  $\Phi_2$  skiljer sig från motsvarande  $w$  i  $\Phi_1$  på följande vis: först beräknas cosinuslikheten mellan varje enskild embedding  $\{a_1, a_2, \dots, a_n\} \in \Phi_1$  av  $w$  med varje enskild embedding  $\{b_1, b_2, \dots, b_n\} \in \Phi_2$  av samma  $w$ .<sup>24</sup> Cosinuslikheten är ett värde mellan -1 och 1 som anger graden av likhet mellan de jämförda vektorernas vinklar. Om vektorerna ”pekar ungefär åt samma håll” anses  $w \in \Phi_1$  och  $w \in \Phi_2$  vara semantiskt lika. I nästa steg beräknar man cosinusavståndet ( $d$ ) från cosinuslikheten såsom  $1 - \text{cosinuslikheten}$  (detta i syfte att undvika negativa tal). Cosinusavståndet varierar således mellan 0 (identiska vektorer) och 2 (diametralt motsatta vektorer). Sist summeras alla cosinusavstånd, och den erhållna summan divideras med antalet embeddings i  $\Phi_1$  multiplicerat med antalet embeddings i  $\Phi_2$  (detta i syfte att beräkna ett medelvärde). APD är alltså ett mått på hur mycket varje  $w$  skiljer sig semantiskt mellan  $\Phi_1$  och  $\Phi_2$ .

$$\text{APD}(\Phi_1, \Phi_2) = \frac{1}{|\Phi_1| \cdot |\Phi_2|} \cdot \sum_{a \in \Phi_1, b \in \Phi_2} d(a, b)$$

Hypotesen bakom APD är att skillnader mellan  $\Phi_1$  och  $\Phi_2$  indikerar semantisk variation. Om variationen beror på polysemi är resultatet således en variation i graden av polysemi.

---

<sup>23</sup>  $n$  står för ”antal”.

<sup>24</sup> med  $\{a_1, a_2, \dots, a_n\} \in \Phi_1$  menar jag ”varje enskild embedding  $a$  av varje enskild  $w$  tillhörande setet  $\Phi_1$ .”

### 3.2.3.2 Word prototype (PRT)

PRT är en mycket mindre dataintensiv algoritm. I PRT räknas först en s.k. *prototype embedding* för varje  $w$  i  $\Phi_1$  och i  $\Phi_2$  som ett medelvärde av alla embeddings i respektive set. Medelvärdena kallas  $\mu_1$  respektive  $\mu_2$ . Hypotesen bakom PRT är att medelvärdena representerar ett slags prototypisk betydelse av  $w$ , och att man genom att beräkna cosinusavståndet mellan  $\mu_1$  och  $\mu_2$  således kan få ett mått på semantisk variation (Periti & Tahmasebi, 2024:4265). Formel sammanfattas på följande vis:

$$\text{PRT}(\Phi_1, \Phi_2) = 1 - d(\mu_1, \mu_2)$$

Observera att vi inte heller här skiljer mellan de olika betydelser som varje lemma kan rymma.

### 3.2.4 Uträkning av APD och PRT per lemma

I stället för att tillämpa APD på ett lemmas alla ordformer på en gång valde jag att tillämpa detta på de olika ordformerna separat, enligt den tidigare beskrivningen, och sedan beräkna ett medelvärde för hela lemmat. Resultaten viktas genom att för varje ordform summera antalet embeddings i  $C_1$  och  $C_2$ , multiplicera det erhållna sammanlagda antalet embeddings med APD-resultaten för ordformen (viktning), summera alla dessa viktade resultat för respektive ordform för att få totalen för varje lemma och sist dividera det med det sammanlagda antalet embeddings för hela lemmat. Formeln skulle kunna sammanfattas på följande vis (formeln gäller även för PRT):

$$\text{APD}_{\text{lemma}(\Phi_1, \Phi_2)} = \left( \frac{\sum_{w \in \text{lemma}} (|\Phi_1, w| + |\Phi_2, w|) \cdot \text{APD}_w}{\sum_{w \in \text{lemma}} (|\Phi_1, w| + |\Phi_2, w|)} \right)$$

Eftersom vi inte hämtar alla ordformer på en gång behövs ingen dimensionsreduktion med PCA, särskilt med tanke på att vi ibland tampas med rätt låga ordfrekvenser. Jag tolkar nämligen den bakomliggande hypotesen bakom användningen av PCA hos Keidar et al. (2022) som att man vill reducera den observerade variationen som föreligger mellan ordformerna för att få ett slags ”förprototypisk” representation av lemmat i fråga, alternativt att man vill förstärka effekten av polysemi i det undersökta  $w$ , innan man tillämpar APD eller PRT, något som alltså inte behövs med min algoritm.

Den teoretiska utgångspunkten för algoritmen är antagandet att latinets olika ordformer tenderar att förekomma i olika semantiska konstruktioner, vilket redan påpekats ovan i fallet *corpore*. Det är således mer logiskt att jämföra APD- respektive PRT-resultaten för *corpore* i  $C_1$  med dem för *corpore* i  $C_2$  och vikta resultaten i förhållande till hur mycket denna jämförelse bidrar till att förklara lemmat *corpus* sammanlagda variation (dvs hur frekvent *corpore* är i båda korpora).

### 3.2.5 Avgränsning

Efter att material och metod valts återstår att göra en. Att jämföra en 768-dimensionell vektor för varje ordform med samma ordforms samtliga vektorer i två korpora utgör som vi har sett en oerhörd påfrestning för en dator, och ett fullständigt korpusdrivet tillvägagångssätt med de till buds stående medlen omöjliggörs därmed. Utan att fördenskull svika vår datadrivna ansats måste alltså en avgränsning ske, detta efter två aspekter.

### 3.2.5.1 Ordklasser och frekvenser

En rimlig utgångspunkt är att begränsa studien till substantiv och adjektiv, då dessa två ordklasser förväntas i högre utsträckning vara föremål för diakron semantisk variation än exempelvis verb. Detta ställer krav på att metadata i form av lemma och ordklass i så stor utsträckning som möjligt skall extraheras från de undersökta korporan.

Keidar et al. har i sin studie observerat en korrelation mellan frekvenser av slangord som *duckface*, och polysemi (2022). Schlechtweg et al. har också observerat detta och noterar att modeller tenderar att räkna ut större semantisk variation med lågfrekventa ord och med ord som uppvisar stor variation i ordfrekvenser mellan två korpora (2020:10). Ordfrekvens är alltså en viktig variabel för vår studie och dessa tidigare resultat visar vikten av att sätta en minimifrekvens och dokumentera de observerade frekvenserna i  $C_1$  och  $C_2$ . Vi behöver även definiera ett tak på hur många ordformer vi rent datatekniskt kan jämföra med varandra. Utöver en minimifrekvens måste vi också säkerställa att de undersökta lemmerna inte råkar förekomma endast i några få texter. Detta gäller exempelvis mycket specifika ord som är tämligen frekventa i en viss pjäs av Plautus men knappt hittas någon annanstans.

Jag sätter en något godtycklig minimifrekvens på 50, range 15, maxfrekvens 1 000 i LA. Dvs varje lemma, för att inkluderas i undersökningen, behöver företrädesvis minst 50 olika ordformer i minst 15 olika texter i  $C_1$  och i  $C_2$  men max 1 000 i LA.

### 3.2.5.2 Tidsaspekten

För att besvara  $F_2$  delas PL i två subkorpora: PL1 med alla texter i PL som skrivits mellan 250 och 600 e.Kr., och PL2 som innehåller alla andra texter i PL.<sup>25</sup>

---

<sup>25</sup> N.B.: Burtons *terminus post quem*, nämligen 200 e.Kr., efterlevs inte, eftersom det förekommer ytterst få texter mellan 200 och 250 e.Kr. i PL (jfr. diagram 2).

## 4 Resultat

Resultaten redovisas i två steg. Först framtagningen av undersökningens mätinstrument XPLatinBERT, och sedan dess tillämpning för att besvara forskningsfrågorna.

### 4.1 Framtagning av XPLatinBERT

Först kommer jag att presentera korpusen som byggts ur *Patrologia Latina*, sedan hur träningsprocessen genomfördes, och avslutar med olika tester av den framtagna modellen.

#### 4.1.1 Initiala överväganden

Jag har från början bestämt mig för att extrahera de meningar och deras sent\_id så som de hittas i XML-filerna, efter att de rensats från metadata och korrigerats för eventuella felaktigheter och oegentlig data. Ett av syftena var att behålla interpunktionen, då den till viss del kan bidra med semantisk information, även om den i sig är en redaktionell efterhandskonstruktion.

Det verkar dock föreligga några parsningsfel i de ursprungliga XML-filerna, men inte av en sådan omfattning att det skulle ha någon nämnvärd effekt på träningen av XPLatinBERT. Det handlar dels om meningar som tokeniserats fel i det att meningsgränsen hamnat på fel ställe, dels om ord som lidit ett liknande öde och slagits ihop med angränsande element (t.ex. *Christi732sunt*). Eftersom jag utgår från de i XML-filerna tokeniserade meningarna accepterar jag dem som de är och filtrerar i görligaste mån automatiskt bort sådana som uppenbarligen innehåller fel, där felen inte kunnat rättas eller pareras på annat vis.

En liknande företeelse är det jag kallar ”ofullständiga meningar”. Några meningar består av endast ett par skiljetecken, andra är ofullständiga då källtexten uppenbarligen innehöll *lacunae*, vilka translittererats med ”...”, medan ytterligare inte avslutas med skiljetecken. Alla dessa typer av ofullständiga meningar undantas från studien, då risken finns att de kan snedvrider träningen något. Vidare undantas meningar som består av färre än tre ord, främst för att fånga upp parsningsfel som leder t.ex. till att en mening består av ett enda skiljetecken. Jag håller med om att detta är en godtycklig avgränsning och att mycket korta meningar förvisso på sitt sätt bidrar till att placera orden i det vektorfyllta rummet. I meningen *Salve Paule!* har *Paule* (förutom vokativändelsen) visserligen ingen särskild semantisk relation till *salve*, och det hade lika gärna kunnat stå *Salve Sulpicia!* Men att *Paule* eller *Sulpicia* (eller för den delen *Marce*) dyker upp efter *Salve* i en kort mening bidrar till den semantiska informationen ”*Salve* kan följas av en s.k. *named entity* (med ändelsen *-e* om lemmat är maskulint och har en grundform på *-us*) och ett skiljetecken i en kort mening”. Det bör dock noteras att dessa korta meningar till sin natur motsvarar en försvinnande liten del av korpusen och därmed är denna godtyckliga gränsdragnings effekt försumbar.

Där meningar korrekt taggats som rubriker i XML-koden, undantas de från träningen, men notera att det förekommer rubriker som taggats fel. Vidare förekommer många grekiska citat, vilket gör tokeniseringen inför träningen utmanande, och fascinerande nog undantagsvis även andra språk, som några meningar på en äldre variant av tyska under den anmärkningsvärda rubriken *Abrenuntiatio diaboli operumque ejus et superstitionum*:

8376:4;7,7: *End ec forsacho allum dioboles wercum, und wordum thuna eren de woden ende saxtonte, ende albem them unholdum, the hira genotas sint.*<sup>26</sup>

Där dessa språk inte utgjort merparten av texten har jag behållit dem. Detta gäller alltså inte *Versio Bibliorum* av Ulfilas Moeso-Gothorum samt *Fragmenta alia Gothicae linguae*, som är helt skrivna på gotiska, samt *Excerpta de poemate de morte* av Helinandus Frigidi Montis, som till övervägande del är skriven på en äldre variant av franska. I texten *Praefationis altera pars* av Nicolaus Claraevallensis har jag manuellt trunckerat slutet motsvarande uppskattningsvis en femtedel<sup>27</sup>, som bestod av en ed på en äldre variant av franska. I valida grekiska meningar har jag maskerat grekiska ord med en specialmarkör [GREEK] och en etikett (5), så att modellen lär sig att det finns ett grekiskt ord utan att tränas på just det.<sup>28</sup> Genom att hämta alla *hapaxlegomena* upptäckte jag vidare att en specifik text bestod av dikter som hade återgivits helt utan modern editering, varpå hela verser tokeniserats som ett enda ord, som t.ex. *AVGVSTOETFIDEICHRISTISVBLEGEPROBATA* i position 7062:17;2,1. Den texten tillför inget av värde till XPLatinBERT och undantas därför från studien.

En ur filologisk synpunkt viktig iakttagelse är att det inte alltid görs någon skillnad i XML-filerna mellan utgivarens företal, om sådant finns, och själva texterna. Detta gör det omöjligt för oss att extrahera endast själva texten, utan företalet följer i förekommande fall med, såtillvida det inte XML-taggs som rubrik och därmed undantagits av ovan angivna skäl. Eftersom Migne ofta givit ut texterna i befintligt skick, vet vi inte heller vem denne utgivare är (Migne eller utgivaren av den tryckta upplagan som Migne gav ut). Detta är problematiskt om man vill ägna sig åt ett kvalitativt filologiskt arbete med korpusen, men det finns flera skäl till att detta inte är problematiskt för syftet i denna uppsats. För det första finns det inte alltid ett företal och där det finns upptar det en mycket liten del av texten eller så är det taggat som rubrik och faller därmed bort. För det andra, om företalet tar upp företeelser som är typiska för kristet latin, bidrar de till den semantiska information som vi vill ladda XPLatinBERT med, och annars bidrar de ändå inte med data som skulle snedvrída resultaten på något märkbart sätt.<sup>29</sup> Det kan dock medges att det hade varit bättre att skilja företalet från alla texter.

#### 4.1.2 Deskriptiv statistik

Efter dessa filter finns alltså 5 255 dokument kvar i korpusen, av ursprungligen 5 277 dokument enligt informationen på *Corpus Corporum*. I tabell 2 visas fördelningen över antal ord och meningar i olika kategorier. De meningar som utgör träningsmaterialet för XPLatinBERT hamnar i kategorin ”valida”, medan fragmentariska är sådana som innehåller *lacunae*, icke-meningar avslutas inte med skiljetecken och korta meningar består av färre än tre ord/tecken. Som grekiska meningar räknas sådana som består av minst fem gånger fler ord skrivna med grekiska bokstäver än ord skrivna med latinska.<sup>30</sup> Sammanlagt extraheras 3 626 722 valida meningar och 78 023 175 ord ur 5 112 texter, vilket kan jämföras med de 85 539 587 ord som *Patrologia Latina* på *Corpus Corporum* anges bestå av.<sup>31</sup> Med tanke på att jag sällar bort rubriker, olika typer av icke-meningar och grekiska meningar, tycks det som att huvuddelen av materialet ändå har kunnat extraheras. Att 22 texter inte producerar några valida meningar alls beror på att deras format inte motsvarar det förväntade: vissa består av enbart rubriker, andra har inte parsats,

<sup>26</sup> Alla meningar ur materialet föregås av sitt sent `_id`. För en lista över alla dokument `_id` och motsvarande titlar, se Lafage (2025a).

<sup>27</sup> Från 11259:2;1,1 till 11259:2.3;3,16.

<sup>28</sup> Som vi snart skall se har detta resulterat i en undervariant av XPLatinBERT som visar sig undermålig.

<sup>29</sup> Man kan knappast tänka sig att det i ett företal skulle förekomma många polysema ord med endast klassisk betydelse, då ett företal i regel ju är en metatext om texten i fråga och svårligen tar upp andra teman än de i texten förekommande.

<sup>30</sup> Märk att romerska siffror följer med i beräkningen, likaså transkriberingar från andra språk.

<sup>31</sup> En konvention är att räkna antal ord med python-metoden `.split()`, som jag själv använt mig av. Då vi dock inte vet hur antalet ord på *Corpus Corporum* räknats behöver vi vara försiktiga vid jämförelser.

eller så består de av endast en mening, som inte uppfyller kriterierna för att räknas som valid. Jag hade visserligen kunnat själv parsa dessa texter och därigenom utvinna ytterligare ord, men de extraherade meningarna är tillräckliga i antal för att träna XPLatinBERT och jag föredrar att lägga min tid på träningen istället. En översikt hittas i tabell 3.

meningar	n texter	n meningar	n ord	n latin	n grek.	$\bar{x}$ ord/mening	% grek.	% extrah.	% texter
valida	5 112	3 626 722	78 023 175	77 913 375	18 590	21,51	0,02	98,18	97,28
fragm.	1 343	11 258	233 878	233 091	88	20,77	0,04	0,29	25,56
icke-men.	2 057	24 364	439 832	438 355	111	18,05	0,03	0,55	39,14
korta men.	3 909	197 192	447 411	437 747	140	2,27	0,03	0,56	74,39
grek. men.	162	9 733	322 515	2 375	320 031	33,14	99,26	0,41	3,08
total	5 255	3 869 269	79 466 811	79 024 943	338 960	20,54	0,43	100,00	100,00

Tabell 3. Fördelning över antal ur *Patrologia Latina* extraherade ord och meningar. Notera att antal ord (*n ord*) räknas med .split()-metoden medan antal ord skrivna med latinska respektive grekiska bokstäver (*n latin*, *n grek.*) räknas med var sitt Regex-uttryck, vilket leder till skillnader i resultaten.

Tabell 3 visar att 98,18% av allt det som extraheras går vidare till träningen och att 97,28% av texterna producerar dessa valida meningar. 25% av alla texter innehåller *lacunae* (*fragm.*), medan grekiska meningar hittas i endast 3% av texterna och motsvarar ca 0,43% av korpusen. I valida meningar är endast ca 0,02% av orden skrivna med grekiska bokstäver. Observera att medan de "latinska" meningarna i genomsnitt består av ca 20 ord (förutom korta meningar, då dessa ju består av max 3 ord), består de grekiska av ca 30 ord.

De extraherade valida meningarna bildar korpusen PL. Diagram 1 ger en visuell översikt av författare och texter i PL.

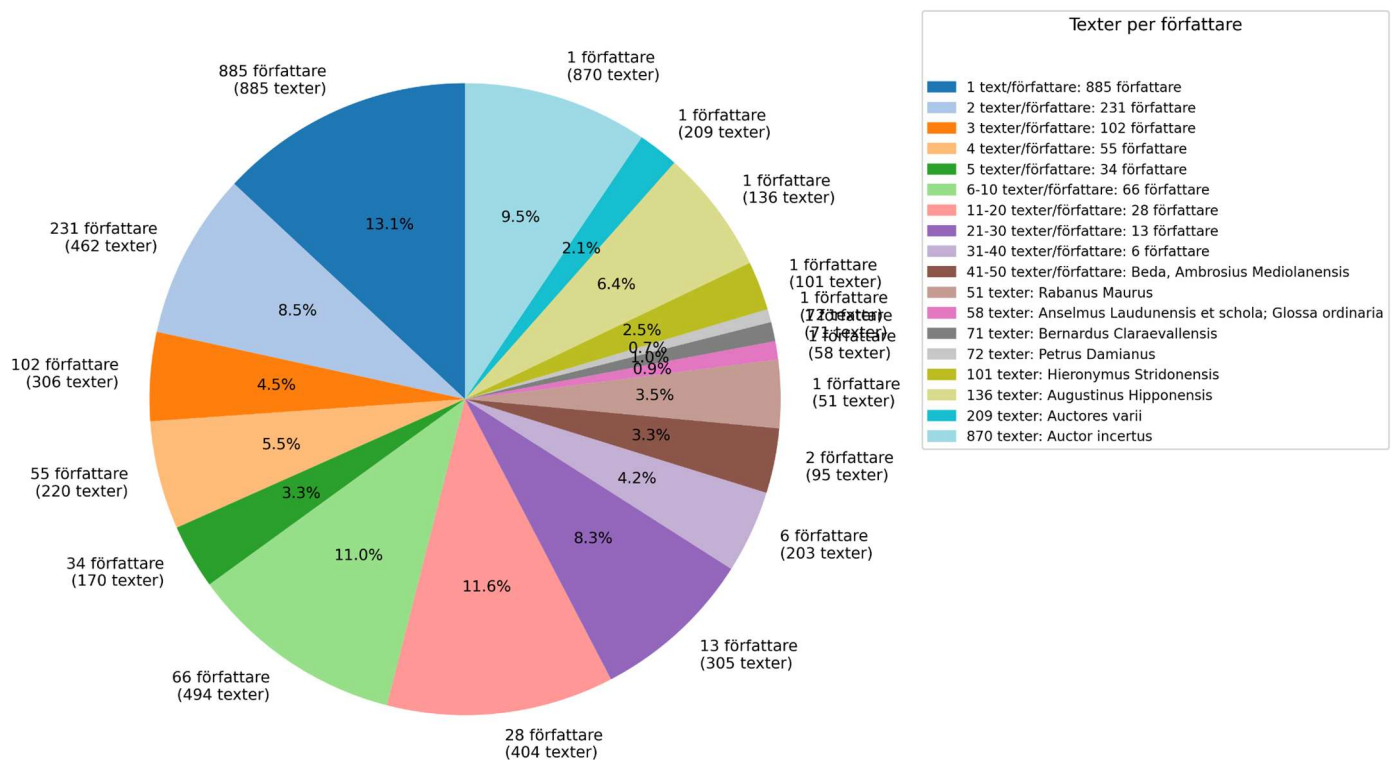


Diagram 1. Fördelning texter/författare i olika kategorier baserade på antal texter per författare i PL. Tårtbitarnas storlek står för antal ord och procentsatserna anger andelen ord på hela korpusen.

Sammanlagt extraheras valida meningar från 1 430 olika författare, och således har vi ett genomsnitt av 3,57 texter per författare, men som vi snart skall se varierar detta mycket. Diagram 1 visar att knappt 10% av materialet är skrivet av "Auctor incertus", och givetvis handlar det om *olika* okända författare. Ytterligare 2% är skrivet av "Auctores varii". Utöver det kommer drygt 13% av orden i PL från texter som vi bara har en av per författare. Detta gäller författare som Hugo Lugdunensis, Albinus Eremita eller en viss Frigobertus. I andra ändan av spektrumet har vi författare som vi har många texter av. Hela 6,4% av materialet består av meningar extraherade ur 136 texter av kyrkofadern Augustinus Hipponensis. Om vi plockar ut de mest representerade författarna kan vi se att drygt 18% av korpusen består av texter skrivna av endast 8 författare (Auctores incerti undantagna, förstås). Vidare kan vi konstatera att storleken på texterna uppenbarligen varierar avsevärt. Således bidrar Rabanus Maurus 51 texter med mer material än Bedas och Ambrosius Mediolanensis 95 texter. Dessa två kan dock trösta sig med att ha bidragit med nästintill lika många ord som de 34 olika författarna i kategorin "5 texter per författare" som tillsammans skrivit hela 170 texter.

Innan tidsaspekten för datan i PL visualiseras i diagram 2, vill jag i tabell 4 visa att många texter faktiskt saknar datumstämpel.

	med datum	utan datum	total	%
<b>författare</b>	656	774	1 430	45,87
<b>ord</b>	67 933 804	10 089 371	78 023 175	87,07
<b>texter</b>	3 774	1 338	5 112	73,83

Tabell 4. Antal författare, ord och texter med och utan datumangivelse i metadata.

