Barbara McGillivray*, Daria Kondakova, Annie Burman, Francesca Dell'Oro, Helena Bermúdez Sabel, Paola Marongiu and Manuel Márquez Cruz

A new corpus annotation framework for Latin diachronic lexical semantics

https://doi.org/10.1515/joll-2022-2007

Abstract: We present a new corpus-based resource and methodology for the annotation of Latin lexical semantics, consisting of 2,399 annotated passages of 40 lemmas from the Latin diachronic corpus LatinISE. We also describe how the annotation was designed, analyse annotators' styles, and present the preliminary results of a study on the lexical semantics and diachronic change of the 40 lemmas. We complement this analysis with a case study on semantic vagueness. As the availability of digital corpora of ancient languages increases, and as computational research develops new methods for large-scale analysis of diachronic lexical semantics, building lexical semantic annotation resources can shed new light on large-scale patterns in the semantic development of lexical items over time. We share recommendations for designing the annotation task that will hopefully help similar research on other less-resourced or historical languages.

Keywords: annotation; LatinISE corpus; Latin lexical semantics; semantic change

***Corresponding author: Barbara McGillivray**, Department of Digital Humanities, King's College London, London, UK, E-mail: barbara.mcgillivray@kcl.ac.uk. https://orcid.org/0000-0003-3426-8200
Daria Kondakova, Faculty of Classics, University of Oxford, Oxford, UK, E-mail: daria.kondakova@classics.ox.ac.uk. https://orcid.org/0000-0002-6634-2762
Annie Burman, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden, E-mail: annie.burman@lingfil.uu.se. https://orcid.org/0000-0003-2876-729X
Francesca Dell'Oro, Helena Bermúdez Sabel and Paola Marongiu, Institute of Language Sciences, University of Neuchâtel, Neuchâtel, Switzerland, E-mail: francesca.delloro@unine.ch (F. Dell'Oro), helena.bermudez@unine.ch (H. Bermúdez Sabel), paola.marongiu@unine.ch (P. Marongiu). https://orcid.org/0000-0002-8343-356X (F. Dell'Oro). https://orcid.org/0000-0002-8627-1367 (H. Bermúdez Sabel). https://orcid.org/0000-0002-5060-3307 (P. Marongiu)
Manuel Márquez Cruz, Department of Linguistics, Complutense University of Madrid, Madrid, Spain, E-mail: manmarqu@ucm.es. https://orcid.org/0000-0001-9040-169X

1 Introduction

Evidence of the phenomena of polysemy, lexical semantic change and variation is widely found in historical and current languages (e.g. Koptjevskaja-Tamm 2002). Over time, new lexemes enter the lexicon, others become obsolete. Existing lexemes acquire new senses, lose senses, or change, expand or restrict their semantic scope. These changes can often be traced back to the close relationship between culture and language. Understanding the conceptual schemes underlying a language gives us insights into how the speakers of that language conceive the world and how this changed over time. For example, the Latin word *uirtus* originally meant something close to 'manliness', acquiring part of the semantic range of the Greek word ἀνδρεῖα. With the conversion of the Latin-speaking world to Christianity, the semantics of this word included Christian virtue and miracles (Heim 1991). Similarly, *passio* originally meant 'suffering' or 'experience' and later extended its range to mean 'emotion' and the suffering and death of Christ and the martyrs (Auerbach 1937).

Anthropological studies in linguistics (Boas 1911: 59–73; Sapir 1912), cultural commentators and conceptual historians (e.g. Kuukkanen 2008; Richter 1995; Williams 1976) have all recognised that studying the phenomenon of lexical semantic change provides a way to understand the dynamics of cultural, social and political systems. Since the 1980s, scholars have tracked the semantic histories of individual words via philological methods (e.g. Kenny 1995; Wierzbicka 1997). In parallel, and sometimes in conversation with the conceptual and intellectual history tradition, linguists from a variety of backgrounds (historical-philological, structuralist, generativist, neo-structuralist and cognitive semantics) have focussed on the language-internal aspects of this phenomenon in relation to linguistic systems (e.g. Geeraerts 2010; Grondelaers et al. 2007; Koch 2016). These include the differentiation between semasiological or meaning-related innovations (i.e. the creation of new meanings within the range of a lexical item) and onomasiological or name-related innovations (i.e. the association of a concept with a new or alternative lexical item), and within the former category, the further distinction between specialisation and generalisation, or metonymy and metaphor.

The relatively recent availability of large-scale corpora of historical languages, coupled with new advances in computational linguistics, is opening new avenues for the study of diachronic phenomena, and particularly diachronic lexical semantics. Ancient languages offer us the opportunity to study long-term lexical semantic change at a relatively large scale. Among ancient languages, Latin is in an especially favourable position because several digital libraries, annotated corpora and other lexical resources are available to the scholarly community working on this language (for an overview, see Mambrini et al. 2020; McGillivray

2014). The past 15 years have seen an increase in the number of Latin annotated corpora, thanks to the contribution of projects such as Perseus Digital Library (Bamman et al. 2008), PROIEL (Haug and Jøhndal 2008) and the *Index Thomisticus* Treebank project (Passarotti 2019). These corpora are provided with morphological and syntactic annotation, and a portion of them is also enriched with semantic-role annotation. Over the last few years, the number of digital language resources available for Latin has increased further, also thanks to the *LiLa: Linking Latin* project (Passarotti et al. 2019), which has produced, among other resources, an extended and corrected version of Latin WordNet, a network of lexical concepts organised hierarchically into sets of synonyms, and a set of Latin sentiment lexicons. Despite its prominent position among ancient languages, Latin still lacks large-scale corpora annotated at the *lexical* semantic level.[1] The *Lexicon Translaticium Latinum* (Buccheri et al. 2021) is one of those digital resources, which supplements the information provided in Latin traditional dictionaries. It is a collaborative digital lexicographic project, which is aimed at developing an online open-access lexicon of metaphors in Latin. Its main objective is to capture large-scale metaphoric patterns documented in the Latin semantic system (Fedriani et al. 2020: 106) and to organise their meanings. At the moment, most of the metaphors collected belong to the domain of basic emotions (e.g. love, anger, hate, fear) and the searches are restricted to the source domains (e.g. opponent), emotional terms (e.g. anger) and image schemas (e.g. location). This electronic resource will allow users to study the relationship between metaphors (e.g. "love is a master" derives from "love is a living entity").

In this work we describe and analyse the results of the first annotation task aimed at the lexical semantic annotation of text passages from the Latin corpus LatinISE (McGillivray and Kilgarriff 2013). We focus on the semantics of a subset of Latin lexical units, offering a resource that can be used for preliminary analyses of the semantic shifts associated with late antiquity and the advent of Christianity. Over the course of its long history, the Latin language underwent various diachronic evolutions, including those related to the origin of Romance languages, linguistic borrowings of elements of its lexicon by other languages, and a series of semantic shifts that affected a large part of its vocabulary. As it became the official religion of Rome, the influence of Christianity on Roman society, and particularly on Western Roman society, is undeniable and is reflected in the areas of religion, ethics, and philosophy, but also in its institutional organisation and, crucially, in

1 A small-scale dataset containing the lexical semantic annotation of the Ancient Greek words μῦς, ἁρμονία and κόσμος in the Diorisis Ancient Greek corpus is now available (Vatri et al. 2019) and was used to test the accuracy of the Genre-Aware semantic change (GASC) system for ancient Greek (McGillivray et al. 2019; Perrone et al. 2019).

its language. The influence of Christianity left a deep mark on the Latin lexicon as a result of linguistic loan translations or calques[2] and semantic shifts.[3] Semantic shifts sometimes equipped terms from the pagan terminology with new meanings, as is the case, for example, with the terms *basilica* and *ecclesia* as opposed to *templum* (López Silva 2003: 122; Ortuño Arregui 2016: 61).[4]

The focus of this study is on the methodological lessons learnt from the experience of a recent corpus annotation task and on its implications for Latin lexical semantics studies. After introducing the features of Latin as a corpus language for what concerns the study of lexical semantics and the history of this language from antiquity onwards (Section 2), and considering semantic change associated with the spread of Christianity and other socio-political changes of late antiquity (Section 3), we will present the annotation task (Section 4). We will outline its challenges and provide a series of recommendations for designing and conducting similar annotation tasks for other ancient languages and further studies, based on qualitative and quantitative analysis of the annotation feedback and data (Section 5). A diachronic semantic analysis will focus on the word-specific patterns found in the annotation dataset and will present some preliminary results on the contribution of corpus-based lexical semantic annotation for Latin lexical semantics (Section 6). We will conduct an in-depth case study on the level of vagueness of some annotated words (Section 7) before the final discussion and conclusion (Section 8). All code produced for this paper is available in McGillivray et al. (2022) and the data are available in McGillivray (2021).

2 The study of lexical semantics with a corpus language

Before addressing the context and the details of the corpus annotation task, we want to draw the reader's attention to some characteristics of Latin which distinguish it from living languages, and which had important implications for the design and the implementation of the annotation task itself.

2 Examples of calques or loan translations from ancient Greek are *carnalis* from σαρκικός and *salvator* from Σωτῆρ (Bastardas 1973: 7; López Silva 2003: 121–122; Ortuño Arregui 2016: 61). Examples of loan translations from Hebrew are *sabbatum*, *pascha*, *satanas*, and *amen*.

3 Examples include *fides*, *sacramentum*, *confessio* and *spiritus* (López Silva 2003: 122; Ortuño Arregui 2016: 61).

4 The term *ecclesia*, borrowed from Greek, was used in Latin with the meaning of 'assembly' and designated both the buildings of Christian religious worship and the congregation of the Christian faithful.

Latin is often described as a dead language, as it is a language without living native speakers, but the phrase *dead language* is not unproblematic. It is easily confused with the concept of language death, which is the process by which a language ceases to be spoken (Crystal 2002: 1; Petrollino and Mous 2010: 208). Langslow has argued that "dead languages need not arise through language death, and language death often yields not a dead language but no language at all" (Langslow 2002: 24). Given these terminological issues, we believe it is more accurate to call Latin a *corpus language* (Cuzzolin 2019; Mayrhofer 1980; McGillivray 2014: 14–15; Untermann 1983). Using this term allows us to acknowledge the broader scope of corpus languages both in terms of time and space. It also serves to emphasise the similarities between the philological study of corpus languages and the approaches of modern corpus linguistics (Langslow 2002: 23–24).

Latin is a prime example of a language that is not only ancient, but has been actively used long after the end of antiquity. There is no conventional end-date for Latin as a living language. The question of when Latin became the Romance languages is nigh impossible to answer, as it is in fact two questions – when what we would recognise as Romance was sufficiently differentiated from Latin to be seen as different languages, and when the speakers started seeing the language they spoke as different from Latin.[5] Regardless of the answers to these questions, this was not the end of Latin as a living language. A scholarly tradition of Latin as a spoken and written language developed. Well into the Early Modern period, the sons of European elites were taught Latin from an early age, and the language remained in use in many educational and religious institutions.

Periodisation of a language, in particular one with as long and complex a history as Latin, is going to be reductive, and different scholars will use the same term in different ways. As a result, attaching dates to periods can be hard. The most commonly used periodisation for Latin identifies the following steps: Archaic Latin, Old Latin, Classical Latin, Late Latin, Mediaeval Latin and Neo-Latin (see Clackson 2011b: 4; Penney 2011; Sidwell 2015).

However, other dimensions exist. For instance, the term Vulgar Latin sometimes corresponds directly to what this periodisation calls Late Latin, but is sometimes used of non-elite Latin during what is seen as the Classical phase. Attempts have also been made to identify other specific forms of Latin. Schrijnen (1932) and the Nijmegen school have argued for the existence of Christian Latin. Although some proponents argued that the Latin of the Christians was an independent phenomenon within Latin (e.g. Loi 1978), other supporters have focused instead on particular Christianisms and Christian Latin as group jargon (e.g.

5 Much has been written on this question. See for instance Wright (1982: 1–4), Janson (1991), Varvaro (1991), McKitterick (1991), Zaccarello and Maiden (2003).

Mohrmann 1950–1951).[6] It is undeniable that Christianity brought with it new vocabulary, much of it borrowed or calqued from Greek, e.g. *episcopus* 'bishop' (Greek ἐπίσκοπος 'overseer, bishop') and compounds with *tinguo* 'dip, dye' such as *intinguo* taking on the meaning 'baptise' (Greek βαπτίζω 'dye') (cf. *TLL* 5,2:676, 7,2:20). The Greek influence is not surprising, as Christianity has its roots in the Greek-speaking Eastern parts of the empire, and Christians in the Western parts overwhelmingly spoke Greek for much of the first two centuries CE (see Lampe 2004: 27).

With its two and a half millennium history, the scope of Latin is incredibly broad, and the distribution of sources is not always balanced. Newer material is more likely to be preserved and available, not only on account of the passage of time and selective preservation, but also in regards to the longevity of certain writing materials.[7] Most of the extant Latin literature written before the spread of Christianity comes from the Imperial period, while the material from the Roman Republic is biased in favour of the 1st century BCE, and even more so, in favour of one individual: the politician, rhetorician and philosopher Marcus Tullius Cicero.[8] As a result, our material for antiquity is centred at certain points in time and with certain actors, which is less the case with post-Classical Latin. This is also the case in the material of this study. In total, there are 175 named individual authors in the dataset. Figure 1 shows the ten most common authors in terms of number of annotated passages.[9] A total of 2,398 passages were used in the annotation task. Of these, the ten most prolific authors had contributed 1,304, corresponding to

6 For an overview on the concept of Christian Latin, see López Silva (2003). For some of the most common criticisms of this theory, see Holford-Strevens (1981), and for methodological issues, Burton (2011).

7 Papyrus, the most common writing material in Roman antiquity for literary works, survives poorly in non-dry environments (see Frösén 2011).

8 As a means of illustration, the Loeb Classical Library contains 190 volumes of Latin. Sixty are by Republican authors. Of these, half – thirty – contain the works of Cicero. These statistics were compiled based on the Excel spreadsheet listing the complete series (Loeb Classical Library 2020). This does not take into account the thickness of the volumes, which range from 320 (the second volume of Varro's *De Lingua Latina*) and 768 pages (a volume of Plautus' play), nor any material not yet published, but gives a rough estimate of how much of the extant Republican literature is from Cicero. As a result, Cicero is our only source for the usage of some words and concepts (see e.g. Tracy 2008–2009).

9 The analysis was based on the information, retrieved 12 April 2021, given under "author" in the LatinISE corpus. Names have been standardised before being counted to avoid the same author with different name-forms (e.g. Peter Abaelardus and Petrus Abaelardus) being counted as two separate people. Unknown and uncertain authorships (a total of 245 passages, marked in LatinISE as "auctor incertus", "unknown", "[Anonymous]" and "No Author", as well as, in the case of six works, where no author is known and the title is given in place of an author's name) are not included in the graph, as they are a group, rather than an individual author. Congregations of the

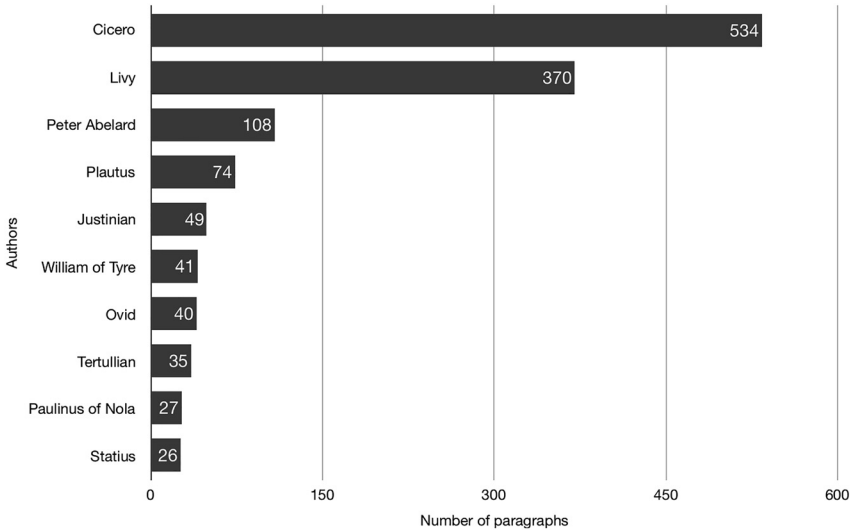

Figure 1: The ten authors whose work has contributed the most passages to the annotation.

54.4%. Only three of these ten authors – Peter Abelard, William of Tyre and Justinian – were active after the collapse of the Western Roman Empire. Taken together, they contribute only 198 passages, less than half the number from Cicero.

It is not only the nature of the material that makes Latin semantic annotation special, but also the annotators themselves. Using native speakers for the Latin annotation, as for the Swedish, German and English annotation, is not an option, for obvious reasons. To fully understand the approach taken by the annotators, it is worth considering how Latin is taught.

In most Western schools and universities, the teaching of Latin was noticeably different from the teaching of modern languages. Frequently, it continued to follow the Grammar-Translation Method (though to differing degrees), where the goal of language acquisition is to be able to read literature in the original (Richards and Rodgers 2005: 5–6).[10] There was a great focus on grammar, and vocabulary

Roman curia and material with multiple unnamed authors (a total six different groupings with a total of 29 annotated passages) have also been excluded.

10 López de Lerma (2015) and López de Lerma and Ambrós (2016) comprise a survey of Latin teachers in Spain, the majority of whom do not teach following the Grammar-Translation principles but hybrid methods. Macías Villalobos (2012, 2015) defends a hybrid methodology in which traditional Latin learning is combined with an inductive and contextual method. Experiences of the inductive method of Latin teaching are described in Antonini and Díaz Pereyro (2020) and Márquez Cruz and Fernández-Pampillón Cesteros (2019). See also Cardinaletti et al. (2016) and

was taught through wordlists, text glosses and dictionaries. Although in recent years there has been a push to include conversation and listening comprehension in the teaching of Latin and ancient Greek, this is a relatively recent trend (Hunt 2016: 2; Saffire 2006: 160). All this has consequences on how students of Latin built up and processed their vocabulary. Where modern language teaching encourages an instinctive understanding of words, the focus in Latin was on translation. This is particularly an issue when vocabulary is introduced by glossing, as students often retain the first translation they see of a word, making it hard to grasp nuances (Deagon 2006: 35; Scott Morrell 2006: 141). A sufficiently advanced student may develop a more nuanced understanding of vocabulary, but some of these patterns may persist.

