SSL

Exploring Latin epigraphy with Distributional Semantic Models: A pilot study

Lucia Tamponi, Alessandro Bondielli

Abstract

In the last few years, Distributional Semantic Models have been successfully applied to the analysis of both modern and ancient languages. In particular, Neural Language Models proved themselves to be a reliable tool to measure semantic relationships between words or documents based on their distributional properties. However, up to the time of writing distributional models have not been applied to the analysis of Latin inscriptions. In this paper, we describe a pilot study on two datasets of inscriptions from Rome and Southern/Central Italy and Sardinia included in the CLaSSES database of non-literary Latin texts (*http://classes-latin-linguistics.fileli.unipi.it*). Our results show that the model can identify both macro-classes and subclasses of inscriptions, thus contributing to the refinement of the classification already proposed in large epigraphic databases.

Keywords: Natural Language Processing, historical linguistics, Latin epigraphy, distributional semantics, corpus linguistics.

1. *Distributional semantics and Neural Language Models*

Distributional Semantics and the advent and growth of Neural Language Models (NLMs) in Natural Language Processing (NLP) and Computational Linguistics (CL) have revolutionized the field time and time again in the last ten years, by enabling computers to represent and interpret the meaning of words and documents in a continuous vector space.

The *Distributional Hypothesis* was first introduced in the 50's by the seminal works of Harris (1954) and Firth (1957), postulating the idea that semantically similar elements in a language tend to share the same, or similar, contexts (Sahlgren, 2008). The earliest attempts at

Received: February 2024
Accepted: July 2024

SSL LXII (2) 2024, pp. 31-56
doi: 10.4454/ssl.v62i2.395

translating the distributional hypothesis into practice involved creating co-occurrence matrices from large corpora. The mathematical definition of a matrix is a rectangular array of numbers, symbols or expressions that provide a mathematical representation of an object or one of its properties. Such objects or elements are arranged in rows and columns. A co-occurrence matrix M is a square matrix used to represent the frequency with which pairs of words co-occur within a specified context window in a given corpus. Let $V = \{w_1, w_2, \ldots, w_n\}$ be the set of all distinct words in the corpus, where n is the number of unique words. The co-occurrence matrix M is an $n \times n$ matrix where each entry M_{ij} represent the number of times word w_i and word w_j appear within a specified context window of each other in the corpus. The context window can be defined as a fixed number of words c that surround the target word. Note that in this case, each word in V is considered both a target word (i.e., the rows) and a context word (i.e., the columns). Each entry M_{ij} can contain either a boolean value (True or False), to indicate whether or not w_i and w_j co-occur, a frequency count, to indicate how many times w_i and w_j co-occur in the corpus, or other statistical measures typically derived from frequency (e.g., TF-IDF, Pointwise Mutual Information) indicating how salient the context word w_j is for the target word w_i. The co-occurrence matrix can then be used to assess the similarity of target words among each other by looking at how similar their distribution in the text is, i.e., how similar are the rows that represent each target word. Recall that for the distributional hypothesis, words with similar distributions tend to have similar meanings.

To assess the similarity of words, they can be compared based on a given similarity metric in the distributional space. One of the most widely adopted metrics is cosine similarity. Words with the highest cosine similarity among each other will have similar, or related, meanings.

Cosine similarity can be defined as follows. Assume that v_i and v_j are the word vectors corresponding to words w_i and w_j (i.e., the list of values in each row). Cosine similarity is given by the dot product of the two vectors divided by their magnitude. Cosine similarity ranges from -1 to 1, where a value of 1 means the vectors are identical, a value of 0 means that there is no similarity, and a value of -1 means that

the vectors are diametrically opposed. In the case of language, values typically range from 0 to 1. A graphical representation of a simplified 3-dimensional distributional space and cosine similarity among few words is shown in Figure 1. In the Figure, the words 'Cat', 'Dog', 'Snake', 'Chair', and 'Human' are used as target words, while 'Legs', 'Speak' and 'Breath' are context words. The direction of each vector represent the relationship between the target word and the considered context words, that are represented by the axes. For example, 'Human' will be equally distant from all the axes, as it has legs, it can speak and can breath; conversely, the 'Snake' vector will be much closer to the 'Breath' axis than to the other ones, and 'Chair' will be closer to the 'Legs' axis as it do not breath and cannot speak, and so on. The cosine of the angle ϑ between each target word vector represents their degree of similarity, i.e., a smaller angle means that two words are similar to each other. For example, 'Cat' and 'Dog' are very similar to each other in this specific context.

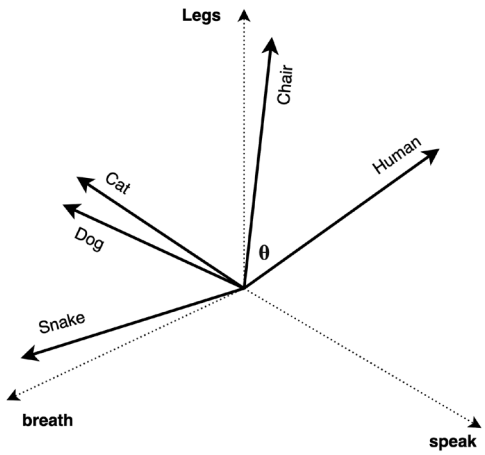

Figure 1. *An example vector representation.*

These early attempts laid the groundwork for later developments in computational linguistics via Neural Language Models. In the case of Neural Language Models, the distributional representation of words, called 'embedding', is not directly obtained from the co-oc-

currence matrices but rather learned using *Deep Neural Networks* on large corpora (Bengio *et al.*, 2000). The resulting word embeddings capture semantic relationships between words based on their distributional properties in text, enabling a wide range of NLP tasks such as semantic similarity measurement, text classification, and machine translation. The earliest and most prominent attempt at using Neural Language Models for creating word embeddings can be identified with Word2Vec (Mikolov *et al.*, 2013). The model learns the distributed representation of words in order to predict a word given its surrounding context or vice-versa.

Doc2Vec, an extension of Word2Vec proposed by Le and Mikolov (2014), extends the same principles to documents, learning fixed-length representations, or embeddings, for entire documents. This capability has made Doc2Vec invaluable for tasks such as document classification and information retrieval.

In the last few years, Transformer-based models have shown impressive capabilities at language modelling leveraging the *Attention mechanism* (Vaswani *et al.*, 2017), which enables capturing contextual relationships between words in a sequence. In essence, these models are language learners endowed with the capability to comprehend contextual nuances and predict subsequent elements within a given linguistic context. Models like BERT (Devlin *et al.*, 2018) and GPT (Radford *et al.*, 2019; Brown *et al.*, 2020) and subsequent developments such as ChatGPT, have achieved remarkable performance across a range of NLP benchmarks, owing to their ability to capture contextual information.