Av tabell 4 framgår att knappt hälften av alla författare i korpusen har en datumstämpel på sina alster, medan nästan tre fjärdedelar av alla texter motsvarande 87% av alla ord har det. Detta betyder alltså att många texter utan datumangivelse tillhör författare som vi har få texter av och att dessa texter tenderar att vara relativt små. Dessa datumlösa författare och texter hamnar av praktiska skäl vid tidsaxelns begynnelse på år 0 i diagram 2, vilket alltså inte betyder att Christi födelse skulle ha föranlett en

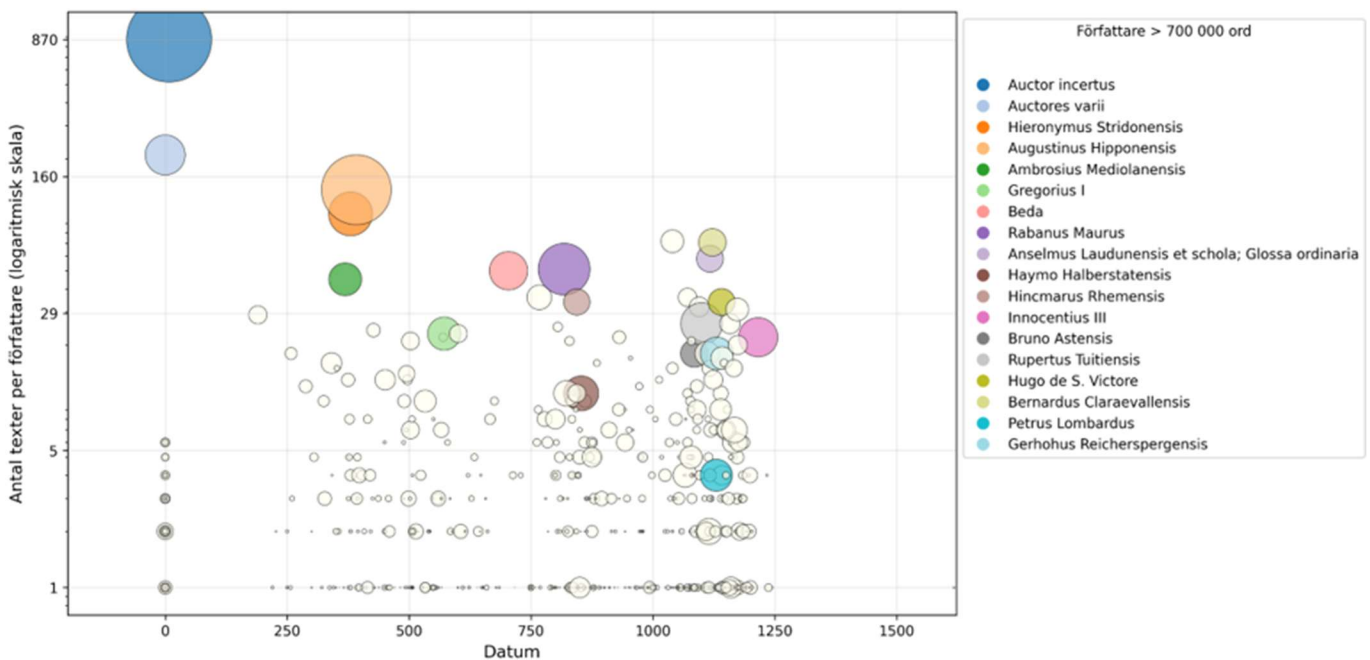


Diagram 2. Antal texter per författare över tid i PL. Bubblornas storlek står för antal ord per författare.

explosion av kristen litteratur. Notera att de författare som bidragit med över 700 000 ord även markeras ut i legenden.

Diagram 2 visar att det mesta av materialet spänner en period på ca 1 000 år mellan ca 250 e.Kr. och 1 250 e.Kr. Vidare skulle vi kunna dela in datan i tre huvudsakliga perioder: en tidigmedeltida (250 – 600) med bjässarna Hieronymus, Augustinus, Ambrosius och Gregorius, en karolingisk period (700 – 900) som domineras av bl.a. Beda och Rabanus Maurus, samt en högmedeltida (1100 – 1250). Av dessa verkar den senare samla flest datapunkter och mest material.

Slutsatsen är alltså att korpusen ser ut att vara långtifrån homogen och balanserad, men att den verkar bestå av många olika texter skrivna av olika författare från ett större tidsspann. Tyvärr är XML-filerna inte annoterade med avseende på geografiskt ursprung, genre eller texttyp, men stickprov tagna ur olika delar av materialet visar att vi har med allehanda genrer att göra (traktat, krönikor, brev, poesi, m.m.) och att de handlar om kristna teman.

Dessa initiala iakttagelser är viktiga då vi vill säkerställa att vårt urval (PL) är representativt för populationen ”kristet latin”. Enligt McGillivray är den bästa metoden för att bilda en korpus *slumpmässigt urval* (2014:16–17). Om vi går tillbaka till historien bakom tillblivelsen av *Patrologia Latina*, kan vi konstatera att slumpen har varit en bidragande faktor. Att specifika texter följt med medan andra strukits, särskilt små sådana vars författare vi bara har en text av, beror delvis på revolutionärernas slumpartade samlande och historiens nycker, branden på L'imprimerie catholique som förstörde många texter, vilka fick rekonstrueras, till viss del kanske Mignes smak, tillgängligheten av äldre editioner versus manuskript, problem med OCR-läsningen, Philipp Roellis missar i parsningen, filer som inte kunde packas upp. Andra texters närvaro i PL är allt från slumpmässig, och detta gäller i högsta grad de stora författare som nämnts ovan, såsom Augustinus eller Hieronymus. Att de ingår och med sitt omfång bidrar till att göra korpusen mindre balanserad beror förmodligen på att de är så centrala för idén om kristet latin att deras utelämnande inte kunde motiveras. Å andra sidan eliminerar mitt val att inkludera hela *Patrologia Latina* som den är all vinkling. Av skäl angivna ovan är det ytterst svårt om ens möjligt för oss latinister att skapa korpora som är helt representativa för en viss varietet, men jag utgår från hypotesen att det relativt stora antalet meningar i PL utgör ett tillräckligt stort urval för att vara representativt för den population som vi ämnar studera. I min bedömning av att texterna är representativa för kristet latin är jag dock till syvende och sist tvungen att förlita mig på Mignes uttalade syfte med samlingen.

### 4.1.3 Träning av XPLatinBERT

Eftersom den typen av träning som vi avser att göra kräver mycket stor datakapacitet, låter den sig inte göras på en vanlig CPU (Central Processing Unit), vilket är den allra vanligaste typen av processorer i dagens PC. Dessa processorer bearbetar data seriellt (alltså i tur och ordning) och skulle ta orimligt lång tid (flera veckor eller månader, om de överhuvudtaget klarar det utan att krascha). Istället brukar man arbeta på en annan typ av processor, antingen GPU (Graphics Processing Unit)<sup>32</sup>, vilka bearbetar data parallellt och därmed är mycket snabbare än CPU, eller ännu bättre TPU (Tensor Processing Unit), vilka är specialanpassade för att effektivisera maskininlärningsalgoritmer. Jag använder mig av en av GU:s två GPU-serverar, närmare bestämt mltgpu-2, som Robert Adesam och Kaj Ailomaa har ställt till mitt förfogande under hela terminen.

#### 4.1.3.1 Inledande reflektioner

Fördelningen mellan tränings-, validerings- och testset enligt avsnitt 3.2.2 visas i tabell 5.

---

<sup>32</sup> Dessa datorer utvecklades ursprungligen för att förbättra upplösningen i dataspel, varav namnet.

	n ord	n dok.	$\bar{x}$ ord/dok.	% ord
träningsset	74 122 016	4 873	15 211	95,00
valideringsset	1 950 564	132	14 777	2,50
testset	1 950 595	107	18 230	2,50
total	78 023 175	5 112	15 263	100,00

Tabell 5. Fördelning över antal ord och texter i träningssetet, valideringssetet och testsetet.

Som framgår av tabell 5 har randomiseringsfunktionen resulterat i färre dokument i testsetet än i valideringssetet, men detta har ingen större betydelse. Två miljoner ord i 87 381 meningar fördelade över 107 slumpmässigt valda texter borde ändå utgöra ett någorlunda balanserat urval ur olika delar av PL. Urvalet hade visserligen varit mer representativt för korpusen om meningar hade slumpmässigt valts ut från korpusens alla texter, men risken är att modellen i sådant fall känner igen meningsbyggnader och framför allt teman ur lästa texter när det tillämpas på testsetet, t.ex. om textspecifika ord såsom ortnamn eller personnamn förekommer i både träningsset och testset. Detta skulle ge XPLatinBERT en skevande fördel över LatinBERT när modellerna jämförs.

LatinBERT tränas om genom en s.k. *unsupervised MLM-task* utförd på träningssetet. Vid varje epok maskeras slumpmässigt 15% av alla tokens med undantag för specialmarkörerna [CLS] och [SEP], vilka läggs till i början respektive slutet av en mening. Därefter försöker modellen gissa de maskerade orden och vikterna justeras sedan genom backpropagation (Eisenstein, 2019:55–56).

Då detta var min första maskininlärningsuppgift undgick jag inte att göra talrika nybörjarmisslag, som jag lyckligtvis tror mig ha lärt mig mycket av, men som har fördröjt arbetet. Jag tränade färdigt sammanlagt två modeller, efter att jag långt senare upptäckte att den första, XPLatinBERT\_1 kallad, hade tränats på ett oväntat sätt. För att exkludera grekiska ord, vilka som vi har sett genomsyrar PL, från träningen, hade jag nämligen skapat en specialmarkör [GREEK] och justerat hela ordbokens indexering för att garantera ett indexnummer för den. Detta gjorde att LatinBERT vid första epoken hämtade fel vikter, men som vi snart skall se bidrog den stora textmängden som PL består av och de olika epoker som modellen tränades på till rika tillfällen att uppdatera vikter från det felaktiga utgångsläget. Jag rättade sedan felet och tränade om LatinBERT igen utan att beakta grekiska ord och utan att ändra indexeringen av orden i ordboken, vilket resulterade i en XPLatinBERT\_2. Från och med denna punkt i uppsatsen kallas de tre modellerna LB (LatinBERT), XPL1 (XPLatinBERT\_1) och XPL2 (XPLatinBERT\_2)

De för uppsatsen relevanta hyperparametrar som jag efter varje epok eventuellt justerat för att anpassa träningen är tre: antal epoker, *learning rate* (LR) och *batch size*, dvs hur många meningar modellen tränas på åt gången. Under arbetets gång har jag lärt mig ackumulera batchdata innan vikterna uppdateras, vilket sparar minne. Genom ackumulering kunde jag uppnå en betydligt högre effektiv batch size utan att behöva göra batcharna större och därmed överbelasta minnet. Detta förklarar varför jag började med en blygsam batch size men kunde sedan höja den. Meningarna batchas med hjälp av en dataloader som även blandar dem så att de inte kommer ut i tur och ordning (så att modellen inte lär sig anpassa vikterna efter specifika stilar). Notera att jag vid träningen av XPL2 helst hade valt en större batch size, men att jag hindrades av att gpu-servern då hade mycket mindre ledigt minne, då fler studenter höll på att träna sina modeller inför terminslutet.

En lägre batch size innebär att LR särskilt i början inte får vara för hög, då det finns för lite data i varje batch att korrekt träna modellen på och vi därför måste uppdatera parametrarna i små steg, eftersom ett högt LR vid liten batch size riskerar att leda till stora uppdateringar, vilket gör att parametrarna oscillerar i det vektorieella rummet snarare än konvergerar. Vad gäller LR skriver Bengio ”The optimal learning rate is usually close to (by a factor of 2) the largest learning rate that does not cause divergence

of the training criterion” (2012:5). Detta betyder att det mest optimala LR hittas genom att först hitta vilket LR som orsakar att cross-entropy loss börjar gå upp efter en epok och halvera det. I praktiken får man testa sig fram, men jag börjar med ett LR på  $5e-5$  (0,00005), vilket är rätt vanligt för en korpus av den storleken.

Målet är att för varje epok få ner cross-entropy loss på träningssetet så mycket som möjligt, utan att lossen på valideringssetet går upp eller än värre börjar bli högre än i träningssetet. Dessa två företeelser är nämligen tecken på att modellen är på väg att överanpassa sig till materialet (*overfitting*).

#### 4.1.3.2 Resultat av träningen

Varje epok tog ca 22 timmar och jag har genomfört totalt 9 epoker för XPL1 och 3 för XPL2. Detaljerna visas i tabell 6.

epoker	b_s XPL1	l_r XPL1	tr XPL1	val XPL1	b_s XPL2	l_r XPL2	tr XPL2	val XPL2
1	16	0,00005	0,56554	0,45862	16	0,00002	0,32420	0,30402
2	64	0,00002	0,46869	0,44776	16	0,00001	0,32139	0,30481
3	32	0,00002	0,45220	0,42915	16	0,00002	0,32153	0,30595
4	32	0,00001	0,43629	0,41238				
5	32	0,00001	0,42599	0,40335				
6	32	0,00001	0,41667	0,39917				
7	32	0,000005	0,40936	0,39444				
8	32	0,000005	0,40348	0,38550				
9	32	0,000005	0,39866	0,38570				

Tabell 6. Cross-entropy loss på tränings- respektive valideringsset per epok och för modellerna XPL1 och XPL2. b\_s = batch size, l\_r = learning rate, tr = resultat på träningsset, val = resultat på valideringsset.

Tabell 6 visar att jag valt en batch size som är avsevärt lägre än den som Gururangan (2020) och Bamman&Burns (2020) använde, nämligen 256. Gururangan arbetade dock på mycket större specialiserade korpora på flera miljarder ord (2020:3), ävenså Bamman&Burns (2020:2)<sup>33</sup>, som dessutom tränade modellen från grunden medan jag avser att göra en DAPT, och som begränsade meningarna till 256 Wordpiece tokens, medan jag har satt värdet till 512. I tabellen kan man se hur förändringstakten minskar för varje epok, även när batch size och LR inte justeras. Detta tolkar jag som ett tecken på att modellen är på väg att konvergera. XPL2 visade sig inte behöva mycket mer än en epok för att fullständigt konvergera. Detta beror på att LatinBERT:s vikter korrekt kunde hämtas redan vid första epoken. Vid tredje epoken visade modellen tidiga tecken på att överanpassa sig, varpå jag avbröt träningen. Jämför Periti och Montanelli (2024:18) som noterar att fem epoker brukar vara ett maximum, men att detta beror på korpusens storlek och domänen. I diagram 3 visualiseras utvecklingen av cross-entropy loss per epok.

<sup>33</sup> 642,7 miljoner ord, alltså ca 9 gånger fler ord än i Pat\_Lat\_val.

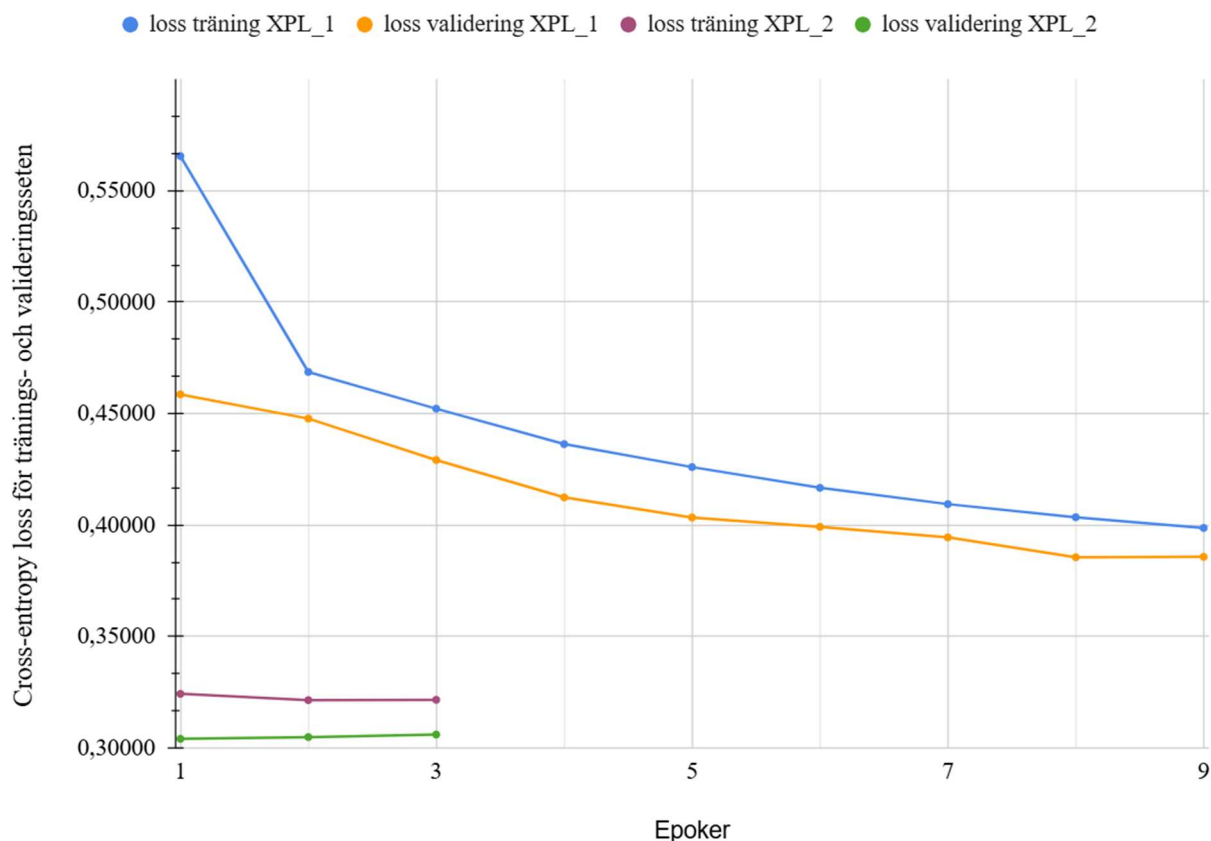


Diagram 3. Cross-entropy loss för tränings- och valideringsset per epok.

De fel som uppdagades i XPL1 föranleder oss att tolka modellens initialt branta kurva för träningssetet mellan epokerna 1 och 2 som ett tecken på att den utgått från en mindre fördelaktig position än XPL2, och vi kan ana en ännu brantare hypotetisk kurva mellan epokerna 0 och 1. Modellen har sannolikt tränats om från början, medan XPL2 troligen har anpassat redan existerande semantiska relationer i LB. XPL1 skulle således ha konvergerat de ord som modellen stött på mest, dvs de mest frekventa orden, medan mindre frekventa ord hypotetiskt inte hittat sin slutdestination i det vektoriella rummet.

#### 4.1.4 Testning av XPLatinBERT

Alla tre modeller testas enligt planen i 3.2.2 och resultaten följer nedan.

##### 4.1.4.1 Test 1 – kvantitativ bedömning på testsetet

Modellerna sätts i evaluation mode, dvs de kommer inte att kunna justera vikterna under tiden de går igenom de ca 2 miljoner ord som testsetet består av. Den genomsnittliga cross-entropy lossen beräknas därefter för varje modell när den tillämpas på okänt material, som testsetet utgör. Notera att LB visserligen delvis tränats på *Patrologia Latina* (som vi har sett i avsnittet om denna korpus med stora OCR-fel), men att korpusen endast utgör ca 4,56% av det material som LB totalt tränats på. Resultaten visas i tabell 7.

	LB	XPL1	XPL2
cross-entropy loss	1,42666	0,35371	0,28195

Tabell 7. Genomsnittlig cross-entropy loss för testsetet med alla tre modeller.

En första iakttagelse är att båda XPL-varianterna gissar ord i okänt material från *Patrologia Latina* i genomsnitt bättre än LB, vilket *a priori* indikerar att träningens syfte har uppnåtts. Förutsatt att modellerna inte överanpassats till PL, skulle detta kunna vara ett tecken på att det i testsetet finns ord, semantiska konstruktioner och återkommande teman som liknar dem som förekommer i träningssetet, och i förlängningen att de semantiska egenheter som är typiska för PL mycket riktigt skulle kunna anses utgöra en domänspecifik form av latin.

Vidare kan vi konstatera att XPL2 verkar prestera något bättre än XPL1, vilket var förväntat: förmodligen presterar XPL2 bättre med mindre frekventa ord, men XPL1 har förmodligen korrekt fångat upp de viktigaste särigheterna som PL består av.

#### 4.1.4.2 Test 2 – kvalitativ bedömning med MLM-task

Nästa test gick ut på att välja ut meningar som XPL inte tränats på eller skapa helt nya meningar, maskera ett ord åt gången och låta modellen ge oss förslag på ord som bäst skulle kunna passa i den semantiska kontexten. För detta använder jag en softmax-funktion, som omvandlar alla logits till sannolikheter och väljer ut dem fem högsta.<sup>34</sup> Låt oss ta en enkel mening ur testsetet, som XPL alltså inte tränats på. Notera att LB skulle kunna ha tränats på just denna mening.

Mening A: *Si fides tua dormit in corde tuo, tanquam in navi tua dormit Christus: quia Christus per fidem in te habitat.*<sup>35</sup>

Vi maskerar fyra ord (ett åt gången) och låter båda modeller gissa dem utifrån kontexten. Svaren sammanställs i tabell 8.

maskerat ord	LB	prob.	XPL1	prob.	XPL2	prob.
HABITAT	operatur	0,3089	habitat	0,5655	habitat	0,4527
	habitat	0,2035	est	0,1511	est	0,0980
	est	0,1525	loquitur	0,0293	operatur	0,0974
	sedet	0,0687	erat	0,0202	surrexit	0,0500
	ascendit	0,0390	uenit	0,0200	manet	0,0358
FIDEM	fidem	0,8657	fidem	0,4819	fidem	0,8954
	dilectionem	0,0574	se	0,0994	dilectionem	0,0320
	te	0,0211	dilectionem	0,0687	gratiam	0,0122
	baptismum	0,0145	te	0,0318	te	0,0085
	gratiam	0,0064	omnia	0,0280	charitatem	0,0066
SI	si	0,1802	quia	0,1705	si	0,1716
	cum	0,1740	et	0,1482	cum	0,0859
	quando	0,1227	nam	0,0893	sed	0,0772

<sup>34</sup> Logits är tal som talar om hur "säker" modellen är på att varje enskilt ord i ordboken är det efterfrågade. För att kunna tolkas behöver de först omvandlas till sannolikheter med hjälp av en softmaxfunktion.

<sup>35</sup> "Om din tro sover i ditt hjärta, är det som att Kristus sover i din båt: ty Kristus bor i dig genom tron." Alla översättningar i uppsatsen är mina egna.

	ergo	0,0642	si	0,0837	nam	0,0630
	nam	0,0589	uel	0,0456	quando	0,0629
NAVI	domo	0,3388	domo	0,3164	domo	0,3702
	petra	0,0821	anima	0,0704	umbra	0,0975
	morte	0,0762	morte	0,0629	anima	0,0802
	manu	0,0638	umbra	0,0624	morte	0,0564
	umbra	0,0392	dextera	0,0398	petra	0,0523

Tabell 8. Förslag på maskerade ord i mening A, givna av XPL1, XPL2 respektive LB, med probabiliteter. En probabilitet på 0,5655 kan läsas som ”56,55% probabilitet”.

Ur tabell 8 ser vi att de ord som alla tre modeller föreslår oftast är rimliga: den förväntade ordklassen verkar i regel gissas korrekt, och även genus samt kasusändelsen. Även om modellerna inte föreslår *nai*, förmodligen eftersom ordet inte förekommer i sitt vanliga sammanhang, dvs tillsammans med ord som har med skepp att göra, föreslås feminina ord i ablativ (på grund av närheten till *in* och *tua*). I övrigt verkar XPL2 prestera ungefär lika bra som LB förutom vad gäller *habitat*, vilket är det enda tillfälle där XPL1 överträffar de övriga modellerna. Låt oss nu testa hur modellerna reagerar om vi eliderar *tua* i *nai tua* (dvs [...] *tanquam in nai dormit Christus [...]*).

maskerat ord	LB	prob.	XPL1	prob.	XPL2	prob.
NAVI	somno	0,2950	somno	0,2124	te	0,1546
	te	0,2035	nocte	0,2085	lecto	0,1459
	est	0,1525	sepulcro	0,0865	somno	0,1157
	sedet	0,0687	die	0,0499	dormie	0,0838
	ascendit	0,0390	lecto	0,0402	petra	0,0511

Tabell 9. Förslag på det maskerade ordet *nai* givna av XPL respektive LB, med probabiliteter.