A very common element of Latin students' toolbox are bilingual dictionaries, which give translations of individual words from Latin into a target language. In order to make dictionaries easier to navigate, different senses are isolated, sometimes tiered into subcategories. Take, for instance, the definition of *uirtus* in Lewis and Short (1879), given here without examples and attestations:

> I. *manliness, manhood; strength, vigor; bravery, courage; aptness, capacity; worth, excellence, virtue,* etc.
>
> I. In gen.
>
> A. Lit[eral]
>
> B. Transf[ered] *goodness, worth, value, power, strength, etc.*
>
> II. In partic.
>
> A. In the phrase deum uirtute, usu[ally] with dicam, *by the aid* or *merit of the gods,* i.e. *the gods be thanked*
>
> B. *Moral perfection, virtuousness, virtue*
>
> 1. Lit[eral]
>
> 2. Transf[ered] *Virtue,* personified as a deity
>
> C. *Military talents, courage, valor, bravery, gallantry, fortitude,* etc.
>
> D. *Obstinacy*

Iovino (2019) on inclusive pedagogical approaches for teaching dyslexic students Latin. Adema (2019) gathers a wide range of works which focus on the learning and teaching of Latin from different methodological and theoretical points of view.

This way of structuring the entries eases the use of dictionaries, but it is not without issues (see Adamska-Sałaciak 2014: 7). It invites us to see the senses as isolated from one another. Examples do not reoccur in more than one category, meaning that nuances, ambiguities and wordplay may be lost. It also means that we may lose the sense of interconnected meanings. The concept of "goodness, value" is connected to "manliness, manhood", which is connected to "power, strength", and the latter two are connected to "military talents, courage", which are connected to "value". There is no easy way of illustrating these complexities in a two-dimensional written medium such as the dictionary entry displayed above. Furthermore, the meanings and their relationships to one another may depend on the author, the context and the genre. Even the same example may be interpreted differently by different readers. If we speak of a general's *uirtus*, we might be referring to the fact that he is a talented and brave military man, or we may be sarcastic, and in fact mean to imply that he is obstinate and self-important. The ambiguity may in fact be the purpose and excluding one meaning in favour of another may impoverish the interpretation of the text.

Lexicographic practice is involved with three main aspects: the nature of the referent, the perception of the referent by the culture under study, and the verbal formalisation of the conceptualisation of the reality that determines the language of study (Fuertes-Olivera 2017: 332). As dictionary sense definitions and translations were the basis for the annotation of this study, this came with further challenges. In Latin dictionaries, definitions are often presented as equivalences between Latin and the language of reference of the dictionary, in the example above, English. Therefore, whenever the semantic scopes of the English and Latin terms do not completely overlap, the semantic development and impact of a term is not fully captured. In the analysis described in Section 8, we have attempted to quantify these grey areas through the concept of vagueness. In addition to the limitations outlined above, there are further challenges arising from the annotation task, which we will cover in Section 5.

3 Semantic change in light of the spread of Christianity and end of antiquity

Two major types of changes occur in the material of this study, which is summarised in Table 1. Some words have changed their meanings altogether, such as *pontifex* going from meaning 'priest' to meaning 'bishop, pope'. Other words add a meaning to the one that already exists. The most common type of semantic addition in this material is that which broadens the semantic window of a word.

For instance, *fidelis* adds the meaning 'Christian' to its earlier meaning 'faithful, safe', but it does not lose this meaning. Neither is it the only extension – in Mediaeval Latin, *fidelis* is used as a word for 'vassal', a lord faithful to the king (Habel 1959: 152).

Another aspect to be considered is the different domains of these words. The change of the words *dolus* and *itero* is not due to new religious beliefs or a changing political situation, but to an internal semantic development. During the Middle Ages, the verb *itero*, which in Classical Latin meant 'repeat', took on the meaning 'travel' by analogy to *iter* 'journey, road', side-lining the Classical deponent *itineror* (Dinkova-Bruun 2011: 291; Stotz 2000: 179). *Dolus* 'deceit' underwent a similar change, taking on the meaning 'pain' (Classical Latin *dolor*). Augustine of Hippo, who was active in the later 4th and early 5th centuries, discusses the use of *dolus* as meaning 'pain' (August. *In Evang. Iohan.* 7.18), but indicates strongly that it is a usage that has no prestige. This meaning later became generally accepted in Mediaeval Latin, but, as it is clear from Augustine's opposition, it is not inherently Christian (see Stotz 2000: 59). With the rejection of Mediaeval Latin by the humanists during the Renaissance, meanings such as this were phased out.

Three other categories can be identified. Some words are associated with secular power structures, such as *ciuitas, cohors, consul, dux* and *imperator*. Others are connected to the Christian church: *pontifex, potestas, sacramentum, sanctus, scriptura*. Yet others are not overtly Christian, but have connections to the Christian faith: *beatus, credo, fidelis, uirtus* and *humanitas*.

Words referring to secular power structures naturally shift when those structures change and this can take on different forms. Some are first generalised and then narrowed, such as *cohors*, whose meaning 'imperial court' is not derived from the military meaning, but rather from the more general meaning 'company'. *Cohors amicorum* is used by Suetonius to refer to the closest circle around the emperor (Suet. *Calig.* 19, *Ner.* 5, *Galb.* 7), making the shift to *cohors* meaning 'court' on its own easier. When a new aristocratic system crystallised during the early mediaeval period, it was easy for words such as *dux* 'leader' to be narrowed to mean a specific type of aristocrat – the word *duke* is ultimately derived from *dux* through Old French.[11] In antiquity, *ciuitas* meant 'citizenship', and by extension 'the citizenry, the body-politic'. Rarely, it was metonymically used to mean 'city' (e.g. Tac. *Hist.* 1.54, Sen. *Ben.* 6.32.1), which in the Middle Ages became the most common meaning (Gy 1975). This was helped by the fact that citizenship became less politically important and a less exclusive category.

11 The office of *dux* in late antiquity and the early mediaeval period has been exhaustively described by Zerjadtke (2019).

Although much of the political power of the consuls was transferred to the emperor at the beginning of the imperial period, it remained an honoured institution. It continued even after the traditional date of the fall of the Western Roman empire in 476. During this time, the office of consul was held by members of a few aristocratic families, who used it as a way of consolidating their power. The office was discontinued in both East and West in the 6th century (Cameron and Schauer 1982: 137–141). Once the consulate was gone, the word *consul* could be applied to a wide number of positions of power, from 'royal advisor' and 'count' to 'counsellor' and 'head of a guild' (Habel 1959: 82; Stotz 2000: 35–36, 102). The change of a word such as *pontifex*, a religious office but nonetheless one with political power, has more in common with these words.

Other words with changes associated with Christianity, *potestas, sacramentum, sanctus* and *scriptura,* gain new additional meanings referring to new religious contexts. Before Christianity, there was no such thing as a sacrament in the Christian sense of the word, just as there were no saints. *Scriptura* is not used in reference to holy texts previous to the Christian usage, as Roman polytheism did not include this as a concept. Other words are directly influenced by Greek. In the Vulgate, *scriptura* is used to translate Greek γραφή (e.g. Matt. 21:42), which was used to refer to the Torah by Greek-speaking Jews (Joseph. *Ap.* 2.4.45). Similarly, *potestas* meaning a type of angel is a direct translation of Greek ἐξουσία, literally 'power' (Eph. 3:10).

In the final category, the cases of *credo* and *fidelis* will serve as examples. Both words take on an additional meaning, 'to believe in God' and 'Christian' respectively, as the way in which divinity is understood in Christianity differs from how it is conceptualised in polytheistic Roman beliefs. A polytheistic Roman would not use the word *credo* to explain their devotion to Jupiter or Juno or any other god for that matter. While gods may be pleased through worship and sacrifice and therefore grant you favours, it was never a question of confiding, trusting or believing in them. The concerns of mortals were only of interest to the gods if they were offered something in return. By contrast, the connection between the Christian God and the Christians is of paramount importance. The act of trusting in the Christian God and devoting oneself to Him is a central part of the religion.

The change seen in *fidelis* is connected to a similar idea, the way in which God takes on the role of a king in Christianity. Being *fidelis* to a god in a Roman polytheistic context would make no sense, at least for those who were not priests or part of a mystery cult. Titles associated with power were often assigned to gods, ranging from Greek ἄναξ, used of gods in Homer (e.g. *Il.* 7.23) but of real-world kings in the Linear B tablets (Shelmerdine and Bennet 2008: 290), to Latin *rex*, used for human rulers and of Jupiter (e.g. Verg. *Aen.* 7.46, 1.65, respectively). At the same

time, mortal and divine kinghood was seen as separate and fully compatible. Christianity builds upon the Jewish concept of God as kings standing above the kings of Assyria, Babylon and Persia (Moore 2009: 162) through mentions of Christ as king (even βασιλεὺς τῶν βασιλευόντων 'King of Kings', 1 Tim. 6:15), and the Kingdom of God (βασιλεία τοῦ θεοῦ, Matt. 6:33) as superseding any earthly institution. In the 4th century, around the conversion of Constantine, the image of Christ as ruler of the universe appeared in earnest (Beskow 1962: 11–12). This conceptualisation of Christ as King means that words previously used for the fidelity to a king were now used for fidelity to God.

The words of the annotation task that show semantic change have undergone it in different ways and for different reasons, but many of these changes are inextricably linked to the temporal and social context where they took place.

4 Annotation task: background and design

The background to the design of the annotation task we describe here can be traced back to research corpus-based and computational lexical semantics. Word Sense Disambiguation (WSD) is a research area within computational semantics which aims to develop systems for automatically assigning the correct sense to each word occurrence in a text. Early WSD research relied on a fixed inventory of senses and a unique correct sense for each word usage instance in text (Navigli 2009; Weaver 1949). This led to very influential lexical semantic annotation projects such as SemCor (Langone et al. 2004) and OntoNotes (Hovy et al. 2006). As discrete approaches to WSD have been shown to be problematic since Kilgarriff (1997), graded approaches have gradually gained popularity, allowing for more than one sense to be explicitly associated with a given textual instance of a word usage (Erk et al. 2009, 2013; McCarthy and Navigli 2009). The semantic web community has also studied this phenomenon in the context of knowledge representation models (e.g. Stavropoulos et al. 2016; Wang et al. 2011).

Connected with WSD research, and largely sharing distributional semantic methods with it, over the past decade computational linguistics research into the automatic detection of lexical semantic change has made significant advances.[12] A variety of different systems have been proposed, which rely on various approaches, including topic models, graph models and distributional semantics models. Most proposed models focus on English and on relatively recent time

12 For comprehensive surveys on this topic, see Kutuzov et al. (2018) and Tahmasebi et al. (2021).

periods, most commonly the 19th and 20th century. However, some research has also been done on ancient languages, notably ancient Greek (McGillivray et al. 2019; Perrone et al. 2019) and Latin (Eger and Mehler 2016; Ribary and McGillivray 2020; Sprugnoli et al. 2019). The shared task 1 "Unsupervised lexical semantic change detection" (https://competitions.codalab.org/competitions/20948 [accessed 3 March 2022]) ran as part of the 2020 edition of SemEval, a series of international workshops aiming to advance the current state of the art in computational semantic analysis. The aim of the task was to produce a multilingual dataset that could serve as a gold standard to assess the accuracy of different systems performing the automatic detection of lexical semantic change (Schlechtweg et al. 2020). The task produced annotations for four languages (English, German, Latin and Swedish) from diachronic corpora. The task was divided into two subtasks: a binary task and a ranking task. The binary task consisted in deciding, for a given list of lexemes, which ones lost or gained sense(s) between two pre-defined time periods. The ranking task consisted in ranking a given list of lexemes according to their degree of lexical semantic change between the same two time periods.

While the annotation of the modern languages (English, German, and Swedish) was on 19th and 20th century corpora and involved native speakers, the Latin annotation task required some specific adaptations. First, the time span covered was much larger, covering the period from the 2nd century BCE and the 21st century CE. We used LatinISE (McGillivray and Kilgarriff 2013), a large-scale Latin diachronic corpus compiled primarily from texts published in the Latin portion of the IntraText digital library (http://www.intratext.com [accessed 3 March 2022]). LatinISE texts have been semi-automatically lemmatised combining the morphological analyser of the PROIEL project (https://www.hf.uio.no/ifikk/english/research/projects/proiel/ [accessed 3 March 2022]) and Quick Latin (http://www.quicklatin.com/ [accessed 3 March 2022]). A study to measure the accuracy of the lemmatisation of LatinISE based on a sample of texts by Cicero's *De Officiis* and Rutilius Taurus Aemilianus Palladius' *Opus agriculturae* against the PROIEL treebank as a gold standard showed an accuracy of 92.77 and 80.96%, respectively. The most frequent lemmas in the corpus were further corrected manually. The corpus was also part-of-speech tagged with TreeTagger (https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [accessed 3 March 2022]; see Schmid (2003 [1997]), trained on the *Index Thomisticus* Treebank (Passarotti 2019), the Latin Dependency Treebank (Bamman and Crane 2007) and the Latin treebank of the PROIEL Project (Haug and Jøhndal 2008). For each text, LatinISE displays metadata related to the author, title,

time (century and date of composition, if known), as well as a basic genre classification of the texts into prose and poetry.

The choice of the set of lexemes for the annotation was based on an initial process of lexical selection and pre-annotation, carried out by a team member. A first list of terms was elaborated from Clackson (2011a) and Clackson and Horrocks (2007), selecting lexical units in which a change of meaning is observed in relation to Christianity. Other lexicographic sources were also used, such as the *Oxford Latin Dictionary* (Glare 1997 [1982]) and Lewis and Short's *Latin Dictionary* (Lewis and Short 1879). We limited the analysis to those terms that underwent lexical semantic change in relation with the changes more broadly associated with the late antiquity period, including the spread of Christianity in the Roman world.

For each lexeme, we identified the macro-sense(s) that the literature has associated with late antiquity. For example, 'blessed' for *beatus* (vs. the classical sense 'happy') or 'imperial court' for *cohors* (vs. 'cohort'). The pre-annotation verified whether the corpus displayed evidence of both the late antiquity senses and the previous senses and whether, indeed, the late antiquity sense appeared in the later texts only and the classical senses in the earlier texts (though they may also have occurred in later texts). Once the list of potential lexical units was selected, we searched each lemma in the LatinISE corpus using the Sketch Engine query tool (https://www.sketchengine.eu/ [accessed 3 March 2022]). We conducted an analysis of a selection of all the corpus concordance lines of each lemma. To keep the task manageable, we limited the analysis to the texts dated between the 3rd century BCE and the 9th century CE.

For example, for the lemma *beatus* 1,230 concordance lines were selected, extracted from texts dated between the 1st century BCE and the 8th century CE. Between eight and ten examples per century were analysed, spanning different literary genres and authors, for a total of 138 instances. Twelve usages of this word in classical texts were identified as displaying the senses 'happy' (11) or 'fortunate' (1). Among the 125 usages in texts dated from late antiquity, 65 displayed the sense 'happy', 35 the "new" sense 'blessed', 15 'fortunate', 8 'rich', and 2 'rewarded'. Therefore, this lemma was retained for the full annotation.

The final set of words selected is shown in Table 1. This list contains 17 so-called changed words and 23 stable words. "Changed words" were selected in the pre-annotation phase because they are known in the literature as having undergone lexical semantic change associated to the late antiquity period, while "stable words" were chosen because they are not known as having undergone lexical semantic change associated to the period of late antiquity. The nature and type of this change can, of course, vary, as described in Section 3.

Table 1: List of lemmas selected for the annotation.

Lemma	Type	PoS	Senses	Frequency
acerbus	stable	ADJ	harsh to the taste, harsh, unripe; (of people), rough, violent; (of things) heavy, sad, bitter	488
adsumo	stable	V	take to oneself; receive	294
ancilla	stable	N	maidservant; someone servilely devoted to anything	684
beatus	changed	ADJ	happy → blessed	3,193
ciuitas	changed	N	citizenship → city	8,693
cohors	changed	N	cohort → imperial court	1,243
consilium	stable	N	determination; judgement; council	6,602
consul	changed	N	consul → municipal official	6,389
credo	changed	V	to lend; to commit or consign something to one; to trust to or confide in a person or thing; to trust; to believe; to think, to suppose → to believe in God	9,876
dolus	changed	N	deceit → pain	762
dubius	stable	ADJ	fluctuating; uncertain; dangerous	2,437
dux	changed	N	leader → duke	6,043
fidelis	changed	ADJ	faithful → Christian	2,911
honor	stable	N	honour, repute; official dignity, office, post; honorary gift; magistrate, office-holder	5,949
hostis	stable	N	stranger; enemy	5,065
humanitas	changed	N	humanity → benevolence	672
imperator	changed	N	general → emperor	11,353
itero	changed	V	to repeat → to journey	282
ius	stable	N	broth; juice; right, justice, duty; court of justice; Justice; authority	7,810
licet	stable	V	it is permitted; it is possible; though; yes	6,891
necessarius	stable	ADJ	unavoidable; connected	3,325
nepos	stable	N	grandson; nephew; a favourite; a spendthrift, prodigal	1,260
nobilitas	stable	N	celebrity, fame; noble birth; the nobles	767
oportet	stable	V	it is necessary; it is proper/it is becoming	4,711
poena	stable	N	punishment; pain	4,191
pontifex	changed	N	one of the college of priests having supreme control in matters of public religion in Rome → bishop	1,601
potestas	changed	N	power → angels	4,687
regnum	stable	N	kingship, royalty; dominion, sovereignty, rule, authority; a kingdom; a territory, estate, possession	8,019
sacramentum	changed	N	civil suit; military oath; oath; secret; mystery → a sacrament	1,707

Table 1: (continued)

Lemma	Type	PoS	Senses	Frequency
salus	stable	N	health, welfare; greeting, salute, salutation; Salvation, deliverance from sin and its penalties	4,224
sanctus	changed	ADJ and N	sacred → saint	10,452
sapientia	stable	N	good sense, discernment; wisdom; practical wisdom; philosophy	2,967
scriptura	changed	N	writing → Holy Scripture	1,570
senatus	stable	N	senate; council	4,747
sensus	stable	N	perception; feeling; reasoning; opinion; moral sense; idea, notion; sentence	3,938
simplex	stable	ADJ	simple; frank, straightforward	2,275
templum	stable	N	a space marked out; a consecrated or sacred place; a small timber	3,569
titulus	stable	N	a superscription, inscription; title of honour; repute; pretext; title of a book	1,523
uirtus	changed	N	manliness → Christian virtues	7,005
uoluntas	stable	N	desire; meaning	3,717

Each lemma (first column) is either a "changed" or "stable" word, as indicated in the second column. The third column contains the part of speech and the last column shows the frequency in the LatinISE corpus. In the fourth column we listed the senses of the lemmas as a list. For changed words, we indicated the senses acquired in the late antiquity era following the arrow signs (for example 'saint' as the later sense of *sanctus)*, although this is not to be intended as indicating that the previous senses were replaced or disappeared.