Despite the widespread success of Transformer-based LMs, some limitations that may apply in the specific context of the present work should be pointed out. First, such models are resource and data-hungry. Largest LMs training data is in the order of magnitude of trillions or tens of trillions of tokens. Smaller-sized LMs such as BERT can be trained with fewer data, but still require a large number of tokens to be trained effectively. For this reason, their application to low-resourced languages such as closed-corpus ones may not achieve the same performances as others based on higher-resourced languages (Sprugnoli and

Moretti, 2019; Bamman and Buns, 2020; Straka *et al.*, 2020; Sommerschield *et al.*, 2023). Second, their application on short and fragmentary texts (such as the Latin inscriptions examined in § 5) provides specific challenges that general models may have trouble facing. In these cases, more tailor-made solutions may provide the best descriptive prowess for the domain's unique syntactical and semantical characteristics.

2. *Diachronic distributional semantics*

Despite being low-resourced languages, in the last few years, distributional models were successfully applied to diachronic studies of Classical languages. Indeed, such examinations showed that it is possible to track semantic change via distributional models both in living languages such as English (cf. Sagi *et al.*, 2011; Hamilton *et al.*, 2016) and in closed-corpus languages such as Greek or Latin, including Neo-Latin (i.e., the Latin language as it was used from the post-medieval period onwards; Bloem *et al.*, 2020)[1]. As for Ancient Greek, for example, Rodda *et al.* (2017) examined Pre-Christian and Christian texts included in the *Thesaurus Linguae Grecae*, exploring the application of *Representational Similarity Analysis* to a large corpus of Ancient Greek spanning over 1000 years. Two semantic spaces were constructed based on Pre-Christian and Christian Greek corpora and then compared, and the results reflect the known evolution of the Greek language. Notable differences were found, especially in terms that acquired new Christian meanings (e.g., παραβολή "parable") or technical meanings in geometry and philosophy (e.g., πνεῦμα "spirit"). Furthermore, nearest neighbours analysis showed the capability of distributional semantics to capture semantic changes, such as the case of μοῖρα "part, portion", whose neighbours in the Christian context were specific to astronomy and geometry, indicating a specialized technical usage. As for Latin, the model was applied by McGillivray and Nowak (2022) to the LatinISE corpus. By examining

[1] For a more detailed evaluation of word embedding models for semantic change in Latin and Ancient Greek, see Perrone *et al.* (2021).

the nearest neighbours of the lemma *ciuitas*, the scholars were able to detect its semantic change from the meaning "citizenship" to "city". For example, among the nearest neighbours of *ciuitas* up to the 1st century BCE *gens*, *libertas* and *status* can be found, whereas in the 4th-9th century CE lemmas such as *urbs*, *villa* and *castellum* are detected. However, despite the interesting results achieved by the application of distributional models to Classical texts, they were not applied to Latin inscriptions up to the time of writing. Indeed, we believe that such an analysis could shed light on some peculiar features of epigraphic corpora, e.g., by highlighting similarities and differences among text types. Such regularities might escape the human eye, especially if a large number of texts is examined. For this reason, we decided to perform a pilot study on two datasets of inscriptions from Rome and Southern/Central Italy and Sardinia, spanning over several centuries – from the 4th century BCE to the 7th century CE.

3. *The corpus*

For our analysis, we examined the sections *Sardinia* and *Rome and Italy* of the CLaSSES corpus (*http://classes-latin-linguistics.fileli.unipi.it*) developed by the Department of Philology, Literature and Linguistics of the University of Pisa. CLaSSES is a resource which gathers non-literary Latin texts (inscriptions, writing tablets, letters) of different periods and provinces of the Roman Empire. Each text is annotated with extralinguistic and linguistic information, in order to examine variation phenomena in the texts (Marotta *et al*., 2020). This corpus was chosen for two main reasons. Firstly, it is entirely lemmatized; thus, it is particularly suitable for applying the methodology described in § 1. Secondly, each text is tagged with extra-linguistic information such as dating and text type, thus allowing us to take into account these variables. Although CLaSSES is structured in four sections (i.e., *Rome and Italy*, *Roman Britain*, *Egypt and the Eastern Mediterranean* and *Sardinia*) we chose to restrict our analysis to the sections *Sardinia* and *Rome and Italy* since they display a comparable number of tokens, al-

though belonging to different geographic areas and time frames. More in detail, the section *Sardinia* consists of 1184 inscriptions from the island (14413 words), dating between the 1st century BCE and the 7th century CE; the section *Rome and Italy* gathers 1250 inscriptions (11804 words) dating between the 4th century BCE and the 1st century CE[2]. In these sections, six text types are available, which were annotated primarily following the traditional classification proposed by *Corpus Inscriptionum Latinarum* and Warmington (1940). Thus, the following epigraphic genres are included: inscriptions dedicated to public figures and/or on public monuments (labelled as *tituli honorarii*), epitaphs and memorial texts (*tituli sepulcrales*), inscriptions dedicated to deities by people holding public offices and communities (*tituli sacri publici*) or by private citizens and brotherhoods (*tituli sacri privati*), texts carved on domestic and movable tools (*instrumenta domestica*) and military diplomas (i.e., personal legal documents on bronze tablets that contain a copy of imperial constitutions by which Roman citizenship and *conubium* was granted to veterans of the auxiliary army units, the fleet and the Praetorian Guard; Speidel, 2015; Rowe, 2015).

4. *Methodology*

Our goal was to understand whether some regularities emerge based on the classification of inscriptions, i.e., whether inscriptions from the same category appeared more similar to each other than to inscriptions categorized differently and whether we could identify similarities and differences among categories.

In order to do so, we chose to leverage a Doc2Vec model. As pointed out in § 1, using a pre-trained LM may not be ideal, especially in the specific content taken into consideration, as the inscriptions in the CLaSSES corpus are often short and possibly incomplete, containing many out-of-vocabulary items that a pre-trained LM may not model well. Thus, we argue that a document-based model specifically trained

[2] For a more detailed illustration of these sections, see MAROTTA *et al.* (2020) and TAMPONI (2022).

on our data may enable a better description of its composition, making similarities and differences more visible.

First, we trained a Doc2Vec model on each portion of our Dataset (i.e., *Rome and Italy* and *Sardinia*). Each text was preliminarily integrated and fully lemmatized. This step was essential to reduce the sparseness caused by the peculiar text type under analysis. Unlike other sources, epigraphic texts display a high number of abbreviations, fragmentary words and non-standard variants, whose presence could have likely resulted in suboptimal model performance.

The model's implementation is based on the Gensim Python Library[3]. We empirically selected training parameters and used the same parameters for training both models. Training parameters are as follows: alpha=0.025, min_alpha=0.00025, min_count=1, dm=1, epochs=200, vector_size=100.