Nu gissar LB inte alltid rätt ordklass medan XPL-modellerna fortsatt föreslår ord i ablativ. En intressant iakttagelse är ordet *dormie*. I Favre (1883-1887) hittas uppslagsordet *dormia* med betydelsen ”anima” och även bibetydelsen ”cadaver, corpus exanime” med anteckningen ”a veteri Gallico *dormie*”. Detta skulle alltså kunna vara ett tecken på en liten överanpassning av XPL2 till PL, som ju som vi sett delvis består av medeltida galliska texter. För att undersöka det vidare laddar jag modellens checkpoint från första och andra epoken (a respektive b) och ber dessa undervarianter av XPL2 gissa samma ord.

maskerat ord	XPL2_a	prob.	XPL2_b	prob.	XPL2	prob.
NAVI	te	0,0864	somno	0,2436	te	0,1546
	lecto	0,0857	lecto	0,1815	lecto	0,1459
	nocte	0,0834	te	0,1152	somno	0,1157
	petra	0,0677	nocte	0,0939	dormie	0,0838
	somno	0,0458	dormie	0,0587	petra	0,0511

Tabell 10. Förslag på maskerade ord i mening A, givna av XPL2 efter epok 1 (XPL2\_a), epok 2 (XPL2\_b) och 3 (XPL2), med probabiliteter.

Vi kan konstatera att probabiliteten för att *dormie* föreslås går upp för varje epok, vilket hänger samman med vår intuition om anpassning till materialet. Vidare verkar inget förslag sticka ut efter första epoken, medan XPL2\_b och XPL2 uppger några högre probabiliteter, bl.a. *somno*, som också verkar vara

toppförslaget i LB och XPL1. Vi noterar detta och går vidare, men vi kommer att återkomma till detta senare i avsnittet.

Låt oss nu skapa en egenodlad latinsk mening för att utesluta att modellerna skulle ha kunnat möta mening A. För att göra meningen ”klurigare” för modellerna skjuter jag till en adverbial bisats inledd med *cum*. Det ”kluriga” ligger i att konjunktionen till formen sammanfaller med prepositionen *cum* samt att bisatsen är inskjuten.

Mening B: *Est in templo ara, quae cum sacerdos sacrum fecit ardere uidetur.*<sup>36</sup>

maskerat ord	LB	prob.	XPL1	prob.	XPL2	prob.
TEMPLO	media	0,1707	ea	0,1099	templo	0,0980
	alta	0,0918	alia	0,0547	ea	0,0657
	iovis	0,0480	eadem	0,0336	altari	0,0527
	aedibus	0,0474	ipsa	0,0336	aere	0,0320
	sinistra	0,0457	altari	0,0279	eadem	0,0314
ARA	flamma	0,6902	lucerna	0,0462	lucerna	0,3113
	ara	0,1089	charitas	0,0461	ecclesia	0,0522
	ignis	0,0465	ignis	0,0440	flamma	0,0460
	lucerna	0,0383	arca	0,0286	domus	0,0438
	faces	0,0160	fides	0,0270	domini	0,0380
SACRUM	signum	0,1262	templum	0,3120	se	0,0784
	verba	0,1174	se	0,0720	altare	0,0767
	sacra	0,1085	tabernaculum	0,0376	hoc	0,0346
	sacrum	0,0971	eos	0,0280	eam	0,0267
	incendium	0,0928	altare	0,0267	eum	0,0260
VIDETUR	coepit	0,4281	templum	0,2499	coepit	0,0978
	uidebatur	0,1641	altare	0,1191	templum	0,0616
	solet	0,0873	corpus	0,0643	fecit	0,0551
	uidetur	0,0736	sacrificium	0,0554	sacrificium	0,0476
	incipit	0,0640	tabernaculum	0,0516	facit	0,0468

Tabell 11. Förslag på maskerade ord i mening B, givna av XPL1, XPL2, respektive LB, med probabiliteter.

I det här exemplet verkar LB oftast gissa rätt, förutom i fallet *templo*. Notera att LB ofta föreslår ord som kan knytas till den hedniska religionen (*Iovis, ara, sacrum*), medan XPL1 och XPL2 föreslår ord som *altari, charitas, ecclesia, domini, tabernaculum*. Observera att XPL1 gissar fel ordklass för *uidetur*, medan LB verkar vara bäst på det. Överlag kan vi se att XPL2 återigen överträffar XPL1, vilket var förväntat. Nu ändrar vi meningen och gör den mer ”kristen”:

Mening C: *Est in ecclesia altare, quod cum episcopus sacrum fecit ardere uidetur.*<sup>37</sup>

<sup>36</sup> ”Det finns i templet ett altare, som anses brinna var gång en präst frambär ett offer.”

<sup>37</sup> ”Det finns i kyrkan ett altare, som anses brinna var gång en biskop frambär ett offer.”

maskerat ord	LB	prob.	XPL1	prob.	XPL2	prob.
ALTARE	simulacrum	0,1627	altare	0,0411	altare	0,2485
	altare	0,1600	illud	0,0330	templum	0,0943
	incendium	0,1331	sanctorum	0,0305	sua	0,0316
	signum	0,0489	christi	0,0301	eius	0,0181
	sepulcrum	0,0338	hoc	0,0234	sancti	0,0163
ECCLESIA	medio	0,2428	altari	0,1588	ea	0,1299
	ecclesia	0,0811	eo	0,0919	eo	0,1192
	summo	0,0551	ea	0,0515	ecclesia	0,1138
	altari	0,0325	ecclesia	0,0322	medio	0,0449
	area	0,0287	loco	0,0254	loco	0,0320

Tabell 12. Förslag på ord i mening C, givna av alla tre modeller, med probabiliteter.

Här gissas *altare* och *ecclesia* något bättre av XPL-modellerna än av LB, även om LB:s förslag överlag är goda och skillnaderna är små. Återigen överträffar XPL2 sin föregångare XPL1. Detta skulle kunna vara ytterligare ett tecken på att träningen har fått avsedd effekt: XPL är känsligare för ord i kristna sammanhang. Sist testar vi modellen på en påhittad mening som kanske hade kunnat stå i en senrepublikansk text.

Mening D: *Catilina cum coniurationem fecisset, rem publicam magnum in periculum obiecit.*<sup>38</sup>

maskerat ord	LB	prob.	XPL1	prob.	XPL_2	prob.
CATILINA	qui	0,2726	qui	0,4363	qui	0,3126
	nam	0,1349	hic	0,1005	is	0,1256
	sed	0,1034	sed	0,0439	hic	0,1251
	itaque	0,0993	is	0,0307	nam	0,1233
	quorum	0,0370	et	0,0298	itaque	0,0300
CONIURATIONEM	multa	0,1864	haec	0,2545	multa	0,2130
	haec	0,1672	hoc	0,1380	magna	0,0875
	id	0,1018	id	0,0527	male	0,0713
	ludos	0,0656	autem	0,0440	pacem	0,0700
	hoc	0,0616	ita	0,0358	haec	0,0612
REM	rem	0,9678	rem	0,7618	rem	0,9628
	libertatem	0,0053	causam	0,1000	causam	0,0054
	uitam	0,0052	iniuriam	0,0156	se	0,0039
	causam	0,0050	uim	0,0074	pecuniam	0,0037
	pecuniam	0,0046	poenitentiam	0,0071	libertatem	0,0033
PUBLICAM	publicam	0,8939	sibi	0,3397	publicam	0,5945
	romanam	0,0184	ei	0,0607	suam	0,0880
	sibi	0,0182	quoque	0,0381	se	0,0348
	illi	0,0057	suam	0,0267	sibi	0,0299
	magnam	0,0047	tam	0,0240	magnam	0,0287

<sup>38</sup> "När Catilina gjorde sin sammansvärjning, försatte han staten i stor fara."

OBIECIT	adduxit	0,8392	intulit	0,1208	duxit	0,2460
	coniecit	0,0542	accepit	0,1112	misit	0,1704
	uocauit	0,0496	recepit	0,0776	uenit	0,1136
	accepit	0,0132	traxit	0,0641	dedit	0,0815
	duxit	0,0099	induxit	0,0381	traxit	0,0438

Tabell 13. Förslag på ord i mening D, givna av alla tre modeller, med probabiliteter.

Det intressanta här är dels att LB inte gissar *Catilina* och *coniurationem*, vilket skulle kunna tydas som att modellen inte är överanpassad till någon tidsperiod eller textkorpus, något resultaten i tabell 7 i övrigt antyder, dels hur de tre modellerna gissar frasen *rem publicam*. Om *rem* maskeras tycks alla tre modeller gissa rätt, dock med en högre probabilitet i LB och XPL2. Likväl gissar XPL1 inte *publicam* när det maskeras, medan LB gör det utan problem och XPL2 intar en ställning mitt emellan. Detta skulle kunna tolkas som att *rem publicam* är ett sällsynt uttryck i PL och att en modell som XPL1, som ju i princip bara tränats på PL, därmed inte kan skilja det från andra vanliga uttryck med *rem* som uttrycket *rem sibi facere*, medan XPL2 bevisligen bevarat ett visst ”minne” av LB, som den tränats ur.

Låt oss avslutningsvis testa modellerna mot en s.k. benchmark, i syfte att avgöra vilken som kan anses vara bäst på att mäta diakron semantisk variation och med vilken algoritm.

#### 4.1.4.3 Test 3 – kvantitativ benchmarkstudie

Avslutningsvis testar vi modellerna mot en goldstandard på en GCD (Graded Change Detection, jfr avsnitt 3.2.3), enligt de kriterier och formler som beskrivits i nämnda avsnitt.

En för vår uppgift användbar (och ganska unik) goldstandard är den som publicerades i kölvattnet av SemEval2020, som ju dedikerades åt unsupervised LSC (Lexical Semantic Change detection, se ovan).<sup>39</sup> De tävlande presenterades med två uppgifter som alltså skulle lösas med hjälp av valfri modell som tränats utan etiketter (*unsupervised training*, se avsnitt 3.2.2), där deluppgift 2 gick ut på att mäta graden av semantisk förändring mellan en förkristen korpus  $C_1$  och en senare korpus  $C_2$ , samt ordna lemmarna efter resultaten. För latinets vidkommande valdes 40 lemmarna ut. Notera att skiljelinjen mellan  $C_1$  och  $C_2$  är Jesu födelse och att materialet för  $C_2$  inte enbart består av texter ur *Patrologia Latina*, vilket innebär att vi i  $C_2$  kommer att hitta texter som i den här uppsatsen hamnar i den förkristna korpusen.

Deluppgift 2 görs med samma metodologi som den beskriven i avsnitt 3.2.3, nämligen APD och PRT. Först bearbetar jag de två korpora och extraherar endast de meningar där något av de 40 undersökta lemmarna (hädanefter kallade  $t$  för *targets*) förekommer, detta i syfte att reducera energiåtgången. Sedan tokeniserar jag datan och parar varje token till dess lemma. Notera att jag misslyckas med enstaka token, på grund av att jag redan i detta steg tillämpar samma wordtokenizer som XPL tränats på, som i vissa fall kan tokenisera enklitiska ord fel (särskilt *-ne* som ibland förväxlas med ablativändelsen i substantiv på *-tio*, se avsnitt 3.2.2). Meningar som inte *alinjerar*, dvs där en token inte kan kopplas till ett lemma, undantas från studien. Observera också att det i materialet på några få ställen finns hela paragrafer utan interpunktion. Inget försök görs för att rätta detta, och de resulterande mycket långa meningarna tas bort från studien eftersom de överstiger gränsen på 512 WordPiece-tokens, som gäller för BERT. Tabell 14 sammanfattar den extraherade datan.

<sup>39</sup> Korpora och metadata hittas på webbsidan McGillivray et al. (2020), och en rapport av arbetet med de under tävlingen framtagna modellerna finns hos Schlechtweg et al. (2020).

	C <sub>1</sub>			C <sub>2</sub>		
	n meningar	n tokens	$\bar{x}$ tok/mening	n meningar	n tokens	$\bar{x}$ tok/mening
utan <i>t</i>	74 321	1 346 490	18,12	359 155	7 182 623	20,00
korrekta	21 748	753 079	34,63	103 746	4 014 374	38,69
inkorrekta	5	174	34,80	47	2 523	53,68
total	96 074	2 099 822	21,86	462 948	11 199 761	24,19

Tabell 14. Antal meningar och tokens i C<sub>1</sub> respektive C<sub>2</sub> i SemEval2020, fördelade mellan meningar utan *t*, korrekt extraherade och inkorrekt extraherade meningar.

Som vi kan se gick det att para varje ord med dess lemma i de allra flesta meningarna. Att de meningar som saknar *t* i genomsnitt innehåller färre tokens beror förmodligen på att korta meningar helt enkelt har en lägre probabilitet att innehålla *t*. Vi kan också konstatera att de meningar som går vidare i studien ("korrekta") är fem gånger fler i C<sub>2</sub> än i C<sub>1</sub>.

Efter att APD och PRT beräknats för varje *t* i enlighet med instruktionerna i Schlechtweg et al. (2020) och med specialanpassningen beskriven i avsnitt 3.2.4, beräknas slutligen hur väl den erhållna rankingen stämmer överens med ground truth med hjälp av Spearmans rangkorrelation (2020:8). En ground truth är inom maskininlärning helt enkelt ett "facit". Resultaten varierar mellan -1 (fullständig överensstämmelse med motsatt ranking, dvs rankingen är fullkomligt fel) och 1 (fullständig överensstämmelse med föreliggande ranking, dvs rankingen stämmer helt överens med ground truth), och möjliggör en jämförelse av modellerna sinsemellan samt av modellerna med övriga medtävlandes i SemEval2020. Resultaten redovisas i tabell 15.

		LB	XPL1	XPL2
APD	rho	0,5403	0,4990	0,6465
	p	0,00032	0,00105	0,00001
PRT	rho	0,4641	0,3942	0,5644
	p	0,00256	0,01184	0,00015

Tabell 15. Spearmans rangkorrelation (rho) på resultaten för algoritmerna APD och PRT och med modellerna LB, XPL1 respektive XPL2, med p-värde.

Den första observationen är att APD verkar bättre fånga semantisk variation än PRT, detta med alla modeller. Vidare tycks APD tillämpad på data från alla tre modeller överträffa det bästa resultatet av alla 23 inlämnade i SemEval2020 för latinets vidkommande (0,412)<sup>40</sup> och att APD levererar de bästa resultaten när algoritmen tillämpas på data hämtad med XPL2, sämre med LB och sämst med XPL1. Resultaten för APD med LB eller XPL2 överträffar också genomsnittet för alla undersökta språk i SemEval2020 (0,527).<sup>41</sup> APD med XPL2 överträffar faktiskt det bästa resultat för alla språk i tävlingen förutom tyska (där hamnar vi i femte position). Notera också det för XPL2 mycket låga p-värdet ( $p < 0,0001$ ), vilket indikerar att probabiliteten att rankingen byggd med APD på data framtagen med XPL2 beror på slumpen är nästintill obefintlig. Även det lägsta p-värdet, nämligen för PRT tillämpad på XPL1 ( $p < 0,05$ ), ligger inom vetenskapligt vedertagna signifikansnivåer (Dror et al. 2020:7), vilket betyder att alla resultat i tabellen är signifikanta, även om det finns viss variation. Att signifikansen är lägre för XPL1 beror nog på det lägre rho-värdet: när skillnaderna mellan min ranking och ground truth är större blir det förmodligen svårare att utesluta effekten av slumpen då datan är mer oregelbunden.

<sup>40</sup> Alla resultat från SemEval2020 i den här paragrafen är hämtade från Schlechtweg (2020:9).

<sup>41</sup> Engelska, tyska, latin och svenska.

I sin rapport noterade Schlechtweg et al. att token-baserade modeller presterade sämre än type-baserade modeller i SemEval2020, och de spekulerar i att detta skulle kunna bero på att token-baserade modeller är förtränade och inte kan tränas exklusivt på de relevanta historiska resurserna (2020:10), men XPL2 är just tränad på för kristet latin någorlunda relevanta historiska resurser. Såsom Schlechtweg et al. beskriver uppgiften i sin rapport tolkar jag det som att deltagarna inte fick träna sin modell på något annat än tävlingsmaterialet om de använde en type-baserad modell (statiska embeddings), men att de fick använda en förtränad token-baserad modell (kontextuella embeddings):

Participants were asked to train their models only on the corpora described in Table 2 [samma korpora som de som utgör materialet för uppgiften], though the use of pre-trained embeddings was allowed as long as they were trained in a completely unsupervised way, i.e., not on manually annotated data. (2020:2)

*Stricto sensu* är XPL2 en omtränad modell som tränats på en unsupervised task, men inte sällan används termen "förtränad modell" för omtränade modeller (vissa personer gör alltså ingen skillnad mellan "förtränad" och "omtränad"), vilket är logiskt om man tänker att båda modeller tränas på ett annat set än det de tillämpas på, till skillnad från type-baserade modeller.<sup>42</sup> Det kan alltså vara så att XPL2 presterar bättre än andra modeller i SemEval2020 därför att den tränats på att just lösa den typen av uppgift, men notera att det ju var just det vi for efter med träningen. Notera också att även en generisk modell som LatinBERT tillsammans med min variant av APD överträffar det bästa resultatet för latinet i SemEval2020 (för både type- och tokenbaserade modeller), vilket alltså är ett kvitto på att algoritmen *per se* är bra i jämförelse med andra algoritmer i tävlingen, och alltså att de goda resultaten inte enbart beror på omträningen. Det skulle kunna vara så att dimensionsreduktionstekniker som PCA, en bra intuition som i princip borde resultera i att ordformerna tydligare bildar kluster baserade på betydelse (i fall av polysemi i  $C_1$ ), faktiskt jämnar ut skillnaderna mellan ordformerna i fall där polysemi saknas eller är otydlig och därmed försvårar APD-uträkningen. Det skulle också kunna vara så att det ligger något i min intuition om att olika ordformer av samma lemma tenderar att förekomma i olika syntaktiska och semantiska konstruktioner, och att det därmed snedvrider resultaten något att jämföra vektorer för två olika  $w$  av  $l$  mellan två korpora. En annan spontan förklaring är att polysemi till följd av semantisk variation kanske inte alltid sker lemmavis. Vissa ordformer av ett lemma kanske utvecklar en utvidgad betydelse före andra.

Sammanfattningsvis tolkar jag resultaten i test 3 som att vi har uppnått vårt syfte med omträningen av LB till XPL2 och att den anpassade APD-algoritmen lämpar sig relativt bra för att mäta semantisk variation under kristendomens inflytande.

#### **4.1.4.4 Sammanfattning av testresultaten**

Vi har alltså sett att XPL2 i de två kvantitativa testerna överträffar LB och XPL1, och i test 3 att modellen överträffar alla andra modeller för latinet i SemEval2020. Med det kvalitativa testet kunde vi visualisera goda resultat för LB och XPL2, och att XPL2 kan ha lättare att generalisera ord som hamnar i semantiska sammanhang som är typiska för det som ovan definierats som "kristet latin". Vi har visserligen noterat en potentiell tendens mot en liten överanpassning av XPL2 till PL, men test 1 och 3 visar tydligt att detta inte hindrar modellen från att prestera bra på okänd data (test 1) och när den skall jämföra semantiska konstruktioner med en förkristen korpus (test 3). Det skulle tvärtom kunna vara så att just modellens känslighet för kristet material underlättar jämförelser mellan kristet/icke-kristet, och därmed är syftet med träningen av XPL uppnått. Låt oss nu, stärka av dessa resultat, använda XPL2 i kombination med

---

<sup>42</sup> En sanning med modifikation för latinets del, eftersom vi arbetar med i jämförelse med andra språk exceptionellt små korpora, vilket innebär att test- och träningsset i praktiken lätt överlappar varandra.

APD, eftersom algoritmen i de ovan redovisade testerna visat sig vara träffsäkrare än PRT, för att besvara forskningsfrågorna i uppsatsen. Först behöver data från den förkristna korpusen extraheras.

## 4.2 Extraktion av data i LA

Texterna bearbetas med smärre anpassningar, liknande de som gjorts för *Patrologia Latina*. Nämnvärda detaljer är att vissa texter, till skillnad från PL, är sämre meningsparsade, vilket innebär att jag behövt lägga på en egen sentence tokenizer ovanpå den befintliga. Jag har tagit bort *fragmenta*-delen ur texten *Petronius\_Satyricon-Fragmenta-et-Poemata.POS.xml*, vilken var problematisk ur olika aspekter om än kort. Vidare har jag tagit bort alla inledningar på engelska som vidlådde de terentianska filerna.

Alla ord och lemman extraheras – så långt det går. Vissa texter uppvisar stora lemmatiseringsproblem, som exempelvis Catullus *Carmina*, där ord som *\*libellumarido*, *\*solebasmeas* eller *\*nugas,iam* tillhörande lemmat ??? – för att endast nämna några få ur samma måhända kända vers – visar att parsningen och därigenom också lemmatiseringen inte alltid gått rätt till. Andra lemmatiseringsproblem är t.ex. att Columellas grekiska ord translittererats till latinska bokstäver med resulterande nonsense-meningar till följd, som ”*ou)d' a)\n bou=s a)po/loit' ei) mh\ gei/twn kako\s ei)/h*”.<sup>43</sup> Vidare fungerar CLTK:s Wordtokenizer sämre med bl.a. transkriberade grekiska ord i ackusativ, vilka t.ex. Propertius poesi visar prov på: dessa ords ackusativvändelse på *-n* tolkas i regel felaktigt som ett enklitiskt *-ne*.<sup>44</sup> Ett liknande öde har drabbat ord som *comest*, vilket felaktigt analyseras som *\*come + est* när det i själva verket tillhör lemmat *comedo*.<sup>45</sup>

Dessa felaktigheter har, där de har förekommit, antingen orsakat att ordet i fråga eller i värsta fall hela meningen inte kunnat mappas upp, och att den därmed fallit bort. Tidsbrist har omöjliggjort en justering av tokeniseraren, men visst ligger förbättringspotential just där. Jag har dock i görligaste mån parerat parsningsfel genom att bygga ett dataprogram som identifierar och hanterar dem, och där detta inte räckte har jag efter bästa förmåga manuellt justerat XML-filerna. Dessa tidsödande ansträngningar till trots avgjorde till sist XML-filernas ursprungliga kvalité, inte litterär smak, att t.ex. stora texter såsom Plinius d.ä.:s *Naturalis Historia* levererade 23 123 korrekt parsade meningar och endast 3 inkorrekt parsade, medan förhållandet i Lucretius *De Rerum Natura* var 1 576 respektive 43. Därutöver saknade en icke oansenlig andel korrekt parsade ord av olika skäl ett lemma, främst där ordformen varit lågfrekvent, och vissa gånger visade det sig även att ord ibland knutits till fel homonym/polysem. En viktig iakttagelse är att de texter som utgör LA, eftersom de härrör från olika projekt, följer olika standarder avseende lemmatiseringen, särskilt PoS. Ett signum i Perseus-materialet är t.ex. att de allra flesta förekomster av *cum* taggas som preposition, även när det används som konjunktion.<sup>46</sup>

En viktig uppgift är att det förekommer en hel del lemman som annoterarna betecknat med två eller fler lemman, som exempelvis: *iouis|iuppiter*, *puer|puera*, *merces|merx*. I vissa fall ser beteckningen ut att härröra från en icke-eftersarbetad maskinannotering, där programmet inte kunnat avgöra vilket lemma som bäst passar, som i: *lucas|lucus|lux*, *mens|menta|mentum* eller *uerus|uer|ueru|uerum*. Jag har valt att betrakta dessa missbildningar som en naturlig avgränsning av ett annars mycket omfångsrikt och dataslukande material och har undantagit dem från studien. För att ge en fingervisning bestod ca 14,8% av alla extraherade substantiv av ett sådant i vårt sammanhang obrukbart lemma.

Notera också apropå lemmatiseringen några enstaka fall som strider mot en intuitiv föreställning, som exempelvis i fallet *unus*, som i LA vid sidan om ordformer som *una* och *uno* även verkar inbegripa *primus*, *prima* och *semel*. Allt detta hade varit ett stort problem om XPL hade tränats med hänsyn tagen till lemmatiseringen, men lemmatiseringsdatan används först *a posteriori* för att filtrera resultat.

<sup>43</sup> 12656:1.4;2,10.

<sup>44</sup> Således tokeniseras *Daphnin* i 12716:2.46;2,7 som *\*Daphni + -ne*.

<sup>45</sup> Detta ord är särskilt vanligt hos Plautus. Jfr. 12727:3.1;17,2.

<sup>46</sup> *Cum* som konjunktion kan även i enstaka fall felaktigt taggas som adverb.