For each lemma in the list, we extracted the primary senses from the Latin portion of *Logeion Online Dictionary*. Specifically, we used the content of Lewis and Short (1879), and Lewis (1890), complemented with Du Cange et al. (1883–1887 [1678]). In some cases, we simplified the sense inventory or shortened the text of the definitions while keeping the major distinction between senses. It is important to underline that the selection and the description of the senses of a word, which mainly depend on previous compilators' work and do not conform to a previously agreed-upon standard, significatively affects the annotation task. For each lemma we also randomly extracted 60 text snippets. Using the existing subcorpora in LatinISE, i.e. the BCE portion of LatinISE (consisting of 1.7 million word tokens) and the CE portion of the corpus (9.4 million word tokens), we extracted 30 snippets from each subcorpus. The annotation task consisted in making a judgement regarding each text snippet and each dictionary sense for each lemma,

following a variation of the DuRel annotation framework (Schlechtweg et al. 2018). The judgements were done on a four-point scale: "1" indicated that the usage of the lemma in the text was *unrelated* to the dictionary sense, "2" indicated a *distant relation* between the two, "3" a *close relation* and "4" was used to indicate that the usage in the text *completely overlapped* with the dictionary sense. The label "0" was used when the annotator was not able to make a decision. This graded approach to sense annotation leads to a much more nuanced description of the lexical semantics of the lemmas, as it allows for more than one sense to be associated (possibly to different degrees) with the same corpus instance. The annotators were not given access to the metadata information (including the datings) about each text to avoid any bias in their judgements.

Table 2 shows an example of two usages for *beatus*. The first usage is from a mediaeval text, *De libero arbitrio* by Robertus Grossetest, dated from the 12th to the13th century CE, while the second text is from a classical text, Cicero's *Tusculanae disputationes* (46 BCE). The senses presented to the annotators were: (a) 'blessed', (b) 'rich', (c) 'fortunate', (d) 'happy' and (e) 'rewarded'. The first usage was annotated as (a) 4, (b) 1, (c) 3, (d) 3, (e) 2 and the second as (a) 1, (b) 1, (c) 3, (d) 3, (e) 2.

Table 2: Two annotated usages for *beatus*.

Left context	word	Right context
Probat autem Augustinus in libro Unde malum; Deum fecisse mundum ex nihilo eo quod sibi sufficiens est, et ideo non sit ab aliqua creatura adiutus, sic dicens: "Nec quisquam de Deo optime existimat, qui non eum omnipotentem credit, rectorem quoque iustissimum omnium, quae creauit, nec ulla adiutum esse natura in creando, quasi qui sibi non sufficeret. Ex quo fit, ut de nihilo creauerit omnia." Eadem itaque ratione solus facit ominia, nulla adiutus natura. Horum autem obiectorum solutio haberi potest ut uidetur ex uerbis	beati	Bernardi sic dicentis: "Ipsa gratia Liberum arbitrium excitat, cum seminat cogitatum. Sanat, cum mutat affectum; roborat, ut perducat ad actum; seruat, ne sentiat defectum." Sic autem ista cum libero arbitrio operatur, ut tantutm in primo illud praeueniat in ceterisque comitetur: ad hoc utique praeueniens, ut iam sibi in ceteris cooperetur, ita tamen, quod a sola gratia coeptum sit pariter ab utroque perficitur, ut mixtim, non singillatim, simul, non uicissim per singulos profectus operentur, non partim gratia partim liberum arbitrium
Quid? ad recte honeste laudabiliter, postremo ad bene uiuendum satisne est praesidi in uirtute?- Certe satis.- Potes igitur aut, qui male uiuat, non eum miserum dicere aut, quem bene fateare, eum negare beate uiuere?- Quidni possim? nam etiam in tormentis recte honeste laudabiliter et ob eam	beata	uita, quaeso, relinquitur extra ostium limenque carceris, cum constantia grauitas fortitudo sapientia reliquaeque uirtutes rapiantur ad tortorem nullumque recusent nec supplicium nec dolorem?- Tu, si quid es facturus, noua aliqua conquiras oportet; ista me minime mouent, non

Table 2: (continued)

Left context	word	Right context
rem bene uiui potest, dum modo intellegas, quid nunc dicam "bene." Dico enim constanter grauiter sapienter fortiter. Haec etiam in eculeum coiciuntur, quo uita non adspirat beata.- Quid igitur? solane		solum quia peruulgata sunt, sed multo magis, quia, tamquam leuia quadam uina nihil ualent in aqua, sic Stoicorum ista magis gustata quam potata delectant. Velut iste chorus uirtutum in eculeum impositus imagines constituit ante oculos cum amplissima dignitate, ut ad eas cursim perrectura nec eas beata uita

The first column shows the left context of the usage, the second column the word form of *beatus*, the third column the right context.

We recruited eight annotators with a strong knowledge of Latin, ranging from undergraduate students to PhD students, post-doctoral researchers, and more senior researchers. Given the labour-intensive nature of the task, each word was assigned to only one annotator. *Virtus* was selected for the purposes of calculating the inter-annotator agreement between four annotators. The average pairwise agreement calculated as Spearman correlation coefficient was 0.69. This value is comparable with the inter-annotator agreement for the modern languages used in the task, i.e. 0.57 for Swedish, 0.59 for German and 0.69 for English (Schlechtweg et al. 2020).

5 Challenges of the annotation task and recommendations

In this section we examine the challenges faced by the annotators and make a series of recommendations that will hopefully help future similar annotation projects for ancient languages.

5.1 Annotators' feedback on the task

The task presented annotators with several challenges, which can be categorised into two main groups: background challenges, mainly originating from the nature of the task, and word or annotator-specific challenges, arising from the features of the word or context in question and the annotator's disposition. The main

background challenge was the need to annotate word meanings in a language that is not one's mother tongue. This leads to more caution in annotating and less reliance on language intuition, which is strengthened by the traditional methods of learning Latin, as we discussed in Section 2.

To assess the difficulties that should be considered when working with the annotation data, annotators were asked to fill in a survey regarding their experiences. The survey contained five open-ended questions, asking the annotators to (i) describe their annotation process; (ii) identify the most significant challenges they faced; (iii) comment on the relative difficulty of annotating different words ("which word(s) did you find most difficult and why?"); (iv) explain their interpretation of the scale 0–4; and (v) comment on the quality, number, and scope of the definitions used. The annotators' answers and the annotation data associated with them were anonymised. In the following, different annotators are designated by codes from A1 to A7.

Based on the feedback from the annotators, we can distinguish several types of issues they faced: (i) working with post-classical Latin; (ii) problems with the dictionary definitions; (iii) the difficulty of being consistent.

In the next sections, we give an overview of the challenges in each group, how the annotators said they had dealt with them, and what implications this might have for the resulting data.

5.1.1 Post-Classical Latin

Although the dates of the texts were unknown to the annotators, most of them noticed the presence of post-Classical Latin contexts in the annotation tasks. Some of the annotators felt less sure about assessing word meaning in those contexts due to the lack of experience with this period of Latin literature. On the other hand, it seems that estimating that a text was late felt easier and could then influence the decision about the meaning of the word.[13] This was notably less difficult in the cases where a word had a specific Christian or post-classical meaning, referring to a new historical concept, for example *sacramentum*: 'military oath' vs. 'mystery'; *pontifex*: 'high priest' vs. 'bishop'.[14]

13 "…it was notably easy to decide if a sentence was late or not…" [A3]; "I often felt that I'm inferring more from the context that I actually am able to base my decision on (such as 'this sounds like a Christian text, so it's more likely to be a Christian meaning')" [A2].

14 "…there were often definitions which only applied in certain situations (e.g. *pontifex* as a Christian bishop rather than a pagan priest), and it was easy to exclude these meanings when they could not apply" [A3].

5.1.2 Definitions

As has already been shown in Section 2, the alignment of the meaning of the word in a context with the given set of dictionary definitions is the standard way of working with a text for most people who study Classics. In the context of the annotation task, this yielded some difficulties concerning the suitability of the set of definitions.

When there was no definition suitable for a given context, the annotators either resorted to the next best definition or, more rarely, suggested a change in the set of definitions. This usually meant splitting the given list of meanings contained in one definition into two. The possibility of there being more than one, or no suitable definition was also flagged in the feedback.[15] This manifested itself in the following scenarios: (a) the definitions are too fine-grained, more than one suits the given context; (b) it is unclear whether the author wanted to use one or the other meaning of the word and reading more context does not help; (c) the definitions evolved from one another, and one cannot decide on which stage of semantic development the given context should stand. For many, the solution to the problem of multiple suitable definitions was to use the intermediate values on the scale ("2" or "3").

Some of the annotators reflected on the suitability of the approach in general: (i) "*uirtus*, by contrast, had a range of meanings which often partially overlapped, many of which the author seemed to want to be present – in these cases, the definitions seemed slightly more arbitrary with a native reader naturally understanding the various connotations that together form the word, while this task often focused on dividing the meanings more forcefully" (A3); (ii) "some of the definitions are based on our modern interpretation of what 'a feeling' is, whereas I wasn't sure it would be the same for an author writing in Latin" (A2).

5.1.3 Scale 0–4 and consistency

The interpretation of the annotation scale was given at the beginning of the task. The feedback and the annotation data suggest that most annotators tried to avoid answering "0" ("Cannot decide") and they attempted to be as consistent as possible. The values "3" and "2" were used in cases of doubt, where more than one definition seemed to be possible. The annotators could then indicate their preference by marking the most likely meaning with a "4" or leave all of them at "3". It is worth noting that the value "2" ("distantly related") was interpreted differently

15 "the suggested options did not appear always as fitting or enough fine-grained" [A7]; "one meaning sometimes blurs into another one and it is not that easy to separate the two" [A5].

by different annotators. Compare the following statements from the annotators' survey: (i) "…it was hard to establish the difference between '2' and '3'" (A6); (ii) "'2' and '3' offered a range of meaning, which I used for the more abstract passages where the connotations of several definitions seemed to be entailed." (A3); (iii) "I interpreted '3' as being semantically related, and […] where it was possible, if not ideal, to use the meaning in question in a translation. '2' meant 'not unrelated, but by no means close'." (A1).

While annotator A6 reports difficulties with deciding between "2" and "3", annotator A1's answer seems to indicate a much larger gap in their assessment of the two values. This part of the feedback is crucial for assessing the data that have been collected, because it shows that the middle values, and "2" in particular, were not interpreted in the same way by all annotators.

Of all ratings that reflect a certain degree of suitability for a sense ("2", "3" and "4"), "2" was used in 16.84% of the cases, "3" in 28.72% of the cases, and "4" in 54.44% of the cases (Figure 2). Thus, it is important to keep in mind that different interpretations and approaches of the annotators could have led to inconsequent usage of the rating "2" in the data.

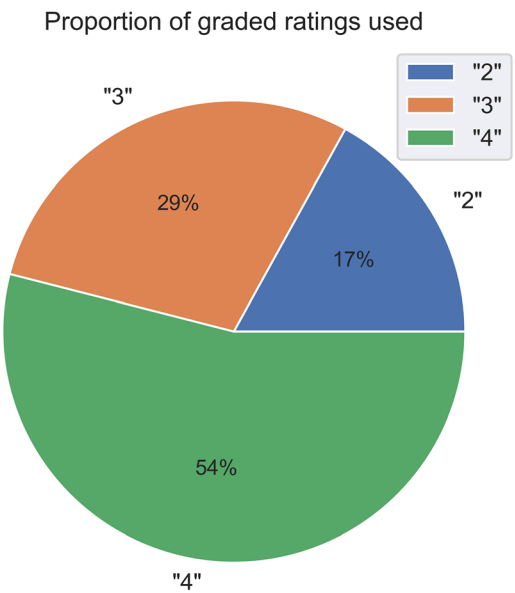

Figure 2: Proportion of graded ("2", "3" and "4") ratings used across the annotation task.

To compare annotations of words with different numbers of senses, we reduced the annotation patterns to sets, counting only unique occurrences of each rating. For example, annotations "1–1–1–4" and "1–4" are recorded as a set {1, 4}, and "2–2–1–3–4" and "1–2–4–3", as {1, 2, 3, 4}. Figure 3 shows the frequency of all sets accounting, cumulatively, for 97.22% of all annotations. The sets containing zero ratings fell below the threshold of 1% and are thus absent.

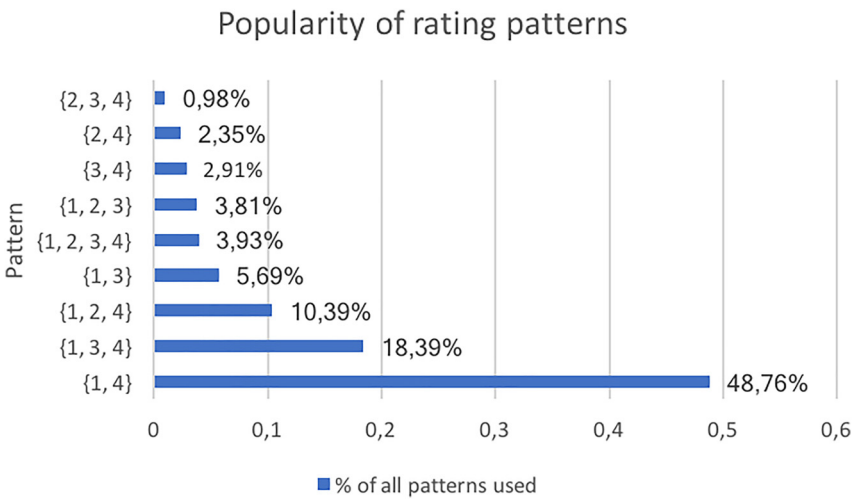

Figure 3: Top nine standardised patterns of annotation (sets), sorted by popularity (97.22% of all patterns).

Overall, all annotators assigned at least one "4" in most cases. The first two most popular sets, {1, 3, 4} and {1, 4}, amount to 67.15% of all sets used (2,338 in total), and neither uses "2" at all. The remaining 18 sets not in Figure 3 amount to only 2.78% of all annotated contexts, which makes them negligible for a large-scale analysis, but shows that the annotators made full use of the flexibility offered by the annotation scheme.

5.2 Recommendations on annotation

The 0–4 annotation system proved easy to use for the annotators, as it implied a simple ranking approach. The difficulties in analysing the data stem from different strategies that the annotators employed. For example, if an annotator was unsure which of the two meanings was more suitable for the given context, they could

either annotate it as "3–3", "4–4", or "3–4", introducing a scale that reflects the relative suitability of the meanings to each other, and not the absolute suitability of each of them for the context.

A further challenge presented by the annotation scheme was that it does not discriminate between the confidence of the annotator about a certain meaning and the relevance of the meaning for the set context. This could be mitigated by the introduction of an additional column, in which the annotators could assess their confidence or flag a context that proved particularly challenging. This was done to an extent by the "comments" column in the annotation task, but a more formalised and quantifiable approach to measuring confidence could prove easier for further analysis.

Considering the analysis of the data and the annotators' feedback, we suggest the following recommendations for further research that involves annotation of ancient languages such as Latin or ancient Greek by specialist annotators, using the approach followed in our study:

1. The researchers should be aware of the challenges that result from an annotation mainly based on bilingual dictionary entries.
2. The scale 0–4 provides necessary granularity and is easy to implement. It would benefit from additional comments in the instructions describing possible situations and how the annotator is expected to deal with them, in order to avoid conflicting interpretations.
3. Measures should be taken to distinguish between the personal confidence of the annotator and the suitability of a meaning to the given context, in the form of a better explanation of the scale and possibly the introduction of a formalised way to flag concerns.

5.3 Analysis of annotation confidence

The setup of the annotation task meant that each word was only annotated by one person, apart from *uirtus*. To account for the potential differences between the individual annotators, we conducted a quantitative analysis of the annotated data. The objectives of the analysis were: (i) to work with data from individual annotators to find out whether there is a personal *style* of annotation that would affect further analysis of the data; and (ii) to look for features of the words themselves that could influence the annotators' decisions. This section will briefly outline the strategy used and its most significant results. A more extensive description is available in the "confidence analysis" folder of the documentation (McGillivray et al. 2022).

We define *confidence score of a context* as the proportion of "4" ratings (n_4) out of all non-zero and non-1 values (n_2, n_3, n_4), expressed as a number in the range

from 0 to 1. This value reflects the level of uncertainty involved in the annotation of a given context.[16] We take the average of confidence scores of all contexts for a given word as the confidence score of the word.

$$confidence = \frac{n_4}{n_2 + n_3 + n_4}$$

To estimate the differences in annotation styles, we analysed the data provided by the annotation of the test word *uirtus* with the data from the rest of the annotation task (Figure 4). The confidence of individual annotators in the test task ranges from 0.27 to 0.47 with the mean value of 0.35. The same confidence in the rest of the annotation task ranges from 0.34 to 0.87 with the mean value of 0.56. There is an increase in average confidence for all annotators in the rest of the annotation compared to the test task, probably due to the fact that the test task served as a

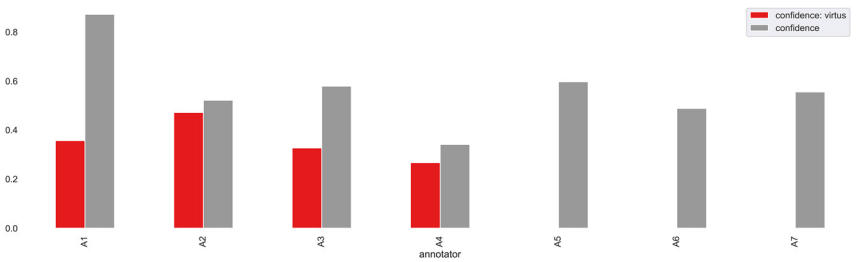

Figure 4: Comparison of annotator confidence between test task (*uirtus*) and annotation task. Annotators A5, A6, and A7 did not complete the test task.

training task, after which the annotators became more confident in the rest of the annotation, combined with the difficulty of the word *uirtus*.