After training, we extracted from the model the embeddings of all the documents. Each document embedding has a length of 100.

Second, to visualize the data in a smaller-dimensional space, we use the t-distributed stochastic neighbour embedding (t-SNE) transformation (Van der Maaten and Hinton, 2008). It is a popular dimensionality reduction algorithm particularly suited for reducing high-dimensional data into 2 or 3 dimensions. The algorithm consists of two main phases: first, a probability distribution is constructed that assigns a high probability value to each pair of points in the original high-dimensional space if the two points are similar, and a low probability value if they are dissimilar; then, a second probability distribution analogous to this is defined in the reduced-dimensional space. The algorithm minimizes the Kullback-Leibler divergence of the two distributions through gradient descent, reorganizing the points in the reduced-dimensional space. We chose to use two dimensions and used 23 as our random seed value for initialization. Note that the resulting dimensions, used as x and y axes in plots for the rest of the paper, do not express a specific meaning, as they are merely the result of the application of the t-SNE algorithm to the vector space obtained with Doc2Vec.

[3] See *https://radimrehurek.com/gensim/* (last accessed: 21.02.2024).

5. *The data*

5.1. *Sardinia*

The results of our analysis of the section *Sardinia* are reported in Figure 2, where the resulting multi-dimensional space has been reduced to two dimensions (§ 4).

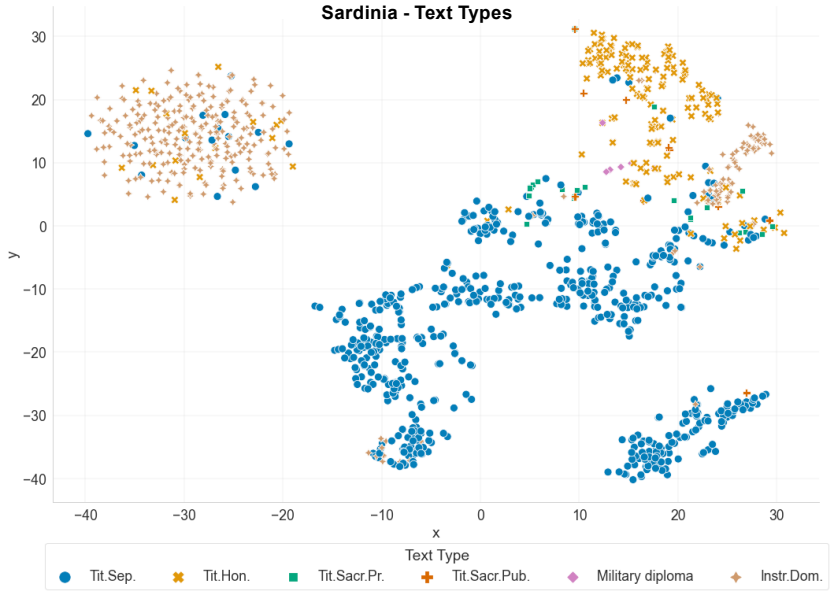

Figure 2. *Vector Space Model: section* Sardinia.

In the Figure, each point in the space corresponds to an inscription; the shape and shade of grey in the plot refer to its assigned text type. The annotation of text types adopted in CLaSSES partially follows the traditional classification of epigraphic texts (see e.g., Warmington, 1940). As was mentioned in § 4, this section includes both public and private inscriptions, as well as funerary ones and military diplomas.

In Figure 2, we can immediately spot separate clusters of inscriptions. A first interesting datum concerns the difference between public and private texts: most *instrumenta domestica*, carved on domestic and movable objects, form an independent cluster, located on the left

of the space (marked with the cross-shaped symbol and highlighted in light grey in the Figure). This configuration is mainly due to the characteristic brevity of the texts, which consist of a single word in most cases. This is the case, for example, of the seals on oil lamps *ILSard* II 471 a-b, c-d (text: *Pulla*(*e*)*ni* / *Pullaenoru*(*m*)) and *ILSard* II 394 a-d (text: *Agri*), mentioning the name of the workshop. However, a second group of *instrumenta domestica* is closer to public inscriptions and is located on the right side of the semantic space (highlighted with the light grey cross-shaped symbol in Figure 2). Indeed, these texts show a different layout from the other *instrumenta domestica*, since they display longer texts with typical features of dedicatory inscriptions, i.e., the indication of the social status of the figure mentioned or the presence of the preposition *ex* in the mention of the workshop. This is the case, for example, of the seals from Acte's workshops (*CIL* X 8046, 9a-e: *Actes Aug*(*usti*) *lib*(*ertae*))[4], and vases and tiles mentioning the producers' workshop (*CIL* X 8046, 7: *Ex figlinis Lucillaes* / *Quartionis*, *CIL* X 8046, 11: *Ex figlinis* / *Avidi Quieti*).

It may be objected, however, that the restricted group of *instrumenta domestica* consisting of only one word is not entirely comparable with the other longer texts included in the corpus. Although very short inscriptions can be problematic for our models, *instrumenta domestica* constitute a substantial category of texts for epigraphic studies, and we preferred not to exclude them to preserve an adequate number of texts as well as textual variability. However, the medium length of the other text types is 15 words for inscriptions. Thus, this feature might affect our results. However, a clear grouping based on text type is detected even if we exclude *instrumenta domestica* from our analysis, as is shown in Figure 3.

[4] As is known, the presence of Nero's freedwoman Claudia Acte in Sardinia is attested mostly by public and private epigraphic material (see e.g., the epistyle of the temple situated in Olbia *ILSard* I 309; Mastino, 2005: 129). The various Sardinian stamps on roof tiles from Olbia, Bolotana, Casteldoria and Macomer also testify to the existence of Acte's workshops on the island.

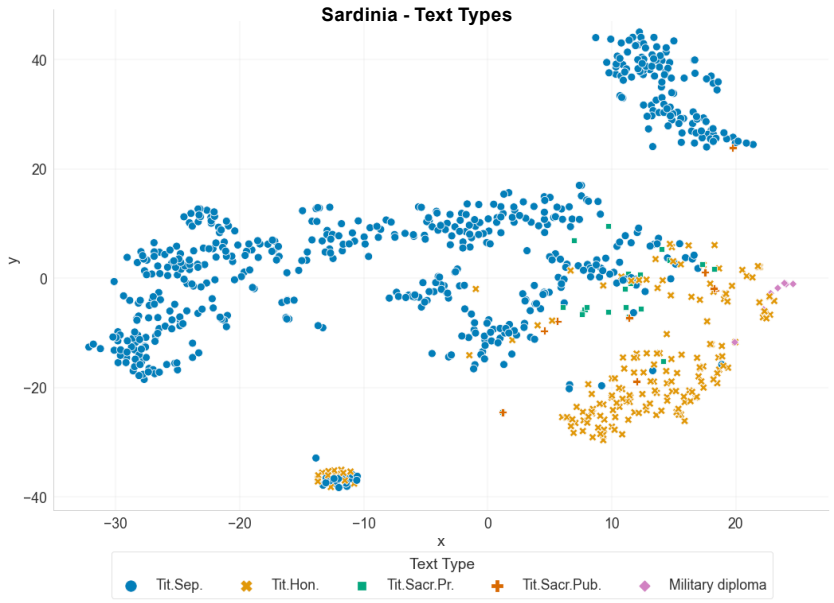

Figure 3. *Vector Space Model: section* Sardinia *(excluding* instrumenta domestica*).*

