Contemporary Voices in Ancient Tongue: Integrating Papal Encyclicals into the LiLa KB

Aurora Alagni[1,*,†], Federica Iurescia[1] and Eleonora Litta[1]

[1]Università Cattolica del Sacro Cuore, CIRCSE Research Centre, Largo Gemelli, 1, 20123 Milan, Italy

Abstract

This paper presents the integration of a new textual resource—the Papal Encyclicals corpus—into the LiLa: Linking Latin Knowledge Base. The inclusion of three recent Encyclicals authored by Pope Francis (*Lumen Fidei*, *Laudato si'*, and *Fratres omnes*) significantly enriches the LiLa Knowledge Base by extending its chronological coverage and introducing contemporary Latin vocabulary. The linking process involved automatic tokenisation, part-of-speech tagging, and lemmatisation using the LiLa Text Linker, followed by manual validation and disambiguation. The newly added lemmas fall into three categories: Latinized anthroponyms and toponyms, ethnic adjectives, and neologisms. These lexical additions reflect both a modernising trend in Vatican Latin and diverse morphological and semantic processes, including borrowing, calquing, and analogy-based reconstruction. The resource also opens avenues for analysing the stylistic and rhetorical features of Papal Encyclicals as a genre.

Keywords

Linked Open Data, Latin, textual resources

1. Introduction

1.1. LiLa

LiLa (Linking Latin) is a Linked Open Data (LOD) Knowledge Base (KB).[1] LiLa has been built to foster interoperability across textual and lexical resources for Latin [1]. The LiLa KB relies on two primary components:

- the Lemma Bank,[2] a collection currently comprising approximately 230,000 Latin lemmas (canonical citation forms of lexical items) published as LOD;[3]
- several language resources for Latin published as LOD and interconnected through the Lemma Bank, including corpora, lexica, and dictionaries.[4]

The LiLa KB employs several ontologies to represent both the data and metadata of the interlinked linguistic resources, such as POWLA for corpus data [2], OLiA for linguistic annotation [3], and Ontolex-Lemon for lexical data [4]. LiLa is an open-ended Knowledge Base: as new resources are integrated, the Lemma Bank is expanded. Lexical resources are linked to the Lemma Bank by connecting their lexical entries to their canonical forms. The single word occurrences (tokens) in textual resources are connected to the corresponding lemma in the LiLa Lemma Bank.

1.2. Papal Encyclicals

Among the textual resources in LiLa, a recent addition is the Papal Encyclicals corpus, which comprises pastoral letters dealing with Catholic doctrine written by Popes to Roman Catholic bishops.[5] In its first release, the corpus includes three encyclicals written by Pope Francis, namely *Lumen Fidei* (LF), *Laudato si'* (LS) and *Fratres omnes* (FO).[6] LF was published in 2013 and explores faith as a divine light illuminating human existence. LS was published in 2015 and advocates for a holistic response to ecological crises. FO was published in 2020, and pleads for universal fraternity and social friendship in the pursuit of a just and peaceful society. The source of the Latin text is their digital version on the Vatican site.[7] The integration of this resource enhances the coverage of the texts included in the LiLa Knowledge Base, extending both the chronological range and the diversity of textual

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author.

† This paper is the result of the collaboration between the authors. For the specific concerns of the Italian academic attribution system, Federica Iurescia is responsible for section 1; Eleonora Litta for section 2; Aurora Alagni for section 3. Section 4 was collaboratively written by all authors.

aurora.alagni@outlook.it (A. Alagni);
federica.iurescia@unicatt.it (F. Iurescia);
EleonoraMaria.Litta@unicatt.it (E. Litta)
0000-0001-5100-5539 (F. Iurescia); 0000-0002-0499-997X (E. Litta)

[1] http://lila-erc.eu

[2] http://lila-erc.eu/data/id/lemma/LemmaBank

[3] The collection of lemmas in the Lemma Bank originates from LEMLAT 3.0, a morphological analyzer [5].

[4] The list of resources interlinked in LiLa is available at https://lila-erc.eu/data-page/.

[5] http://lila-erc.eu/data/corpora/PapalEncyclicals/id/corpus.

[6] At the moment of writing, the encyclical *Dilexit nos*, published in 2024, was not available.

[7] https://www.vatican.va/content/francesco/la/encyclicals.index.html#encyclicals.