Upon inspection of the data for single words, annotators' confidence showed considerable fluctuation.

We thus looked for factors that could explain the differences in the annotation confidence. A major source of influence that transpired was the number of senses of the annotated word, since the likelyhood of "2"s and "3"s increases with it. This influence can be seen from Figure 5.

16 The reason behind excluding the "1" ratings lies in their prevalence as the number of meanings grows: for a word with four meanings, a context annotated with the highest confidence will yield "1–1–1–4", with the proportion on "1" ratings being 75%, whereas for a word with two meanings, in the annotation "1–4" the proportion of "1" ratings is 50%, in spite of the fact that both cases display a single closely related sense, with all other senses assessed as distantly related. Using our definition, the *confidence score* for both annotations, however, would equal 1.0, as we would expect.

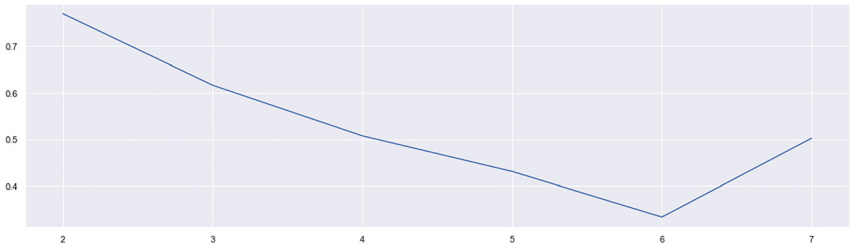

Figure 5: Average confidence of words' annotation by number of their senses.

Apart from the two words that have seven meanings, *the average confidence drops steadily as the number of senses of the annotated word increases*. This suggests that the average confidence of an annotator might be connected with the nature of the words they were allocated. So, annotator A1, whose confidence more than doubled in the annotation task compared to the test task, only had words with two or three senses assigned to them. To account for this, we weighted the confidence of each word by the number of its senses and produced new, weighted data for the annotators' confidence. The average confidence of a word with three senses

Table 3: Weighting coefficients for words with 2, 3, 4, 5, and 7 senses.

Number of senses (*n*)	Average confidence for *n* senses	Weighting coefficient (*k*)
2	0.77	0.80
3	0.62	1.00
4	0.51	1.21
5	0.43	1.43
6	0.56	1.09
7	0.50	1.23

was taken as a baseline, and its weighting coefficient was set as 1. Weighting coefficients can be seen in Table 3.[17]

17 For the calculation of weighting coefficients, see the folder "confidence analysis" in McGillivray et al. 2022. In measuring the confidence for a word with six senses, the outlier *ius* has been taken out of the consideration in order not to skew the picture. Upon inspection, it turned out that the annotation of *ius* had a large number of contexts where no "4" has been given, which led to an overall low confidence measure for the word. This, in turn, influenced the average confidence measure for all words with six senses, as there were only three of them. The average confidence of the remaining two words is approximately 0.56. After calculating the coefficient based on this value and applying it to the remaining words with six senses, weighted confidence scores

Table 4 shows an example of weighting. *Acerbus*, a word with three senses, does not change, while the confidence score of *senatus* (two senses) drops and the confidence score of *uoluntas* (five senses) rises. The confidence score of *sacramentum* (six senses) also rises slightly.

After weighting the confidence score of each word, we used the new data to calculate the confidence scores of the annotators. Table 5 provides statistics and confidence scores for each of the annotators A1–A7 before and after weighting, including the original and weighted data for *uirtus* for those who annotated it.

Table 4: Example of weighting the confidence score of the word by taking into account the number of its senses.

Annotator key	Word	Number of senses	Confidence	*k*	Weighted confidence
A1	*acerbus*	3	0.77	1.00	0.77
A3	*senatus*	2	0.97	0.80	0.77
A7	*uoluntas*	5	0.57	1.43	0.81
A2	*sacramentum*	6	0.51	1.09	0.56

Table 5: Overview of the annotators' work and their confidence scores. Annotators A5, A6 and A7 did not do the test task (*uirtus*), hence the presence of "n/a" (not applicable) in some of the cells.

Annotator	Words annotated	Cells in total	Confidence: *uirtus*	Weighted confidence: *uirtus*	Confidence	Weighted confidence
A1	4	600	0.36	0.39	0.87	0.79
A2	14	3,472	0.47	0.52	0.52	0.61
A3	11	2,100	0.33	0.36	0.58	0.67
A4	4	960	0.27	0.29	0.34	0.52
A5	2	780	n/a	n/a	0.60	0.69
A6	1	180	n/a	n/a	0.49	0.49
A7	3	660	n/a	n/a	0.56	0.64

The analysis of the annotators' confidence showed that it is possible to analyse the data resulting from the annotation task to look for linguistic features influencing it. While the personal differences between annotators are impossible to eliminate fully, the resulting confidence scores were seen to be mostly

amounted to 0.56 for *sacramentum* and 0.71 for *potestas*, putting them into the expected confidence bracket.

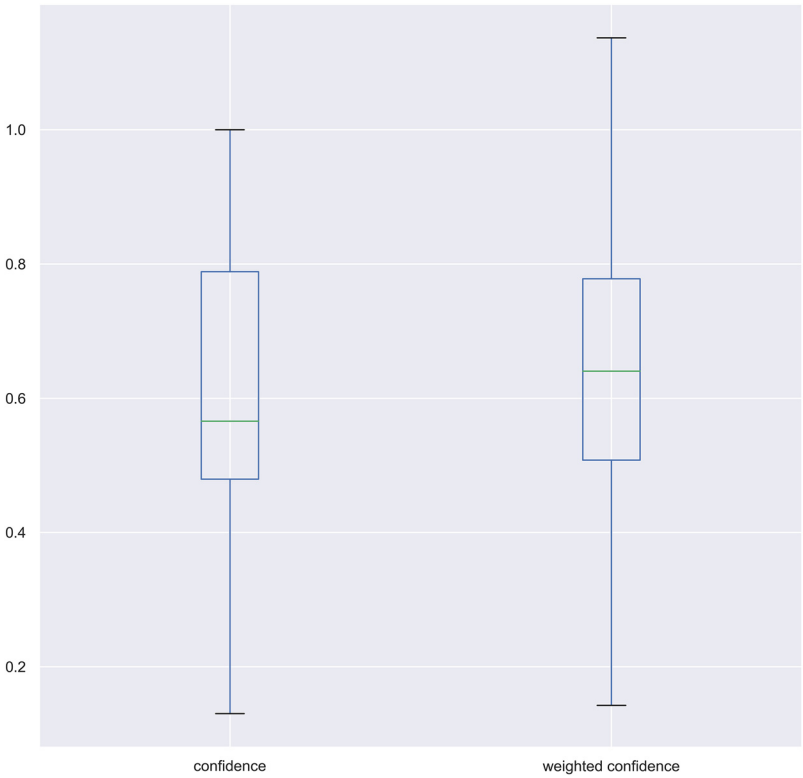

Figure 6: Distribution of confidence scores before and after applying the weighting coefficient. Two 4-meaning words with an exceptionally high non-weighted confidence (*consul* and *cohors*) are responsible for the cases of weighted confidence > 1.

influenced by the qualities of the annotated words and not by individual 'annotation styles'.

Weighting by the number of senses also led to a more consistent picture of the confidence scores across all words (Figure 6). After weighting, most words fell into the bracket of confidence between 0.5 and 0.8, with some outliers remaining.

It is now worthwhile to look at the words behind this scatter plot (Figure 7). As was expected, not all differences have been eliminated by weighting. For example, annotator A4 shows a dramatic difference in confidence while annotating two words with a similar task structure: their annotations of *cohors* and *regnum* (four senses each) have confidence values of 0.9 and 0.27, respectively.

The analysis has shown that the most important feature that influenced the confidence of the annotation was the number of senses of the annotated word. This is true for words with two, three, four or five senses. The data for words with six

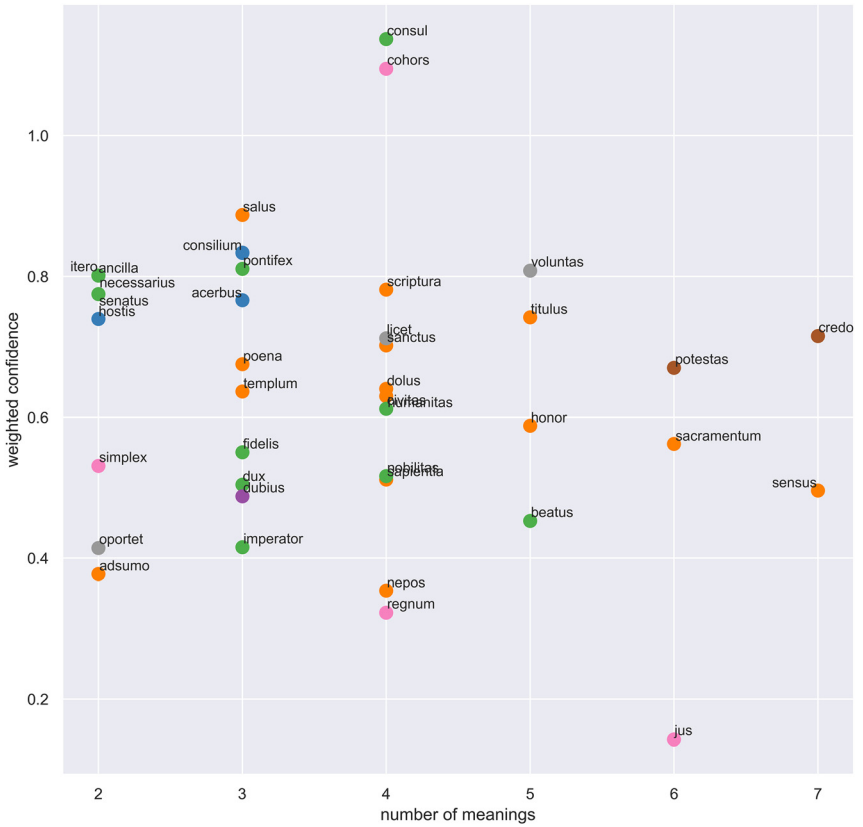

Figure 7: Weighted confidence for all words from the annotation task, arranged by the number of meanings. Words annotated by the same person are marked with the same colour.

(three words) and seven (two words) meanings is not sufficient to make general claims. Moreover, it cannot be excluded that the annotation strategy changed as the meanings were more granular: the absolute number of different annotation patterns peaks for words with four meanings but remains stable or even drops in words with five, six and seven meanings (Figure 8).

The analysis of annotation confidence summarised here shows that the discrepancies in the confidence scores between different annotators can to a large extent be attributed to the nature of the words they annotated, and thus supports the presented methodology as a way to acquire knowledge about Latin semantics. The relationship between the number of senses of the word and its confidence score has been identified as an important factor influencing the confidence score.

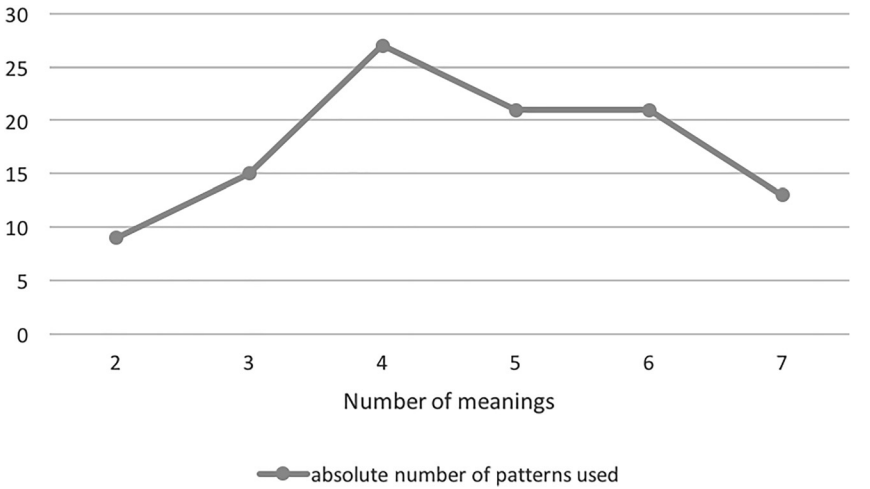

Figure 8: Number of annotation patterns used for a word with *N* meanings.

When the differences between individual words and sets of words given to a single annotator are accounted for, the remaining discrepancies should be attributed either to the inherent semantic qualities of the given word or to an idiosyncrasy of a specific annotator in a specific case, which can never be excluded altogether. The results of this analysis have to be taken into consideration during the inspection of individual words, which will be presented in the next section.

6 Lexical semantic analysis

In this section we report on the preliminary results of a series of diachronic analyses of the annotation data.

6.1 Diachronic analysis

In order to analyse the annotation data diachronically, we arranged the annotation contexts by extracting composition dates from the metadata. The dates reflect the century in which the passage has been written and range from *–2.0* for 2nd century BCE to *20.0* for 20th century CE. Where metadata gave a range of centuries, the middle point was taken; so, "2–3 cent. CE" was registered as 2.5.

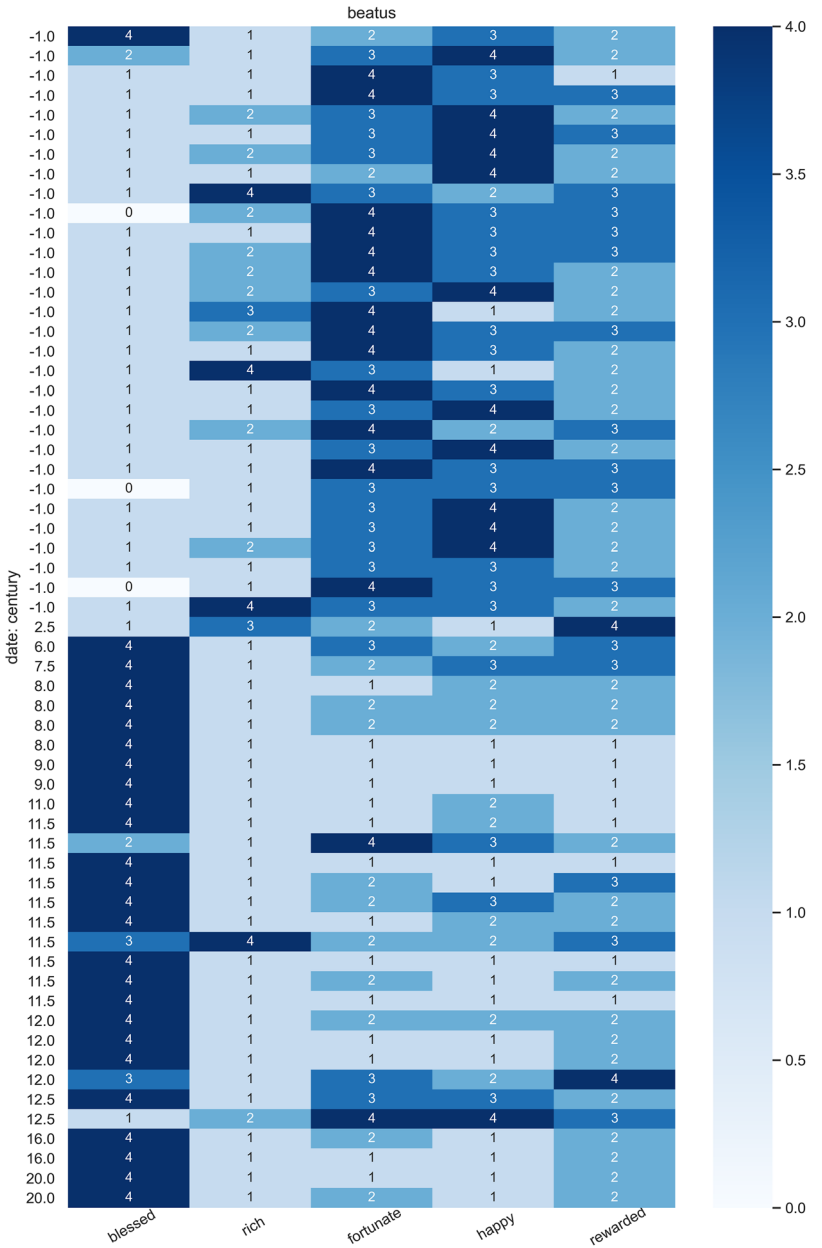

Figure 9: Annotation of all contexts for the word *beatus*, sorted chronologically by century (rows). The columns correspond to the senses 'blessed', 'rich', 'fortunate', 'happy', and 'rewarded'.

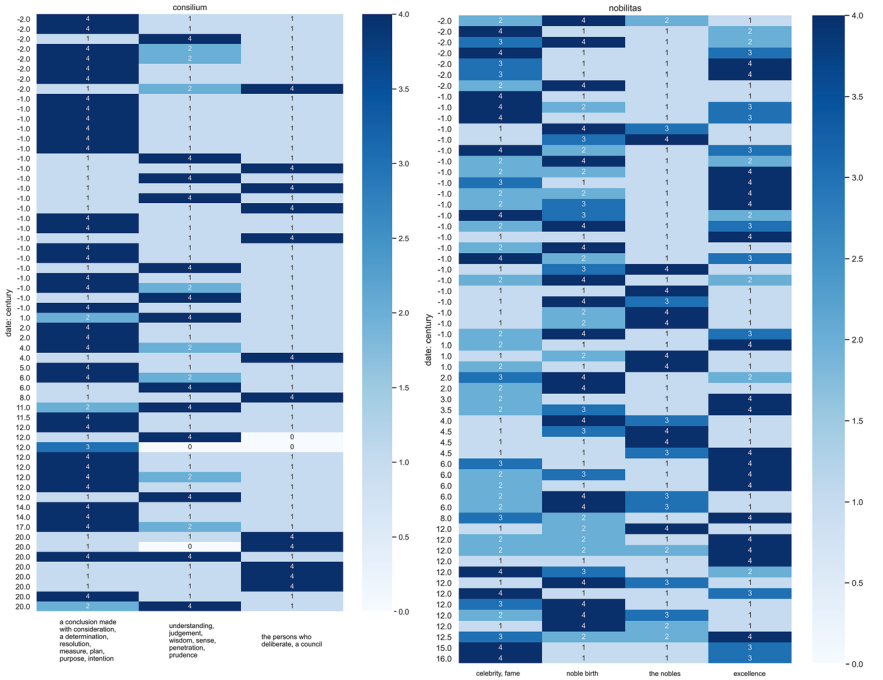

Figure 10: Annotations of the words *consilium* and *nobilitas* exhibiting different levels of confidence, as shown by the prevalence of "1" and "4" ratings for *consilium* (left) and the higher presence of "2" and "3" ratings for *nobilitas*.