Interestingly, four quite distinct clusters can be identified. A group of public inscriptions (highlighted in light grey and marked with a cross-shaped symbol) can be detected. The few sacred texts available also concentrate on the central part of the space (square symbol); however, their low number prevents us from making reliable observations. The clusters are thus distinguishable but not highly homogeneous, with overlapping elements, particularly between funerary and public inscriptions. These two categories often share lexical components, such as proper names and public offices – referring to the deceased in funerary inscriptions and to the honoree in public ones – along with dates and numerals. In funerary inscriptions, numerals typically indicate the age of the deceased, while in public inscriptions, they are used for dating purposes. A more interesting datum, however, concerns the configuration of funerary inscriptions (highlighted in dark grey and marked with a circle-shaped symbol), which seem to form two distinct clusters. To explain this variability, it is vital to take into account the dating of the texts. As we mentioned in

§ 3, the section *Sardinia* covers a broad time span, from the 1st century BCE to the 7th century CE. As is known, the formulae adopted in funerary inscriptions changed considerably during the Christian era all over the Empire. For example, the invocation to the Manes (*Dis Manibus*) is a typical *incipit* of funerary texts dating to the first three centuries of the Empire, whereas it is absent in Christian texts. The first attestations of the formula were identified by Solin (1971) in Rome belong to the age of Augustus (e.g., *CIL* I^2 761, 1273), and it became more and more frequently adopted during the Imperial age (Tantimonaco, 2013). However, the formula *Dis Manibus* was later subjected to *damnatio*, so that it is very rarely attested in Christian inscriptions (Caldelli, 1997; Carletti, 1997; Tantimonaco, 2013). Our inscriptions perfectly conform to this trend, with 260 texts over 310 dating before 400 CE displaying a reference to the Manes[5]. On the other hand, formulae such as *requiescit in pace* or *bonae memoriae* (followed by the name of the deceased) are only or mainly found in Sardinian inscriptions from the 4th century onwards, which contributes to the differentiation between the texts[6].

This difference is by no means exclusive to Sardinia: however, this result deserves our attention since our methodology allows us to identify some diachronic variation within the same text type and area even in a relatively small epigraphic corpus.

Finally, the small cluster of inscriptions located in the lower part of the semantic space consists of outliers, i.e., of 'anomalous' fragmentary inscriptions and of texts consisting of only one or two words (e.g., *ANRW* B31: *M. Aelius*). The presence of this cluster seems to be due to an inescapable limit to the application of such a methodology to closed-corpus languages. Such systems would need a much higher amount of data to perform successful training: however,

[5] See e.g., *ILSard* I 260 (Porto Torres, 2nd-3rd c. CE): *D*(*is*) *M*(*anibus*) / *Heracula* / *vix*(*it*) *ann*(*os*) *XXX* / *VI mens*(*es*) *IIII* / *d*(*ies*) *XVII fec*(*it*) *Sa*/*turnina con*/*iugi b*(*ene*) *m*(*erenti*); *CIL* X 7957 (Porto Torres, 51-250 CE): *D*(*is*) *M*(*anibus*) / *Proculus* / *colonus bixit* / *annis XXXV fe*/*cit uxor bene* / *merenti*.

[6] See e.g., later texts such as *CIL* X 7757 (Cagliari, 4th-5th c. CE): *B*(*onae*) *m*(*emoriae*) / *Fortunatus* / *qui vixit annis* / *pl*(*us*) *m*(*inus*) *XL quievit* / *in pace d*(*ie*) *N*(*onarum*) *No*(*vem*)*br*(*ium*); and *CIL* X 7768 (Cagliari, 304-500 CE): *B*(*onae*) *m*(*emoriae*) *Proiectus* / *qui vixit an*(*nos*) *XXIII* / *recessit* / *d*(*ie*) *VII K*(*a*)*l*(*endas*) *Feb*(*ruarias*) *in pace*.

the limited number of available texts does not allow this condition to be fulfilled – especially after the elimination of *instrumenta domestica*, which covered one-fourth of the corpus. However, we reserve to tackle this issue shortly, by training the system on a larger corpus.

5.2. *A case study:* ICS *FTR-003* - Forum Traiani*, OR, 301-500 CE*

Our methodology can also be of use for the classification of 'hybrid' texts, such as funerary/sacred Christian inscriptions. Such inscriptions are not easily classified in the traditional framework proposed above: should they be categorised by function (funerary vs sacred) or by type of monument (tomb vs altar, etc.)? These categories overlap, and it is often problematic to decide how to classify the texts. As for the function, epitaphs could have the double purpose of identifying the individual's burial place and indicating the location for rituals; they could also appear on a variety of monuments, such as altars, *stelae*, *cippi*, etc. (see e.g., Cooley, 2012: 127 ff.).

From a theoretical perspective, this problem can be connected to the well-known broader theoretical issue concerning the limitations of categorizing entities into closed categories, which often fails to account for the inherent complexity and fluidity of real-world phenomena. By overlooking the nuanced and multifaceted nature of entities, such a classification system may lead to potential misinterpretations, especially in complex fields like Latin epigraphy. As is well-known, several attempts to overcome the limits of strict categorization (known as the 'Classical theory of categorization', see e.g., Locke, [1697[1]] 1960; Carnap, 1932) have been proposed by scholars belonging to different fields, including psychology and cognitive linguistics. One of the most well-known proposals is the prototype theory of categorization (Rosch, 1973; 1975; 1978), positing that categories are not defined by fixed boundaries but by a set of prototypical examples that exhibit the most representative features of that category. This approach highlights the inherent fluidity and contextual nature of categorization, as categories often overlap and blend into one another. This theoretical framework is particularly relevant when considering the categorisation challenges posed by the 'hybrid' Latin inscriptions described above

since they can incorporate elements from multiple contexts, blending commemorative language with religious symbolism in ways that defy rigid categorisation as either funerary or sacred. Just as prototypes in the prototype theory embody the core attributes of a category while allowing for peripheral variations, these inscriptions represent a convergence of funerary and sacred themes, underscoring the complexity and nuanced nature of ancient textual artefacts.