Table 1
Tokens per letter

	total
LF	17,051
LS	35,801
FO	36,611

Table 2
Match results per letter

	1:1	1:N	1:0
LF	11,507	1,251	3,058
LS	25,351	1,688	1,744
FO	26,225	1,346	1,407

genres represented. Moreover, the addition of this corpus not only expands the Lemma Bank with new lemmas but also enables the study of lexical innovation strategies employed to express modern concepts in Latin.

2. Linking

2.1. Linking

The initial phase of the linking process involved the acquisition of plain-text versions of the three texts, retrieved from the official Vatican website. Collectively, these texts comprise 89,463 tokens, including punctuation and numerical elements associated with verse numbering and biblical references.

Tokenisation, sentence segmentation, part-of-speech (PoS) tagging, and lemmatisation were carried out automatically using the LiLa Text Linker—an NLP tool specifically designed for Latin. Table 1 displays the number of tokens per letter, excluding punctuation and numbering. Developed as part of the user-oriented services provided by LiLa [6], the Text Linker not only performs linguistic annotation but also establishes links between the annotated output and corresponding entries in the Lemma Bank. For PoS tagging and lemmatisation, the system relies on a UDPipe model trained on customised data. The linking procedure operates as follows: whenever the lemmatisation of a token yields a lemma–PoS pair that exactly matches a corresponding entry in the LiLa Lemma Bank, the system returns the URI of the matched lemma. These cases are referred to as 1:1 matches. In instances where the same lemma–PoS combination corresponds to multiple entries in LiLa, the system returns all relevant URIs, constituting 1:N matches. Conversely, when no entry in the Lemma Bank corresponds to the lemma–PoS pair produced during lemmatisation, the system returns no URI. These instances are classified as 1:0 matches. The output of this task is in Table 2.

Inevitably, the output of the lemmatisation process was not definitive. The accuracy of the 1:1 matches amount to around 97%. However, in certain cases, incorrect URIs were assigned.[8] One common source of error was the lemmatiser's assumption that any word beginning with a capital letter should be classified as a proper noun (PROPN). As a result, nouns occurring at the beginning of a sentence were sometimes misclassified, leading to erroneous matches when a proper noun homograph exists for a regular noun (e.g., *Amor*, the Roman god of love, versus *amor*, the common noun for 'love'). Another frequent error involved the lemmatisation of *quod*, which was uniformly tagged as a pronoun (PRON), despite its potential to function as a subordinating conjunction (SCONJ) or determiner (DET), depending on its syntactic role in the sentence. Similarly, *quam* was consistently tagged as a subordinating conjunction (SCONJ), although it could also serve as a pronoun (PRON) or a determiner (DET). Errors can arise for various reasons. As a result, the lemmatisation output was subjected to systematic manual review and correction by trained annotators, as well as disambiguation of 1:N matches.[9] Some of the one-to-zero matches also resulted from errors in lemmatisation or tokenisation. In particular, it was necessary in all instances to segment tokens containing enclitics, such as *-que*, *-ne*, and *-ue*, in order to enable accurate matching. For example, in tokens like *socialemque* 'and (something/someone) social', *eritne* 'will it be', *licetne* 'is it allowed', and *practicumue* 'or (something) practical', proper token splitting was required so that appropriate URIs could be assigned to both the first token (noun, verb, or adjective)

[8] Manual intervention was required in approximately 3% of the 1:1 matches subset.

[9] For an explanation of why 1:N matches regularly arise in the process of linking a textual resource to the LiLa LB, see the detailed report on how homography was handled during the integration of the LASLA corpus into the LiLa Knowledge Base [7, pp. 30–31].

and to the enclitic.

3. Papal Encyclicals in LiLa: Adding New Lemmas

Following the disambiguation process, several lemmas remain unlinked to LiLa, as they are not yet represented in the Knowledge Base. A thorough analysis of the 1:0 match types is necessary before considering their inclusion in the Lemma Bank. A subset of these unmatched lemmas corresponds to non-Latin words, which are not intended to be integrated into the Knowledge Base. These include: non-Latinized anthroponyms, such as *Nietzsche*, *Dostoevskij* (LF), *King*, and *Al-Tayyeb* (FO); words transliterated into the Latin alphabet from other languages, e.g., *emûnah* from Hebrew or *didachés* from Greek (LF); acronyms such as *DNA*, *OGM* (LS); and compound words joined by a hyphen or other special characters, such as *Deo-Amen* (LF) or *Rio+20* (LS).