Dessutom avser vi med studien att fokusera på ord som det finns gott om belägg för, vilket minimerar effekten av felaktigt lemmatiserade ord. I slutändan kan vi inte förvänta oss att all data och metadata är felfri. Någonstans måste man ändå visa tacksamhet för det oerhörda arbete som en rad människor lagt ner på att framställa och tillgängliggöra denna data.

Efter dessa inledande utmaningar kunde cirka 97,25 % av totalt 4,6 miljoner ord korrekt extraheras ur LA (155 dokument), med information om lemma och PoS där detta fanns tillgängligt. I dessa meningar hade 454 029 ord som en konsekvens av ovannämnda lemmatiseringsproblem inget lemma, vilket motsvarar 10,2% av alla valida meningar. PoS-taggningsen har förenklats så att lemmarna samlas under huvudkategorier, istället för många olika underkategorier<sup>47</sup>, vilket minskar uppenbara fel och dubletter som ett mer komplext system kan ge upphov till. Den totala extraherade textmängden motsvarar i huvudsak vad man kan hitta i korpusen LatinISE för motsvarande tidsperiod (McGillivray & Kilgariff, 2013). En sammanställning visas i tabell 16.

meningar	n texter	n meningar	n ord	n latin	n grek.	$\bar{x}$ ord/mening	% grek.	% extrah.	% texter
valida	155	226 549	4 444 535	4 439 096	4 991	19,62	0,11	97,23	100,00
fragm.	35	1 071	14 708	14 684	21	13,73	0,14	0,32	22,44
icke-men.	104	2 336	32 360	32 303	52	13,85	0,16	0,71	66,67
korta men.	152	25 899	53 095	53 026	59	2,05	0,11	1,16	97,44
grek. men.	22	352	6 432	1 014	5 397	18,27	84,18	0,14	14,10
lem. prob.	56	309	19 840	19 719	87	64,21	0,44	0,43	35,90
total	155	256 516	4 570 970	4 559 842	10 607	17,82	0,23	100,00	100,00

Tabell 16. Fördelning över extraherade lemmatiserade ord och meningar ur LA.

Notera att enstaka ord (449 bland valida meningar) inte kunnat parsas som latinska eller grekiska, exempelvis där de innehåller oväntade tecken, eller då de felaktigt transkriberats med både latinska och grekiska bokstäver. Detta förklarar varför summan latinska och grekiska ord inte exakt överensstämmer med det totala ordantalet.<sup>48</sup> Observera också det genomsnittliga antalet ord per mening i de meningar som hamnat i kategorin ”lem. prob.” (lemmatiseringsproblem), som är tre gånger så hög som i valida meningar, vilka här som i PL bekräftar att det latinska språket i genomsnitt har ca 20 ord per mening. Detta visar att där det skett lemmatiseringsproblem, så har det varit främst i meningar som parsats fel (och därmed buntats ihop).

Sammanfattningsvis ges en visualisering av LA i diagram 4.

<sup>47</sup> Kategorin *v:sup:abl* (för ”verb, supinum, ablativ”) förenklas således till *v*, för att nämna ett exempel.

<sup>48</sup> Några exempel är tal i ovanligt format såsom *d\_c\_c\_c\_* (12675:5.5;3,8), förkortningar som *deosum*<sup>o</sup> (12168:163;2,5) eller latinskgrekiska hybrider som *ἀκροατικῆ* (12728:22.5;3,3). Till detta bestiariums präktigaste exemplar hör kanske ordet *non63met415mita* (12616:2.1,16).

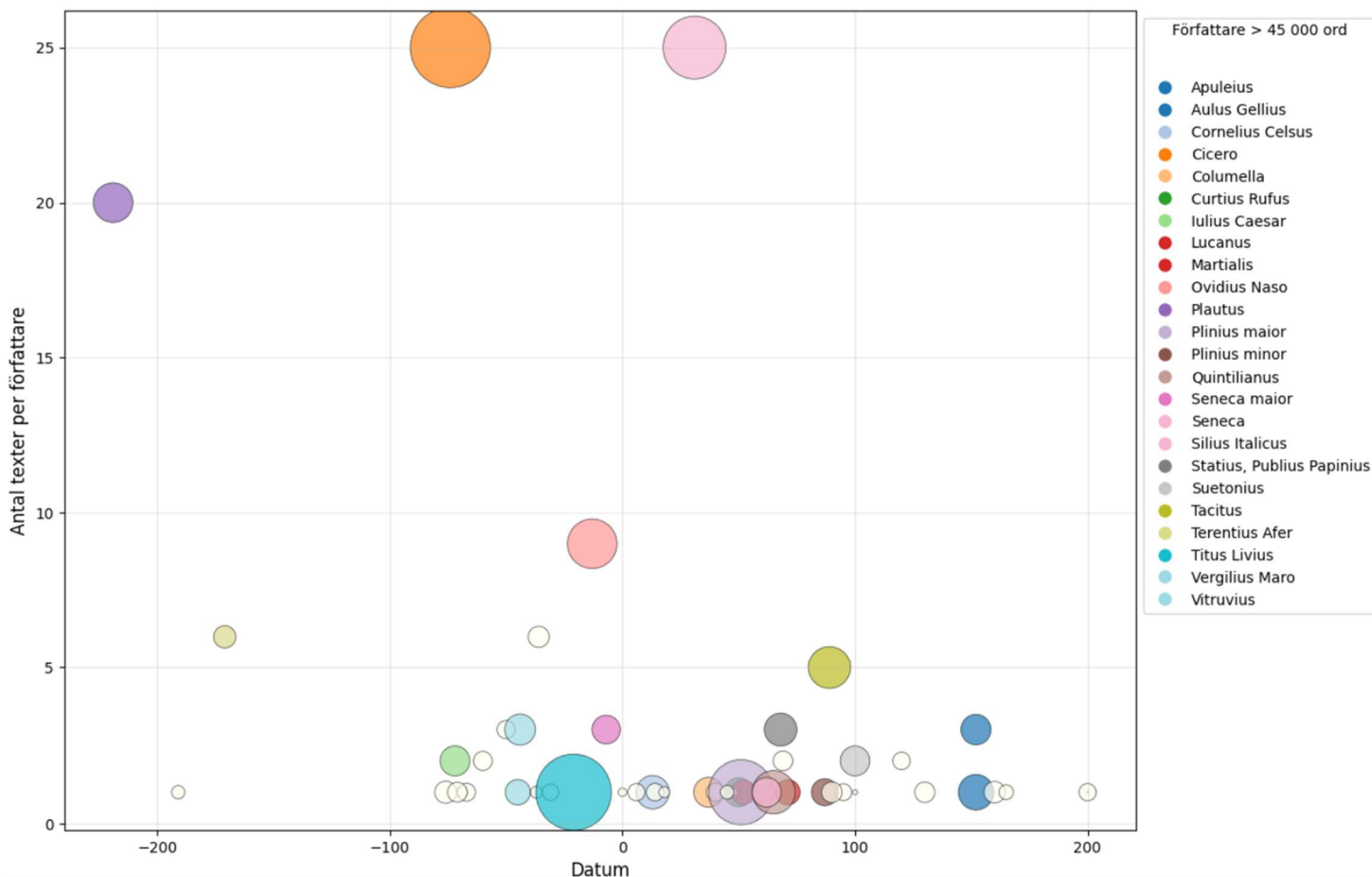


Diagram 4. Antal texter per författare över tid i LA\_val. Bubblornas storlek står för antal ord per författare. Författare vars material består av fler än 45 000 ord markeras även ut i legenden.

Diagrammet visar att LA består av det klassiska och förklassiska latinets gamla garde: om vi fokuserar på antal ord är de största bidragarna Cicero, Livius, Plinius d.ä., Seneca d.y., Tacitus och Plautus. Märk dock skillnaderna i antal texter, där Livius endast med sitt *Ab urbe condita* leverar nästan lika många ord som Ciceros ca 25 texter, likaså Plinius d.ä:s *Naturalis Historia* som Senecas samtida pjäser, brev och traktat. Här kan konstateras att vi i brist på material återigen tvingas bygga en korpus med få texter av mycket olika storlekar, vilket gör att vissa författare eller genrer kan vara överrepresenterade och därutöver saknas andra mer eller mindre centrala författare: den uppstådda skenbara variationen är i kraft av att vara ”det vi har” endast representativ för den *traderade* klassiska och förklassiska litteraturen. Men vi kan också konstatera att all denna ändå icke oansenliga textmängd utgör en bra referenskorpus för det som kristet latin, såsom definierat ovan, onekligen inte är.

### 4.3 Extraktion av embeddings

Samma information om lemmen och PoS som i LA extraheras också ur PL, och embeddings genereras för varje token med modellen XPL2. I samma kod har jag lagt till en sentence-tokenizer för att säkerställa att alla meningar omvandlades till samma format som de meningar som XPL2 förväntade sig, och en extra kontroll för att varje token korrekt attribuerats rätt PoS och lemma efter den data som fanns i XML-filerna. För de tokens där PoS- och lemmaattribueringen av olika skäl inte kunde göras har lemmat markerats som okänt, och där programmet inte kunde ta ut hela meningen på grund av att för

mycket metadata inte alinjerade med PoS och lemman, har meningen exkluderats från studien. Dessa aborterade meningar är dock relativt få till antalet, vilket framgår vid en jämförelse av tabellerna 3, 16 och 17 nedan. Eftersom PL är mycket större än LA har ingen anpassning av koden gjorts för att parera eventuella korpus-specifika fel i PL, vilket resulterat i att fler meningar i just den korpusen inte tagit sig igenom hela pipeline, och ett lägre antal meningar korrekt kunnat extraheras med avseende på lemman och PoS: sammanlagt kunde 99,97% av orden i valida meningar i LA generera embeddings, och endast 97% för PL.

För att ge en överblick över alla tokens vars embeddings extraherats presenteras tabell 17. Notera att tokeniseraren lägger till två tokens i varje mening: [CLS] i början av en mening och [SEP] i slutet av densamma.<sup>49</sup>

	LA	PL	PL1	PL2
n meningar	226 521	3 633 148	823 131	2 810 017
n tokens	5 689 209	97 978 689	20 875 675	77 103 014
n icke-ord	1 245 893	22 318 745	4 938 692	17 380 056
n ord	4 443 316	75 659 944	15 936 983	59 722 958
n lemman	23 168	26 759	21 433	26 092
n okänt lem	453 946	3 214 073	531 550	2 682 523
% okänt lem	10,22	4,25	3,34	4,49
% icke-ord	22	23	24	23
tokens / lemna	245,56	3 661,52	974,00	2 955,04

Tabell 17. Statistik över de tokens vars embeddings genererats med XPL2 i LA och i PL (även uppdelat mellan PL1 och PL2).

Vi kan bl.a. notera att en större andel ord saknar lemma i LA (10,22%) än i PL (4,25%). Inom PL är detta något lägre i PL1 än i PL2. En intressant iakttagelse är antalet olika lemman. Trots det mycket större antalet tokens i PL än i LA finns det endast 3 591 fler lemman i PL, och i PL1 finns det till och med färre olika lemman än i LA, trots att PL1 innehåller 3,5 gånger fler ord än LA. Detta skulle kunna tolkas som att LA och PL är någorlunda balanserade och saturerade, men inte PL1. Som vi kan se i diagram 2 dominerar PL1, som ju består av texter författade mellan 250 och 600 e.Kr., av jättarna Augustinus, Hieronymus och Ambrosius, vilket förmodligen bidrar till denna brist på variation.

## 4.4 Diakron semantisk variation

Alla extraherade embeddings motsvarar trots alla komprimeringar sammanlagt ca 800 GB av data. Låt oss nu använda denna data och beräkna APD för alla *targets* ( $t$ ) enligt avgränsningen beskriven i avsnitt 3.2.5 och med samma anpassade algoritmer som den jag använde i SemEval2020-studien. Alla  $t$  klarar uppgiften förutom enstaka  $w$  av *spiritus* (substantiv) och *sanctus* (adjektiv), vilka var så frekventa i PL att den resulterande matrisen fick min session att krascha när APD skulle räknas ut.

Notera att vissa  $w$  av  $t$  då och då inte hittas i LA eller PL, vilket innebär att den sammanlagda frekvensen för  $t$  ibland inte lever upp till tröskelnivån på 50 token. Detta gäller även för  $t$  i PL, som

<sup>49</sup> Notera att jag inte tränat XPL2 på [CLS] och [SEP], vilka är vektorer som skall sammanfatta hela meningen och som kan användas i andra typer av uppgifter än den föreliggande. Skulle man behöva använda XPL2 i en sådan uppgift får man i sådant fall själv räkna ut ett medelvärde (*average sentence embedding*).

alltså i vissa fall är mer frekventa i LA. Därför filtrerar jag resultaten så att endast *t* med en absolut frekvens över 50 i LA och i PL redovisas.

Först presenteras resultaten där APD räknats ut genom att jämföra LA med PL (alltså hela *Patrologia Latina*) för substantiv och därefter adjektiv, och sedan kommer en tröskelnivå att uträknas för potentiell semantisk utvidgning. Sist skall jag använda inferentiell statistik för att avgöra om det finns en statistiskt säkerställd skillnad mellan att använda sig av hela PL och att dela upp det i PL1 och PL2.

Resultaten av mätningarna gjorda med PL, PL1 och PL2 hittas i tabellform på Github (Lafage, 2025a).

#### 4.4.1 APD-resultat för PL

Resultaten i det här avsnittet utgår från konceptet om kristet latin som ett icke-tidsbegränsat fenomen. Låt oss först analysera resultaten för substantiv. Observera att APD-resultaten härstammar från en GCD vars syfte är att ungefär avgöra hur mycket ett lemma har förändrat sig semantiskt mellan två korpora. APD är alltså en trubbig proxy för ”grad av förändring”, som inte säger mycket om det observeras i isolering. Således måste resultaten betraktas som relativa och inte absoluta: de lemmen med det högsta APD-värdet har sannolikt genomgått en semantisk förändring mellan LA och PL, medan de lemmen längst ner i tabellerna förväntas ha varit mest stabila, men att lemma *x* har ett APD-värde på 0,25169 är i sig intetsägande.

Som referens används fyra ordböcker, nämligen Blaise & Chirat (1967, förkortad B&C), Blaise (1975, B), Souter (1949, S) och Favre (1883-1887, F). Jag refererar till dessa ordböcker som *ref\_litt\_xp*. Med *ref\_litt* menar jag ordböckerna Lewis&Short (1933, L&S), Gaffiot (1934, G) och Ahlberg (1966, A). Alla ordböcker förutom A hittas på webbplatsen Database of Latin Dictionaries, som Göteborgs Universitet för närvarande har licens till. För att undvika talrika fotnoter i löpande text med referenser till artiklar i *ref\_litt\_xp* och *ref\_litt*, refererar jag till ordböckerna med förkortningarna ovan i huvudsak utan fotnot. Således bör läsaren tolka frasen ”lemmat *cuspis* hittas endast i B&C i *ref\_litt\_xp*” som ”lemmat *cuspis* {substantiv} hittas i Blaise & Chirat (1967) under uppslagsordet *cuspis*, *-idis*, *f* {substantiv}, men inte i Blaise (1975), Souter (1949) eller Favre (1949), när en sökning görs i Database of Latin Dictionaries”, detta om frasen dyker upp i avsnittet om substantiv (ersätt annars {substantiv} med {adjektiv}). Där fler uppslagsord förekommer använder jag sedvanliga referenser. Lemman är nämligen uppslagsord i ordböcker, varav referenser lätt blir tautologiska och riskerar att uppta onödigt mycket plats i uppsatsen.

##### 4.4.1.1 Substantiv

I kvantitativa studier som denna är det viktigt att inte bortse från kvalitativa inslag. Låt oss därför först göra en kvalitativ helhetsbedömning av resultaten, för att avgöra om de överhuvudtaget följer intuitionen om diakron semantisk variation under kristendomens inflytande.

En översiktlig analys av de tjugo lemmen med högst respektive lägst APD-värde visar att de senare omedelbart för tankarna till klassiskt latin, medan lemmen med högt APD ofta kan sättas in i ett kristet sammanhang. En översikt ges i tabell 18. Låt oss gå igenom dessa lemmen och tolka de extraherade APD-värdena på ett kvalitativt sätt.

Högst APD							Lägst APD								
lemma	w	a	fr LA	r_fr LA	a fr PL	r fr PL	APD	lemma	w	a	fr LA	r_fr LA	a fr PL	r fr PL	APD
proximus	5	219	4.93	9325	12.32	0.40616		commilito	7	88	1.98	450	0.59	0.14895	
missa	6	116	2.61	9604	12.69	0.3926		zephyrus	6	100	2.25	106	0.14	0.14857	
par	5	69	1.55	639	0.84	0.37868		euphrates	5	86	1.94	573	0.76	0.14855	
solum	1	200	4.5	2216	2.93	0.37667		biennium	3	124	2.79	723	0.96	0.14854	
nihilum	3	453	10.2	5668	7.49	0.36806		tarquinius	6	186	4.19	190	0.25	0.14827	

censura	4	99	2.23	2912	3.85	0.35539	amphora	6	74	1.67	339	0.45	0.14736
testamentum	6	436	9.81	18167	24.01	0.34998	agamemnon	5	93	2.09	59	0.08	0.14709
natalis	7	132	2.97	4426	5.85	0.34877	tibicen	5	61	1.37	69	0.09	0.1469
cinis	1	290	6.53	853	1.13	0.34805	necessitudo	8	117	2.63	745	0.98	0.14672
manis	4	241	5.42	531	0.7	0.34261	poples	7	63	1.42	246	0.33	0.14583
passus	4	559	12.58	2751	3.64	0.33983	formica	6	84	1.89	283	0.37	0.14535
sedes	7	864	19.44	32876	43.45	0.33704	centaurus	7	64	1.44	54	0.07	0.14402
sus	5	126	2.84	913	1.21	0.33448	triennium	3	77	1.73	725	0.96	0.14355
appellatio	5	128	2.88	6264	8.28	0.33016	proauus	7	79	1.78	311	0.41	0.1415
comes	8	763	17.17	21514	28.44	0.32766	cuspis	7	195	4.39	200	0.26	0.14142
dilectus	6	139	3.13	1551	2.05	0.32627	capricornus	4	60	1.35	126	0.17	0.14072
leo	7	359	8.08	14555	19.24	0.3255	triclinium	7	106	2.39	236	0.31	0.13976
remus	7	266	5.99	970	1.28	0.32508	gracchus	3	207	4.66	82	0.11	0.13798
sacramentum	6	128	2.88	28806	38.07	0.32404	demosthenes	5	163	3.67	140	0.19	0.13367
humanitas	5	215	4.84	10344	13.67	0.32215	mithridates	5	103	2.32	115	0.15	0.12857

Tabell 18. 20 lemmen med högst respektive lägst APD-värde. *w* står för antal olika ordformer, *a\_fr* för absolut frekvens och *r\_fr* relativ frekvens (per 100 000 ord).

De mest stabila substantiviska lemmena ser alltså ut att vara egennamn, antingen historiska personer som *Mithridates*, *Demosthenes* eller *Gracchus*, eller platser som *Euphrates*. Vi hittar också en del ord som kan kopplas till den hedniska mytologin, såsom *capricornius*, *centaurus* och *zephyrus*. Att dessa ord är semantiskt stabila är väntat: egennamn och platsnamn har alltid en unik referent, vilket är ett hinder för att polysemi och därmed semantisk variation skall uppstå. Andra substantiv är ord som okontroversiellt kopplas till den förkristna kulturen, som *triclinium* eller *amphora*. Övriga lemmen antingen upptas sparsamt i *ref\_litt\_xp* och i så fall med samma betydelse som i *ref\_litt*, som *cuspis*, som endast förekommer med en referens i B&C, *proauus* (endast i F), *poples* (endast i B&C), *tibicen* (endast i B&C) eller inte alls, som *triennium*. *Commilito* hittas i B&C och i B, med den något fördjupade betydelsen ”Kristi vapenbroder”, vilket dock inte skiljer lemmat nämnvärt från *commilito* i *ref\_litt*. *Formica* finns i S, B och F, men i samma betydelse som i *ref\_litt*.

Det enda undantaget jag hittar är *biennium*. Lemmat upptas i F antingen som variant av *bidaem* (bl.a. i betydelsen ”kar med vatten som driver en mjölkvarn”), eller som variant av *biennum*, synonym till *corvatae* (modern franska *corvée*), dvs ”dagsverke”. Jag är faktiskt förvånad över att denna sista betydelse inte verkar ha varit övervägande i de 723 förekomsterna av lemmat i PL, där en del av materialet kan förväntas spegla den feodala kulturen. Enligt resultaten finns det alltså ingen nämnvärd semantisk skillnad mellan dessa och de 124 förekomsterna i LA, vilka enligt *ref\_litt* förväntas betyda ”tvåårsperiod”. Vi hämtar därför dessa 723 förekomster och tar stickprov.<sup>50</sup> Efter ett par timmar av närläsning av en större andel av dessa meningar, hittar jag inte en enda kandidat för denna specialbetydelse, utan alla de förekomster jag möter har den klassiska betydelsen och dyker upp tillsammans med temporala uttryck som *decurso*, *delapso*, *transacto*, *euoluto*, *expleto* (ablativi absoluti), *post*, *ante*, *prior*, *ultra*, *triennum*, *biduo*, *tempore*, *iam*, *usque ad*, etc.

<sup>50</sup> Märk att *sent\_id:n* kan skilja sig något med dem som dokumenteras i databasfilerna, vilket kommer sig därav att meningarna hittas i detta format strax innan de ytterligare sentence-parsas (i de få fall detta behövs) för att motsvara det förväntade formatet inför extraktionen av embeddings. Där ytterligare parsning sker läggs ett nummer till de berörda meningars ”undermeningar” för att alla *sent\_id:n* skall vara unika.

De lägsta APD-värdena, som troligen kommer från ord som inte har genomgått semantisk förändring, ligger ändå inte på 0. Detta tyder på att dessa ord i PL förekommer i sammanhang som ändå skiljer sig något, alltså tillsammans med ord och konstruktioner som är ovanliga i LA.

Om vi nu riktar uppmärksamheten till de 20 lemmorna med högst APD hittar vi ord som ofta skulle kunna placeras i ett kristet sammanhang. Enligt L&S betyder substantivet *proximus* ”granne” eller ”närmast släkting”<sup>51</sup>. I ref\_litt\_xp hittas den kristna betydelsen ”din näste” endast i B&C<sup>52</sup> (jfr spanskans *prójimo*). Nedslag i de 9 325 förekomster av lemmat i PL visar att det i mycket få fall handlar om ett lemmatiseringsfel, utan det handlar om den förväntade betydelsen ”din näste”, som i:

10085:2.2;2,4: noli ergo despiciere proximum, si non uis deo despiciabilis apparere.<sup>53</sup>

Algoritmen har alltså korrekt identifierat en förändrad betydelse. Notera att det relativt höga APD-värdet kanske påverkas av några förväxlingar mellan adjektiv och substantiv i LA.<sup>54</sup>

För nästa lemma, *missa* (116 förekomster) handlar det olyckligtvis om ett annoteringsfel, där *missa* i de flesta fall i LA är perfekt particip av *mitto*. Sålunda noterar algoritmen en skillnad med den i PL vanligare betydelsen ”gudstjänst” (B&C, 6). Enligt S kommer substantivet från *missa est* + feminint substantiv för ”församlingen”, och betecknar något slags avslut i gudstjänsten och i förlängningen själva gudstjänsten. Man kan förmoda att *missa* användes som perfekt particip i LA och först i PL gick över till att användas som substantiv. Annoteringsfelet till trots har *missa* alltså mycket riktigt fått en ny, kristen betydelse.

Gällande *par* (69 förekomster i LA) verkar det ofta röra sig om egennamnet Paris (i t.ex. 12594:10.27,1,8 och 12713:9.5;6,13) och några gånger om adjektivet *par*, och den observerade semantiska variationen beror här alltså på ett annoteringsfel. I fallet *solum*, som endast består av en form, *solo*, handlar det också i PL om både substantivet och adjektivet om vartannat (annoteringsfel). I fallet *Nihilum* ser annoteringen ut att stämma. I tabell 18 visas frekvenserna och APD-värdena för de olika *w* av lemmat *nihilum*.

lemma	token	abs frek LA	rel frek LA	abs frek PL	rel frek PL	apd
nihilum	nihili	49	1,10	285	0,38	0,18587
nihilum	nihilo	373	8,39	3578	4,73	0,43193
nihilum	nihilum	31	0,70	1805	2,39	0,26377

Tabell 19. Absolut frekvens för de olika formerna av lemmat *nihilum* i LA och i PL, samt APD-värde.