A case in point is the inscription *ICS* FTR-003 from *Forum Traiani* (OR, 301-500 CE; Figure 4).

ICS FTR-003
Forum Traiani (OR), 301-500 CE
(*H*)*ic effusus est sangu*(*is*)
beatissimi martyris
Luxuri celebratur
natale eius XII C(*a*)*l*(*endas*) *S*(*e*)*p*(*tem*)*b*(*re*)*s*
Renobatu sup temporibus Helia ep(*is*)*c*(*o*)*p*(*i*)

Figure 4. ICS *FTR-003 (image from* EDCS *database).*

The inscription is located on a marble slab built into the south side of San Lussorio church (almost 1.5 km from Fordongianus, province of Oristano). The sanctuary is located on the place of the saint's martyrdom (dated to the 4th century CE), which is thought to coincide with the saint's burial place. The text dates to the 6th century CE and commemorates the renewal of the sanctuary. The incipit (*hic*) corresponds to the typical starting of funerary inscriptions (*hic situs est*) – although in this case, it refers to the place of martyrium (Zucca, 1988: 23). A further analogy with typical funerary texts is to be found in the indication of the *dies natalis* of the martyr, corresponding to the day of Luxurius' death, which parallels with the indication of the time of death of the deceased in funerary texts. However, also a sacred element is found due to the reference to martyrdom; furthermore, in analogy with sacred inscriptions, reference is made to the bishop who took care of the renewal (*renobatu sup temporibus Helia ep*(*is*)*c*(*o*)*pi*).

As one might expect, the classification of this inscription is problematic. On the one hand, it could be categorized as a sacred text based

on its function and the reference to the *renovatio*; however, it can also be regarded as a funerary inscription. A clear classification is generally not proposed in the literature – not even by updated epigraphic databases such as the Epigraphic Database Roma (EDR number EDR081951)[7] or the Epigraphic Database Clauss-Slaby (EDCS-ID: EDCS-05200298)[8]. In the CLaSSES database, we classified the inscription as a *titulus sacer*, giving more prominence to its sacred element, with a certain degree of uncertainty.

In such cases, our methodology could shed some light on the possible categorization of the text. As is shown in Figure 5 below, in our semantic space the inscription (highlighted by the grey circle) clearly clusters with the other later funerary texts, although it is situated at the external periphery of the cluster, which might suggest a possible classification as a funerary inscription.

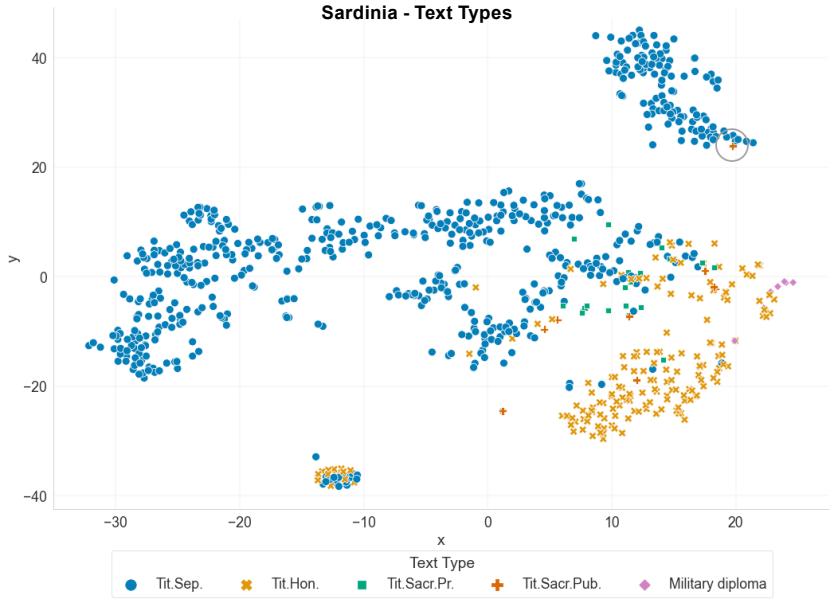

Figure 5. ICS *FTR-003 in the semantic space.*

[7] See *http://www.edr-edr.it/default/index.php* (last accessed: 21.02.2024).
[8] See *http://www.manfredclauss.de/* (last accessed: 21.02.2024).

This clustering is semantically justified: the incipit features the lemma *hic*, as in the funerary formula *hic situs est*; reference is made to lemmas such as *sanguis* and *episcopus*, which can be found in funerary Christian inscriptions. The sacred elements included in the text – absent in 'prototypical' funerary inscriptions – could also explain its location at the edges of the cluster. Acknowledging the complexity of inscriptions merging commemorative formulas with religious symbolism, our methodology seems to suggest that this text displays some of the core features of a 'prototypical' funerary inscription while allowing for some peripheral variation represented by its sacred elements. It can thus be categorized as a 'non-prototypical' funerary inscription. In conclusion, we believe that this case study could provide examples for reflection to epigraphists and a refinement of the classification of the texts proposed in epigraphic databases, in the light of contemporary epigraphic uses.

5.3. *Rome and Italy*

Before delving into the discussion of the inscriptions from Rome and Italy, a methodological issue needs to be addressed. Whereas the Sardinian material belongs to a delimited geographical and political area, the section *Rome and Italy* gathers inscriptions from various locations, i.e., Rome (for the majority of the texts), Central Italy and Southern Italy. Furthermore, due to the low number of available texts, we could not subdivide urban and non-urban inscriptions in CLaSSES. As is known, this is an important variable for the linguistic analysis of Latin texts, since literary texts and metalinguistic comments of ancient authors can – with due caution – provide testimony that Roman citizens were aware of the contrast between *urbanitas* and *rusticitas* from early times, considering the Latin spoken in Rome superior to that of non-urban places. For instance, this viewpoint is evident in Plautus' comparison (*Truc*. 687-91) between the Praenestine word *conea* with the Roman *ciconia* (Adams, 2007: 121). Also Cicero's rhetorical treaties provide notable instances of such references, distinguishing between *urbanitas* and *rusticitas*. Of course, caution is due when interpreting such comments, since the authors might confuse diatopic variation with other deviations from the standard, so that

stigmatized features adopted by the Roman citizens might be misleadingly defined as rural (Adams, 2007: 146-147; see e.g., the various interpretations proposed for Cic. *Orat*. 161 on the mission of final *-s*; Marouzeau, 1911; Belardi, 1965; Adams, 2007; Weiss, 2009; Pezzini, 2015; Marotta and Tamponi, 2019). Still, the testimonies by authors such as Plautus, Cicero and Varro described above hint at an opposition between an urban, more prestigious variety, as opposed to a non-urban, more 'vulgar' one[9]. While acknowledging the relevance of the linguistic opposition centre vs periphery in ancient Rome, we could not assess this variable in our corpus, since the resulting subcorpora would have been too small for our methodology to be applied.