We visualised the annotated data for each word in the form of a heatmap, with the most secure annotation ("4") coloured in dark blue and the least secure ("0") coloured in white, as in Figure 9. All heatmaps produced for this paper can be seen in the folder "heatmaps" in McGillivray et al. (2022).

The heatmaps provide an intuitive overview of the annotation structure and patterns. For example, the Christian meaning 'blessed' of *beatus* (first column in Figure 9) takes hold in post-Classical texts (starting from 6th cent. CE) and dominates over the other senses.

The heatmaps also show the level of confidence with which annotators marked different meanings (Figure 10). For example, *nobilitas* has a high number of "2" and "3" ratings along with "1" and "4" ratings, whereas *consilium* shows a higher prevalence of "1" and "4" ratings.

A higher presence of "2" and "3" ratings indicates a difficulty in disambiguation of multiple senses of the word. This, in turn, can illustrate the coexistence of several senses at a given time, as well as the rises and falls in their prevalence. We

therefore conducted an analysis of ratings received by different senses of a word to locate the senses that are shown to coexist in the sample.

6.2 Coexistence of meanings across time

Due to the nature of the annotation data, there is an overrepresentation of texts from the 1st century BCE along with several gaps in later centuries. In order to account for this, we calculated the average of the sense's representation in a given century out of all "2", "3" and "4" ratings it received. We then added the averages together and normalised by dividing by the sum of all the senses annotated in the given time period. The resulting data represents the coexistence of a set of meanings at a given point in time and their development. Consider the following examples:

i. *Scriptura*: the sense 'Holy Scripture' first appears in the annotated sample in the 2nd century CE and coexists with the original meaning 'writing'; the senses 'tax' and 'will' are marginal; 'tax' reappears in the 6th century, 'will' does not occur in the sample after 1st century BCE (Figure 11).

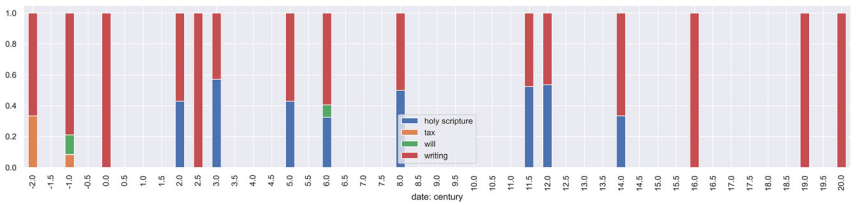

Figure 11: Diachronic distribution of the senses of *scriptura* in the annotation data.

ii. *Oportet*: both senses, 'it is necessary' and 'it is proper/it is becoming' coexist without prevailing over one another throughout the sample (Figure 12).

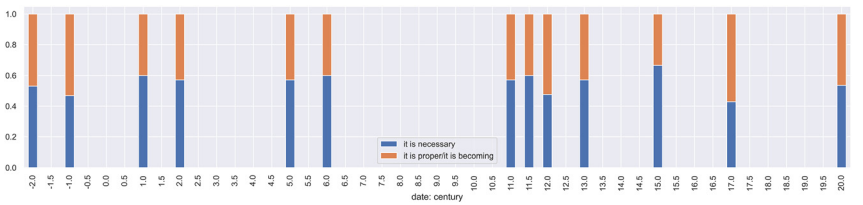

Figure 12: Diachronic distribution of the senses of *oportet* in the annotation data.

iii. *Voluntas*: two prominent senses, 'will, freewill, wish' and 'disposition towards a person or a thing' coexist without major fluctuations throughout the sample. Of the remaining three senses, two occur sporadically and one is not attested (Figure 13).

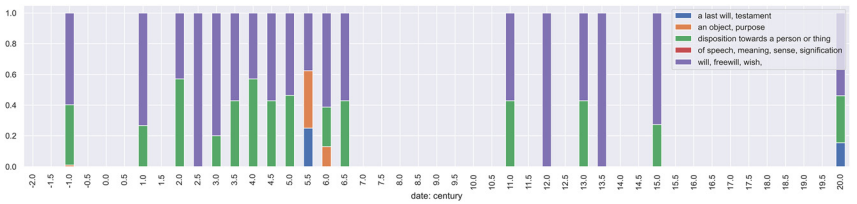

Figure 13: Diachronic distribution of the senses of *uoluntas* in the annotation data.

iv. *Pontifex*: the original meaning 'pontiff' disappears and is replaced by a Christian sense 'bishop' and an unmarked sense 'priest' (Figure 14).

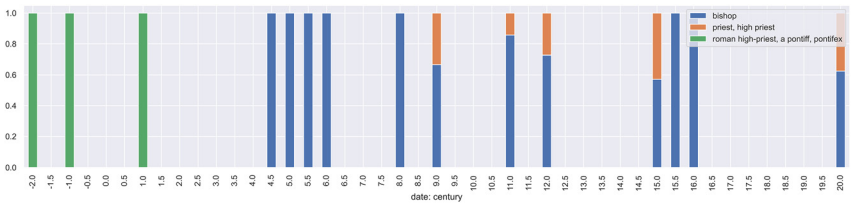

Figure 14: Diachronic distribution of the senses of *pontifex* in the annotation data.

6.3 New senses and newer texts

Additionally, we conducted an analysis to assess whether the annotated data show that the changed words underwent the diachronic development that we expected. We looked for evidence of an association between the senses of changed words that, according to the literature and to historical dictionaries, emerged with late antiquity and Christianity, and more recent texts in our corpus.

For each of the 17 changed words we identified the sense or senses that emerged in the late antiquity era according to the literature, as shown in Table 1, and which we will refer to as "emerging" here. Then, for each passage, we calculated the average of the ratings assigned to its emerging sense(s). We also calculated the average rating assigned by the annotators to the other senses (which we will refer to as "previous" here). We took the difference between these two

averages for each passage, thus obtaining a series of 60 values for each word (59 for *scriptura*). For example, the changed word *credo* has the following senses:

i. 'to give as a loan, to loan, lend';
ii. 'to commit or consign something to one';
iii. 'to trust to or confide in a person or thing, to have confidence in, to trust';
iv. 'to trust one in their declarations, to believe';
v. 'to believe a thing, hold or admit as true';
vi. 'to think, to suppose';
vii. 'to believe in God'.

Sense vii was the sense identified as emerging because it is the only one specifically linked to Christianity and, according to the dictionaries collected in *Logeion*, it is associated with ecclesiastical Latin (cf. *Logeion Online Dictionary*, s.v. *credo*). In our analysis we refer to senses i–vi as "previous". By way of example, let us consider the following passage from the 1st century BCE:

(1) *Quae prius quam perficerentur, Longinus omnem suum equitatum emisit; quem magno sibi usu fore credebat, si pabulari frumentarique Marcellum non pateretur, magno autem fore impedimento, si clausus obsidione et inutilis necessarium consumeret frumentum. (BAlex* 61)
'But before these could be completed, Longinus sent out his entire cavalry force, in the belief that it would stand him in very good stead if it stopped Marcellus from collecting fodder and corn, whereas it would prove a great handicap if, shut up by blockade and rendered useless, it used up precious corn.' (Way 1955: 109, 111)

The annotator assigned the following ratings to this passage, showing that it displays an instance of the previous sense v:

i. 1: Unrelated
ii. 1: Unrelated
iii. 1: Unrelated
iv. 1: Unrelated
v. 4: Identical
vi. 2: Distantly Related
vii. 1: Unrelated

Therefore, in this case, the average rating for the emerging sense (number vii) is 1 and the average rating for the other senses is $(1 + 1 + 1 + 1 + 4 + 2)/6 = 1.67$. The difference between these values is $1 - 1.67 = -0.67$. We proceed with this approach for each passage. When the annotators' ratings for emerging senses are higher than their ratings for previous senses this distribution has positive values, and it has

negative or zero values in the other cases. Figure 15 shows the histogram of this distribution for the word *beatus* and Figure 16 shows this distribution (*y* axis) over time (*x* axis). The plots for all words can be found in the folder "histograms" in McGillivray et al. (2022).

In Figure 15, negative values point to the cases in which this word was annotated as displaying the emerging sense and positive values point to the cases in which this was annotated as displaying the previous senses.

Inspecting plots such as that displayed in Figure 16 is useful to spot the first time point in which the distribution takes positive values. In the case of *beatus* this happens in the 6th century CE, which indicates the emergence of the Christian sense 'blessed' in our corpus.

In order to see if there was any diachronic trend in this distribution, we analysed its correlation (i.e. its strength of association) with the sequence of centuries

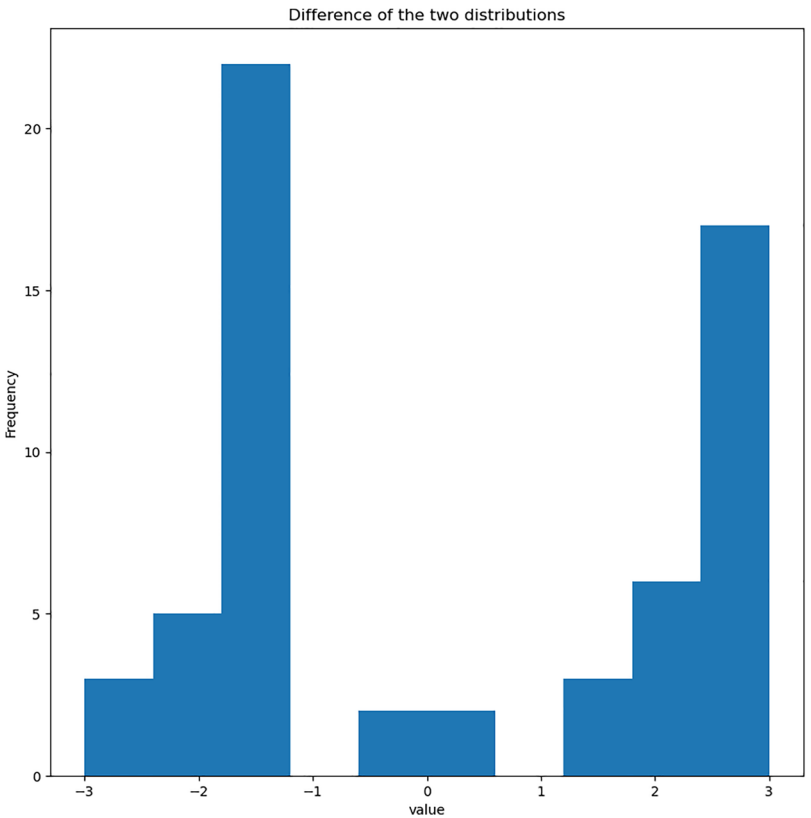

Figure 15: Histogram of the distribution of the difference between the average ratings of the emerging sense 'blessed' of *beatus* and the average ratings of all its other senses.

associated to each annotated passage via Kendall's Tau rank correlation. In the example above, the value for "century" would be –1.

Table 6 contains the results of the correlation analysis, with the values for Kendall's Tau rank coefficient and the corresponding *p*-values of the correlation test for each word. A statistically significant result occurs when the *p*-value of the test is below 0.05. From Table 6, we see that we found a statistically significant positive correlation for eight words (*beatus*, *fidelis*, *imperator*, *pontifex*, *potestas*, *sacramentum*, *sanctus*, *and scriptura*). This means that for these eight words there is a less than 5% chance of observing this correlation result if there is no underlying relationship between the two variables. We can therefore focus on the statistically significant results for our interpretation. The values of the Kendall's Tau coefficient for the eight words are all above 0.30. This means that for all these words we have found a strong correlation: as the centuries progress, the difference between the

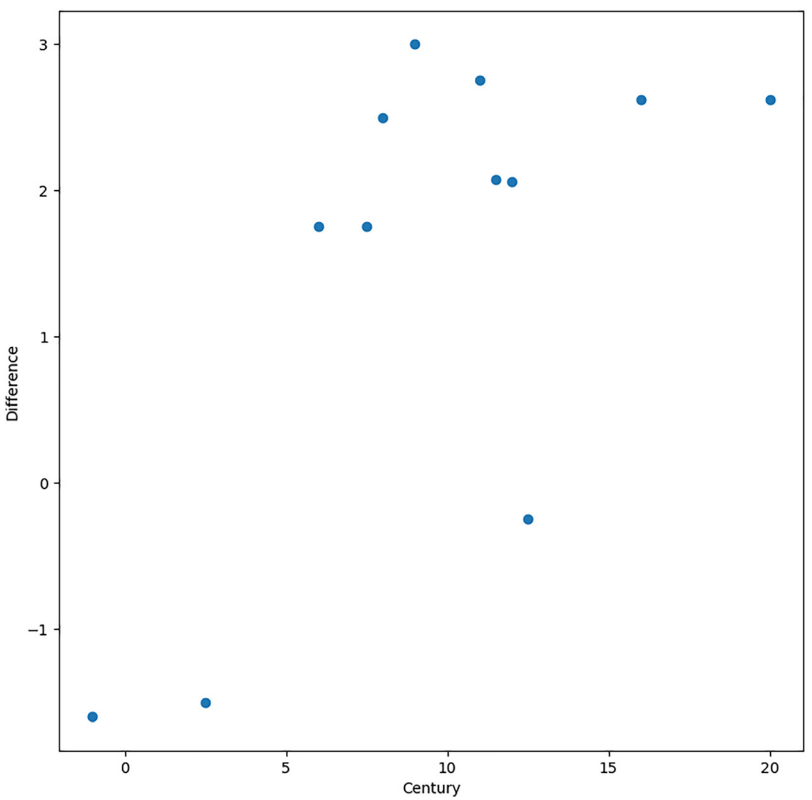

Figure 16: Distribution of the difference between the average ratings of the emerging sense 'blessed' of *beatus* and the average ratings of all its other senses.

Table 6: Kendall's rank correlation coefficient Tau with its *p*-value on the correlation of the difference between average ratings for emerging senses and average ratings for previous senses. Statistically significant coefficients at the 0.05 level are marked with an asterisk.

Changed word	Kendall's Tau	Kendall *p*-value
beatus	0.58*	<0.05
ciuitas	0.20	0.079
cohors	0.03	0.799
consul	−0.18	0.136
credo	0.03	0.813
dolus	−0.17	0.106
dux	0.02	0.899
fidelis	0.47*	0.000
humanitas	−0.25*	0.028
imperator	0.58*	0.000
itero	0.04	0.726
pontifex	0.61*	0.000
potestas	0.47*	0.000
sacramentum	0.63*	0.000
sanctus	0.31*	0.005
scriptura	0.54*	0.000
uirtus	0.20	0.061

average ratings of their emerging sense(s) and the average ratings of all the other senses also goes up. Of course, we do not expect a clear monotonic trend in the distributions, i.e. a trend which is consistently upward or downwards through time, as we know that there is likely to be a point when the distribution switches from mainly positive to mainly negative values, but that it can take negative values also later on, in correspondence to usages of the previous senses in later time periods. Moreover, words' semantic development is more complex and interacts with other factors, such as the texts' genre, or where and by whom they were produced. However, this preliminary analysis shows the potential for the annotation data to highlight patterns for further exploration.

After analysing the data of the annotation task from a quantitative point of view, it is worthwhile to look more in detail into the semantics of the lemmas as outlined by the selection of English meanings offered to the annotators. In order to do this, we tried to interpret the three possible scenarios concerning definitions described above (cf. 5.1.2 Definitions) in terms of semantic vagueness. The focus of the next section is on the possible correlations between the inherent vagueness of a lemma and the results of the annotation: is an annotation "more confident" because the meanings of a lemma are inherently more distinguishable (i.e. less

vague) than others and vice versa? How to measure how vague a word is and, on the other hand, how to measure how certain an annotation can be?

7 An assessment of vagueness

In this section we offer an analysis of the annotation patterns which is complementary to the analysis of confidence scores presented in Sections 5 and 6.

7.1 Theoretical introduction to the analysis

We start from the hypothesis that the clear identification of the presence of a meaning is based on many factors, as explained above, and that one of those factors is the inherent polysemy of a word. Such polysemy can be completely or only partially disambiguated thanks to the context or not disambiguated at all, giving rise to annotation patterns which reveal the presence of multiple fitting senses such as "4–4", "4–3", "3–4", "3–3". We analyse the annotated words in terms of *vagueness*. Vagueness is a property of any polysemic word form whereby its exact meaning remains undetermined supposedly without this being an obstacle to the communication between speaker and addressee addre or writer and reader. Consider this dialogue:

(2) A: *Where is John?*
B: *He is at his aunt's place.*

Whether the aunt in question is John's father's sister (sense 1) or John's mother's sister (sense 2) or the partner of his parent's sibling (sense 3) is irrelevant for A to be able to understand B. If a specific context requires this type of information, the speaker or writer can make it clearer by using some specifications. Example (2) shows two important aspects tied to any disambiguation task which is based on the interpretation of a word in a given linguistic context:

- communication can be successful, despite the presence of vagueness and sometimes even ambiguity (Wasow 2015), and A manages to know where John is;
- the interlocutors can have pieces of information which are not evident to people who hear (or read the transcription of) their dialogue and are not the intended addressees addr of the message. It is for instance impossible for us to know whether A knows John's aunt and where she lives so that A can try to go and see John. It is not only the context that helps to disambiguate the word, but also the dialogue's extra-linguistic context and the extra-linguistic knowledge.

This section has four goals: (i) to introduce the concepts of vagueness, polysemy and ambiguity and to present our view of the annotation task in relation to these issues; (ii) to outline a hypothesis of correlation between vagueness score and closeness between the meanings of a lemma; (iii) to present some of the most interesting cases we encountered in our analysis; (iv) to draw some conclusions on vagueness in relation to the annotation. All the 18 words we analysed can be consulted in McGillivray et al. (2022), in the file Annex.pdf in the folder "Vagueness".