*Nihilo* är både den mest frekventa formen och har det högsta APD-värdet. *Nihilum* har nämligen annoterats som substantiv i uttryck där ablativ är vanligt förekommande, som *nihilo minus*. I PL noterar vi ett påfallande högt antal uttryck som *ex nihilo*, *de nihilo*, *pro nihilo*, etc, särskilt i teologiska texter, och dessa uttryck är ovanliga i LA.

*Censura* är ett intressant fall. I ref\_litt noterar vi betydelsena ”censorsämbetet” och något liknande ”bedömning, avgörande, dom” (jfr L&S, G). Det är dessa betydelser vi huvudsakligen möter i LA. I PL förekommer den förra betydelsen knappast. Samtidigt verkar *censura* inte sällan kopplas till det gudomliga (jfr det frekventa uttrycket *diuina censura*), och utan att ha närläst alla tusentals passager noterar jag en tydlig tendens mot betydelsen ”straff/dom”:

<sup>51</sup> L&S s.v. *propior* II.A.1 och II.B.3.

<sup>52</sup> s.v. *proximus*, 4.

<sup>53</sup> ”Förakta alltså inte din näste, om du inte vill framstå som föraktlig inför Gud.”

<sup>54</sup> Ordet är faktiskt ett substantiverat adjektiv.

8224:36.3;2,2: quisquis uero conatus fuerit tentare prohibita, sentiet censuram sedis apostolicae minime defuturam.<sup>55</sup>

Skillnaden med LA skulle alltså kunna vara att det inte längre handlar om människors fria vilja utan Guds dom, men man skulle behöva närläsa alla dessa passager och relevant vetenskaplig litteratur. Ändå tyder resultaten klart på en betydande semantisk förändring.

En liknande semantisk glidning ser vi med *testamentum*. Lemmat har betydelsen ”testamente” i ref\_litt, men i ref\_litt\_xp ser vi att den utvidgats. Det handlar bl.a. om en pakt, särskilt mellan Gud och människor (*διαθήκη*, jfr B&C) eller löfte. Det återkommande uttrycket *testamentum nouum (et uetus)* bidrar sannolikt till denna förändrade betydelse.

*Natalis* påverkas förmodligen av att det i PL förekommer listor över vilka martyrs födelsedagar man firar, liksom i filen 030\_Auctor-incertus-(Hieronymus-Stridonensis)\_Martyrologium.POS. I sådana filer upprepas det substantiviska *natalis* med ord som *sancti*, *sanctorum*, etc. I LA förekommer även egennamnet *Natalis* (jfr L&S) men inga kristna referenser (*sanctus*, *martyres*, etc). Lemmat har alltså inte fått en ny betydelse, men används konsekvent i ett jämfört med LA annorlunda semantiskt sammanhang.

*Cinis* är ett annat annoteringsfel, där nominativformen av *cinis*, *cineris*, *m* konsekvent kopplats till lemmat *cinis*, medan övriga former kopplats till lemmat *ciner*. Eftersom detta även gjorts i PL är APD-värdet ändå relevant för formen *cinis*. I LA handlar det helt enkelt om aska, antingen i samband med offer eller i olika preparat (hos t.ex. Plinius d.ä och Columella). I PL handlar det däremot ofta om uttryck som *cinis et puluis es* och varianter därav (*uermis*, *putredo*). Jag observerar också att citatet *quid superbis terra et cinis?* förekommer många gånger och i olika texter i PL. Kopplingen *terra – cinis* stöter man också ofta på i andra meningar. Här handlar det alltså inte om en betydelseförändring utan om något som liknar det Burton kallar kanoniserad metafor.

Vidare har *manes* i PL den klassiska betydelsen ”de dödas själar”, men i flera fall handlar det i PL om Manes/Manis, alltså manikéismens grundare. Ordformerna är identiska, men man kan tycka att annoteringen borde ha skilt på egennamn (s.k. NER eller *Named Entity Recognition*, jfr Eisenstein, 2019:175). Jfr *Natalis* ovan.

*Passus* är ett annoteringsfel: i LA är lemmat mestadels korrekt annoterat och används i måttangivelser, särskilt i texter över militära kampanjer. I PL avser lemmat ofta perfektparticipformen *passus* av *patior* och syftar nästan alltid på Kristi lidelse.

*Sedes* är däremot ett exempel på korrekt semantisk förändring, vilket vi såg i exemplet ovan (*censuram sedis apostolicae*). De klassiska betydelseerna ”säte” och ”plats” har tydligen fördjupats till något som vi idag kallar ”(biskops)säte” och liknande (ofta i uttrycket *prima sedes*), något vi hittar i ref\_litt\_xp.

*Sus* är också ett annoteringsfel i PL (huvudsakligen korrekt annoterat i LA). Av alla fem former verkar nämligen endast *sue* registrera ett högt APD, men ordformen är egentligen en medeltida variant av adjektivet *suae*:

10790:1.4;2,4: ob quam causam ipsa andouera perpetuo a rege est separata, et in monasterio cum ipsa sua filia uelamine uelata usque ad finem uite sue permansit.<sup>56</sup>

Vidare *appellatio*: detta lemma har i ref\_litt betydelseerna ”tilltal” och ”benämning”, men i B&C citeras den specifikt juridiska betydelsen ”överklagan (inför kyrkodomstol, kejsaren, etc)”, vilken faktiskt påträffas ganska ofta i PL, men uppenbarligen specifikt i några enstaka texter, som 055\_Leo-I\_De-haeresi-et-historia-Eutylichiana.POS.xml, vilken uppskattningsvis innehåller över hundra sådana

<sup>55</sup> ”Men den som försöker sig på det förbjudna, kommer att utstå den apostoliska stolens dom, som visst inte kommer att utebli.”

<sup>56</sup> ”Därför skildes Andovera för alltid från kungen och spenderade resten av sitt liv i ett kloster, beslöjad, tillsammans med sin dotter.”

förekomster. Detta illustrerar effekten av genre och ämnesområde: den semantiska utvidgningen förekommer med högre sannolikhet i juridiska texter eller i texter som behandlar juridiska tvister.

Från ”följeslagare”, ”kamrat” verkar *comes* ha använts senare i latiniteten för att beteckna olika typer av ämbetsmän (jfr B&C), en betydelse vi kan se i många passager i PL. Både *appellatio* och *comes* är exempel på ord som fått nya betydelser utan direkt koppling till kristendomen, och kan snarare ses som uttryck för ”medeltida latin”.

*Dilectus*, som med alternativstavningen *delectus* i ref\_litt betyder ”urval” eller ”trupprekrytering”, har i ref\_litt\_xp en betydelse som liknar ”kärlek”, ”vänskap”. I PL förekommer uttrycket *dilectus eius* frekvent, vilket kan ligga bakom det högre APD-värdet.

7794:55;2,19: si non praecessisset dilectus tuus, non sequeretur amor meus.<sup>57</sup>

Notera att det i många fall är svårt att avgöra om vi har med substantivet eller med ett particip att göra (alltså ”din vänskap” eller ”din älskade”).<sup>58</sup>

*Leo* och *remus* beror förmodligen på de högre frekvenser av Romulus bror i LA i förhållande till PL, där lemmat huvudsakligen har betydelsen ”åra” (notera att vi återigen tampas med NER-relaterade fel), respektive det påvliga namnet *Leo*, som dyker upp i PL. Notera dock att det sistnämnda ofta används bildligt i PL. Bilden av lejonet associeras både med positiva (mod, Kristus) och negativa företeelser (djävulen, fara).

Sist återstår två i litteraturen över kristet latin ofta citerade lemman, nämligen *sacramentum* och *humanitas*. Den förra går från ”(fan-)ed” till olika kristna betydelser som kretsar kring det vi kallar ”sakrament” och den senare från ”mänsklighet” till ”filantropi” (*φιλανθρωπία*, jfr B&C). Dessa är alltså också fullträffar.

#### 4.4.1.2 Adjektiv

I tabell 20 visas de 20 lemman med högst respektive lägst APD.

Högst APD							Lägst APD								
lemma	w	a	fr LA	r_fr LA	a fr PL	r fr PL	APD	lemma	w	a	fr LA	r_fr LA	a fr PL	r fr PL	APD
imus	11	686	15.44	18374	24.28	0.41236	transmarinus	10	58	1.31	643	0.85	0.15002		
beatus	17	547	12.31	61598	81.41	0.37504	centenus	5	54	1.22	192	0.25	0.14906		
c	2	89	2.0	3615	4.78	0.34454	inmensus	9	136	3.06	91	0.12	0.14825		
natalis	8	87	1.96	3084	4.08	0.34119	equinus	11	59	1.33	125	0.17	0.14737		
sanctus	10	250	5.63	119730	158.25	0.34088	stoicus	11	329	7.4	488	0.64	0.14736		
notus	11	651	14.65	7024	9.28	0.3356	octoginta	11	111	2.5	1769	2.34	0.14695		
mundus	8	132	2.97	9135	12.07	0.33346	tempestius	11	121	2.72	188	0.25	0.14631		
dilectus	8	94	2.12	3950	5.22	0.32869	aestius	11	193	4.34	580	0.77	0.14549		
praesens	9	870	19.58	44406	58.69	0.32458	torquatus	7	109	2.45	110	0.15	0.14457		
futurus	12	471	10.6	27897	36.87	0.32296	septingenti	10	82	1.85	513	0.68	0.14433		
salutaris	8	193	4.34	6739	8.91	0.31973	missilis	5	69	1.55	122	0.16	0.14279		
putus	5	129	2.9	936	1.24	0.3193	aequoreus	10	89	2.0	128	0.17	0.14181		
m	2	173	3.89	113	0.15	0.31472	sediciosus	11	75	1.69	322	0.43	0.14136		
paulus	4	462	10.4	1634	2.16	0.31207	intempestius	11	71	1.6	165	0.22	0.14109		
i	2	312	7.02	3808	5.03	0.30844	uectigalis	7	94	2.12	242	0.32	0.14011		

<sup>57</sup> ”Om din vänskap inte hade gått före, hade min kärlek inte blivit till.” Termer som ”kärlek” och ”vänskap” har specifika konnotationer och det är inte möjligt att översätta dem korrekt utan att närläsa texterna i fråga, vilket tyvärr inte låter sig göras i denna uppsats.

<sup>58</sup> Obs! Inte att förväxlas med *electus* i frasen ”Guds utvalde”.

oratorius	10	91	2.05	2315	3.06	0.30449	plerusque	1	264	5.94	466	0.62	0.13969
ii	1	451	10.15	6319	8.35	0.30348	octingenti	11	105	2.36	403	0.53	0.13443
iustus	10	297	6.68	44755	59.15	0.302	improvisus	9	115	2.59	1010	1.33	0.13419
silus	10	77	1.73	400	0.53	0.30139	diutinus	11	91	2.05	799	1.06	0.12937
fidelis	10	144	3.24	43631	57.67	0.30085	pyrenaicus	7	65	1.46	146	0.19	0.12294

Tabell 20. 20 lemmen med högst respektive lägst APD-värde. Notera att *sanctus* inte kunde leverera alla ordformer då lemmat är extremt högfrekvent i PL.

Bland de lemmen med högst APD känner vi igen *beatus*, *sanctus* och *fidelis* från tidigare listor över semantisk utvidgning under kristendomens inflytande. Okontroversiellt hittar vi bland de lemmen med lägst APD monosema ord med otvetydiga referenter, såsom tal (*octingenti*, *septingenti*, *octoginta*, *centenus*). Notera också att vi hittar tal i romerska bokstäver bland de högsta APD-värdena (*C*, *I*, *II*, *M*). Eftersom dessa står i olika kombinationer uppstår det en naturlig variation och därför är de irrelevanta för studien. Bland de mest stabila lemmena hittar vi geografiska beteckningar som *pyrenaicus*, *transmarinus*, och till viss del *aequoreus*. Alla andra ord har antingen samma betydelse i *ref\_litt* som i *ref\_litt\_xp* (exempelvis *torquatus*, *intempestivus*, *tempestivus*, *uectigalis*) eller saknas helt i *ref\_litt\_xp* (*sediciosus*, *maestus*, *stoicus*).

Om vi fokuserar på ord med högst APD igen noterar vi *imus*. De vanligaste formerna i LA är *imo* och *ima*, vilket vi kan se i tabell 21.

lemma	token	abs frek LA	rel frek LA	abs frek PL	rel frek PL	apd
imus	imo	193	4,34	15 788	20,87	0,44554
imus	ima	171	3,85	1 507	1,99	0,24488
imus	imus	52	1,17	249	0,33	0,26274
imus	imi	15	0,34	20	0,03	0,20906
imus	imis	88	1,98	560	0,74	0,23974
imus	imum	80	1,80	163	0,22	0,21212
imus	imos	25	0,56	23	0,03	0,20173
imus	imas	27	0,61	16	0,02	0,19834
imus	imam	21	0,47	41	0,05	0,19646
imus	imae	13	0,29	4	0,01	0,24003
imus	imorum	1	0,02	3	0,00	0,15943

Tabell 21. Olika *w* av *imus* med respektive APD-värden.

I PL ser vi däremot att *imo* har en ovanligt hög relativ frekvens (fem gånger så hög som i LA). Nedslag i materialet ger vid handen att adjektivet är korrekt annoterat i LA men att det i PL råkat förväxlas med alternativstavningen av adverbet *immo*. Återigen en s.k. falsk positiv. För *natalis* se 4.4.1.1.

Man skulle kunna tro att *notus* lidit samma lemmatiseringsöde, men faktum är att de flesta orden ser ut att vara korrekt annoterade, även om jag observerar någon förekomst av sunnanvinden *Notus*. Fokuserar man på olika *w* är det *notum* som sticker ut för PL, både i frekvenser och i APD-värde. Här handlar det förmodligen om den i PL mycket frekventa konstruktionen (*alicui*) *notum* + *sit/factum sit*, som inte sällan inleder en mening:

8629:4.15;2,20: notum tibi sit, rex, quod deos tuos non colimus, et statuam auream quam erexisti, non adoramus.<sup>59</sup>

Denna konstruktion hittar jag inte någonstans i LA.<sup>60</sup> Här rör vi oss i ett annat område: algoritmen har visserligen upptäckt en konstruktion som skiljer LA och PL åt, men det är inte frågan om en semantisk utvidgning, utan en syntaktisk företeelse.

*Mundus* torde vara tydligare. Redan i L&S markeras den kristna betydelsen ”moraliskt ren” från den klassiska ”ren”, ”proper”.<sup>61</sup> Här ett exempel från LA:

12587:12.4;5,18: natura enim homo mundum et elegans animal est.<sup>62</sup>

Tyvär hittar vi också en del förekomster av substantivet *mundus* i LA, som alltså felaktigt annoterats som adjektiv. Däremot verkar betydelsen ”moraliskt ren” vara frekvent i PL, även om vi också där observerar fall av *mundus* som substantiv och av verbet *mando* (i återkommande uttryck som *ab occultis meis munda me*).

Notera i fallet *dilectus* visserligen samma förväxling med substantivet som i LA, men i PL är det mycket tydligt att den vanliga betydelsen är ”kär” i uttryck som ”min käre”. Observera att oerhört många av dessa konstruktioner (ja tusentals) består av superlativformen, vars APD inte kunde beräknas p.g.a. att motsvarande former helt saknades i LA. Således ser vi t.ex. 2 942 förekomster av *dilectissimi* och 737 av *dilectissimo*. Trots detta, kunde algoritmen mycket riktigt finna en semantisk förändring lik den vi noterade i 4.4.1.1. Det är oklart om användningen av *dilectus* i tilltal, särskilt superlativer, är en kristen eller en senantik/medeltida företeelse. Vi kan bara konstatera att uttrycket är oerhört mycket vanligare i PL.

*Praesens* verkar ha två huvudsakliga betydelser i ref\_litt, nämligen ”fysiskt närvarande” och ”omedelbar”. Dessa två betydelser är ungefär det vi möter i LA. I PL hittar vi visserligen samma betydelser, men vi noterar även en tämligen frekvent specialbetydelse, enligt vilken *praesens* kopplas till *caro* och *futurus* till *spiritus*:

6871:5.27;2,2: quod praesens, ad israel carnalem; quod futurum, ad spiritalem attinet.<sup>63</sup>

Således kopplas *praesens* till den temporala, sekulära, dimensionen och *futurus* till ett spirituellt liv efter detta. Denna topos ser av allt att döma ut att vara tämligen frekvent i PL. Den temporala betydelsen är också vanlig, men skillnaden med LA är att *praesens* och *futurus* mycket ofta hamnar i en antites i samma mening i PL:

21450:45.1.14;2,1: ob prophetiae certitudinem profertur in praesenti quod spondetur in futurum.<sup>64</sup>

Detta är förmodligen ytterligare ett tecken på samma etiska narrativ (notera förresten den ca tre gånger högre relativa frekvensen av båda lemmarna i PL jämfört med LA). Denna antites noteras i B&C s.v. *praesens*. I samma ordbok noteras att *futurus* refererar till livet efter detta, vilket är den betydelse vi oftast möter i PL (ofta i kombination med *saeculum* eller *uita*).

<sup>59</sup> ”Du skall veta, konung, att vi inte dyrkar dina gudar, och att vi inte vördar den gyllene staty som du reste.”

<sup>60</sup> Det verkar också som att filen 099\_Auctores-varii\_Chartularium-Werthinense.POS.xml innehåller frasen *notum fieri cupio omnibus, tam praesentibus quam futuris* många gånger, vilket också kan ha bidragit till resultatet.

<sup>61</sup> L&S s.v. *mundus*, *a*, *um*, 1.II.C. respektive 1.I.

<sup>62</sup> ”Till sin natur är människan nämligen ett välvårdat och förfinat djur.”

<sup>63</sup> ”Det närvarande har med den köttslige Israel att göra, det kommande med den andlige Israel.”

<sup>64</sup> ”Genom övertygelsen om profetian framskjuts till nuet det som utlovas i framtiden.”

*Salutaris* är också ett tydligt exempel på semantisk utvidgning. I LA har det huvudbetydelsen i ref\_litt, nämligen ”lyckosam”, ”nyttig”. I PL kopplas det däremot nästan uteslutande till tanken om frälsning. Observera att ordet kan användas som substantiv också (och därmed annoterats felaktigt som adjektiv):

7093:1.6.1;2,3: salutare tuum exspectabo, domine.<sup>65</sup>

*Putus* är tyvärr inkorrekt annoterat i båda korpora: allra oftast handlar det om imperativformen *puta* av *puto*, som enligt t.ex. G används nästan adverbialt och i uttrycket *ut puta* i betydelsen ”antag att”.<sup>66</sup>

Jag ser ingen betydelseförändring mellan LA och PL i fallet *paulus*. Det som kan ha hänt är att egennamnen Paula och Paulus i vissa (om än inte särskilt många) fall hamnat i denna kategori och förmodligen påverkat resultatet för *paulum*. I tabell 22 visas uppdelningen för varje *w* av *paulus*.

lemma	token	abs frek LA	rel frek LA	abs frek PL	rel frek PL	apd
paulus	paulo	204	4,59	1415	1,87	0,31524
paulus	paulum	250	5,63	102	0,13	0,33683
paulus	paula	7	0,16	71	0,09	0,20373
paulus	paulam	1	0,02	46	0,06	0,19747

Tabell 22. Olika *w* av *paulus* med respektive APD-värden.

Det är alltså *paulum* och *paulo* som förklarar det mesta av variationen i LA. Båda former används huvudsakligen adverbialt (*paulo ante*). I PL möter vi båda former relativt sett mindre ofta (om detta inte beror på ett lemmatiseringsproblem), särskilt *paulum*. Jag kan inte se att lemmat skulle ha fått en ny betydelse. Skillnaden är förmodligen mer syntaktisk, i det att *paulo* oftast dyker upp i temporala uttryck med *ante*, *post* och *superius* i PL, medan det bjuder på mer variation i LA (t.ex. i konstruktionen *haud paulo* + komparativ, som jag inte hittar i PL). Dessa låga frekvenser i PL och ordens alltmer begränsade användningsområden tyder på att ordet förmodligen hamnat ur bruk, vilket en passage ur Augustinus *Sermo 279*, vilken hittas i PL, faktiskt belyser:

7333:7.1.5;2,13–14 : Usum latinae locutionis advertite: quia paulum, modicum dicitur. Paulo post videbo te, paulum hic exspecta; id est, post modicum videbo te, modicum hic exspecta.<sup>67</sup>

I citatet kan vi alltså se att *paulum* ersatts med andra ord som *modicum*. Detta har förmodligen inte skett under kristendomens inflytande och det är oklart om det ligger diastratiska eller diakrona faktorer bakom variationen.

*Oratorius* betyder i ref\_litt endast ”oratorisk” men i ref\_litt\_xp finner vi substantivet *oratorium* i betydelsen ”bönelokal” samt adjektivet *oratorius* i betydelsen ”som har med böner att göra”. I PL är det främst substantivet vi hittar (alltså fel ordklass). Oavsett felet överensstämmer resultatet med förväntningen att hitta både *oratorium* och *oratorius* i PL i en annan semantisk kontext.

*Iustus* betyder i ref\_litt ”rättvis”, ”laglydig”, ”tillbörlig” medan ref\_litt\_xp anger betydelsen ”dygdig”, ”rättfärdig”. Notera att lemmat förekommer nästan tio gånger oftare i PL än i LA, vilket tyder

<sup>65</sup> ”Jag skall vänta ut din frälsning, Herre.”

<sup>66</sup> G s.v. *puto*, *are* 5. Jfr den spanska imperativformen *pon* (av latinets *pono*) i samma betydelse.

<sup>67</sup> ”Observera bruket av det latinska uttrycket. Ty *paulum* betyder *modicum*. ’Jag kommer snart (*paulo post*) att se dig’, ’Vänta här ett ögonblick (*paulum*)’ betyder ’Jag kommer snart (*post modicum*) att se dig’, ’Vänta här ett ögonblick’ (*modicum*).”

på dess centrala roll i det kristna narrativet. Mycket riktigt används den nya betydelsen främst i PL, där adjektivet ofta används substantiviskt:

6956:17.18;2,27: inops peccator, diues est iustus; quia nullo bono deficit, qui dominum semper inquirat.<sup>68</sup>

Skillnaden i betydelsen är alltså en glidning från "laglydig" till "som följer den religiösa lagen".

Egennamnet *Silus* och *Sila* har annoterats som tillhörande adjektivet *silus* i LA, medan ordet *Silo* i PL oftast handlar om orten på Västbanken, som ofta omnämns i Gamla testamentet:

11455:84.1;133,5: silo est ciuitas, in qua erat arca.<sup>69</sup>

*Fidelis* går från "trogen", "pålitlig", "trofast" i LA till "kristen". Lemmat är alltså ca tjugo gånger vanligare i PL än i LA och används oftast substantiverat, vilket vittnar om dess centrala betydelse i den kristna vokabulären.

#### 4.4.2 Sammanfattning av resultaten

Vi kan konstatera att APD beräknat på embeddings hämtade med den omtränade modellen XPL2 verkar leverera resultat som stämmer överens med data i ordböckerna, förutsatt att metadataannoteringen (lemmatisering och PoS-tagging) är korrekt, även om det ibland går att se meningsfulla resultat även där PoS-tagging varit undermålig, särskilt i fall där skillnaden mellan substantiv/adjektiv huvudsakligen är syntaktisk (men inte semantisk).

De framvaskade orden kan placeras i Burtons kategorier för semantisk utvidgning: semantiskt fokus (t.ex. *proximus*, *salutaris*, *iustus*, *fidelis*, *dilectus*, *mundus*), semantisk innovation (t.ex. *testamentum*, *sacramentum*, till viss del *missa*) och kanoniserade metaforer (möjligtvis *cinis*). Förmodligen har åtminstone några av dessa förändringar, som vi har sett i 2.1, sitt ursprung i grekiska eller hebreiska (som *testamentum*).