Despite these shortcomings, however, our results still deserve our attention. Coherently with the inscriptions from Sardinia, a separate cluster of private inscriptions (highlighted in light grey) can also be spotted for the section *Rome and Italy* (Figure 6).

Like Sardinia, the cluster on the left-hand side of the Figure is composed of private texts consisting of one or two words, such as the Campanian vases from Cales and Capua, where the name of the producer is reported (e.g., *CIL* I^2 405,a,b-g, text: *K. Atili(o)*). In this section, however, the category of *instrumenta domestica* displays a higher degree of internal variation, with both small or fragmentary inscriptions and longer texts. This variation is reflected in our semantic space: detached groups of private texts are located on the edges of the central areas of the space (in light grey) and consist of private inscriptions displaying different types of longer texts, such as the augural messages inscribed on the oil lamps (e.g., *CIL* XV 6201-6204: *annu(m) nov(u)m faustu(m) felice(m) tibi / mihi*). Such longer texts blend some of the characteristic features of public and private inscriptions: following the theoretical premises put forward in § 5.2, they could be considered as 'non-prototypical' private inscriptions. Similarly to public texts, they are longer than 'prototypical'

[9] For reasons of brevity, the issues connected with the opposition between *urbanitas*, *rusticitas* and *antiquitas* in Ancient Rome cannot be discussed at length here. However, for an overview of the opposition between *urbanitas* and *rusticitas* and its relevance for (socio-) linguistic analysis, see Mancini (2006), Adams (2007: 121 ff.) and the recent contribution by Rovai (2020) on the distribution of accusative-based and ablative-based forms of frequency adverbs indicating ordinal rank in a sequence.

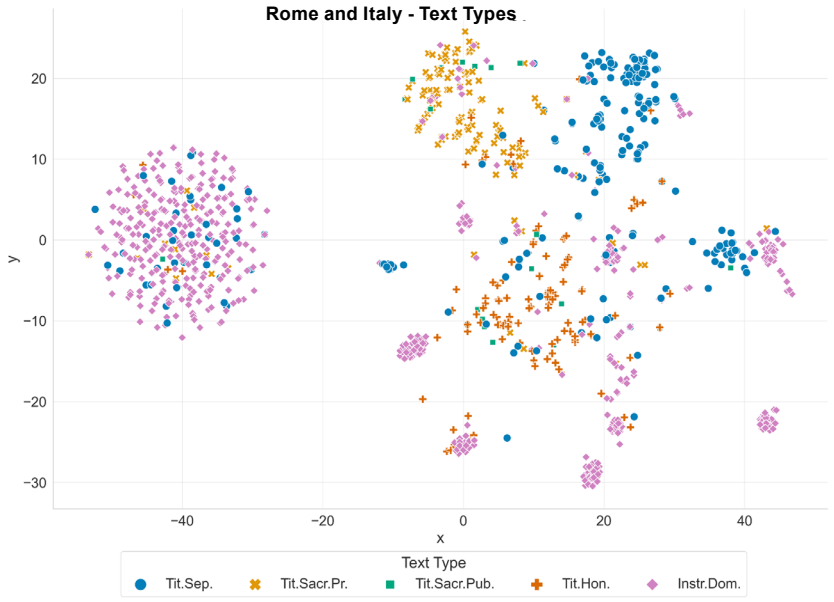

Figure 6. *Vector Space Model: section* Rome and Italy.

private inscriptions; however, in line with private texts, they feature augural messages on movable objects. In this respect, our methodology mirrors the non-prototypical nature of those inscriptions, by collocating them at the fuzzy edge of the two categories.

As was done for Sardinia, we also decided to apply our methodology to a more coherent corpus of non-private inscriptions. The results are displayed in Figure 7.

When private inscriptions are excluded from the analysis, some clusters are still visible, although the data are more sparse than those observed for Sardinia. More in detail, a small group of sacred inscriptions is detected in the upper part of the Figure (highlighted by the black circle), and a cluster of funerary texts is found on the right-hand side of the space (highlighted by the light grey circle). The official inscriptions (highlighted by the dark grey circle) are scattered in the central part of the space. Finally, as was discussed for Sardinia, the small group of texts detached by the other data (located in the lower left corner of the space) consists of outliers, i.e., fragmentary or extremely

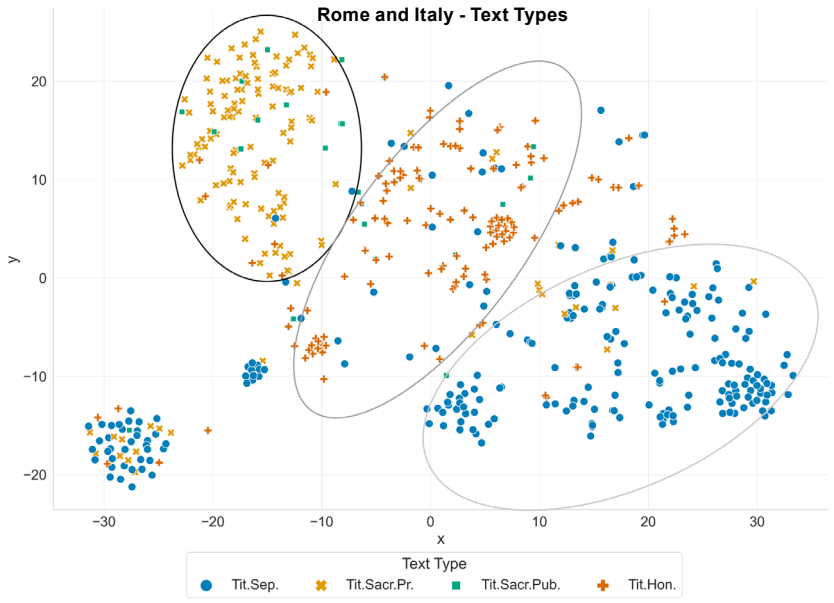

Figure 7. *Vector Space Model: section* Rome and Italy *(excluding* instrumenta domestica*).*

short texts that are categorized as 'anomalous' by the system for the reasons outlined in § 5.1. More generally, however, we also detected a higher data sparseness concerning the Sardinian material. This result is probably due to the above-mentioned geographic and political variety of this section, which gathers inscriptions from various parts of Italy from both urban and non-urban areas. We reserve to address this issue in the near future, by including other inscriptions in our data set so that it will be possible to select a more coherent corpus (e.g., of inscriptions from Rome) with an adequate number of texts.