In addition, a specific subset of the 1:0 matches—consisting of orthographic variants, dialectal forms, or alternative spellings of standardized forms—required targeted handling. In accordance with the OntoLex model used in LiLa, these cases have been incorporated as written representations (`ontolex:writtenRep`) of existing lemmas already present in the Lemma Bank [8, p. 69]. Specifically, these cases result from greater accuracy in transliteration from Hebrew (*Bethlehem*, LF; *Hillel*, FO), from the gemination of the sibilant in the toponym *Assisium* (FO; present in the Lemma Bank as *Asisium*), from the abandonment of a more Hellenising or archaic spelling of *Babilonia* (LS; listed in the Lemma Bank as *Babylonia*), and from a different graphical representation of the consonant cluster [ks] in *exstraneus* (FO; found in the Lemma Bank as *Extraneus*). These examples may reflect a modernising tendency in Latin spelling practices adopted by the Vatican, possibly aimed at aligning Latin orthography more closely with modern Italian spelling conventions (cf. *Assisi*, *Babilonia*, *Estraneo*). The same tendency will be noted again in later parts of the analysis.

The lemmas that have been added to the LiLa Knowledge Base, on the other hand, can be classified into three main categories.

The first category of lemmas added to the Lemma Bank includes Latinised anthroponyms and toponyms. Among the anthroponyms are *Desmondus*, *Martinus Luterus*, *Irenaeus* (FO), *Ludouicus* (LF), the patronymic *Alighierius* (LS), *Teresia*, and *Bonauentura* (LS, LF). These figures, cited in the Encyclicals, can play one of two roles: that of *auctoritas* or *exemplum*. In the case of Dante Alighieri, Saint Bonaventure, Saint Irenaeus, and Ludwig Wittgenstein, Pope Francis primarily refers to their words and works to support his arguments. For example, he cites Canto XXXIII of Dante's *Paradiso*, particularly the verse "l'amor che move il sole e l'altre stelle", to emphasise that "God's love is the fundamental moving force in all created things" (LS, 77).[10] Similarly, he references Wittgenstein's *Vermischte Bemerkungen*, where the philosopher discusses the "connection between faith and certainty" (LF, 27), and Irenaeus of Lyon's *Adversus haereses*, particularly the passage that uses the metaphor of melody to explain how different sounds can come from the same composer, just as each of us comes from the same Creator (FO, 58). By contrast, Desmond Tutu, Martin Luther King Jr., Mother Teresa of Calcutta, and Saint Thérèse of Lisieux serve as *exempla* to be emulated: for their acts of universal brotherhood despite religious differences, their faith in suffering, and their daily gestures of love and peace. As for toponyms, the category includes *Australia*, *Columbia*, *Corea*, *Croatia* (FO), and *Zelandia* (LS), all appearing in the genitive case following *episcopi*, as well as *Congus* (LS) and *Hiroshima* (FO), cited respectively as examples of the importance of preserving land and biodiversity, and of the moral imperative not to forget historical tragedies to which "we must never grow accustomed or inured" (FO, 248).

The second category consists of ethnic adjectives. Of the 15 instances found, 12 appear for the first time in the encyclical *Laudato si'*, two in *Fratres omnes*, and only one in *Lumen fidei*. From a derivational morphological perspective, these adjectives can be divided into three main types. The largest group (10 lemmas) consists of denominal adjectives derived from a toponym with the suffix *-ensis* (*Basileensis*, LS), including its extended form *-iensis* (*Canadiensis*, LS), a suffix typically used in Latin for forming ethnic adjectives [9, p. 439]. The second group includes *Apparitiopolitanus*, *Boliuianus*, *Paraguaianus* (LS), *Nazarethanus* (LS, FO), and *Bonaeropolitanus* (FO), formed with the equally canonical suffix *-anus* [9, p. 410]. A further distinction, intersecting with the previously discussed category of Latinized toponyms, concerns the nature of the geographical names from which these adjectives are derived. Some are adapted borrowings (**Basilea* from *Basileensis*), while others seem to be structural calques [10, pp. 118, 122], such as **Flumenianuarius* (from *Flumenianuariensis*, "of the city of Rio de Janeiro", LS). Some of these calques may undergo an additional morphological process, i.e. compounding with the Greek lexeme *polis*, resulting in forms like **Apparitiopolis* (from *Apparitiopolitanus*, "of the city of Aparecida", LS) and **Bonaeropolis* (from *Bonaeropolitanus*, "of the city of Buenos Aires", FO). Morphologically, the adjectives belong either to the second declension with two endings (first group) or to the first declension (second group), depending on the suffixation process. Semantically, the

[10] The text of this and other encyclicals is available in several languages at: https://www.vatican.va/content/francesco/it/encyclicals.index.html.

adjectives occur in different contexts: some appear in the genitive plural linked to *episcopi* (6); others refer to cities where documents, declarations, or environmental agreements were signed (4); two are characteristic attributes of female saints. *Bonaeropolitanus* refers to the positive influence of Jewish culture in Rio de Janeiro, while *Nazarethanus*, in both instances, occurs in the feminine form, dependent on *familia*.