Där resultaten är s.k. falska positiva handlar det ofta om annoteringsproblem, inte sällan i samband med egennamn, som på latin kan sammanfalla med substantiv, som *Remus* eller *Leo*. Vi har även märkt att APD-värdena ibland kan fånga syntaktiska snarare än semantiska företeelser, förmodligen där lemmat används i högfrekventa konstruktioner, som i fallet *natalis*. Ordet har inte förändrat sin betydelse nämnvärt; det är förmodligen de långa listor av helgons födelsedagar som ger ett högt APD. Likaså konstruktionen (*alicui*) *notum* + *sit/factum sit* och det begränsade bruket av *paulus* i PL.

Vi har också flera gånger konstaterat att den observerade semantiska variationen inte kan skiljas från det latinska språkets diakrona utveckling, som i *appellatio*,  *censura* eller *comes*. Det är alltså uppenbart att det finns fler samverkande faktorer bakom resultaten än just kristendomen, och att vi i PL hittar särdrag som kan tillskrivas sen- och medeltida latin. Vi kan alltså inte garantera att den observerade variationen uppstått med kristendomen, utan endast att vissa betydelse, metaforer eller syntaktiska uttryck är mer typiska, ibland unika, för PL. Metoden tar inte heller hänsyn till närvaro eller frånvaro av polysemi i någondera korpusen, vilket gör att vi i fall av polysemi i LA inte kan skilja mellan olika betydelse.

Lemmatiseringsproblemen skulle kunna bero på att PL åtminstone delvis annoterats automatiskt och att maskinen som gjort detta förmodligen tränats på klassiskt latin. Detta ser vi t.ex. i fall som *oratorius*, som konsekvent annoteras som adjektiv i PL, trots att ordet oftast är ett substantiv.

APD-värdena i sig räcker alltså inte för att avgöra om ett lemma har fått en ny betydelse, men det ser i alla fall ut att kunna finnas en korrelation mellan ett högre APD-värde och en högre probabilitet att en

<sup>68</sup> "Medellös är syndaren, rik är den rättfärdige. Ty inget gott fattas den som alltid söker Gud."

<sup>69</sup> "Shilo är staden, där arken fanns."

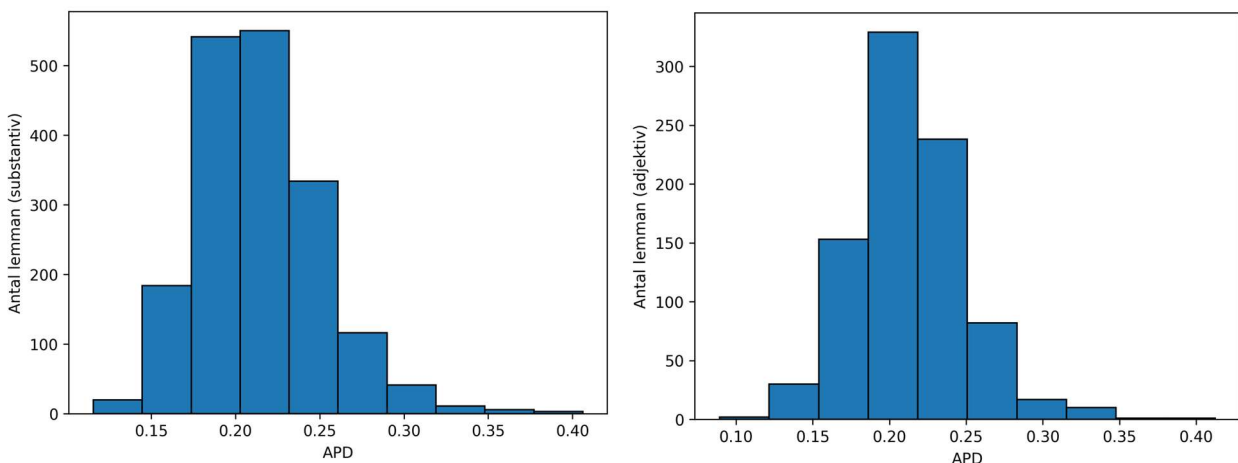
semantisk förändring har skett. Om detta är sant återstår att fastställa var gränsen går. Vi behöver alltså definiera en tröskelnivå för APD-värdena.

#### 4.4.3 Tröskelnivå för semantisk variation

Det finns två sätt att sätta en tröskelnivå för APD-värdena, över vilken vi kan anse att semantisk variation är sannolik. Man kan antingen träna en regressionsmodell på en benchmark lik den i SemEval2020, som består av ord som människor manuellt annoterat som ”förändrad” eller ”oförändrad” samt APD-värdena. Det finns dock flera problem med detta förfarande. För det första består benchmarken av endast 40 lemman (26 annoterade som förändrade och 14 som oförändrade), vilket är för skalt för att träna en modell. Dessutom finns diskrepanser mellan två olika redovisningar av samma data på webbsidan McGillivray et al. (2020) och McGillivray et al. (2022a). Listan med binär data (förändrad / oförändrad) alijnerar nämligen inte med listan över den (manuellt annoterade) graderade semantiska variationen (GSV), på så vis att ett lemma som *consul* med ett GSV på 0,1298862845 anses ha fått en ny betydelse medan ett lemma som *salus* med ett GSV på 0,4695031831 anses inte ha gjort det (notera att det lägsta GSV-värdet i setet är 0 och det högsta 0,9056003337). GSV-värdet har också räknats ut genom att endast 30 slumpmässiga förekomster av varje  $w$  har hämtats från  $C_1$  och  $C_2$  och att man manuellt satt ett betyg på hur mycket man anser att ordet i fråga har en betydelse som liknar ett set av  $x$ -antal olika betydelser hämtade från en ordbok, vilket i sig är problematiskt då ordboksdefinitioner ibland överlappar varandra. Dessutom fick annoterarna utgå från översättningar då man insåg att de inte förstod en del latinska meningar korrekt (McGillivray et al., 2022a:62–63). Det finns alltså en risk att dessa GSV-värden är ungefärliga och därmed att denna ground truth är av sämre kvalitet.

Det andra sättet är att istället utgå från resultaten själva och göra en datadriven beräkning. Detta gjordes bl.a. för latinets del i SemEval2020 av Zhou & Li, som först räknade ut kosinusavståndet mellan samma  $w$  i den förkristna respektive i den ”efterkristna” korpusen, sedan omvandlade fördelningen av kosinusvärdena till en gammafördelning, för att till slut välja ut de ord vars kosinusvärde hamnade över värdet för 75:e percentilen i denna gammafördelning (2020:224). En gammafördelning är helt enkelt en funktion som bäst förklarar spridningen av datapunkterna. Samma metod används av Perrone et al. för att fastställa en tröskelnivå i syfte att bygga en ground truth (2021).

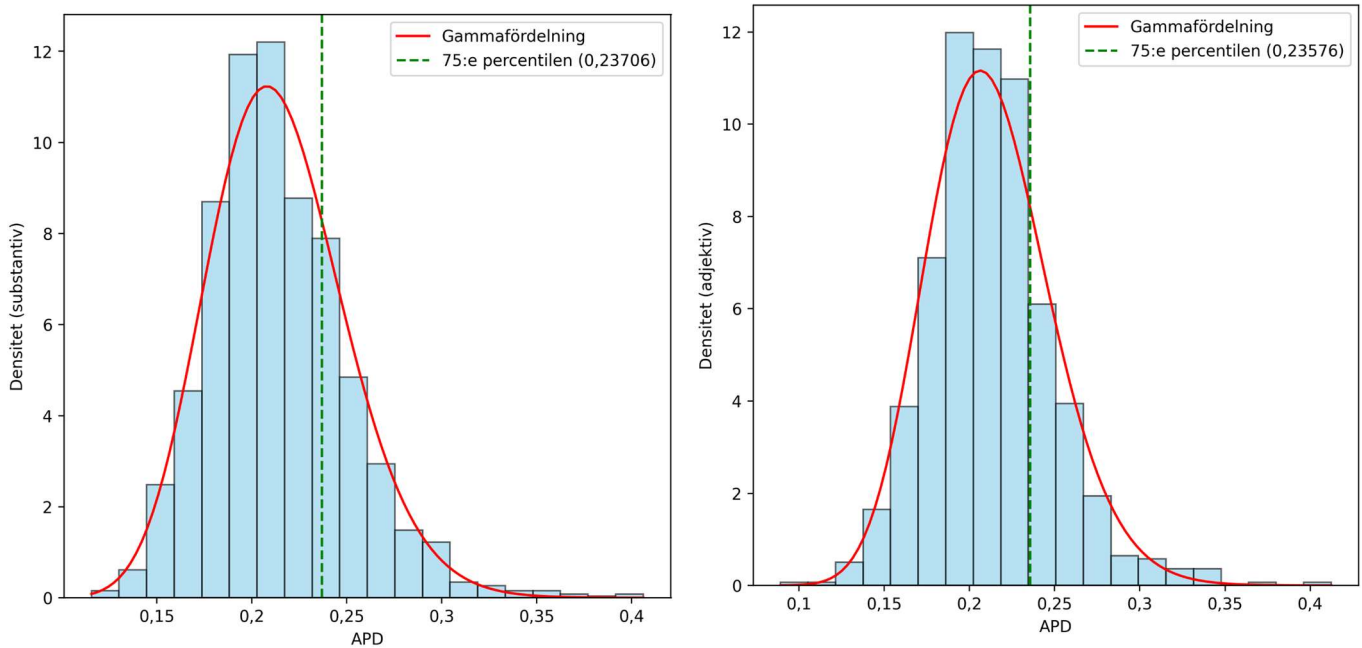
Innan tröskelnivån beräknas enligt Zhou & Li visualiseras spridningen av APD-värdena i ett histogram.



Histogram 1. Spridning av APD-värdet för substantiv respektive adjektiv.

Som vi kan se, får många lemman ett APD-värde mellan 0,18 och 0,25 och några få får ett högre eller ett lägre APD-värde. Denna typ av distribution är typisk i jämförelser av vektoriella embeddings mellan två korpora (jfr Zhou & Li, 2020:224). Vi följer Zhou & Li samt Perrone & al (2021) och antar att de

flesta lemmor är stabila, och att de som har undergått semantisk variation således hamnar i den högra "svansen" i histogrammet. Histogram 2 visar samma data och gammafördelningen.



Histogram 2. Densitet av substantiviska och adjektiviska lemmor i olika APD-kategorier baserat på APD-värdet samt gammafördelning.

Märk att skalan nu är normaliserad, men tanken är densamma och spridningen följer samma mönster. Den således beräknade tröskelnivån för APD efter 75:e percentilen blir 0,23706 för substantiv och 0,23576 för adjektiv. Vi anser alltså att APD-värden över dessa tröskelnivåer tillhör lemmor som har större probabilitet att ha genomgått semantisk variation, men givetvis är probabiliteten högre ju högre APD-värdet är. Sammanlagt hamnar 23,7% av alla substantiviska och 24,5% av alla adjektiviska lemmor över tröskeln, vilket ligger nära 25%. Att det förhåller sig så beror antagligen på att datan inte är så långt från normalt fördelad, vilket syns i histogrammen. Jag har också beräknat en probabilitet utifrån gammafördelningen (mellan 0 och 1) att ett ord har genomgått en semantisk förändring, och bifogar det i den slutgiltiga listan (Lafage, 2025a).

#### 4.4.4 APD-värde för lemmor i litteraturen om kristet latin

Återstår för att besvara  $F_1$  att jämföra resultaten med dem för de lemmor samlade i 2.1.4. Resultaten kompileras i tabell 23.

	$w$	abs frek LA	rel frek LA	abs frek PL	rel frek PL	APD	benchmark
missa	6	116	2,61	9604	12,69	0,3926	
beatus	17	547	12,31	61598	81,41	0,37504	0,816392
passio	2	3	0,07	11221	14,83	0,3642	
sanctus	10	250	5,63	119730	158,25	0,34088	0,425203
sacramentum	6	128	2,88	28806	38,07	0,32404	0,68804
humanitas	5	215	4,84	10344	13,67	0,32215	0,455671
fides	6	1885	42,42	101377	133,99	0,31059	
spiritus	4	405	9,11	49769	65,78	0,30462	

caro	8	463	10,42	89576	118,39	0,30258	
gloria	5	1311	29,5	52591	69,51	0,30172	0,170439
fidelis	10	144	3,24	43631	57,67	0,30085	
uerbum	6	4374	98,44	123493	163,22	0,29994	
uersus	6	832	18,72	7510	9,93	0,29894	
pax	7	1663	37,43	43037	56,88	0,28301	
salus	7	1026	23,09	47455	62,72	0,28094	0,469503
potestas	8	1352	30,43	40918	54,08	0,27991	0,548475
misericordia	4	268	6,03	36695	48,5	0,27785	
gens	8	2409	54,22	52713	69,67	0,2776	
dolus	6	362	8,15	5391	7,13	0,27739	0,176682
scriptura	5	45	1,01	34907	46,14	0,27615	0,516652
plaga	6	428	9,63	6956	9,19	0,26373	
pontifex	9	250	5,63	23730	31,36	0,26272	0,9056
confessio	6	112	2,52	16504	21,81	0,2603	
natio	7	281	6,32	5668	7,49	0,25895	
dux	8	2269	51,07	19286	25,49	0,2584	0,289054
peccatum	3	110	2,48	53791	71,1	0,25415	
lectio	7	100	2,25	12283	16,23	0,24709	
imperator	10	1154	25,97	26626	35,19	0,24496	0,846816
consul	8	4568	102,81	3893	5,15	0,24279	0,129886
uirtus	8	3234	72,78	89814	118,71	0,23797	0,39711
conuersio	5	42	0,95	5913	7,82	0,2379	
ciuitas	9	2394	53,88	48085	63,55	0,22657	0,322392
flagellum	6	87	1,96	4958	6,55	0,22319	
caritas	6	179	4,03	2543	3,36	0,20332	
lauacrum	4	23	0,52	2452	3,24	0,20306	
paganus	6	13	0,29	2593	3,43	0,19535	
cohors	8	853	19,2	1072	1,42	0,18698	0,28083
collyrium	5	68	1,53	174	0,23	0,14421	
					median	0,27677	
					medel	0,272175	

Tabell 23. APD-värden för lemmarna i 2.1.4 och tröskelnivå.

Notera att jag inte inkluderar adjektiven från sammanställningen i tabell 1: *beatus*, *fidelis* och *sanctus* eftersom de ju redan redovisats i tabell 20 med mycket höga APD-värden, *arrepticus* och *ceruicatus* då deras frekvenser är alldeles för låga för att leverera trovärdiga resultat. Vi kan se att de flesta lemmarna hamnar över tröskelnivån, vilket ytterligare tyder på att våra resultat överensstämmer med tidigare kvalitativ och kvantitativ forskning om kristet latin och därmed är valida. Här behöver vi dock försöka reda ut varför några lemmarna inte följer det förväntade mönstret. Jag ser flera förklaringar till detta.

För det första kan två ord påverkas av att för låga frekvenser i LA (under vår gräns på 50 förekomster) jämförs med relativt mer frekventa lemmarna i PL: att jämföra 13 embeddings med 2 593 för *paganus* och 23 med 2 452 för *lauacrum* är riskabelt, då endast ett par förekomster i den ena eller den andra riktningen kan förändra slutresultatet totalt.

Notera också att APD-värdet för *flagellum* och *ciuitas* ligger mycket nära tröskelnivån. Förmodligen ligger dessa "sämre" resultat delvis på att gränsdragningen för tröskelnivån är något godtycklig, men låt

oss inspektera dessa meningar i materialet. I fallet *ciuitas* skriver Burton att det egentligen snarare tillhör senlatin än kristet latin:

Many features of biblical Latin, then, are probably best identified as belonging to a sort of post-Classical koiné rather than to any definitely stigmatized register. For instance, the widespread use of *ciuitas* rather than *urbs* as the standard word for "city" may be seen with hindsight as early evidence for one of the distinctive changes between Latin and the Romance languages; as indeed it is (2011:487).

Just betydelsen "stad" markeras även i t.ex. B&C som "postklassisk", men utöver den understryks i samma ordbok betydelsen "himmelriket". Denna betydelse stöter vi dock sällan på i PL, där man nästan uteslutande hittar betydelsen "stad". Orsaken till det relativt låga APD-värdet är enligt min bedömning att det ofta är svårt att skilja mellan den klassiska betydelsen "samhälle" och den efterklassiska "stad", "borg"<sup>70</sup>. Många av meningarna ser t.ex. ut på följande vis i LA:

12583:90.23;2,16: nulla est ciuitas, quae non et improbos ciues aliquando et imperitam multitudinem semper habeat.<sup>71</sup>

Det är endast i kraft av att meningen är hämtad från Livius som vi kan anta att *ciuitas* här betyder "samhälle", men annars hade betydelsen "stad" fungerat. Jag misstänker alltså att dessa två betydelser ofta ligger för nära varandra för att algoritmen tydligt skall kunna skilja dem åt.

I andra fall rör det sig om metaforer och metonymier. Således används *flagellum*, som i ref\_litt har den konkreta betydelsen "piska", "gissel", metonymiskt i PL i betydelsen "piskslag"<sup>72</sup> och dessutom metaforiskt i betydelsen "plågoris", "straff"<sup>73</sup>, och dessa figurativa betydelser ligger inte långt från den konkreta: alla har med straff att göra, där någon utför straffet och någon mottar det, och alla orsakar smärta.

8319:1.49;31,1: quantum quisque aut in corpore aut in mente flagella sustinet, tantum se in finem remunerari speret.<sup>74</sup>

Ur algoritmens perspektiv är skillnaden mellan den konkreta och de överförda betydelserna att det saknas ett ord för "slag" (från "piska" till "piskslag"), men problemet är att dessa figurativa betydelser ofta förekommer tillsammans med ord som påminner om en konkret fysisk bestraffning (ofta *percussio* eller *percutior* men även *verberor*, osv):

8106:13.194;2,8: his flagella ab hac uita inchoant, et in aeterna percussione perdurant.<sup>75</sup>

Skillnaden mellan den metonymiska och den metaforiska betydelsen är i övrigt mycket svår att tolka från kontexten: således lär algoritmen ha svårt att skilja mellan kroppslig och andlig bestraffning. Notera att den konkreta betydelsen "piska" också förekommer i PL.

*Caritas* förekommer i Mohrmann, dock utan exempel och med betydelsen "amour chrétien" (1977:17), vilket även återfinns i B&C i betydelse 3 som översättning av termen *ἀγάπη* ("kärlek till sin näste"). I ref\_litt hittar vi betydelserna "kärlek" och "dyrt pris". Problemet med specialbetydelsen "kärlek till sin näste" är att den endast skiljer sig från "kärlek" i det att den inte är riktad till en specifik individ utan till alla individer i den kristna gemenskapen, och liksom med *ciuitas* är denna subtilitet svårare för algoritmen att fånga.

<sup>70</sup> Observera att "samhälle" på svenska har båda betydelser.

<sup>71</sup> "Det finns inget samhälle, som inte har både oredliga medborgare ibland och en enfaldig människomassa alltid."

<sup>72</sup> B&C. s.v. *flagellum* 1.

<sup>73</sup> B&C. s.v. *flagellum* 2.

<sup>74</sup> "Lika mycket skall man hoppas på att belönas med i evigheten som man utstår under piskan, antingen kroppsligt eller själsligt." Notera specialbetydelsen "i evigheten" i *in finem* (B&C, s.v. *finis*, 2).

<sup>75</sup> "Med det börjar plågor från det här livet, och håller ut i eviga slag."

*Cohors* tas visserligen upp i McGillivray et al. men har egentligen ingen koppling till kristet, utan rör snarare sen- och medeltida latin (2022a:56). Notera att ordet tillhör en av tre kategorier varur McGillivray et al. hämtar orden till undersökningen, nämligen sekulära maktstrukturer, liksom i fallen *ciuitas*, *consul*, *dux* och *imperator*, vilka i tabell 22 i samtliga fall fått ett lägre resultat. Vid sökning efter *cohors* i PL hittas oftare den klassiska militära betydelsen, vilket nog förklarar varför lemmat får ett lågt APD-värde.

7911:5.11;2,4: inter haec uenit magister militum, et circa ecclesiam per cohortes suos milites ordinauit.<sup>76</sup>

*Collyrium* är också ett fall av ett ord med metaforisk användning. I ref\_litt får lemmat betydelsen "salva" och i LA hittar vi ordet ofta i kopplingar till ögonsjukdomar, vilket även är fallet i PL. Nedan visas ett exempel på den överförda betydelsen i PL:

11552:3.12;4,18: inunge collyrio oculos tuos ut uideas, id est, scrutare scripturam ut ueritatem cognoscas.<sup>77</sup>

Problemet är att vi oftast saknar en förklaring (*id est*) eller andra signaler på att meningen skall tolkas metaforiskt, och i sådant fall kan vi praktiskt taget inte avgöra om vi har med den konkreta eller den överförda betydelsen att göra. Detta är nackdelen med word embeddings, vilka inte tar hänsyn till kontexten utan bara till den aktuella meningen.

Summa summarum kan vi konstatera att metodologin i denna uppsats fångar alla ord som framhävs i litteraturen om kristet latin, förutsatt att den semantiska variationen markeras i texten genom specifika ord och konstruktioner, och att ordfrekvenserna inte är för låga. Där så inte sker, t.ex. med det Burton kallar kanoniserade metaforer, som *flagellum* eller *collyrium*, kan upptäckten vara svårare.

#### 4.4.5 Inferentiell statistik

Vi har nu all data för att besvara den sista forskningsfrågan. Alla beräkningar görs med LA som referenskorpus. Först undersöker vi om det föreligger en skillnad i rankingen av lemmarna beroende på om LA jämförs med PL, PL1 eller PL2. Detta gör vi som ovan med en Spearman-korrelation och resultaten visas i tabell 24.

	PL1	PL2	PL	PL1	PL	PL2
rho	0,94837	0,96343	0,99668			
p	0	0	0			

Tabell 24. Spearmans rangkorrelation (rho) när resultaten mellan PL1 och PL2 jämförs, respektive PL och PL1 samt PL och PL2, med p-värde.