6. *Conclusions*

The application of the Doc2Vec model described in this paper allowed us to examine the distribution of variants in the epigraphic texts. The model appears to be more successful when geographically and po-

litically delimited areas are examined, provided that an adequate number of data is available. More in detail, the model can identify both macro-classes of inscriptions, differentiating between public and private texts, as well as their subclasses, distinguishing between funerary, sacred and official inscriptions. Thus, the application of the model can serve as a good verification tool for the classification proposed for large epigraphic corpora and databases, detecting classification errors that might escape the human eye. Finally, the application of our methodology can provide interesting insights into the classification of problematic texts, such as the funerary/sacred inscription described in § 5.2. For this reason, we believe that such models could offer points for reflection both to linguists and epigraphists and contribute to the refinement of the classification proposed in large epigraphic databases. In the near future, we plan to apply the model to larger epigraphic databases, potentially those containing inscriptions spanning broader time periods and collected from different regions, such as the Epigraphic Database Clauss-Slaby or the Epigraphic Database Roma. On a macro level, this would allow for comparisons between inscriptions from various regions of the Empire, both synchronically and diachronically. On a micro level, comparisons could be made between individual cities with adequate epigraphic density. Finally, we will experiment with training the model on non-lemmatized corpora to assess potential data sparsity. Moreover, we plan to develop a more refined Language Model, by training on additional data and using more recent LMs such as BERT as our starting model, in order to leverage their advantages as well.

Acknowledgements

We express our sincere gratitude to Giovanna Marotta, Alessandro Lenci, and Francesco Rovai for their unwavering support and valuable input from the preliminary phases of this research. We also thank the anonymous reviewers for their insightful suggestions and constructive observations, which have significantly strengthened this paper. Our research was undertaken within the PRIN Project 'Ancient languages and writing systems in contact: a touchstone for language change' (2017JBFP9H).

Conflict of interest and Authorship disclosure

There is no conflict of interest between any of the individuals involved in the publication process. The present paper was conceived and discussed by both authors. For academic reasons only, the scientific responsibility is attributed as follows: §§ 1 and 4 to Alessandro Bondielli, §§ 2, 3 and 5 to Lucia Tamponi; § 6 was jointly conceived by the authors. All authors have approved the final version.

References

Adams, J.N. (2007), *The Regional Diversification of Latin, 200 BC-AD 600*, Cambridge University Press, Cambridge.

ANRW = Sotgiu, G. (1988), *L'epigrafia latina in Sardegna dopo il C.I.L. X e l'E.E. VIII*, in Temporini, H. and Haase, W. (1988, *Hrsg.*), *Aufstieg und Niedergang der römischen Welt (ANRW), II: Principat, 11.1*, De Gruyter, Berlin / New York.

Bamman, D. and Buns, P. (2020), *Latin BERT: A Contextual Language Model for Classical Philology*, arXiv:2009.10053 [available online at *https://arxiv.org/abs/2009.10053*, accessed on 15.04.2024].

Belardi, W. (1965), *Di una notizia di Cicerone (*Orator *161) su* -s *finale latino*, in Schiaffini, A. (1965, *a cura di*), *Studi in onore di Alfredo Schiaffini*, Edizioni dell'Ateneo, Roma, pp. 114-142.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2000), *A neural probabilistic language model*, in Leen, T., Dietterich, T. and Tresp, V. (2000, eds.), *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, MIT Press, Cambridge (MA).

Bloem, J., Parisi, M.C., Reynaert, M., Oortwijn, Y. and Betti, A. (2020), *Distributional Semantics for Neo-Latin*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), Marseille, pp. 84-93.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.,

Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020), *Language models are few-shot learners*, in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F. and Lin, H. (2020, eds.), *Advances in Neural Information Processing Systems (N2020)*, Curran Associates Inc., New York, pp. 1877-1901.

Caldelli, M.L. (1997), *Nota su* D*(*is*)* M*(*anibus*)* *e* D*(*is*)* M*(*anibus*)* S*(*acrum*)* *nelle iscrizioni cristiane di Roma*, in Di Stefano Manzella, I. (1997, *a cura di*), *Iscrizioni dei cristiani in Vaticano*, Musei Vaticani, Città del Vaticano, pp. 185-187.

Carletti, C. (1997), *Nascita e sviluppo del formulario epigrafico cristiano: prassi e ideologia*, in Di Stefano Manzella, I. (1997, *a cura di*), *Iscrizioni dei cristiani in Vaticano*, Musei Vaticani, Città del Vaticano, pp. 143-164.

Carnap, R. (1932), *Uberwindung der Metaphysik durch logische Analyse der Sprache*, in «Erkenntnis», 2, pp. 219-241.

CIL I^2 = Lommatzsch, E. (1918), CIL I^2 2,1: *Inscriptiones vetustissimae*, De Gruyter, Berlin (including the relative *Addendae*).

CIL X = Mommsen, T. (1963 [1883^1]), *Corpus Inscriptionum Latinarum, vol. X Inscriptiones Bruttiorum, Lucaniae, Campaniae, Siciliae, Sardiniae latinae*. Fasc. I, section *Pars posterior inscriptiones Siciliae et Sardiniae comprehendens*, De Gruyter, Berlin.

CIL XV = Dressel, H. (1969 [1899^1]), *CIL XV, 2: Inscriptiones urbis Romae Latinae. Instrumentum domesticum*, George Reimer, Berlin.

Cooley, A.E. (2012), *The Cambridge Manual of Latin Epigraphy*, Cambridge University Press, Cambridge.

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2018), *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805 [available online at *https://arxiv.org/abs/1810.04805*, accessed on 15.04.2024].

Firth, J.R. (1957), *A synopsis of linguistic theory 1930-1955*, in Philological Society (1957, ed.), *Studies in Linguistic Analysis*, Blackwell, Oxford, pp. 1-32.

Hamilton, W.L., Leskovec, J. and Jurafsky, D. (2016), *Cultural shift or linguistic drift? Comparing two computational measures of semantic change*, in Su, J., Duh, K. and Carreras, X. (2016, eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, November 1-5*, Association for Computational Linguistics, Austin, pp. 2116-2122.

Harris, Z. (1954), *Distributional structure*, in «Word», 10, 23, pp. 146-162.

ICS = Corda, A.M. (1999), *Le iscrizioni cristiane della Sardegna anteriori al VII secolo*, Pontificio Istituto di Archeologia Cristiana, Città del Vaticano.

ILSard I = Sotgiu, G. (1961), *Iscrizioni latine della Sardegna (supplemento al* Corpus Inscriptionum Latinarum*, X e all'*Ephemeris Epigraphica*, VIII).* Vol. 1, CEDAM, Padova.

ILSard II = Sotgiu, G. (1968), *Iscrizioni latine della Sardegna*. Vol. 2: Instrumentum domesticum. *I. Lucerne*, CEDAM, Padova.