7.2 Vagueness

Vagueness and ambiguity are concepts inherently entrenched with the notion of polysemy. Polysemy is the phenomenon whereby a word form is associated with more than one meaning (e.g. Matthews 2014 [1997]). According to the milestone analysis of Tuggy (1993), ambiguity and vagueness are the extremes of the continuum of polysemy. At the one end there is ambiguity: ambiguity is best represented by the case of homophonic or homographic words which are not etymologically related, such as English *bank* 'river edge' (Proto-Germanic **bankan-* 'elevation') or 'financial institution' (from Proto-Germanic **banki-* m. 'bench'),[18] English *well* (noun) 'a spring of water' (from Proto-Germanic **wallan-* 'to well up, boil, seethe') and *well* (adverb) 'in a good manner' (likely from Proto-Germanic **waljan-* 'to choose'). Very distant meanings of (etymologically) the same word can also be considered cases of ambiguity: the English noun *spring* can refer to very different entities, such as a 'source', 'one of the seasons' and a 'jump', though the three nouns are closely related from the etymological point of view (see *OED Online*, s.v. *spring, n.*). At the other end of the continuum we find vagueness. This is best represented by cases in which a word has several more or less closely related senses such as *aunt* 'father's sister' and 'mother's sister' which can be 'intuitively united into one, parent's sister' (Tuggy 1993: 273–274). From the point of view of the classification of family relations the kind of relation between a person and her father's sister is the same as the one between the same person and her mother's sister and it comes as no surprise that dictionaries can subsume such senses of *aunt* under a unique entry, cf. *OED Online*, s.v. *aunt, n.* Between the two poles of ambiguity and vagueness, there is any other kind of polysemy, sometimes closer to ambiguity, other times closer to vagueness. Tuggy suggests the case of the

18 This is the example suggested by Tuggy. However, it must be noted that the two Proto-Germanic forms are probably connected. For the etymologies of Germanic words, see Kroonen (2013).

verb *to paint* which, depending on the context, can mean an artistic activity or a utilitarian one. Unlike the case of *bank*, they do not indicate completely distinct kinds of activity, as in both cases some colour is being put on a surface. Unlike the case of *aunt*, the two meanings do not indicate a similar way of painting as you do not paint a wall in the same way as you paint a portrait.[19]

All words of the annotation task are clearly polysemous as shown by the fact that they have received more than one meaning (which in our case corresponds to a translation) and can be put in the polysemy continuum *ambiguity — vagueness*. Apart from one clear case of ambiguity (*ius*, see below), all other words sit on a continuum that goes from more or less high polysemy to vagueness. In order to better understand the work of and the problems encountered by the annotators in making their choices, we chose to focus on vagueness, intended, as explained above, as the partial overlap and/or implication of two or more meanings of a word. For example, sense 4 'territory, estate, possession' of the lemma *regnum* can be seen as implied in sense 3 'kingdom'. In fact, one of the main challenges of the annotation task was to deal with vagueness specifically. At least in principle, a word can be polysemous, but not necessarily vague: this is the case when its meanings are clearly distinguishable, ideally in any given context. In such a case an annotator should be able to easily make a choice. On the contrary, when beside polysemy, there is specifically vagueness, the annotator will encounter more difficulties in choosing.

7.3 Methodology

In order to calculate the vagueness of a word we devised the following methodology. In the word's annotation, an annotation labelled "4" (the word's usage displays the same meaning as the dictionary sense in question) followed by one or more "1"s (the word's usage is unrelated to the sense) implies that the meaning of the word in the given context was clear, i.e. that there is no vagueness (see Table 7). On the contrary, when there is more than one "4", more than one "3" or the presence of both "3" and "4", this means that the meaning of the word in the given context was not clear and that several interpretations were possible, i.e. it is a case of vagueness (see Table 8). The ideal case of absence of vagueness is an annotation in which only one meaning is annotated with a "4" or a "3" and where all the other senses are annotated with a "1" or a "2". The vagueness score aims to represent any deviation from this ideal case.

19 For another approach to ambiguity, see for instance Magni (2020).

Table 7: Example of annotation with no vagueness (lemma: *adsumo*).

Left context	word	Right context	Sense 1: take to oneself	Sense 2: receive
Haec dum in India geruntur, Graeci milites nuper in colonias a rege deducti circa Bactra orta inter ipsos seditione defecerant, non tam Alexandro infensi quam metu supplicii. Quippe, occisis quibusdam popularium, qui ualidiores erant, arma spectare coeperunt et Bactriana arce, quae casu neglegentius adseruata erat, occupata Barbaros quoque in societatem defectionis inpulerant. Athenodorus erat princeps eorum, qui regis quoque nomen	adsumpserat	, non tam imperii cupidine quam in patriam reuertendi cum iis, qui auctoritatem ipsius sequebantur. Huic Biton quidam nationis eiusdem, sed ob aemulationem infestus, conparauit insidias, inuitatumque ad epulas per Boxum quendam Bactrianum in conuiuio occidit.	4: Identical	1: Unrelated

Table 8: Example of annotation with vagueness (lemma: *adsumo*).

Left context	Word	Right context	Sense 1: 'take to oneself'	Sense 2: 'receive'
Extenuat corpus aqua calida, si quis in ea descendit, magisque si salsa est; ieiuno balineum, inurens sol ut omnis calor, cura, uigilia; somnus nimium uel breuis uel longus, per aestatem durum cubile; cursus, multa ambulatio, omnisque uehemens exercitatio; uomitus, deiectio, acidae res et austerae; et semel die	adsumptae	epulae; et uini non praefrigidi ieiuno potio in consuetudinem adducta. Cum uero inter extenuantia posuerim uomitum et deiectionem, de his quoque proprie quaedam dicenda sunt. Reiectum esse ab Asclepiade uomitum in eo uolumine, quod DE TUENDA SANITATE composuit, uideo; neque reprehendo, si offensus eorum est consuetudine, qui cotidie eiciendo uorandi facultatem moliuntur.	3: Closely Related	3: Closely Related

We set that an annotation yielding the maximum confidence score, i.e. 1.0 (see Section 5.3), will have a counterpart vagueness score of 0. It is important to underline that the confidence score and the vagueness score work on the basis of a different principle. The confidence score is lower with the presence of "2" and "3" ratings and the presence of more than one "4" rating. The vagueness score is higher when there is more than one occurrence of "4" and/or "3" ratings. The details of the calculation including an explanation of the formula and the code itself are available in the folder "vagueness" in McGillivray et al. (2022); see especially the Readme file in this folder.

The goal of the vagueness score is to measure the frequency of passages for which annotators have considered that at least two senses could fit (or closely fit) the meaning of the word being annotated, thus it focuses on the co-occurrences of "4" and/or "3" ratings. It should be noted that by grouping "4" and "3" as the values with which an annotator assigns a meaning (or meanings) to a passage, we neutralise certain nuances. For example, a passage of a three-meaning word annotated "4–1–1" has the same vagueness score as one annotated "3–2–2", that is 0, or no vagueness. However, this last example would have a low confidence score.

Another limitation of our approach is that by focusing on the average presence of "3" and "4" ratings, we cancel out the specificities of each passage. There are a few examples where indeed the annotator did not assign any of the proposed meanings to the passage, either because none of the meanings fit the passage or because the annotator was not confident about the meaning of the passage itself. Any passage that does not contain at least a "4" or a "3" lowers the overall vagueness score of the word regardless whether this annotation is related to the distinctness of the meanings or not.[20] This also means that the vagueness score can be negative.

7.4 Analysis and discussion of results

We calculated the vagueness score for each word and an overview of the results can be seen in Figure 17. We then selected the six words with the highest score of vagueness: *oportet, dux, regnum, fidelis, ius, beatus*; six words with a medium score of vagueness:[21] *scriptura, dubius, sacramentum, credo, adsumo, simplex*; and finally, the six words with the lowest score of vagueness: *consilium, ancilla,*

20 For example, an annotator might consider that none of the proposed senses corresponds to the meaning of a specific passage. This type of annotation would decrease the vagueness score of the word without necessarily being related to the vagueness of the passage itself.

21 The words with a medium score of vagueness were randomly selected from the range of values comprised between Tukey's upper hinge and lower hinge, that is, the value of the first quartile (the

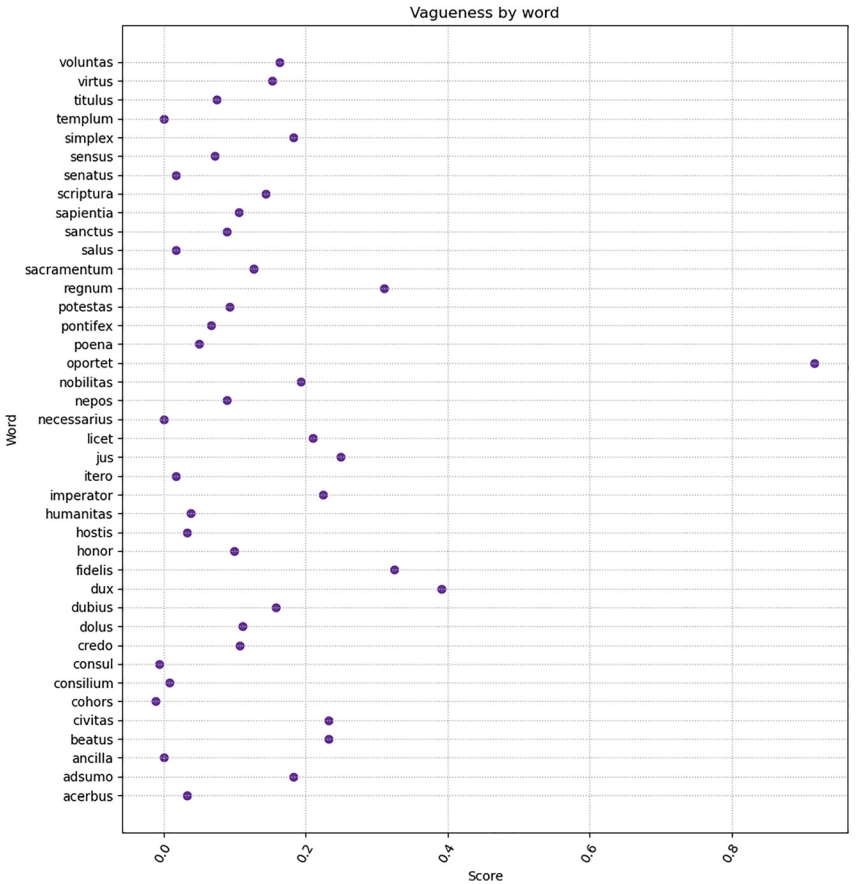

Figure 17: Vagueness score of the 40 words annotated in our study.

necessarius, *templum*, *consul*, *cohors*. For the results for all words, see McGillivray et al. 2022 in the folder "vagueness".

Before presenting and discussing the results, it is worthwhile to mention the special case of the word *ius*. Six senses were available for the annotators: (i) 'broth, soup, sauce', (ii) 'juice, mixture', (iii) 'right, justice, duty', (iv) 'a court of justice', (v) 'justice, justness', 6) 'legal right, power, authority, permission'. As can be inferred by looking at the list of meanings, there are two distinct semantic fields. The homonym meaning 'broth, soup' is related to Sanskrit *yū́ṣ* and maybe Greek ζύμη (Proto-Indo-European **i(e/o)uH-s-* 'broth, soup'); the other derives from the Proto-Indo-

median of the lower half of the dataset), and the value of the third quartile (the median of the upper half of the dataset).

European root *$h_2$oi-u 'vital force' > *$h_2$ieu-os/es (de Vaan 2008: 316). While there are only two senses that belong to the first homonym, the second one has four senses.

The senses of homonyms are not generally implied by one another. Indeed, a closer look at the annotation reveals that the annotators identified the homonyms without difficulties. However, we registered a high score of vagueness (0.25), which makes *ius* the fifth word with the highest score. This result derives from the closeness of the four senses related to the semantic field of 'justice'. In fact, the word *justice* appears as part of the definition of two of the provided senses, which illustrates the close relationship between them. These results can be put in relation with the low confidence score obtained by this word (see Section 5 and, in particular, footnote 17).

The following subsections give an overview of the trends observed in our analysis. We organised them around three main results.

7.4.1 Result 1. Vagueness is not directly correlated with the number of meanings

We could expect that the higher the number of senses of a word has, the higher the probability that we find evidence of vagueness between some of those senses. However, from the distribution of the average vagueness of the complete dataset by number of senses we can see that the number of meanings itself is not the only factor to consider when analysing the vagueness between the meanings of a word (see Figure 18). Thus, words with only two senses may have a higher vagueness

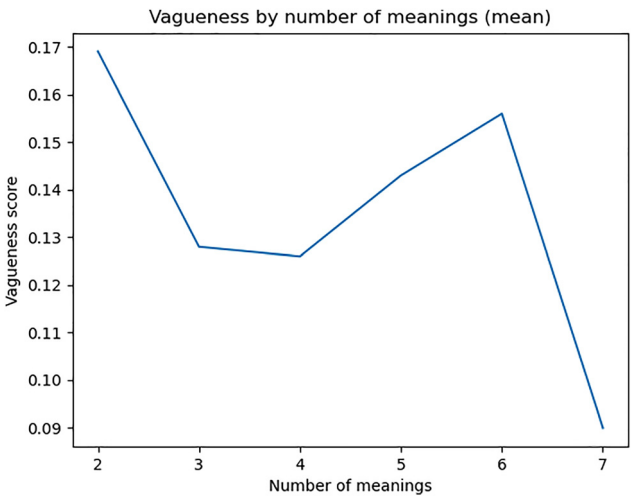

Figure 18: Average vagueness score (arithmetic mean) in relation to the number of meanings.

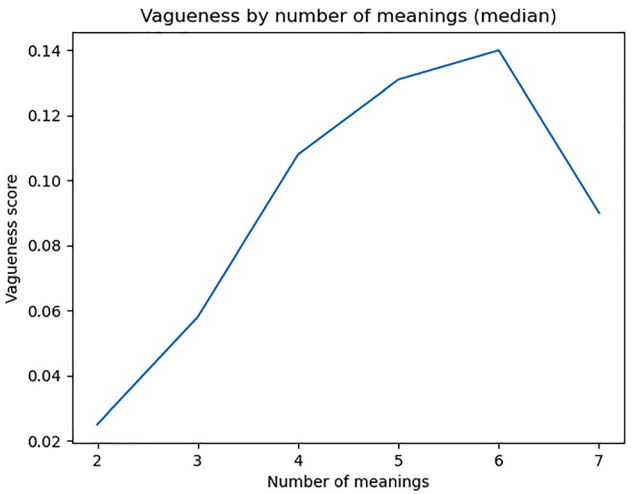

Figure 19: Median vagueness score in relation to the number of meanings.

score than words with four or more senses. This is the case of *oportet*, which shows a high score of vagueness, though it has only two senses, and *consul*, which is ranked as one with the lowest vagueness score, and yet has four senses. This confirms our assumption made earlier in this section that a word can be polysemous, but not necessarily vague. The trends represented in Figure 17 need to be interpreted taking into consideration the effect of outliers. As we could see in Figure 17, *oportet* is an outlier and the reason behind the high score of two-sense words. Another way to represent central tendencies is to calculate the median. Those are the results presented in Figure 19 where the trend is clear: the higher the number of meanings, the higher the vagueness is, with the exception of the two seven-meaning words, which have a relatively low level of vagueness.

The lemma *oportet* shows the highest vagueness score among all the lemmas annotated. This result is due to the fact that in 55 out of its 60 contexts the annotation is "4–3". If 'it is necessary' is annotated with a "4", 'it is proper/it is becoming' receives a "3" and vice versa. This type of annotation led us to the conclusion that the annotator perceived the two meanings as being closely related, tied by a relation of implication that could function in both ways: the "properness" is almost always implied in the notion of "necessity", and the notion of "properness" also seems to imply, to a certain extent, that of "necessity". Analysing the results of the annotation with reference to our modality theory framework established by Nuyts (2016: 34–37) and Dell'Oro (2019: 6–7), we could say that the two meanings presented for *oportet* – 'it is necessary' and 'it is proper/it is becoming' – express, depending on the context, three types of modality: deontic acceptability

(in the values of absolutely necessary and desirable), deontic necessity and dynamic necessity. The meaning 'it is proper/it is becoming' corresponds to the expression of deontic acceptability (desirable), whereas 'it is necessary' could express deontic acceptability (absolutely necessary), deontic necessity or dynamic necessity, depending on the context.

The analysis of the annotation of *oportet* revealed some regularities. As mentioned before, in most cases, when 'it is necessary' is annotated with a "4", 'it is proper' receives a "3". This is the most difficult pattern to analyse, because the type of modality expressed in the passage depends on the context, varying from dynamic/deontic necessity to deontic acceptability with the value of absolute necessity. The expression of these types of modality can entail (to a certain extent depending on the type of modality, cf. *infra*) a relation of implication between the two meanings of deontic/dynamic necessity and properness. The correlation between the type of modality and the presence or absence of implication varies depending on the former. Specifically, the passages where *oportet* expresses deontic acceptability in the value of desirability and dynamic necessity are particularly interesting with respect to the relation of implication. When *oportet* expresses deontic acceptability with the value of desirable, 'it is proper' is always annotated with a "4", and 'it is necessary' receives a "3", without exceptions. This type of annotation establishes a pattern: whenever the passage containing *oportet* expresses this type of modality, a relation of implication between the two meanings, deontic properness and necessity, is always verified. When *oportet* expresses dynamic modality, another pattern is established. There are only five cases of deviation from the annotation "4–3". In all five, 'it is necessary' receives a "4" and 'it is proper' is annotated with "1" (in one case, "2"). This represents, in our perspective, a relation of non-implication between the two meanings of dynamic necessity and properness. In the five passages *oportet* always has a dynamic modal value. Based on the annotation we can hypothesise that, whenever any implication between the two meanings is absent, the passage in which *oportet* appears expresses dynamic modality. Therefore, the absence of implication seems to be a key of disambiguation for the detection of dynamic modality. This is explained by the fact that in those passages the speaker is not performing an evaluation on what is expressed in the state of affairs.

The case of *oportet* shows that this word is inherently vague, as even the context sometimes cannot help in the task of disambiguation. On the other hand, this result is not surprising: it is known that modal markers among other linguistic elements are polyfunctional in some languages such as Latin (Magni 2010: 212), Romance (Le Querler 2004: 652) and Germanic languages (van der Auwera 1999; see van der Auwera and Plungian [1998] for a cross-linguistic overview) and that such markers can remain vague in certain contexts.