The final category of new lemmas linked to the Lemma Bank consists of neologisms. The introduction of new lexical units into the inventory of a language can occur not only through internal resources and mechanisms, but also by drawing on elements from other languages, either through borrowing or calquing [11, p. 281]. In the present case, the linguistic influence is unidirectional, from Italian to Latin, which is unsurprising, given that Italian, although descended from Latin, is a living language with an active speaker community, unlike Latin. However, what is particularly noteworthy is that some of the Italian terms themselves are the result of interference from other languages. These layers of influence have contributed significantly to the enrichment of the Latin lexicon recorded in the LiLa Knowledge Base. Across the three encyclicals of Pope Francis under consideration, 234 neologisms have been identified, though they are not evenly distributed. In the first and shortest encyclical (see Table 1), *Lumen Fidei*, 32 neologisms appear for the first time. In the second, *Laudato si'*, 126 new formations are attested. Finally, in the third and longest encyclical, *Fratres omnes*, 76 neologisms are recorded.

Before proceeding with the analysis of this final category, a preliminary methodological clarification is required. In 1992, the *Libreria Editrice Vaticana* published the *Lexicon Recentis Latinitatis* (hereafter LRL), a lexicon that translates into Latin "many new words introduced by this era", generated "while preserving the norms of philology and the character of the Latin language".[11] This lexicon was fundamental for aligning word forms in the Encyclicals with the corresponding correct lemma. However, its application has also revealed the need for updates. Of the 234 lemmas analyzed, 145 are attested in the LRL. The remaining 89 were manually reconstructed by observing the word forms in their textual context. In some cases, reconstruction was straightforward; in others, it was not possible to determine the lemma with certainty. In these cases, the principle of analogy was applied. For instance, among the 36 neologisms formed with the suffix *-ismus*, half are found in the LRL. Of the remaining 18, only four appear in the nominative case. For the other 14, given the absence of modifying adjectives that could disambiguate gender (and therefore the case, which might otherwise suggest a nominative in *-ismum*), the lemmas were assigned masculine gender and classified as second-declension nouns with nominative in *-us*, based on analogy with the attested forms. For instance, the tokens *dynamismum* (LF, accusative) or *deconstructionismi* (FO, genitive), are entered into the KB as *dynamismus* and *deconstructionismus*, respectively. These reconstructions follow the model of lemmata such as *fatalismus* and *determinismus*, both attested in the LRL, or *anthropocentrismus* (LS), which is already found in the nominative form within the corpus. Another example involves nine lemmas pertaining to the semantic field of "Chemistry and Mineralogy" (see below). Among these, six such as *carbonium* (LS) and *fermentum* (FO) are present in the LRL as Latin equivalents of 'carbon' and 'enzyme' respectively, and are clearly neuter nouns of the second declension. One more, *dioxydum* (LS), appears in the nominative. By analogy, the word forms *cyanido* and *nitrogeni* were reconstructed as *cyanidum* and *nitrogenum* and added to the KB as neuter second-declension nouns. Moreover, the LRL further reflects a modernizing tendency in the lexical choices of the Latin used in the Encyclicals. A number of Italian terms that the LRL renders using periphrasis—in accordance with its assertion that "Latin is less suited (than Greek) to compounding words into one"[12]—reappear in the Encyclicals as single new lexical items. These are often modeled directly on Italian, incorporating morphological adaptations. For example, the Italian noun *totalitarismo* is translated in the LRL as "absolutum civitatis regimen", but appears in both *Lumen Fidei* and *Fratres omnes* as *totalitarismus*. Similarly, the adjective *mammifero*, which is translated in the LRL as "belua mammans", appears in *Laudato si'* as *mammiferum*, clearly modeled on the Italian form. Having established the necessary methodological premises, we can now proceed with an analysis of the neologisms. These may be classified into adjectives, nouns, and verbs.