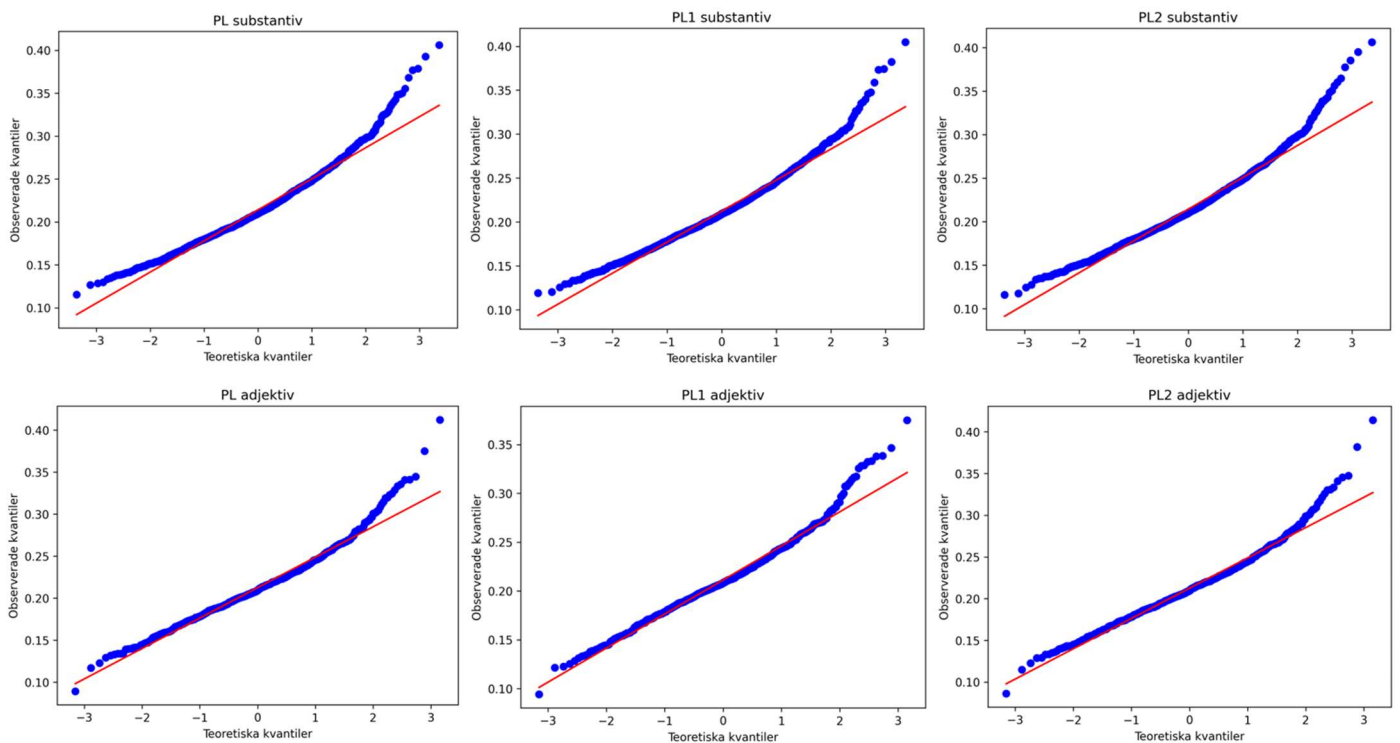
Observera för det första att ett p-värde på 0 är omöjligt: man kan närma sig 0 men aldrig uppnå det, eftersom vi har med probabiliteter att göra. Värdet är dock så nära 0 att python inte kan representera det på annat sätt. Det betyder alltså att rankingarna av lemmarna baserat på APD-värdena, oavsett om jämförelsen görs mellan LA med PL, PL1 eller PL2, är så lika att sannolikheten att likheterna skulle ha uppstått av en slump är nära noll. Rho-värdena, vilka alla ligger nära maxvärdet på 1, visar hur lika rankingarna faktiskt är. Notera att Spearman-korrelationskoefficienten brukar räknas mellan listor där värdena är oberoende av varandra, vilket inte helt gäller här. PL1 och PL2 är nämligen delar av PL, vilket gör att en del av de word embeddings som används i beräkningen av APD-värdet per lemma är desamma mellan PL och subkorporan PL1/PL2, även om de slutgiltiga APD-värdena, vilka här jämförs, alltid skiljer sig. Det är alltså inte förvånande att rho-värdet är extremt högt när PL jämförs med PL2, eftersom PL2 utgör ca 78,9% av PL och PL1 alltså 21,1% (jfr tabell 17). Ändå ser vi att rho-värdena är något lägre mellan PL\_PL1 och PL\_PL2, och att PL1\_PL2, två helt oberoende subkorpora, ger det lägsta

<sup>76</sup> "Under tiden kom infanteriets befälhavare och placerade sina trupper kohortvis runt kyrkan."

<sup>77</sup> "Smörj dina ögon med ögonsalva så att du ser, dvs utforska skriften så att du lär känna sanningen".

rho-värdet. Detta antyder att rankingen av APD-värdena skiljer sig något när man jämför embeddings mellan LA och PL1 jämfört med LA och PL eller PL2. Skillnaderna är dock i sådant fall mycket små, vilket betyder att där en skillnad uppstått i APD-värdet för ett specifikt lemma, så är denna skillnad inte så stor.

Spearman-korrelation är en s.k. icke-parametrisk metod, vilket innebär att datan inte behöver vara normalt fördelad, men för parametriska tester måste den följa något slags normalfördelning. Histogram 1 och 2 antyder att detta verkar vara fallet, men för att säkerställa det visualiserar vi datan med hjälp av det som på engelska kallas Q-Q plot. I figur 1 nedan ser man s.k. z-score för de observerade APD-värdena på y-axeln och z-score för de teoretiska APD-värdena om datan hade varit helt normalfördelad på x-axeln. En z-score representerar ”antal standardavvikelser från medelvärdet”. Om datan är normalfördelad skall t.ex. en viss procentsats av datapunkterna hamna mellan -3 och -2 standardavvikelser och därför talar man om percentiler. Notera att jag för läsbarhetens skull visar APD-skalan på y-axeln, men själva datapunkterna bakom bygger på z-score också (alltså normaliserad data). Den röda linjen visar hur många standardavvikelser från medelvärdet APD-värdena borde ha hamnat om datan hade varit helt normalfördelad. De blå punkterna visar de observerade z-scores för varje mätpunkt (alltså varje lemmas APD-värde).



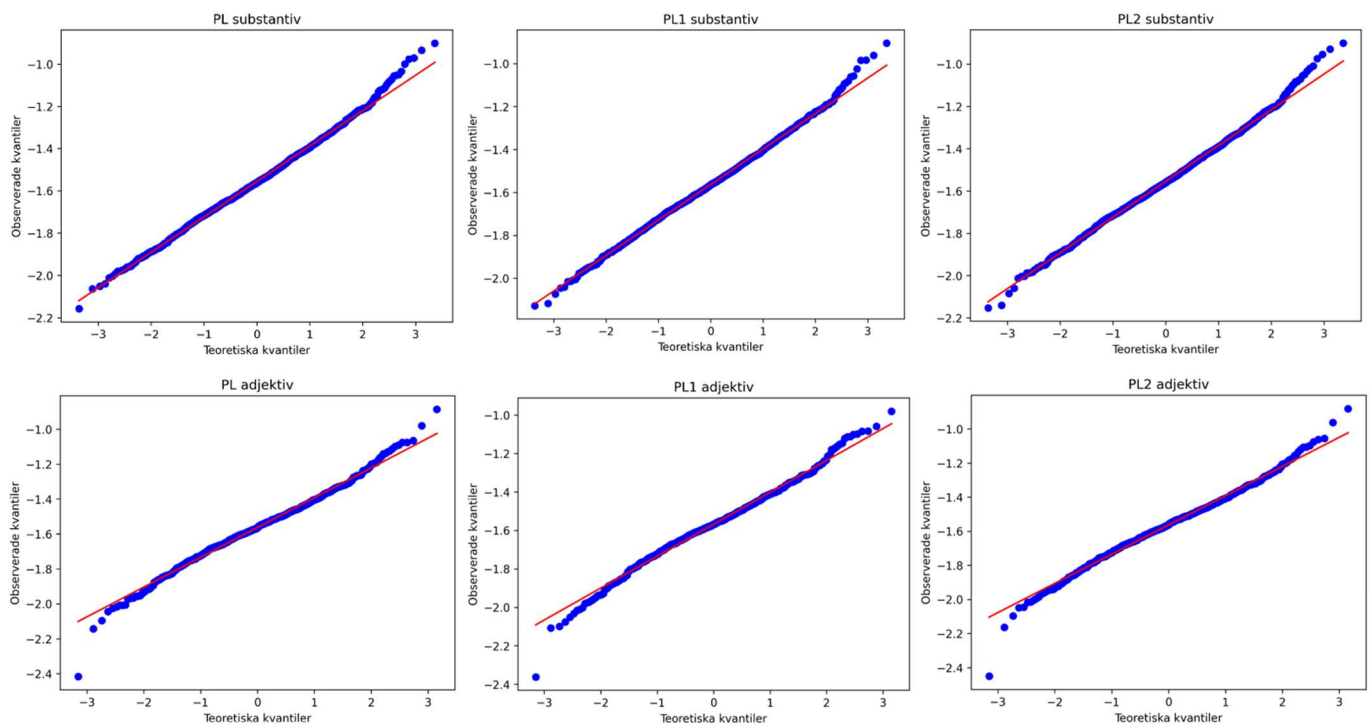
Figur 1. Q-Q plot för APD-värdena när embeddings i LA jämförs med embeddings i PL, PL1, respektive PL2, för substantiv och adjektiv.

Graferna läses på följande vis: i grafen PL substantiv ser vi en mätpunkt runt 0,11 i APD. Denna punkt ligger uppskattningsvis -3,5 standardavvikelser från medelvärdet i datasetet (vilket är ganska extremt). Som vi ser ligger punkten något ovanför den röda linjen. Detta betyder alltså att om de uppmätta APD-värdena vore normalfördelade, skulle den mätpunkten rent probabilistiskt hamnat runt 0,9 istället (alltså på den röda linjen).

Som framgår följer de observerade percentilerna de teoretiska i huvudsak, särskilt för adjektiven, men kurvan är positivt snedfördelad, vilket histogrammen också visade. Det innebär att några enstaka värden har ett högre APD än förväntat om datan vore normalfördelad ("APD-värdena går plötsligt snabbt uppåt mot slutet, när man sorterar dem från lägst till högst"). Att det förhåller sig på det viset i alla grafer tyder kanske på att probabiliteten att få ett högt APD-värde går ner ju högre APD-värdet är,

och att färre lemman når dessa nivåer (s.k. *outliers*), vilket i förbifarten skulle kunna utgöra ett bra datadrivet mått på särskilt intressanta lemman, som ett alternativ (eller en komplettering) till Zhou & Lis metod med 75:e percentilen (2020). Om vi skulle utgå från detta mått och sätta tröskelnivån efter 1,8 standardavvikelser, vilket ungefär motsvarar punkten då datan drar iväg i figur 1, kommer vi fram till en tröskelnivå på 0,28009 för PL (0,27695 för PL1 och 0,28179 för PL2) för substantiv och 0,27923 för PL (0,27843 för PL1 och 0,27911 för PL2) för adjektiv. Om vi tillämpade dessa tröskelnivåer skulle vi sälla fram 85 substantiv och 35 adjektiv med ”särskilt högt APD”, som har större probabilitet att ha något att säga om kristet latin.

Eftersom vi har tämligen många mätpunkter och den observerade datan har konstaterats skev dock ”någorlunda” normalfördelad, måste vi transformera datan så att effekten av extrema värden minimeras när parametriska tester utförs. Stora dataset riskerar annars att ge artificiellt låga p-värden (Dror et al. 2020:32–33). Dror et al. parerar detta genom att rapportera effektstorleken i sina resultat, vilket jag också inkluderar här, men jag använder också log-funktionen i pythonbiblioteket numpy för att minska skevheten i datan. Denna funktion räknar ut den naturliga logaritmen för varje värde i ett dataset. Efter transformationen ser q-q-graferna ut som i figur 2.



Figur 2. Q-Q plot för logaritmerade APD-värden när embeddings i LA jämförs med embeddings i PL, PL1, respektive PL2, för substantiv och adjektiv.

Som vi ser i figur 2 har effekten av outliers dämpats och datan liknar nu en klassisk normalfördelning. Detta gör det möjligt att använda parametriska tester för att avgöra om det föreligger en statistiskt säkerställd skillnad mellan seten beroende på om APD beräknas med utgångspunkt från PL eller PL1. Vi väljer Students t-test (Dror et al. 2020:13–14).

	substantiv			adjektiv		
	PL1	PL2	PL PL1	PL PL2	PL PL1	PL PL2
T-test	-1,60145	1,35335	-0,25283	-0,61666	0,66689	0,04737
p-ttest	0,10937	0,17603	0,80041	0,53754	0,50493	0,96222
Cohens d	-0,05341	0,04513	-0,00842	-0,02979	0,03219	0,00228
Levenes	0,77308	0,3445	0,08658	1,10757	0,78542	0,02794
p-levene	0,37932	0,55728	0,76859	0,29276	0,37561	0,86728

Tabell 25. Students t-test med p-värde (p-ttest) och effektstorlek (Cohens d), samt Levenes test med p-värde (p-levene), för jämförelserna PL1\_PL2, PL\_PL1, och PL\_PL2.

Levenes test i tabell 25, som i alla jämförelser ger ett  $p > 0,05$ , visar att det inte finns något tecken på att varianserna skiljer sig mellan seten, vilket är en av förutsättningarna för att Students t-test skall anses valid. Alla p-värden för Students t-test överstiger också signifikansnivån  $\alpha = 0,05$ , vilket gör att vi inte kan förkasta nollhypotesen om att det inte finns någon verklig skillnad mellan dataseten (Dror et al. 2020:7). De skillnader som observeras i APD när embeddings i LA jämförs med embeddings i PL, PL1 eller PL2 är alltså inte tillräckligt stora för att vara statistiskt säkerställda. Notera dock att det finns viss variation. Varje gång PL jämförs med PL2 får vi nämligen mycket höga p-värden, både i Students och i Levenes test, både för substantiv och för adjektiv. Detta tyder alltså på att det finns en högre probabilitet att PL och PL2 har lika varianser och att de sannolikt är mycket lika varandra, vilket är logiskt med tanke på att ca  $\frac{3}{4}$  av alla word embeddings i PL finns i PL2. När PL jämförs med PL1 är probabiliteten för dessa observationer lägre och lägst när PL1 jämförs med PL2 (två set helt oberoende av varandra), förutom för adjektiv, där p-värdena är lika i Students t-test (men varianserna mellan adjektiven verkar skilja sig mer mellan PL1 och PL2 än mellan PL och PL1). Dessutom är p-värdet för jämförelsen PL1\_PL2 hos substantiven inte så långt från signifikansnivån (dock tillräckligt långt för att nollhypotesen måste förkastas). Det finns alltså skäl att anta att skillnader i resultaten skulle kunna uppstå beroende på om man definierar kristet latin som bestående av texter fram till år 600 e.Kr. eller om man anser det vara ett register som *Patrologia Latina* i sin helhet är representativ av, men att dessa skillnader inte är tillräckligt markerade för att vara statistiskt säkerställda. Observera dock att denna hypotetiska skillnad, även om den skulle vara signifikant, ändå skulle vara tämligen begränsad. Detta ser vi i det att resultaten för Cohens d, vilka visar effekten av den observerade skillnaden, är mycket låga. Detta beror kanske på att algoritmen vi valde (en anpassad form av APD), snarare än modellen XPL2, inte levererar fingranulära skillnader, även om vi har sett att kombinationen APD och XPL2 gav goda resultat i benchmarkstudien. Våra APD-värden går nämligen från ca 0,18 till 0,40 och det vore intressant att hitta en algoritm som ger resultat i ett större spann, utan att fördenskull förlora i precision eller recall.

Detta om vi betraktar alla lemman samtidigt. Givetvis finns skillnader i enskilda fall, och det är kanske dessa som mest intresserar latinisten. För att identifiera de lemman som mest varierar i APD beroende på vilken korpus man jämför LA med, räknar jag den absoluta skillnaden mellan APD-värdena tagna med de olika delarna av PL och normaliserar listan med z-score (eftersom vi handskas med tämligen små skillnader och det är svårt att skilja outliers, vilket är det som intresserar oss). Därefter sorterar jag resultaten på z-score för jämförelsen PL1\_PL2, vilken är som vi har sett den mest markerade skillnaden. Resultaten sammanställs i tabell 26 för de substantiv där  $z > 3$ , vilket är ganska extremt och därmed motsvarar en mycket hög probabilitet att vederbörande lemma får ett mycket annorlunda APD beroende på korpus.

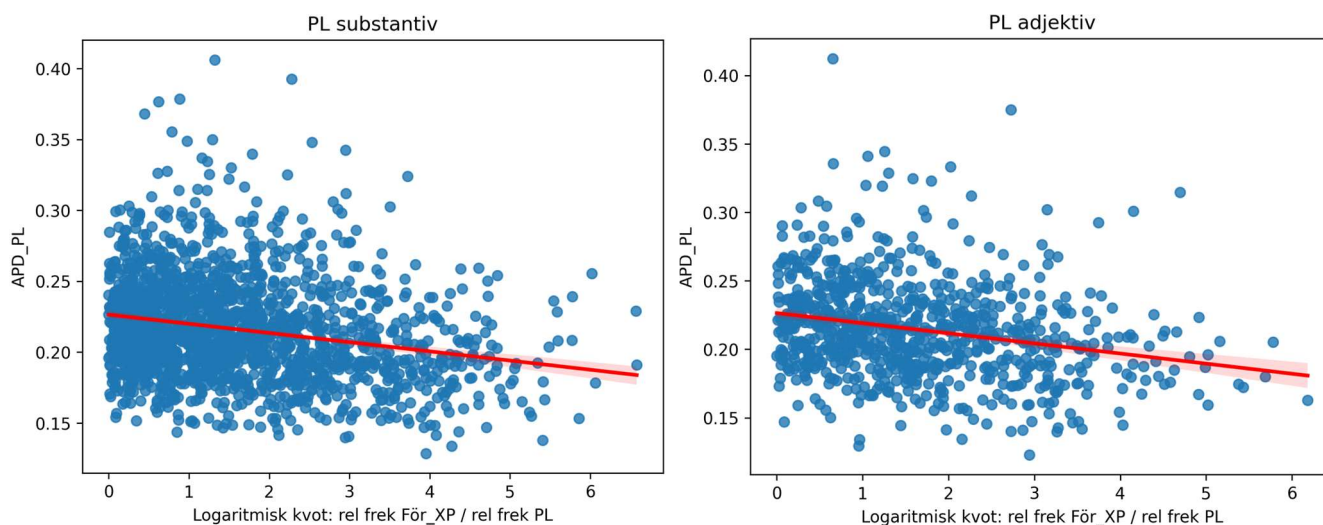
	w	APD PL	APD PL1	APD PL2	z PL PL1	z PL1 PL2	z PL PL2
remus	7	0,32508	0,2333	0,33281	10,12430	9,15842	2,06914
sus	5	0,33448	0,2471	0,34105	9,60483	8,60572	1,68126
natalis	7	0,34877	0,28025	0,35613	7,37815	6,80944	1,94542
seruus	6	0,2837	0,22155	0,29416	6,62608	6,48438	2,98198
comitatus	3	0,29584	0,23176	0,30054	6,85395	6,10365	1,05599
par	5	0,37868	0,31676	0,38519	6,59893	6,06886	1,66120
appellatio	5	0,33016	0,27232	0,3386	6,11723	5,85513	2,30654
manis	4	0,34261	0,37423	0,31569	3,02161	5,08572	8,48576
alexander	4	0,25267	0,20082	0,25829	5,41003	4,97936	1,36361
aquila	6	0,30044	0,33504	0,27949	3,37343	4,78850	6,48955
censura	4	0,35539	0,30856	0,36036	4,81735	4,41572	1,14627
uitis	3	0,24624	0,20502	0,25543	4,15502	4,27755	2,55732
comes	8	0,32766	0,27964	0,32975	4,95785	4,24773	0,18327
patrocinium	6	0,24038	0,1968	0,24598	4,43364	4,15528	1,35692
pagina	6	0,27445	0,23031	0,27909	4,49976	4,11552	1,03592
interdictum	6	0,29088	0,24536	0,29339	4,66269	4,04096	0,32371
translatio	7	0,29996	0,25977	0,30722	4,03341	3,98330	1,91198
feria	5	0,27529	0,23388	0,27862	4,17745	3,71391	0,59790
largitio	7	0,23946	0,19747	0,24189	4,24592	3,68210	0,29696
merx	8	0,23024	0,19498	0,23926	3,45136	3,66819	2,50048
dionysius	5	0,24294	0,2028	0,24701	4,02751	3,66123	0,84533
fabius	7	0,20801	0,22589	0,18204	1,39941	3,62544	8,16810
recordatio	5	0,25447	0,21726	0,25997	3,68158	3,51212	1,32349
dis	5	0,24138	0,20767	0,24999	3,26836	3,47335	2,36339
pratium	4	0,2272	0,18721	0,22909	4,00980	3,42961	0,11640
scriptum	3	0,29669	0,26014	0,3016	3,60366	3,38786	1,12621
inceptum	5	0,17412	0,21341	0,17203	3,92715	3,37991	0,18327
periurium	6	0,1907	0,23021	0,19012	3,95313	3,25167	-0,32163
tus	6	0,25546	0,22253	0,26229	3,17627	3,21887	1,76820
germanus	6	0,27689	0,23979	0,27944	3,66859	3,20793	0,33708
cotta	4	0,23915	0,20995	0,24928	2,73589	3,17612	2,87163
receptus	3	0,20832	0,17081	0,20935	3,71700	3,09759	-0,17116

Tabell 26. Substantiviska lemmorna sorterade efter störst normaliserad (z-score) diskrepans mellan APD mätt på PL1 respektive PL2.

Utän att gå in i för många detaljer kan vi konstatera att det felaktigt lemmatiserade ordet *sus*, tidigare kommenterat, faktiskt får ett högt APD-värde i PL2, men hamnar under tröskelnivån i PL1. Detta stämmer överens med intuitionen om probabiliteten att hitta formen *sue* för *suae*, vilken borde öka med tiden och vara högst efter 600 e.Kr. Samma öde lider *remus*. Hittar vi alltså referenser till Romulus bror i större utsträckning i PL1 än i PL2? *Comitatus* är intressant: enligt B används ordet under medeltiden i specialbetydelsen ”grevskap”. Inget märkvärdigt alltså att dess APD sticker iväg i PL2. Förmodligen gäller detta även *appellatio*. Gällande *manis* observerar vi det omvända: texterna i PL1 handlar förmodligen i högre grad om den maniska religionen, som ju låg närmare i tiden, än de i PL2. Vi kan också ana att de flesta texter i PL1 handlar om just religionen och vi har sett tidigare hur ämnet (topic)

kunde påverka resultaten. För adjektivens del hittar vi 18 lemman med  $z > 3$ , men skillnaderna mellan PL1 och PL2 är mindre markerade.

Slutligen vill jag påpeka att jag inte kan hitta i materialet några tecken på att frekvenserna skulle kunna ha en effekt på APD-värdena, som Schlechtweg et al. påpekade (2020:10–11), vilket framgår i figur 3 och tabell 26.



Figur 3. Logaritmisk kvot mellan de relativa frekvenserna i LA och de relativa frekvenserna i PL, samt motsvarande APD-värde, för substantiv och adjektiv, med linjär regression. ”För\_XP” står för LA.

I figur 3 visas på x-axeln en normalisering av skillnaden mellan de relativa frekvenserna i LA och i PL, som på engelska kallas *log-ratio* och används mycket i korpuslingvistik: först divideras de relativa frekvenserna av varje lemma i LA med de i PL, en logaritm (med bas 2) av kvoten beräknas för att normalisera och reducera effekten av extremvärden, och därefter tas det absoluta värdet.<sup>78</sup> Som vi kan se tyder den linjära regressionen snarare på att skillnader i frekvenser har motsatta effekten på APD: ju större skillnad i frekvenserna, desto lägre APD. Detta kan vi också bekräfta med Pearsons test (Dror et al. 2020:16).

	Substantiv			Adjektiv		
	APD PL	APD PL1	APD PL2	APD PL	APD PL1	APD PL2
Log ratio	-0,22444	-0,23832	-0,22259	-0,24552	-0,26735	-0,24239
p-värde	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000

Tabell 27. Korrelationskoefficienten Pearson  $r$  för den logaritmiska kvoten och APD-värdena, mätta i PL, PL1 och PL2, för substantiv och adjektiv.

Pearsons test visar, med ett p-värde nära 0, dvs med mycket hög signifikans, att det finns en svagt negativ korrelation mellan den logaritmiska kvoten på de relativa frekvenserna och APD-värdena. Pearsons  $r$  ligger mellan -1 (fullkomlig negativ korrelation) och 1 (fullkomlig positiv korrelation). Som vi kunde se i figur 3 finns det relativt få värden där de relativa frekvenserna skiljer sig markant och trenden är inte tydlig. Dessa resultat hade säkerligen påverkats om vi hade inkluderat låga frekvenser och det verkar därför ha varit ett klokt beslut att utesluta absoluta frekvenser under 50.

<sup>78</sup> Resultatet av den logaritmiska kvoten är nämligen negativ eller positiv beroende på vilken av de två jämförda frekvenserna som är större, men i vårt fall är vi intresserade av magnituden, inte riktningen. Därför tar vi det absoluta värdet, som alltid är positiv.

## 4.5 Sammanfattande slutsatser och diskussion

Vi har med hjälp av den omtränade BERT-modellen XPL2 kunnat mäta diakron semantisk variation mellan en förkristen korpus och *Patrologia Latina*. I kombination med en anpassad metod för APD har denna metodologi visat sig överträffa alla andra för latinets del redovisade i Schlechtweg et al (2020). I fortsättningen skall jag besvara forskningsfrågorna och försöka utvidga perspektivet med avstamp i resultaten. Sist skall jag dra metodologiska slutsatser.