7.4.2 Result 2. The impact of a new meaning on the vagueness score depends on the degree of implication of the meanings

As explained in Section 4, some lemmas of the annotation task present a new sense that emerged during the CE period. This can have an influence on the vagueness score, as shown by Figure 20.

The new sense may induce two different behaviours:

- Vagueness increases in CE.
- Vagueness decreases in CE.

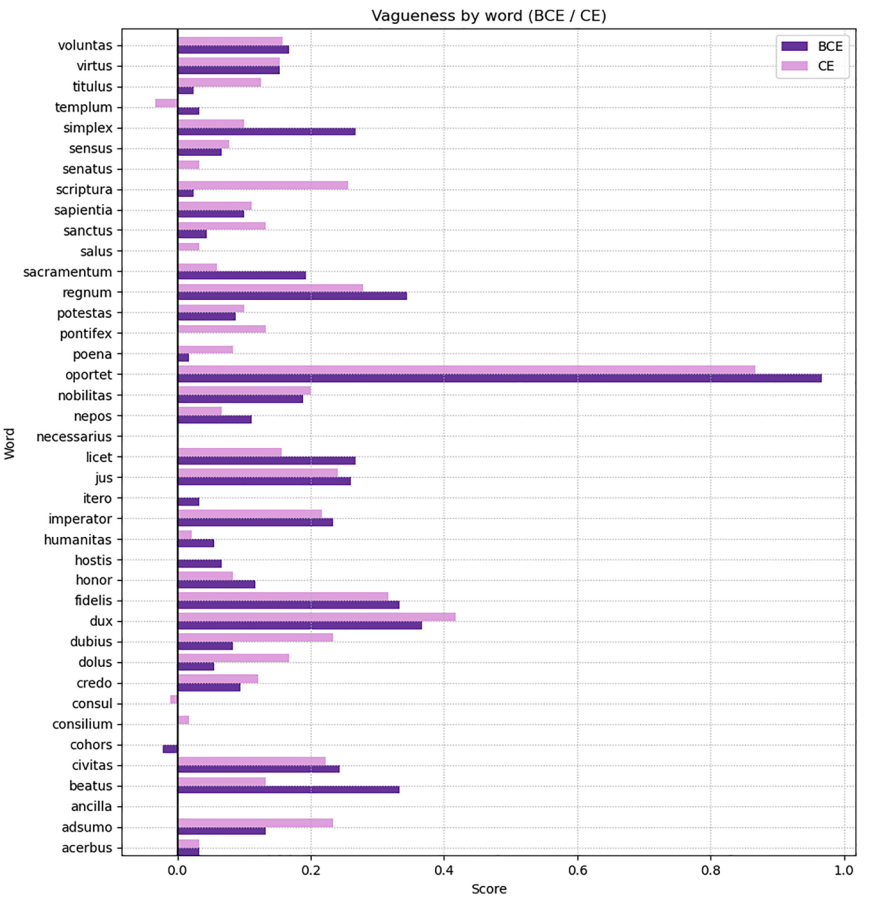

Figure 20: Vagueness score of each lemma in relation to BCE/CE periods.

If one of the older senses is implied in the new meaning, the vagueness score increases. This is the case of *scriptura*, which has four senses: (i) 'writing', (ii) 'tax', (iii) 'will', (iv) 'Holy Scripture', the latter being the new meaning which emerged in the CE period. Looking at the annotation data, 'Holy Scripture' is perceived by the annotator as always implying the meaning 'writing': whenever 'Holy Scripture' is selected with a "4", the meaning 'writing' always receives a "3". This consistently raises the level of vagueness in the CE centuries (see Figure 20). The senses (ii) and (iii) are rarely selected, and in general the relation of implication with 'writing' is not perceived as to be as strong, or is not perceived at all. This keeps the vagueness score low for what concerns the BCE period. The difference in the level of implication between the meanings emerging from the annotation could be due to the fact that, while senses (ii) and (iii) can refer to the content of, respectively, the text of a tax or a law, sense (iv) is rather perceived as a specification of sense (i), and is therefore annotated as strongly implied by it. Interestingly, in some cases the sense 'Holy Scripture' is expressed by the word *scriptura* and an adjective such as *sacra* 'sacred', *sancta* 'holy' or *diuina* 'divine'. In those cases, the annotator selects 'writing' as the exact meaning, and 'Holy Scripture' as being closely related, explaining in a comment that *scriptura* itself carries the meaning 'writing', and the adjective adds the specification 'Holy'. However, the annotator adds that the meaning 'Holy Scripture' should not be discarded, as it is clearly conveyed by the overall context.

On the other hand, a very distinct new sense, with minor implication or with a very clear contextual use, entails a lower score of vagueness in the period in which it emerges. This is what happens for words like *beatus*: even if we can argue that the base meaning 'happy' is implied in 'blessed', the contexts of the latter are easily identified: when the context was a religious one, the annotator identified the meaning 'blessed' and annotated the other meanings as being distantly related or unrelated.

7.4.3 Result 3. Vagueness can be limited to specific contexts

We found that for many words with an average level of vagueness the meaning of most passages is clearly assigned to one of the proposed meanings, and only a number of contexts are difficult to clearly assign to one meaning only. This is the case for words like *simplex* or *adsumo*.

The annotation shows that most of the vague usages of *simplex* are present in rhetorical treatises. This is a context where 'simple, plain, uncompounded' and

'without dissimulation, open, frank, straightforward' can easily be seen as fitting.[22]

The meanings of *adsumo* are: (i) 'take to oneself' and (ii) 'receive'. Three of the contexts annotated as vague are part of *De medicina* by Aulus Cornelius Celsus. In this case, the annotator states in a comment the reason why the two senses are not clearly distinct: medical passages are difficult to annotate because sometimes it is unclear whether the medical patient has a passive or an active role. It seems that this issue can remain undetermined: *adsumo* represents a perfect case of vagueness, because the specific role of the participant can be unknown without affecting the understanding of the passage.

7.4.4 Some observations on the study of vagueness

The main limitation of the study is that the vagueness score was calculated on the basis of only one set of annotations for each word. Considering the annotators' particular styles and the level of inter-annotator agreement discussed in Sections 4 and 5, it must be noted that the vagueness score of some words could show a different value if they were annotated by someone else.

Another factor that needs to be taken into consideration is how the selection of the lexicographic work of reference affects the results. See the case of *beatus*, for which five meanings are provided to the annotator: 'happy', 'fortunate', 'rewarded', 'rich' and 'blessed (religion)'. In Gaffiot et al. (2016 [1934]), the first four meanings get simplified into three which can be paraphrased as follows: 'happy', 'rich (from a materialistic point of view)' and 'rich (in a metaphoric sense)'. The examination of the annotation of this word reveals that the highest level of vagueness concerns the meanings 'happy' and 'fortunate', which are usually annotated as being closely related. Consider that the second of these meanings is not present in dictionaries like Gaffiot et al. (2016 [1934]). We can conclude that the organisation of meanings in the lexicographic work affects the annotation task and, consequently, the calculation of the vagueness score.

It should be noted as well that this study is based on the general analysis of 40 words but on the detailed study of only 18 of those words. We sampled the initial list of words using the vagueness score as criterion. The goal was to obtain a more manageable dataset that was representative in terms of vagueness.

We can compare the weighted confidence per number of meanings (Figure 7) and the vagueness score (Figures 18 and 19). In principle, words with a low score of vagueness are expected to be associated with a high level of confidence, while

22 See as an example: *Defensoris narratio simplicem et dilucidam expositionem debet habere* […] (*Rhet. Her.* 2.3), "The Statement of Facts of the defendant's counsel should contain a simple and clear account" (Caplan 1954: 61).

words with a high score of vagueness are expected to be associated with a low level of confidence. The first point seems confirmed: the six least vague words (*consilium*, *ancilla*, *necessarius*, *templum*, *consul*, *cohors*) are all associated with medium or high levels of confidence (however always >0.6). Notice that *consul* and *cohors* are associated with the highest score of weighted confidence and with the lowest score of vagueness. The second point also seems confirmed: the six most vague words (*oportet*, *dux*, *regnum*, *fidelis*, *ius*, *beatus*) are associated with low or medium scores of weighted confidence (however always <0.6). Note, however, that, for example, the most vague word (*oportet*) received a higher score of weighted confidence than *dux*, *beatus* and *fidelis*. We can say that the vagueness of a word had some influence on the degree of confidence, though it must be underlined that the methods to calculate the two types of score partially overlap and these results are not totally without some bias (vagueness is calculated on the basis of the annotation). On the other hand, it is important to underscore that the correlation between the values of the two types of score cannot be said to be mechanically correlated, as shown by the words with a medium score of vagueness (see for instance *adsumo* and *scriptura*, which are very distant as concerns the weighted confidence).

8 Conclusion

We have presented a new corpus-based resource and methodology for the annotation of Latin lexical semantics. The resource consists of 2,399 annotated passages from the LatinISE corpus, a large diachronic corpus of Latin. We selected 40 Latin lemmas, including nouns, adjectives and verbs. Seventeen of them were chosen because they are known in the literature to have undergone lexical semantic change associated with the cultural and religious changes associated with the religious, cultural, social and political changes linked to the late antiquity; the remaining 22 lemmas were chosen as comparable *stable* words. For each lemma, we annotated 60 passages from LatinISE.[23] Following a variation of the DuRel framework (Schlechtweg et al. 2018), the passages were annotated by associating the corpus usage of each lemma with its dictionary definition according to the degree of relatedness between each corpus usage and each definition, measured on a scale from 1 to 4. This corpus resource is the first one providing lexical semantic annotation for Latin.

23 The exception is *scriptura*, where we only found 59 corpus occurrences conforming to the criteria.

In addition to sharing this resource, we describe the design of the annotation and how it was adapted from the model used for living languages in the context of the SemEval 2020 shared task on unsupervised lexical semantic change detection. We also analysed the annotation data to measure each annotator's style and the agreement between annotators. We complemented this analysis with a study on the lexical semantics of the 40 lemmas, including their diachronic change and a case study on semantic vagueness. Our dataset, and our analysis, has some limitations, which we discussed in this paper. It is a relatively small dataset, collected for a list of only 40 Latin words, amounting to only 60 corpus instances per word, each of which were annotated by only one person (with the exception of *uirtus*). This was due to the time constraints of the study, which aimed at preparing the annotation on time for the SemEval competition. Due to the corpus composition and the limited number of text passages annotated, we are fully aware of the fact that the annotated instances were too few to lead to conclusive evidence for lexical semantic change in the context of late antiquity. Despite these limitations, our analysis offers a solid methodological basis and the starting point for further analyses.

The novel aspects of our contribution are of a methodological nature. First, we show how the DuReL framework can be successfully applied to the case of a historical corpus language such as Latin. As the availability of digital corpora of such languages increases, and as computational linguistics research develops new methods for large-scale analysis of diachronic lexical semantics, building resources annotated at the level of lexical semantics has the potential to reveal new large-scale patterns on the semantic development of lexical items over time. Therefore, we share a series of recommendations for designing the annotation task that will hopefully serve as a basis for similar research on other less-resourced and/or historical languages. These recommendations include clarifying how the 0–4 rating system can be used to ensure consistency across annotators, and adding a formalised way to flag concerns and the annotator's confidence about individual passages.

Second, based on a quantitative analysis of the annotated data, we propose a "confidence score" which captures the level of uncertainty involved in the annotation of a given corpus context. Apart from the two words that have seven meanings, we find that the annotators' average confidence goes down as the polysemy of the lemmas increases, showing that the data reflects word-specific patterns rather than patterns to be purely associated with each annotator's style. We employ this score to conduct a quantitative analysis of the dataset. This analysis identified a series of outliers to be further analysed with qualitative methods. This approach allowed us to combine a high-level analysis of patterns emerging from the data with a closer inspection of individual lexical items.

Third, we conducted a quantitative diachronic analysis of the semantics of the lemmas of focus. We used metrics based on the annotation data to identify words which have coexisting senses and looked into the development of new senses in later texts.

Fourth, we propose an analysis of the annotations based on the measure of vagueness. This measure is meant to detect the words that were annotated as the vaguest, that is, for which the annotator could detect some overlap or implication between their meanings. In terms of the annotation, in the majority of the contexts presented for these words, more than one meaning was selected by the annotator, or more meanings were annotated as closely related. This analysis revealed that the number of meanings does not have a direct impact on the vagueness of a word. Rather, the degree of implication between its meanings and the context in which it is used can affect its vagueness. Together with the confidence score, metrics related to vagueness are a methodological tool to tease out lemmas that deserve a more detailed semantic analysis.

Acknowledgements: We would like to thank Hugo Burgess and Rozalia Dobos for their work on the annotation.
Research funding: This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. The work and the writing of the "vagueness" section is part of the project "A World of Possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language" funded by the Swiss National Science Foundation under the project no. 176778.
Author contributions: BMcG wrote Sections 1, 3, 6, 8; conceptualisation and scoping of the study, design of the annotation, data processing for annotation, implementation of the inter-annotator agreement analysis, design of the annotator and word-level analysis; design and implementation of the diachronic analysis. AB wrote Sections 2, 3 and 6, and contributed to the diachronic analysis. DK wrote Sections 4, 5 and 6 and conducted the analysis on annotator's data. FDO, HBS and PM wrote together Section 7: FDO conceptualised the vagueness framework, supervised the elaboration of the whole section and mainly wrote part 7.1 and 7.2; HBS elaborated the mathematical formula to calculate the vagueness score and mainly wrote part 7.3, the analysis of the words *beatus*, *simplex* and *adsumo*; PM mainly wrote the analysis of the words *oportet* and *scriptura*. MMC wrote Sections 1 and 2 and conducted the pre-annotation analysis.

References

Adamska-Sałaciak, Arleta. 2014. Bilingual lexicography: Translation dictionaries. In Patrick Hanks & Giles-Maurice De Schryver (eds.), *International handbook of modern lexis and lexicography*, 1–11. Berlin: Springer.

Adema, Suzanne. 2019. Latin learning and instruction as a research field. *Journal of Latin Linguistics* 18(1/2). 35–59.

Antonini, Sergio & Pereyro Verónica Díaz. 2020. Otra vía al latín: Testimonios y recursos. *Revista exlibris* 9. 97–115.

Auerbach, Erich. 1937. Remarques sur le mot 'passion'. *Neuphilologische Mitteilungen* 38(3). 218–224.

van der Auwera, Johan. 1999. On the semantic and pragmatic polyfunctionality of modal verbs. In Ken Turner (ed.), *The semantics/pragmatics interface from different points of view*, 49–64. Oxford: Elsevier.

van der Auwera, Johan & Vladimir Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2(1). 79–124.

Bamman, David & Gregory Crane. 2007. The design and use of a Latin Dependency Treebank. In Jan Hajič & Joakim Nivre (eds.), *The Fifth International Treebanks and Linguistic Theories Conference (TLT 2006)*, 67–78. Prague: Institute of Formal and Applied Linguistics.

Bamman, David, Marco Passarotti, Roberto Busa & Gregory Crane. 2008. The annotation guidelines of Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds.), *Sixth International Conference on Language Resources and Evaluation (LREC 2008). May 28–30, 2008, Marrakech, Morocco, 71–76*. Paris: European Language Resources Association.

Bastardas, Juan. 1973. El latín de los cristianos: Estado actual de su problemática. *Boletín del Instituto de Estudios Helénicos* 7(2). 5–17.

Beskow, Per. 1962. *Rex Gloriae: The kingship of Christ in the Early Church*. Stockholm: Almqvist & Wiksell.

Boas, Franz. 1911. *Handbook of American Indian languages*. (Bureau of American Ethnology Bulletin 40). Washington, DC: Government Printing Office.

Buccheri, Alessandro, Irene De Felice, Chiara Fedriani & William M. Short. 2021. Semantic analysis and frequency effects of conceptual metaphors of emotions in Latin. From a corpus-based approach to a dictionary of Latin metaphors. *Journal of Latin Linguistics* 20(2). 163–189.

Burton, Philip. 2011. Christian Latin. In James Clackson (ed.), *A companion to the Latin language*, 485–501. Oxford: Wiley-Blackwell.

Cameron, Alan & Diane Schauer. 1982. The last consul: Basilius and his diptych. *Journal of Roman Studies* 72. 126–145.

Caplan, Harry (ed.). 1954. *Cicero: Rhetorica ad Herennium* [Cicero: Rhetoric for Herennius]. Cambridge, MA: Harvard University Press.

Cardinaletti, Anna, Giuliana Giusti & Rossella Iovino. 2016. *Il latino per studenti con DSA: Nuovi strumenti didattici per la scuola inclusiva*. Venice: Cafoscarina.

Clackson, James (ed.). 2011a. *A companion to the Latin language*. Oxford: Wiley-Blackwell.

Clackson, James. 2011b. Introduction. In James Clackson (ed.), *A companion to the Latin language*, 1–6. Oxford: Wiley-Blackwell.

Clackson, James & Geoffrey Horrocks. 2007. *The Blackwell history of the Latin language*. Oxford: Wiley-Blackwell.

Crystal, David. 2002. *Language death*. Cambridge: Cambridge University Press.

Cuzzolin, Pierluigi. 2019. Qualche riflessione per la costituzione di un corpus di latino tardo. *Rhesis* 10(5). 5–18.

Deagon, Andrea. 2006. Cognitive style and learning strategies in Latin instruction. In John Gruber-Miller (ed.), *When dead tongues speak: Teaching beginning Greek and Latin*, 27–49. Oxford: Oxford University Press.

Dell'Oro, Francesca. 2019. WoPoss guidelines for annotation. *Zenodo*. https://doi.org/10.5281/zenodo.3560951.

de Vaan, Michiel. 2008. *Etymological dictionary of Latin and the other Italic languages*. Leiden: Brill.

Dinkova-Bruun, Greti. 2011. Medieval Latin. In James Clackson (ed.), *A companion to the Latin language*, 284–302. Oxford: Wiley-Blackwell.

Du Cange, Charles du Fresne, Pierre Carpenter, G. A. Louis Henschel & Léopold Favre. 1883–1887 [1678]. *Glossarium mediæ et infimæ latinitatis* [Glossary of Middle and Low Latin]. Niort: Favre.

Eger, Steffen & Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In Katrin Erk & Noah A. Smith (eds.), *The 54th Annual Meeting of the Association for Computational Linguistics. vol. 2 (Short Papers)*, 52–58. Stroudsburg, PA: Association for Computational Linguistics.