As for adjectives, there are a total of 99. From a morphological perspective, 68 are first-class adjectives; 3 are first-class adjectives ending in *-ius* (*communitarius*, *consumptorius*, *fragmentarius*); 27 are second-class adjectives with two endings; 1 is a second-class adjective with a single ending (*globalizans*, present participle of **globalizo*). From a derivational standpoint, first-class adjectives are typically denominal, formed using the suffixes *-icus* (*atomicus*) and *-osus* (*gasiosus*), which are commonly employed in Latin for this type of morphological construction [9, p. 1125]. The nouns from which these adjectives are derived originate from Ancient Greek (*agnosticus*),[13] Classical Latin (*Prometheicus*), Medieval Latin

[11] Author's translation from Latin: "multa verba nova, quae haec aetas induxit" and "servatis normis disciplinae philologae et indole linguae Latinae" [12, p. 7].

[12] Author's translation from Latin: "linguam Latinam minus aptam esse (quam Graecam) ad componenda verba ita ut in unum coalescant" [12, p. 7].

[13] From this point on, only one example per source language or variety is cited. The list is not intended to be exhaustive; this editorial choice was made for space reasons. The linguistic analysis con-

(*inclusiuus*), Scientific Latin (*electricus*), Modern Latin (*aestheticus*), and Ecclesiastical Latin (*encyclicus*). They also result from interlinguistic influence between Italian and modern languages such as French (*acusticus*), English (*romanticus*), Czech (*roboticus*), German (*nazistus*). The second-class adjectives with two endings are formed either through suffixation with *-alis/-aris* (*structuralis*, *polaris*) or with *-bilis* (*renouabilis*). These are derived from nouns of various origins: Greek (*theologalis*), Classical Latin (*optionalis*), Medieval Latin (*interdisciplinaris*), Late Latin (*exsistentialis*), Scientific Latin (*molecularis*), Legal Latin (*solidalis*), and modern languages, such as English (*internationalis*). Remaining within the scope of derivation, it is particularly noteworthy that many of the neologisms exhibit prefixal or compositional structures prior to suffixation. These include prefixoids such as *inter-* (as in *interdisciplinaris*, *internationalis*), *multi-* (*multilateralis*, *multinationalis*, *multipolaris*), and *trans-* (*transgeneticus*, *transnationalis*). Other frequent compositional elements include bases such as *anthropo-* (*anthropocentricus*, *anthropologicus*), *auto-* (*autonomus*, *autotestimonialis*), and *techno-* (*technocraticus*, *technologicus*). Particularly prominent is the suffixoid *-logicus* (*methodologicus*, *oecologicus*, *technologicus*), which highlights how these adjectival neologisms respond to the growing need for terminology that addresses the study of the human being, its place within an increasingly interconnected world, the technologies it produces, and the discourse surrounding it.

As for new nouns, there are 131 in total. Morphologically, the majority belong to the second declension (64, of which 22 are neuter), followed, at a significant distance, by the third declension (33, of which 4 are neuter and 1 masculine), the first declension (32, with only 1 masculine noun, *asceta*), and finally, just 2 nouns belong to the fourth declension. Particularly interesting data emerge from the derivational morphological analysis of these nouns: 36 are denominal nouns formed with the suffix *-ismus*, which is used to create abstract nouns referring to religious, political, social, philosophical, literary, or artistic doctrines and movements (*dualismus*, *ascetismus*, *absolutismus*, *populismus*, *materialismus*, *romanticismus*), as well as attitudes, trends, collective or individual traits (*fanatismus*, *localismus*, *globalismus*), behaviors or actions (*fatalismus*), and even conditions or qualities, including moral or physical defects and harmful habits (*egoismus*, *narcissismus*). The high number of neologisms formed with this suffix clearly demonstrates not only the increasing need for its use but also its overuse in contemporary language.[14] There are also 11 nouns ending in *-tas*, all abstract and conveying a positive meaning, such as *actuositas*, *biodiversitas*, *solidarietas*, as well as nouns related to the sphere of the individual, such as *sacralitas*, *responsalitas*, *intimitas*, and *sexualitas*. Another noteworthy suffix is *-tio*, used to form deverbal nouns denoting actions, such as *dissentio*, *immigratio*, and *globalizatio*. Among the most common combining forms is *-logia* (from the Greek *logos*, and also the basis for the suffixoid *-logicus*, see above), which forms nouns such as *ideologia* and *oecologia*. Also worth noting is that, in the case of nouns as well, some of foreign origin have entered the Latin lexicon via Italian. Examples include *imanus* from Arabic (*imam*); three chemistry-related terms from French: *methanum*, *nitrogenum*, *dioxydum*; *mangrouia* from English; three terms with the combining form *gen-*, *genetica*, *genoma*, *genum*, from German. There are also nouns derived from Classical Latin (*uniuersalismus*), Late Latin (*reciprocitas*), Legal Latin (*solidalitas*), Medieval Latin (*represalia*), and Scientific Latin (*gasium*). Finally, particularly interesting from a derivational point of view are several structural calques from other languages: *tromocratia*, with its derivative *tromocratus*, from French *terrorisme* (from *terreur* + *-isme*); *autocinetum* or *autoraeda* from French *automobile*; *caeliscalpium* and *interrete* from English *skyscraper* and *internet*; and *ferriuia* from German *Eisenbahn*.