### 4.5.1 F<sub>1</sub>

Med hjälp av Zhou & Lis metod med den 75:procentilen har vi kunnat bekräfta de allra flesta ord som i litteraturen anges som exempel på kristet latin, och formulerat hypoteser kring varför några enstaka hamnar under tröskelnivån: för låga frekvenser, ord som egentligen speglar medeltida latin snarare än kristet latin. I andra fall har vi konstaterat att metodologin kan vara mindre effektiv för att upptäcka semantisk variation i metaforer, och i ord där den semantiska förändringen är subtil, detta i båda fall eftersom markörer i form av ord kan utebli och meningarna förstås av en människa med hjälp av kontexten, vilket modellen inte tar hänsyn till och därmed förbiser. Metoden verkar också ibland fånga upp förändringar som är mer av syntaktisk än av semantisk art. Med samma metod har vi levererat en lista över ord som har högre probabilitet att ha genomgått en diakron semantisk förändring och föreslår en kompletterande datadriven metod för att hitta dem som mest sticker ut.

En kvalitativ analys med nedslag i materialet och jämförelser med referenslitteraturen har visat att de föreslagna lemnarna i huvudsak kan motiveras, förutsatt att inga systemiska lemmatiseringsfel föreligger. Detta har tyvärr visat sig vara ett stort aber i denna studie. Således står sig de föreslagna orden inte utan en *post-hoc* kvalitativ analys med nedslag i materialet. Ett annat problem har varit att den observerade variationen inte alltid kan motiveras ha skett under kristendomens inflytande, utan fler faktorer kan ligga bakom.

### 4.5.2 F<sub>2</sub>

Vi har visserligen inte hittat någon statistiskt säkerställd effekt av att avgränsa *Patrologia Latina* med årtalet 600 e.Kr. som skiljelinje, men vi har kunnat *skönja* en viss påverkan, både kvantitativt på hela datasetet och kvalitativt på enstaka ord. Notera dock att resultaten förmodligen påverkas av att de ordlistor som jämförs med varandra är tämligen stora. Denna observerade skillnad beror tveklöst på att det i *Patrologia Latina*, som sträcker sig över ett årtusende, förekommer en viss diakron variation som är orelaterad till kristendomen. Allt annat än det hade varit förvånande, eftersom språk ju tenderar att utvecklas över tid, även om man kan förvänta sig att en genuin språklig utveckling hindras av att latin med tiden övergår från modersmål till att i stort sett fungera endast som skriftspråk. Man kan också fundera på om variationen beror på att många texter skrivna mellan 250 och 600 e.Kr. är enhetligare sinsemellan än senare texter, i det att de i stor utsträckning tillhör kyrkofäderna. Ändock är endast 32 lemmarna med markerad variation beroende på om LA jämförs med PL1 eller med PL2 (jfr tabell 26) ett relativt klen resultat. Detta betyder alltså antingen att skillnaderna i APD mellan PL1 och PL2 överlag är små, vilket den statistiska analysen pekat på, eller att algoritmen har svårt att fånga upp skillnader på den nivån, eller en kombination av dessa två faktorer.

### 4.5.3 Diskussion

Om vi väljer att förkasta hypotesen om att en indelning före och efter 600 e.Kr. har en mätbar effekt på resultaten, måste vi dra den provisoriska slutsatsen att de semantiska företeelser som vi stöter på i litteraturen om kristet latin i huvudsak inte är bundna vid en viss tid utan förekommer i princip i all kristen litteratur. Enligt ett sådant synsätt betraktas kristet latin som ett slags register. På samma sätt som vi i en text om IT förväntar oss att *mus* är ett pekdon medan samma ord i en naturvetenskaplig tidskrift med högre probabilitet refererar till däggdjuret, kan vi förvänta oss att ord som *salus*, *proximus* eller

*fidelis* i en utpräglat kristen text betyder en sak, men en annan i en text som skrivs i ett annat register. Kyrkofädernas latin skulle i sådant fall ha varit som en katalysator: då tar de nya semantiska betydelseformerna form för att beskriva alla aspekter av den nya religionen, och repliceras sedan i liknande texter under hela latiniteten. Detta förutsätter att *Patrologia Latina* är representativ för kristet latin, vilket jag vill påstå som nu arbetat med korpuserna i flera månader. Dels är den tämligen stor, hela 15 gånger större än det material jag med stor möda lyckas skrappa ihop från den förkristna tiden, dels någorlunda varierad i fråga om genrer (även om jag inte tycker mig ha sett några teaterpjäser, som i det förkristna materialet), men framför allt sticker en sak ut: få är de meningar i PL som inte visar någon koppling till kristendomen. I ett sådant perspektiv skulle den hypotetiska bristen av förkristna betydelser i andra register efter kyrkofädernas tid kunna förklaras som en effekt av att de flesta texter vi möter från medeltiden är just skrivna i det kristna registret, eller kanske rentav behandlar kristna teman. Frågan är: hur ser det i sådant fall ut i t.ex. vetenskapliga texter från renässansen och senare?

Brist på tydliga bevis betyder dock inte att effekten saknas, bara att vi inte kunnat mäta den på ett tillfredsställande sätt. Förmodligen skulle vi kunna få fram tydligare resultat med en bättre anpassad algoritm, som ger APD-värden i en större skala. Vi behöver också en bättre metodologi för att skilja medeltida latin från kristet latin, om detta överhuvudtaget låter sig göras. Om det verkligen skulle finnas en skillnad, hur skulle vi kunna tolka den? Är den resultatet av att materialet i PL1 förmodligen till stora delar består av teologiska traktat skrivna av förhållandevis få författare, i vilket fall det är frågan om en skillnad i temat eller genrer? Eller finns det en verklig semantisk skillnad som kan förklaras diakront, alltså att PL2 i större utsträckning fångar upp medeltida företeelser än PL1, vilket vi har kunnat se i enstaka nedslag? Enligt detta andra synsätt summerar kristet latin alla de semantiska egenheter som vi i huvudsak hittar hos kyrkofäderna, och som liksom andra språkliga företeelser inte står tidens tand utan utvecklas med tiden.

Om vi skall utgå från *Sondersprache*-hypotesen bör vi studera företeelsen från ett sociolingvistiskt perspektiv, något som svårligen låter sig göras med det material som valdes för studien. Det är dock oklart om något sådant överhuvudtaget är möjligt. En möjligt första steg vore att annotera texterna med ytterligare metadata, såsom geografiskt ursprung och inte minst genre, något tidigare forskning har påpekat effekten av (McGillivray et al. 2019:898).

#### 4.5.4 Metodologisk diskussion

Studien är bland de första att analysera kristet latin från ett datadrivet perspektiv, dvs utan att utgå från på förhand givna uppslagsord, och resultaten fördjupar det Sprugnoli et al. med en liknande ansats påbörjade (2020).

Jag menar att en kvalitativ analys av de i studien framtagna lemmarna över 75:e percentilen, för att skilja falska positiva från riktiga resultat, eventuellt parad med en annotering enligt riktlinjerna i McGillivray & al. (2022a), skulle kunna producera en användbar goldstandard som framtida modeller skulle kunna testas mot. Med en sådan metodologi skulle man kunna ta fram nya modeller och algoritmer, som bättre kunde belysa det vi kallar kristet latin. Ett sådant arbete skulle förslagsvis kunna utföras inom ramen för en kandidatuppsats.

Jag insåg ganska tidigt i analysen att man med fördel skulle kunna analysera diakron semantisk variation av enskilda ordformer, utan koppling till ordklass (PoS) och lemma. Man skulle visserligen inte kunna skilja mellan homonymer som *cura* (substantiv) och *curo* (imperativ av *curo*), men å andra sidan skulle man kunna undvika lemmatiseringsproblemen. Det bästa hade förstås varit att förbättra våra lemmatiserare. I samband med den här uppsatsen har jag också insett vikten av s.k. NER: att person- och platsnamn dyker upp i vårt material som semantiskt förändrade är inte intressant, särskilt då romerska kognomina i kraft av sin etymologi inte sällan lemmatiseras som substantiv. En caveat är dock att en analys baserad på ordformerna skulle vara datakrävande. Jag har givit detta ett försök men valde av etiska skäl att avbryta det, då jag insåg att beräkningarna skulle ta flera veckor på gpu-servern och därmed kosta stora mängder el, och jag ansåg att jag trots allt hade samlat redan tillräckligt med material för studien.

Ett alternativt tillvägagångssätt skulle också vara att i stället för att bunta ihop ett polysemt lemmas olika betydelser beräkna graden av semantisk variation av med hjälp av lämplig WSD-teknik åtskilda

betydelser. På så vis skulle man kunna studera vilka specifika betydelser av ett ord som expanderar eller uppstår. Som Tahmasebi et al. skriver är denna metod utmanande (2021:34), och den låter sig nog inte göras i en explorativ korpusdriven ansats som i denna uppsats, men skulle väl kunna ingå i en korpusbaserad studie av utvalda ord.

Ett relaterat problem i min studie är de tekniska tillkortakommandena, som gjorde att jag tvingades begränsa studien till lemman inom ett visst frekvensspann. Således filtrerades lemman med en absolut frekvens över 1 000 godtyckligt bort. En förbättring vore att hitta ett sätt att analysera högfrekventa lemman. Jag har provat att batcha APD-beräkningarna, men minnet blir ändå fullt med ord som *sanctus* och *spiritus*, som av förklarliga skäl är oerhört frekventa i PL. Jag planerar att bygga en egen gpu-dator, men det är oklart om jag kommer att kunna överträffa minnet på GU:s gpu-server. En egen GPU skulle dock ge större kontroll över resurserna och möjliggöra anpassning till dessa beräkningar. Man kan också fundera på nyttan av en explorativ studie över ännu fler ord som under kristendomens inflytande förändrats semantiskt. Det kanske vore klokt att utgå från resultaten i denna studie och liksom påpekats ovan sälla fram de mest markanta exemplen, som skulle kunna fungera som markörer och användas i andra studier, snarare än försöka beskriva hela latinitetens semantiska variation.

Något som jag tidigt valt att bortse från är att det finns fler LLM än LatinBERT, t.ex. RoBERTa Latin (Ströbel, 2022). Man skulle med fördel kunna utgå från andra modeller och göra liknande studier.

## 5 Förkortningar och termer

APD	Average Pairwise Distances. Metod vid jämförelse av embeddings mellan två korpora.
BERT	Bidirectional Encoder Representations from Transformers. Typ av LLM.
CBOW	Continuous Bag of Words. Metod som går ut på att gissa $w$ utifrån ett fördefinierat kontextfönster runtom det.
DAPT	Domain-Adaptive PreTraining. Omträning av LLM med ett domänspecifikt material.
DH	Distributiva Hypotesen.
GCD	Graded Change Detection. Uppgift som går ut på att räkna ut hur mycket ett ord har förändrats semantiskt mellan två korpora.
Gold standard	Manuellt annoterat dataset som används för att träna eller utvärdera maskininlärningsmodeller.
Ground truth	”Facit”.
GSV	Graderad Semantisk Variation.
LA	<i>Latinitas Antiqua</i> .
LLM	Large Language Model. Maskininlärningsmodell för NLP.
LR	Learning Rate. Hyperparameter vid maskininläring.
MLM	Masked Language Modeling. Typ av uppgift vid träning av LLM.
NER	Named Entity Recognition. Metoder för upptäckt av egennamn, orter, etc.
NLP	Natural Language Processing. Forskningsfält inom språkteknologi, i vilket algoritmer och verktyg utvecklas för automatisk eller semiautomatisk bearbetning av språklig data.
PL	<i>Patrologia Latina</i> .
PCA	Principal Component Analysis. Dimensionsreduktionsmetod.
PoS	Part of Speech. Ordklass.
PMI	Point-Wise Mutual Information. Används för att sälla bort funktionsord.
PRT	Word Prototype. Metod vid jämförelse av embeddings mellan två korpora.
SDSC	Sense-Differentiated Sense Change.
SG	Skip-Gram. Metod som går ut på att gissa kontexten givet $w$ .
SVD	Singular Value Decomposition. Metod för dimensionsreduktion.
TEI	Text Encoding Initiative. Standard för annotering av XML-filer.
WLSC	World-Level Sense Change.
WSD	Word Sense Disambiguation. Fält inom NLP där syftet är att automatiskt tilldela en specifik betydelse i fall av polysemi. Även kallad WSI.
WSI	Word Sense Induction. Se WSD.

## 6 Referenslista

### 6.1 Tryckt litteratur

- Ahlberg, A. W., Sörbom, G., & Lundqvist, N. (1966). *Latinsk-svensk ordbok (andra upplagan)*. Esselte studium.
- Bamman, D. & Burns, P. J. (2020). *Latin BERT: A Contextual Language Model for Classical Philology*.
- Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. I Montavon, G., Orr, G., Muller, K.-R. (Red.), *Neural Networks: Tricks of the Trade* (s. 437–478). Springer.
- Blaise, A. (1975). *Dictionnaire latin-français des auteurs du moyen-âge. Lexicon latinitatis medii aevi, praesertim ad res ecclesiasticas investigandas pertinens*. Brepols.
- Blaise, A., & Chirat, H. (1967). *Dictionnaire latin-français des auteurs chrétiens*. Brepols.
- Bloch, R. H. (1994). *God's Plagiarist. Being an Account of the Fabulous Industry and Irregular Commerce of the Abbé Migne*. The University of Chicago Press.
- Burns, P. J. (kommande). *Exploratory Philology: Learning About Ancient Languages through Computer Programming*.
- Burton, P. (2011) Christian Latin. I: Clackson, J. (Red.), *A Companion to the Latin Language* (s. 485–501). Wiley-Blackwell.
- Caffagni, D., Cocchi, F., Mambelli, A., Tutrone, F., Zanella, M., Cornia, M. & Cucchiara, R. (2025). Benchmarking BERT-based Models for Latin: A Case Study on Biblical References in Ancient Christian Literature. I Cornia, M. et al. *Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science, Udine, Italy, February 20-21, 2025*. CEUR Workshop Proceedings 3937.
- Clackson, J., & Horrocks, G. C. (2007). *The Blackwell history of the Latin language*. Blackwell.
- Clackson, J. (2011). *A Companion to the Latin Language*. Wiley-Blackwell.
- Clérice, T. (2022) Antiquité tardive et littératures latines : corpus et perspectives numériques. *Koinōnia = Koinōnia*, (46), 207–215.
- Dahl, D. A. (2023). *Natural Language Understanding with Python: Building Human-like Understanding with Large Language Models*. Packt Publishing.
- Denecker, T. (2018). Among Latinists: Alfred Ernout and Einar Löfstedt's responses to the 'Nijmegen School' and its Christian *Sondersprache* hypothesis. *Historiographia Linguistica*, 45(3), 325–362.
- Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. I Burstein et al. (Red.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1* (s. 4171–4186), Association for Computational Linguistics, Minneapolis.
- Dror, R., Shlomov, S., Reichart, R., & Peled-Cohen, L. (2020). *Statistical Significance Testing for Natural Language Processing*. Morgan & Claypool Publishers.
- Eisenstein, J. (2019). *Introduction to Natural Language Processing*. The MIT Press.
- Favre, L., du Fresne, C., du Cange, D. et al. (1883–1887). *Glossarium mediae et infimae latinitatis*. Niort.
- Firth, J. R. (1962). A synopsis of linguistic theory, 1930-1955. I Firth (red.), *Studies in Linguistic Analysis*. Basil Blackwell, Oxford.
- Gaffiot, F. (1934). *Dictionnaire illustré latin-français*. Librairie Hachette.
- Geeraerts, D., Speelman, D., Heylen, K., Montes, M., De Pascale, S., Franco, K., & Lang, M. (2024). *Lexical Variation and Change: A Distributional Semantic Approach* (1st ed.). Oxford University Press.

- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. I Jurafsky et al. (Red.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (s. 8342–8360). Association for Computational Linguistics.
- Keidar, D., Opedal, A., Jin, Zh. & Sachan M. (2022). Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang. I Muresan et al. (Red.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1* (s. 1422–1442). Association for Computational Linguistics.
- Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1), 151–171.
- Lewis, C. T. & Short, T. (L&S). (1933). *A Latin Dictionary*. Oxford: Oxford University Press.
- Liu, Y., Tilahun, G., Gao, X., Wen, Q. & Gervers, M. (2024). *Comparative Analysis of Static and Contextual Embeddings for Analyzing Semantic Changes in Medieval Latin Charters*. University of Toronto.
- López Silva, X. A. (2003). El influjo del latín de los cristianos en la evolución general de la lengua latina. *Ianua. Revista Philologica Romanica*, (4), 115–126.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge.
- McGillivray, B. and Kilgarriff, A. (2013). Tools for Historical Corpus Research, and a Corpus of Latin. *New Methods in Historical Corpus Linguistics*, 1(3), 247–257. Tübingen: Narr.
- McGillivray, B. (2014). *Methods in Latin computational linguistics*. Brill.
- McGillivray, B., Hengchen, S., Lähteenoja, V. et al. (2019). A Computational Approach to Lexical Polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, 34(4), 893–907.
- McGillivray, B. & Tóth, G. M. (2020). *Applying Language Technology in Humanities Research: Design, Application, and the Underlying Logic*. Springer International Publishing.
- McGillivray, B., Kondakova, D., Burman, A., Dell'Oro, F., Bermúdez Sabel, H., Marongiu, P. & Márquez Cruz, M. (2022a). A New Corpus Annotation Framework for Latin Diachronic Lexical Semantics. *Journal of Latin Linguistics*, 21(1), 47–105.
- McGillivray, B. (2022b). *How to Use Word embeddings for Natural Language Processing*. SAGE Publications Ltd.
- Mohrmann, C. (1977). *Études sur le latin des chrétiens. Tome IV, Latin chrétien et latin médiéval*. Rom: Edizioni di storia e letteratura.
- Ortuño Arregui, M. (2016). Latín de los cristianos: Aproximación lingüística. *ArtyHum Revista de Artes y Humanidades*, (20), 57–65.
- Ortuño Arregui, M. (2019). Los cristianismos lexicológicos y semánticos en la obra literaria de Lactancio. *Revista chilena de estudios medievales*, (16), 20–25.
- Periti, F. & Montanelli, S. (2024). Lexical Semantic Change through Large Language Models: a Survey. *ACM Computing Surveys*, 56(11), artikel 282, 1–38.
- Periti, F. & Tahmasebi, N. (2024). A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change. I *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1* (s. 4262–4282).
- Perrone, V., Hengchen, S., Palma, M., Vatri, A., Smith, J. Q., & McGillivray, B. (2021). Lexical semantic change for Ancient Greek and Latin. I Tahmasebi et al. (Red.), *Computational Approaches to Semantic Change* (s. 287–310). Language Science Press.
- Rao, D. & McMahan, B. (2019). *Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning* (First edition). O'Reilly Media.
- Ribary, M., & McGillivray, B. (2020). A Corpus Approach to Roman Law Based on Justinian's Digest. *Informatics*, 7(4).
- Rodda, M. A., Senaldi, M. S. G. & Lenci, A. (2017). *Panta rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek*. *Italian Journal of Computational Linguistics*, 3(1), 11–24.

- Saussure, F. de, Bally, C., & Sechehaye, A. (1971). *Cours de linguistique générale*. Payot (publicerad första gången 1916).
- Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H. & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. I Herbelot et al. (Red.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (s. 1–23). Barcelona.
- Schrijnen, J. (1932). Charakteristik des altchristlichen Latein. I Mohrmann, C. (1977). *Études sur le latin des chrétiens. Tome IV, Latin chrétien et latin médiéval*. Rom: Edizioni di storia e letteratura (s. 367–404).
- Souter, A. (1949). *A glossary of Later Latin to 600 A. D.* Clarendon Press.
- Sprugnoli, R., Moretti, G., & Passarotti, M. (2020). Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas. *IJCoL*, 6(1), 29–45.
- Tahmasebi, N., Borin, L. & Jatowt, A. (2021). Survey of Computational Approaches to Lexical Semantic Change Detection. I Tahmasebi et al. (Red.), *Computational Approaches to Semantic Change* (s. 1–91). Language Science Press.
- Tahmasebi, N. & Dubossarsky, H. (2023). Computational modeling of semantic change. *arXiv:2304.06337*.
- Versteegh, K. (2017). Religion as a linguistic variable in Christian Greek, Latin, and Arabic. *Philologists in the World: A Festschrift in Honour of Gunvor Mejdell*, 57–99.
- Wang, R., & Choi, M. (2023). Large Language Models on Lexical Semantic Change Detection: An Evaluation. *arXiv:2312.06002*.
- Zhou, J. & Li, J. (2020). TemporalTeller at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection with Temporal Referencing. I Herbelot et al. (Red.), *Proceedings of the 14th International Workshop on Semantic Evaluation* (s. 222–231). Barcelona.

## 6.2 Webbsidor och blogginlägg

- Brepols (2025). *Database of Latin Dictionaries*. <https://databaser.ub.gu.se/database-of-latin-dictionaries-dld/188464>, hämtad 2025-04-21.
- English Corpora, *NOW Corpus*. <https://www.english-corpora.org/now/>, hämtad 2024-11-14.
- Bamman, D. & Burns, P. (u.å.) *latin-bert*. <https://github.com/dbamman/latin-bert>, hämtad 2024-11-08.
- Github, CLTK, *CLTK*. <https://github.com/cltk/cltk/blob/master/notebooks/CLTK%20Demonstration.ipynb>, hämtad 2024-11-27.
- Github, OpenGreekAndLatin project, *Patrologia Latina*. [https://opengreekandlatin.github.io/patrologia\\_latina-dev/](https://opengreekandlatin.github.io/patrologia_latina-dev/), hämtad 2024-11-26.
- Lafage, D. (2025a). *Corpus\_Christi\_data*. [https://github.com/guslafda/Corpus\\_Christi\\_data](https://github.com/guslafda/Corpus_Christi_data), hämtad 2025-05-14.
- Lafage, D. (2025b). *XPL*. <https://github.com/guslafda/XPL>, hämtad 2025-05-14.
- Linnéuniversitetet, *Corpus Methods in Practice*. <https://lnu.se/en/course/corpus-methods-in-practice/vaxjo-international-part-time-spring/>, hämtad 2024-11-08.
- Linnéuniversitetet, *Programming for Digital Humanities*. <https://lnu.se/en/course/programming-for-digital-humanities/vaxjo-distance-exchange-part-time-autumn/>, hämtad 2024-11-08.
- Longree, D. & Fantoli, M. (2023). "LASLAfiles\_Latin\_DATformat", ULiège Open Data Repository, V1, <https://doi.org/10.58119/ULG/27VZID>, hämtad 2024-12-03.
- McGillivray, B., Schlechtweg, D., Dubossarsky, H., Tahmasebi, N. & Hengchen, S. (2020). Zenodo. <https://zenodo.org/records/3992738>, hämtad 2024-11-28.
- NE (Nationalencyklopedin), polysemi. <https://www-ne-se.ezproxy.ub.gu.se/uppslagsverk/encyklopedi/lång/polysemi>, hämtad 2025-05-03.
- OpenAI (2025). *ChatGPT*, version GPT-4-turbo. San Fransisco: OpenAI. <https://chat.openai.com/chat>
- Polycrates (2022, 22 oktober). *Latin enclitic tokenizer broken? #1190*. <https://github.com/cltk/cltk/issues/1190>

Ströbel, P.B. (2022). *RoBERTa Base Latin Cased v1*. <https://huggingface.co/pstroe/roberta-base-latin-cased>, hämtad 2024-04-23.

Universität Zürich, *Corpus Corporum – Repositorium operum Latinorum apud universitatem Turicensem*. <https://mlat.uzh.ch/home>, hämtad 2024-11-08.

Universität Zürich, *Antiquitas Posterior*. <https://mlat.uzh.ch/browser?path=/14004>, hämtad 2025-02-08.

Universität Zürich, *Auctores scientiarum varii*. <https://mlat.uzh.ch/browser?path=/12130>, hämtad 2025-02-08.

Universität Zürich, *Latinitas Antiqua*. <https://mlat.uzh.ch/browser?path=/35>, hämtad 2025-02-08.

Universität Zürich, *Patrologia Latina*. <https://mlat.uzh.ch/browser?path=/38>, hämtad 2024-11-08.