Erk, Katrin, Diana McCarthy & Nicholas Gaylord. 2009. Investigations on word senses and word usages. In Keh-Yih Su, Jian Su, Janyce Wiebe & Haizhou Li (eds.), *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 1*, 10–18. Stroudsburg, PA: Association for Computational Linguistics.

Erk, Katrin, Diana McCarthy & Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics* 39(3). 511–554.

Fedriani, Chiara, Irene De Felice & William Michael Short. 2020. The Digital Lexicon Translaticium Latinum: Theoretical and methodological issues. In Cristina Marras, Marco Passarotti, Greta Franzini & Eleonora Litta (eds.), *Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) La svolta inevitabile: sfide e prospettive per l'informatica umanistica*, 106–112. Milano: Università Cattolica del Sacro Cuore.

Frösén, Jaakko. 2011. Conservation of ancient papyrus material. In Roger S. Bagnall (ed.), *The Oxford handbook of papyrology*, 79–100. Oxford: Oxford University Press.

Fuertes-Olivera, Pedro Antonio. 2017. Cómo abordar en el aula la equivalencia entre lenguas. In Maria José Domínguez Vázquez & Maria Teresa Sanmarco Bande (eds.), *Lexicografía y Didáctica. Diccionarios y Otros Recursos Lexicográficos en el Aula (Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation 115)*, 329–342. Frankfurt am Main: Peter Lang.

Gaffiot, Felix, Gérard Gréco, Mark De Wilde, Bernard Maréchal & Katsuhiko Ôkubo (eds.). 2016 [1934]. *Dictionnaire Latin-Français. Nouvelle Édition Revue et Augmentée*.

Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.

Glare, Peter G. W (ed.). 1997 [1982]. *Oxford Latin dictionary*. Oxford: Oxford University Press.

Grondelaers, Stefan, Dirk Speelman & Dirk Geeraerts. 2007. Lexical variation and change. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford handbook of cognitive linguistics*, 988–1011. Oxford: Oxford University Press.

Gy, Györffy. 1975. *'Civitas', 'Castrum', 'Castellum'. Acta Antiqua Academiae Scientiarum Hungaricae* [Ancient acts of the Hungarian academy of sciences], vol. 23. 331–334.
Habel, Edwin. 1959. *Mittellateinisches Glossar*. Paderborn: Schöningh.
Haug, Dag T. T. & Marius L. Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European Bible translations. In Caroline Sporleder & Kiril Ribarov (eds.), *The Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34. Paris & Marrakesh: European Language Resources Association (ELRA).
Heim, François. 1991. *Virtus. Idéologie Politique et Croyances Religieuses au IVe Siècle*. Berne: Peter Lang.
Holford-Strevens, Leofranc A. 1981. Christian Latin. *The Classical Review* 31(2). 230–233.
Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel. 2006. OntoNotes: The 90% solution. In Robert C. Moore, Jeff Bilmes, Jennifer Chu-Carroll & Mark Sanderson (eds.), *The human language technology conference of the NAACL, Companion Volume: Short papers*, 57–60. Stroudsburg, PA: Association for Computational Linguistics.
Hunt, Steven. 2016. *Starting to teach Latin*. London: Bloomsbury Academic.
Iovino, Rossella. 2019. Rethinking the teaching of Latin in the inclusive school. *Journal of Latin Linguistics* 18(1/2). 85–99.
Janson, Tore. 1991. Language change and metalinguistic change: Latin to Romance and other cases. In Roger Wright (ed.), *Latin and the Romance languages in the Early Middle Ages*, 9–28. London: Routledge.
Kenny, Neil. 1995. Interpreting concepts after the linguistic turn: The example of 'curiosité' in 'Le Bonheur des sages / Le Malheur des curieux' by Du Souhait (1600). In John O'Brien (ed.), *Ré) interprétations: Études sur le Seizième Siècle* (Michigan Romance Studies 15), 241–270. Ann Arbor: University of Michigan.
Kilgarriff, Adam. 1997. What is word sense disambiguation good for? In *Natural Language Processing in the Pacific Rim (NLPRS '97). Phuket, Thailand*, 209–214. Bangkok: National Electronics and Computer Technology Center.
Koch, Peter. 2016. Meaning change and semantic shifts. In Päivi Juvonen & Maria Koptjevskaja Tamm (eds.), *The lexical typology of semantic shifts* (Cognitive Linguistics Research 58), 21–66. Berlin & Boston: Mouton de Gruyter.
Koptjevskaja-Tamm, Maria. 2002. The lexical typology of semantic shifts: An introduction. In Päivi Juvonen & Maria Koptjevskaja Tamm (eds.), *The lexical typology of semantic shifts* (Cognitive Linguistics Research 58), 1–20. Berlin & Boston: Mouton de Gruyter.
Kroonen, Guus. 2013. *Etymological dictionary of Proto-Germanic*. Leiden: Brill.
Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In Emily M. Bender, Leon Derczynski & Pierre Isabelle (eds.), *The 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA*, 1384–1397. Stroudsburg, PA: Association for Computational Linguistics.
Kuukkanen, Jouni-Matti. 2008. Making sense of conceptual change. *History and Theory* 47. 351–372.
Lampe, Peter. 2004. Early Christians in the city of Rome: Topographical and social historical aspects of the first three centuries. In Jürgen Zangenberg & Michael Labahn (eds.), *Christians as a religious minority in a multicultural city: Modes of interaction and identity formation in Early Imperial Rome*, 20–32. London: T&T Clark.

Langone, Helen, Benjamin R. Haskell & George A. Miller. 2004. Annotating WordNet. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL. Boston, MA, USA*, 63–69. Stroudsburg, PA: Association for Computational Linguistics.

Langslow, David R. 2002. Approaching bilingualism in corpus languages. In James N. Adams, Simon Swain & Mark Janse (eds.), *Bilingualism in ancient society: Language contact and the written text*, 23–51. Oxford: Oxford University Press.

Le Querler, Nicole. 2004. Les modalités en français. *Revue Belge de Philologie et d'Histoire* 82(3). 643–656.

Lewis, Charlton T. 1890. *An elementary Latin dictionary*. Oxford: Oxford University Press.

Lewis, Charlton T. & Charles Short. 1879. *A Latin dictionary, founded on Andrews' edition of Freund's Latin Dictionary. Revised, enlarged and in great part rewritten by Charlton T. Lewis, PhD. and Charles Short*. Oxford: Clarendon Press.

Loeb Classical Library. 2020. Spreadsheet: The complete series [MS Excel]. https://www.hup.harvard.edu/resources/booksellers/downloads/Loeb-Classical-Library.xlsx (accessed 3 March 2022).

Loi, Vincenzo. 1978. *Origini e Caratteristiche della Latinità Cristiana*. Rome: Accademia Nazionale dei Lincei.

Logeion Online Dictionary. https://logeion.uchicago.edu/ (accessed 3 March 2022).

López de Lerma, Gala. 2015. *Análisis Comparativo de Metodologías para la Enseñanza y el Aprendizaje de la Lengua Latina*. Barcelona: University of Barcelona PhD thesis.

López de Lerma, Gala & Alba Ambrós. 2016. Enseñanza de la lengua latina: Resultados preliminares sobre las ventajas e inconvenientes en el empleo de diferentes metodologías. *Methodos* 3. 67–83.

López Silva, Xosé Antonio. 2003. El influjo del latín de los cristianos en la evolución general de la lengua latina. *Ianua* 4. 115–126.

Macías Villalobos, Cristóbal. 2012. La aplicación del método inductivo-contextual a la enseñanza del latín en el ámbito universitario: Una experiencia. *Thamyris* 3. 151–228.

Macías Villalobos, Cristóbal. 2015. Algunas consideraciones y materiales para abordar la enseñanza del latín según una metodología híbrida. *Thamyris* 6. 201–300.

Magni, Elisabetta. 2010. Mood and modality. In Philip Baldi & Pierluigi Cuzzolin (eds.), *New perspectives on historical Latin syntax. Volume 2, Constituent syntax: Adverbial phrases, adverbs, mood, tense* (Trends in Linguistics 180), 193–275. Berlin & Boston: Mouton de Gruyter.

Magni, Elisabetta. 2020. *L'ambiguità delle Lingue*. Rome: Carocci.

Mambrini, Francesco, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Marco Carlo Passarotti & Paolo Ruffolo. 2020. LiLa: Linking Latin: Risorse linguistiche per il latino nel Semantic Web. *Umanistica Digitale* 4(8). 63–78.

Márquez Cruz, Manuel & Ana María Fernández-Pampillón Cesteros. 2019. Motivación en el aprendizaje del latín: Evaluación de una nueva metodología didáctica. *ReiDoCrea* 8. 432–441.

Matthews, Peter H. 2014 [1997]. Polysemy. In *The concise Oxford dictionary of linguistics*, 3rd edn, 309. Oxford: Oxford University Press.

Mayrhofer, Manfred. 1980. *Zur Gestaltung des etymologischen Wörterbuches einer 'Grosscorpus-Sprache'*. Wien: Österreichische Akademie der Wissenschaften.

McCarthy, Diana & Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation* 43(2). 139–159.

McGillivray, Barbara. 2014. *Methods in Latin computational linguistics*. Leiden: Brill.

McGillivray, Barbara. 2021. Dataset: Latin lexical semantic annotation. *Figshare*. https://doi.org/10.18742/16974823.v1.

McGillivray, Barbara, Hengchen Simon, Viivi Lähteenoja, Marco Palma & Alessandro Vatri. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities* 34(4). 893–907.

McGillivray, Barbara & Kilgarriff Adam. 2013. Tools for historical corpus research, and a corpus of Latin. In Paul Bennett, Martin Durrell, Silke Scheible & Richard J. Whitt (eds.), *New methods in historical corpus linguistics* (Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache 3), 247–257. Narr: Tübingen.

McGillivray, Barbara, Daria Kondakova & Helena Bermúdez Sabel. 2022. Code for analysing the semantic annotation of Latin data from SemEval 2020 task 1. *Zenodo*. https://doi.org/10.5281/zenodo.6513482 (accessed 3 May 2022).

McKitterick, Rosamund. 1991. Latin and Romance: A historian's perspective. In Roger Wright (ed.), *Latin and the Romance languages in the Early Middle Ages*, 130–145. London: Routledge.

Mohrmann, Christine. 1950–1951. L'étude de la latinité chrétienne: État de la question, méthodes, résultats. *Conférences de l'Institut de Linguistique de l'Université de Paris* 10(1). 25–141.

Moore, Anne. 2009. *Moving beyond symbol and myth: Understanding the kingship of God of the Hebrew Bible through metaphor*. New York: Peter Lang.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys* 41(2). 1–69.

Nuyts, Jan. 2016. Analyses of the modal meanings. In Jan Nuyts & Johan van der Auwera (eds.), *The Oxford handbook of modality and mood*, 31–49. Oxford: Oxford University Press.

OED Online. Oxford: Oxford University Press. https://www.oed.com (accessed 3 March 2022).

Ortuño Arregui, Manuel. 2016. Latín de los cristianos: Aproximación lingüística. *ArtyHum Revista de Artes y Humanidades* 20. 8–65.

Passarotti, Marco. 2019. The project of the Index Thomisticus Treebank. In Monica Berti (ed.), *Digital classical philology. Ancient Greek and Latin in the digital revolution* (Age of Access? Grundfragen der Informationsgesellschaft 10), 299–319. Berlin & Boston: Saur de Gruyter.

Passarotti, Marco, Flavio M. Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini & Paolo Ruffolo. 2019. The LiLa knowledge base of linguistic resources and NLP tools for Latin. In Thierry Declerck & John P. McCrae (eds.), *The 2nd Conference on Language, Data and Knowledge (LDK 2019), Leipzig, Germany, May 21, 2019* (CEUR Workshop Proceedings 2402), 6–11. Aachen: CEUR-WS.

Penney, John. 2011. Archaic and Old Latin. In James Clackson (ed.), *A companion to the Latin language*, 220–235. Oxford: Wiley-Blackwell.

Perrone, Valerio, Marco Palma, Hengchen Simon, Alessandro Vatri, Jim Q. Smith & Barbara McGillivray. 2019. GASC: Genre-aware semantic change for ancient Greek. In Nina Tahmasebi, Lars Borin, Adam Jatowt & Yang Xu (eds.), *The 1st international workshop on computational approaches to historical language change*, 56–66. Stroudsburg, PA: Association for Computational Linguistics.

Petrollino, Sara & Maarten Mous. 2010. Recollecting words and expressions in Aasá, a dead language in Tanzania. *Anthropological Linguistics* 52. 206–216.

Ribary, Marton & Barbara McGillivray. 2020. A corpus approach to Roman law based on Justinian's digest. *Informatics* 7(4). 44.

Richards, Jack C. & Theodore S. Rodgers. 2005. *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Richter, Melvin. 1995. *The history of political and social concepts: A critical introduction*. Oxford: Oxford University Press.

Saffire, Paula. 2006. Ancient Greek in classroom conversation. In John Gruber-Miller (ed.), *When dead tongues speak: Teaching beginning Greek and Latin*, 158–189. Oxford: Oxford University Press.

Sapir, Edward. 1912. Language and environment. *American Anthropologist* 14(2). 226–242.

Schlechtweg, Dominik, im Walde Sabine Schulte & Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In Marilyn Walker, Heng Ji & Amanda Stent (eds.), *The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 169–174. Stroudsburg, PA: Association for Computational Linguistics.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May & Ekaterina Shutova (eds.), *Proceedings of the fourteenth workshop on semantic evaluation*, 1–23. Barcelona: International Committee for Computational Linguistics.

Schmid, Helmut. 2003 [1997]. Probabilistic part-of-speech tagging using decision trees. In Daniel B. Jones & Harold L. Somers (eds.), *New methods in language processing*, 154–164. London: Routledge.

Schrijnen, Joseph. 1932. *Charakteristik des Altchristlichen Latein*. Nijmegen: Dekker van de Vegt.

Scott Morrell, Kenneth. 2006. Language acquisition and teaching ancient Greek: Applying recent theories and technology. In John Gruber-Miller (ed.), *When dead tongues speak: Teaching beginning Greek and Latin*, 134–157. Oxford: Oxford University Press.

Shelmerdine, Cynthia W. & John Bennet. 2008. Mycenaean states: Economy and administration. In Cynthia W. Shelmerdine (ed.), *The Cambridge companion to the Aegean Bronze Age*, 289–309. Cambridge: Cambridge University Press.

Sidwell, Keith. 2015. Classical Latin – Medieval Latin – Neo-Latin. In Sarah Knight & Stefan Tilg (eds.), *The Oxford handbook of Neo-Latin*, 13–26. Oxford: Oxford University Press.

Sprugnoli, Rachele, Marco Carlo Passarotti & Giovanni Moretti. 2019. 'Vir' is to 'moderatus' as 'mulier' is to 'intemperans': Lemma embeddings for Latin. In Raffaella Bernardi, Roberto Navigli & Giovanni Semeraro (eds.), *Sixth Italian Conference on Computational Linguistics* (CEUR Workshop Proceedings 2481). Aachen: CEUR-WS.

Stavropoulos, Thanos G., Stelios Andredis, Marina Riga, Efstratios Kontopoulos, Panagiotis Mitzias & Ioannis Kompatsiaris. 2016. A framework for measuring semantic drift in ontologies. In *Paper presented at SuCCESS'16 – 1st Int. Workshop on Semantic Change & Evolving Semantics, Leipzig, Germany, 12 September*.

Stotz, Peter. 2000. *Handbuch zur Lateinischen Sprache des Mittelalters: II. Bedeutungswandel und Wortbildung*. Munich: Beck.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 1–91. Berlin: Language Science Press.

TLL = Thesaurus Linguae Latinae. Berlin & Boston: Walter de Gruyter.

Tracy, Catherine. 2008–2009. The people's consul: The significance of Cicero's use of the term 'popularis'. *Illinois Classical Studies* 33/34. 181–199.

Tuggy, David. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics* 4(3). 273–290.

Untermann, Jürgen. 1983. Indogermanische Restsprachen als Gegenstand der Indogermanistik. In Eduardo Vineis (ed.), *Le Lingue Indoeuropee di Frammentaria Attestazione. Die Indogermanischen Restsprachen. Atti del Convegno della Società Italiana di Glottologia e della Indogermanische Gesellschaft, Udine, 22–24 settembre 1981* (Biblioteca della Società italiana di glottologia 4), 11–28. Pisa: Giardini.

Varvaro, Alberto. 1991. Latin and Romance: Fragmentation or restructuring? In Roger Wright (ed.), *Latin and the Romance languages in the Early Middle Ages*, 44–51. London: Routledge.

Vatri, Alessandro, Viivi Lähteenoja & Barbara McGillivray. 2019. Ancient Greek semantic annotation datasets and code. *Figshare*. https://doi.org/10.6084/m9.figshare.c.4445420.

Wang, Shenghui, Stefan Schlobach & Michel Klein. 2011. Concept drift and how to identify it. *Journal of Web Semantics* 9(3). 247–265.

Wasow, Thomas. 2015. Ambiguity avoidance is overrated. In Susanne Winkler (ed.), *Ambiguity: Language and communication*, 29–47. Berlin: Walter de Gruyter.

Way, Geoffrey (ed.). 1955. *Caesar. Alexandrian war. African war. Spanish war*. Cambridge, MA: Harvard University Press.

Weaver, Warren. 1949. Translation. In William N. Locke & Andrew D. Boothe (eds.), *Machine translation of languages*, 15–23. Cambridge, MA: MIT Press.

Wierzbicka, Anna. 1997. *Understanding cultures through their key words: English, Russian, Polish, German, and Japanese*. Oxford: Oxford University Press.

Williams, Raymond. 1976. *Keywords. A vocabulary of culture and society*. London: Fontana.

Wright, Roger. 1982. *Late Latin and Early Romance in Spain and Carolingian France*. Liverpool: Francis Cairns.

Zaccarello, Michelangelo & Martin Maiden (eds.). 2003. The early textualization of the Romance languages: Recent perspectives: Atti del Convegno di Oxford 23–24 marzo 2002. Trinity e Pembroke College. [Special issue]. *Medioevo Romanzo* 27(2).

Zerjadtke, Michael. 2019. *Das Amt "Dux" in Spätantike und Frühem Mittelalter*. Berlin & Boston: Walter de Gruyter.