Finally, there are only four verbal neologisms. Of these, two belong to the first conjugation (*obstaculo*, *subordino*), one to the third conjugation (*interconecto*), while the remaining verb, *secumfert*, is classified as anomalous. This is due to its composition: it is formed by the enclitic attachment of the reflexive pronoun *se* to the preposition *cum*, followed by the verb *fero*, which itself is classified as an anomalous verb. From a derivational morphological perspective, three of these new verbs are the result of compounding, having been created by adding a prefix (*sub-*, *inter-*) or a prefixoid (*secum-*) to an already existing Latin verb. In contrast, *obstaculo* has undergone a derivational process, being a denominal verb derived from *obstaculum*, 'obstacle'.

From a semantic perspective, the classification of neologisms pertaining to the three parts of speech was conducted by mapping them to the 41 *domains*, defined as "spheres of activity or knowledge", established by BabelNet - a multilingual semantic network that integrates diverse resources, including WordNet, Wikipedia, the Italian WordNet and Wiktionary [13, p. 4560].[15] Across the three Encyclicals, and counting the occurrences of individual word forms, the neologisms most frequently attested (167 tokens) belong to the domain "Environment and meteorology", even though this domain comprises only nine lemmas. This result is unsurprising, consider-

ducted in this study is primarily based on the *Grande dizionario della lingua italiana*, available at https://www.gdli.it which served as the main reference for determining the historical and etymological origins of the lemmas.

[14] See, for example, the corresponding entry in the Treccani online dictionary at https://www.treccani.it/vocabolario/ismo/.

[15] For the process of identifying and refining domains, see *BabelDomains: Large-Scale Domain Labeling of Lexical Resources* [14].

ing that Pope Francis is widely regarded as one of the Popes most committed to environmental and climate-related issues. Notably, the adjective *ambitalis* alone appears 47 times. Ecology, represented through terms such as *oecologia*, *oecologicus*, and *oecosystema*, is a central theme of his pontificate. Throughout the texts, the Pope repeatedly reminds both global leaders and all people (*geosystema*) of their responsibility to protect and preserve biodiversity (*biodiuersitas*, *biosphaera*). This is followed by neologisms belonging to the domain "Philosophy, psychology and behavior" (58 lemmas, 159 tokens), "Culture, anthropology and society" (33 lemmas, 135 tokens), and "Politics, government and nobility" (21 lemmas, 103 tokens). As previously mentioned, philosophical reflection on the human condition is central to the Encyclicals, and is addressed from psychological (*actuositas*, *creatiuitas*, *egoismus*, *exsistentialis*, *infrahumanus*, *responsalitas*, *uulnerabilitas*), social (*communitarius*, *discriminatorius*, *ethicisticus*, *phyleticus*, *xenophobus*), and political (*absolutismus*, *demagogicus*, *nazistus*, *sinistrorsus*, *technocraticus*) angles. There is a noticeable drop in the number of occurrences for neologisms in the domain "Craft, engineering and technology" (8 lemmas, 39 tokens), which nevertheless reflect the idea of humanity as the primary agent of progress (*biotechnologia*, *nanotechnologia*, *technica*) and technological innovation (*roboticus*, *telegraphum*). At this point, and with the same number of occurrences (38) as those in the domain "Chemistry and mineralogy", appear the neologisms of the domain "Religion, mysticism and mythology" (22 lemmas). This is particularly significant, as one might have expected this to be among the most represented domains. The data instead confirm that the Encyclicals are not intended solely for Christian audiences, but are addressed to people of all faiths, promoting values intrinsic to the notion of humanity, not exclusively of Christianity. In fact, among the lemmas within this domain, only a few are explicitly tied to the Christian faith (*catechumenatus*, *christifidelis*, *christologicus*, *encyclicus*, *liturgia*, *trinitarius*), while others testify to the variety of world religions and belief systems (*agnosticus*, *ascetismus*, *dualismus*, *sacralitas*, *syncretismus*, *theologalis*). For the distribution of domains, see Figure 1 above.

The incorporation of new lemmas of modern and contemporary origin into the Lemma Bank, using the corpus of the three Encyclicals promulgated by Pope Francis between 2013 and 2020, has proven to be highly fruitful from both a quantitative and a qualitative standpoint. Undoubtedly, the efforts involved in the development and maintenance of a project such as LiLa—which was conceived as a network of interconnected language resources specifically for Latin—intersect with those of the Catholic Church, which continues to employ Latin as a universal language of communication. Both share a common goal: "to support the commitment to a greater knowledge and more competent use of Latin".[16]

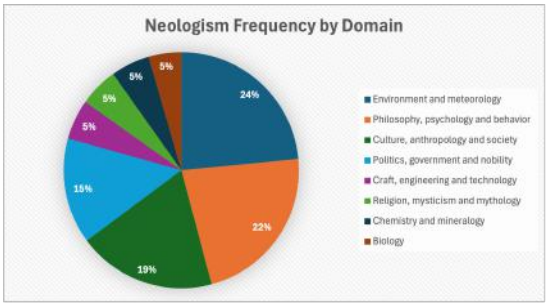

Figure 1: Distribution of neologism occurrences in Papal Encyclicals by Semantic Domain

4. Conclusions and Future Works

This paper has presented the integration of a new textual resource—the Papal Encyclicals corpus—into the LiLa Knowledge Base (KB). Although this is not the first instance of integrating a new corpus into LiLa—recent additions include Augustine of Hippo's *Confessiones*,[17] *de Ciuitate Dei*,[18] *de Trinitate*,[19] and Ovid's *Tristia* and *Epistulae ex Ponto* [15]—this first release of the Papal Encyclicals corpus is the result of a fine-grained manual revision of the automatic output. It constitutes a gold standard, whereas other textual resources linked to LiLa did not benefit from such an accurate manual revision - as in the case of the *Biblioteca Digitale di Testi Latini Tardoantichi*, where the considerably larger size of the corpus posed a limiting factor.[20] Furthermore, the inclusion of the Papal Encyclicals corpus is significant on a more fundamental level. A core assumption about Latin corpora is that they are static, since Latin is no longer a spoken language with native speakers. As a result, existing texts have been the subject of intense and ongoing scholarly investigation. For example, *Confessiones*, *de Ciuitate Dei* and *de Trinitate*, now linked to LiLa, have been studied for centuries from a variety of perspectives, ranging from psychological to strictly philological. Ovid's exilic writings have a long tradition of linguistic, historical and thematic analysis. In contrast, the Latin texts of Papal Encyclicals have not yet been the focus of consistent scholarly study. This means that the work presented in this paper is not built upon an

[16] Citation from the English version of the Apostolic Letter *Latina Lingua*, promulgated by Pope Benedict XVI on November 10, 2012. The full text is available online in eight languages at https://www.vatican.va/content/benedict-xvi/la/motu_proprio/documents/hf_ben-xvi_motu-proprio_20121110_latina-lingua.html.

[17] https://github.com/CIRCSE/AugustiniConfessiones.

[18] https://github.com/CIRCSE/AugustiniDeCiuitateDei.

[19] https://github.com/CIRCSE/AugustiniDeTrinitate.

[20] https://github.com/CIRCSE/digilibLT.

existing body of research, but is instead pioneering and foundational. It lays the groundwork for future studies and opens the door to a renewed consideration of Latin as a living language in specific, ongoing institutional contexts. Within the LiLa framework, the inclusion of a corpus that engages with contemporary concepts and referents significantly enriches the KB along several dimensions. First, the Lemma Bank has been expanded with new lexical items, enabling the study of linguistic strategies employed to create lemmas for concepts that did not exist in antiquity. This opens avenues for investigating the mechanisms of lexical innovation in Latin, particularly in the context of modern discourse. Second, the addition of the Encyclicals corpus offers a valuable opportunity to explore the distinctive linguistic and stylistic features of Papal Encyclicals as a genre. This resource allows for a more nuanced understanding of its rhetorical structures, specialised vocabulary, and register-specific phenomena. Third, the corpus contributes to extending the diachronic coverage of texts represented in the LiLa KB, facilitating longitudinal studies of Latin usage and lexical evolution across time. Future work will focus on expanding this initial integration to include the complete set of Latin Encyclicals authored by all Popes. This will support in-genre, cross-temporal comparisons, enabling scholars to trace linguistic trends and shifts within a consistent textual domain. Additionally, further analysis of unmatched lemmas and their potential inclusion will continue to refine the coverage and connectivity of the KB.

References

[1] M. Passarotti, F. Mambrini, G. Franzini, F. M. Cecchini, E. Litta, G. Moretti, P. Ruffolo, R. Sprugnoli, Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin, Studi e Saggi Linguistici LVIII (2020) 177–212. URL: https://www.studiesaggilinguistici.it/index.php/ssl/article/view/277. doi:10.4454/ssl.v58i1.277.

[2] C. Chiarcos, POWLA: Modeling Linguistic Corpora in OWL/DL, in: E. Simperl, P. Cimiano, A. Polleres, O. Corcho, V. Presutti (Eds.), The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27–31, 2012, Proceedings, number 7295 in Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, Germany, 2012, pp. 225–239. doi:10.1007/978-3-642-30284-8_22.

[3] C. Chiarcos, M. Sukhareva, OLiA – Ontologies of Linguistic Annotation, Semantic Web 6 (2015) 379–386. URL: https://www.semantic-web-journal.net/system/files/swj518_0.pdf.

[4] J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano, The OntoLex-Lemon Model: Development and Applications, in: Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference, Lexical Computing CZ s.r.o., Brno, Czech Republic, 2017, pp. 587–597. URL: https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

[5] M. Passarotti, M. Budassi, E. Litta, P. Ruffolo, The lemlat 3.0 package for morphological analysis of Latin, in: G. Bouma, Y. Adesam (Eds.), Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Linköping University Electronic Press, Gothenburg, 2017, pp. 24–31. URL: https://aclanthology.org/W17-0506/.

[6] M. Passarotti, F. Mambrini, G. Moretti, The services of the LiLa knowledge base of interoperable linguistic resources for Latin, in: C. Chiarcos, K. Gkirtzou, M. Ionov, F. Khan, J. P. McCrae, E. M. Ponsoda, P. M. Chozas (Eds.), Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 75–83. URL: https://aclanthology.org/2024.ldl-1.10.

[7] M. Fantoli, M. Passarotti, F. Mambrini, G. Moretti, P. Ruffolo, Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin, in: T. Declerck, J. P. McCrae, E. Montiel, C. Chiarcos, M. Ionov (Eds.), Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 26–34. URL: https://aclanthology.org/2022.ldl-1.4.

[8] F. Mambrini, F. M. Cecchini, G. Franzini, E. Litta, M. C. Passarotti, P. Ruffolo, LiLa: Linking Latin. Risorse linguistiche per il latino nel Semantic Web (AIUCD 2019), Umanistica Digitale (2020). URL: https://umanisticadigitale.unibo.it/article/view/9975. doi:10.6092/issn.2532-8816/9975, number: 8.

[9] G. Rohlfs, Grammatica storica della lingua italiana e dei suoi dialetti. Sintassi e formazione delle parole, volume 3, Giulio Einaudi editore, Torino, 1969.

[10] G. Gobber, Argomenti di linguistica, ISU Università Cattolica, Milano, 2003.

[11] G. Berruto, M. Cerruti, La linguistica. Un corso introduttivo, 2° ed., UTET Università, Torino, 2017.

[12] F. Latinitas, Lexicon recentis latinitatis, Libreria Editrice Vaticana, Urbs Vaticana, 1992.

[13] R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, F. Cecconi, Ten Years of BabelNet: A Survey, volume 5, 2021, pp. 4559–4567. URL: https://www.ijcai.

org/proceedings/2021/620. doi:10.24963/ijcai.2021/620, iSSN: 1045-0823.

[14] J. Camacho-Collados, R. Navigli, BabelDomains: Large-Scale Domain Labeling of Lexical Resources, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 223–228. URL: https://aclanthology.org/E17-2036/.

[15] A. Alagni, F. Mambrini, M. Passarotti, Lifeless Winter without Break: Ovid's Exile Works and the LiLa Knowledge Base, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 4–12. URL: https://aclanthology.org/2024.clicit-1.2/.