Le, Q. and Mikolov, T. (2014), *Distributed representations of sentences and documents*, in Xing, E.P. and Jebara, T. (2014, eds.), *International Conference on Machine Learning, 22-24 June 2014, Beijing, China*, ML Research Press, Beijing, pp. 1188-1196.

Locke, J. (1960 [1697[1]]), *An Essay Concerning Human Understanding, abr. and ed. by A. S. Pringle-Pattison*, Clarendon Press, Oxford.

Mancini, M. (2006), Dilatandis litteris*: uno studio su Cicerone e la pronunzia 'rustica'*, in Cifoletti, G., Fusco, F., Gusmani, R., Bombi, R., Innocente, L. and Orioles, V. (2006, *a cura di*), *Studi linguistici in onore di Roberto Gusmani*, Edizioni dell'Orso, Alessandria, pp. 1023-1046.

Marotta, G., Rovai, F., De Felice, I. and Tamponi, L. (2020), *CLaSSES: Orthographic variation in non-literary Latin*, in «Studi e Saggi Linguistici», 58, 1, pp. 39-65.

Marotta, G. and Tamponi, L. (2019), *Omission of final* -s *in Latin inscriptions: Time and space*, in «Transactions of the Philological Society», 117, 1, pp. 79-95.

Marouzeau, J. (1911), *Notes sur la fixation et formation du latin classique I-V*, in «Mémoires de la Société de Linguistique de Paris», 17, pp. 266-280.

Mastino, A. (2005), *Roma in Sardegna: l'età imperiale*, in Mastino, A. (2005, *a cura di*), *Storia della Sardegna antica*, Il Maestrale, Nuoro, pp. 125-163.

McGillivray, B. and Nowak, K. (2022), *Tracing the semantic change of socio-political terms from Classical to early Medieval Latin with computational methods*, in *Latin vulgaire – latin tardif XIV. 14th International Colloquium on Late and Vulgar Latin. September 5-9, 2022, Ghent University*, Book of Abstracts, Ghent University [available at *https://www.lvlt14.ugent.be/wp-content/uploads/2022/09/LVLT14-Book-of-abstracts.pdf*, accessed on 06.03.2024].

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781 [available online at *https://arxiv.org/abs/1301.3781*, accessed on 15.04.2024].

Perrone, V., Hengchen, S., Palmac, M., Vatrid, A., Smith, J.Q. and Mcgillivray, B. (2021), *Lexical semantic change for Ancient Greek and Latin*, in Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y. and Hengchen, S. (2021, eds.), *Computational Approaches to Semantic Change*, Language Science Press, Berlin, pp. 287-310.

Pezzini, G. (2015), *Terence and the Verb 'To Be' in Latin*, Oxford University Press, Oxford.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019), *Language models are unsupervised multitask learners*, in «OpenAI blog», 1, 8.

Rodda, M.A., Senaldi, M. and Lenci, A. (2017), Panta rei: *Tracking Semantic Change with Distributional Semantics in Ancient Greek*, in «Italian Journal of Computational Linguistics», 3, 1, pp. 11-24.

Rosch, E. (1973), *Natural categories*, in «Cognitive Psychology», 4, 3, pp. 328-350.

Rosch, E. (1975), *Cognitive reference points*, in «Cognitive Psychology», 7, pp. 532-547.

Rosch, E. (1978), *Principles of categorization*, in Rosch, E. and Lloyd, B.B. (1978, eds.), *Cognition and Categorization*, Lawrence Erlbaum Associates, Hillsdale / New York, pp. 27-48.

Rovai, F. (2020), Consul tertium *o* consul tertio*? Dubbi metalinguistici, sincretismo e variazione nelle formule di iterazione delle cariche pubbliche*, in «Studi e Saggi Linguistici», 58, 2, pp. 33-63.

Rowe, G. (2015), *The Roman state: Laws, lawmaking, and legal documents*, in Bruun, C. and Edmondson, J. (2015, eds.), *The Oxford Handbook of Roman Epigraphy*, Oxford University Press, Oxford, pp. 3-20.

Sagi, E., Kaufmann, S. and Clark, B. (2011), *Tracing semantic change with latent semantic analysis*, in Allan, K. and Robinson, J.A. (2011, eds.), *Current Methods in Historical Semantics*, Mouton de Gruyter, Berlin / New York, pp. 161-183.

Sahlgren, M. (2008), *The Distributional Hypothesis*, in «Italian Journal of Linguistics», 20, 1, pp. 33-53.

Solin, H. (1971), *Beiträge zur Kenntnis der griechischen Personennamen in Rom*. Vol. 1, Societas Scientiarum Fennica, Helsinki.

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I. and De Freitas, N. (2023), *Machine learning for ancient languages: A survey*, in «Computational Linguistics», 49, 3, pp. 703-747.

Speidel, M.A. (2015), *The Roman army*, in Bruun, C. and Edmondson, J. (2015, eds.), *The Oxford Handbook of Roman Epigraphy*, Oxford University Press, Oxford, pp. 319-344.

Sprugnoli, R. and Moretti, G. (2019), *Word Embeddings for Processing Historical Texts*, poster presented at the Digital Humanities Conference 2019 (DH2019), Utrecht, the Netherlands 9-12 July, 2019 [available online at *https://dh-abstracts.library.virginia.edu/works/10077*, accessed on 15.04.2024].

Straka, M. and Straková, J. (2020), *UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings*, arXiv preprint arXiv:2006.03687 [available online at *https://arxiv.org/abs/2006.03687*, accessed on 15.04.2024].

Tamponi, L. (2022), *Variation and Change in Sardinian Latin*, Pisa University Press, Pisa.

Tantimonaco, S. (2013), *La formula* Dis Manibus *nelle iscrizioni della* Regio X, in «Polymnia: Collana di Scienze dell'Antichità. Studi di Archeologia», 5, pp. 261-278.

Van der Maaten, L. and Hinton, G. (2008), *Visualizing data using t-SNE*, in «Journal of Machine Learning Research», 9, 11, pp. 2579-2605.

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017), *Attention is all you need*, in Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (2017, eds.), *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Curran Associates, Inc., New York, pp. 5999-6009.

Warmington, E.H. (1940), *Remains of Old Latin*. Vol. 4: *Archaic Inscriptions*, Harvard University Press / Heinemann, Cambridge (MA) / London.

Weiss, M. (2009), *Outline of the Historical and Comparative Grammar of Latin*, Beech Stave Press, New York.

Zucca, R. (1988), *Le iscrizioni latine del* martyrium *di* Luxurius *(*Forum Traiani*,* Sardinia*)*, Editrice S'Alvure, Oristano.

Lucia Tamponi
Dipartimento di Filologia, Letteratura e Linguistica
Università di Pisa
Via Santa Maria 36
56126 Pisa (Italy)
lucia.tamponi@fileli.unipi.it

Alessandro Bondielli
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo 3
56127 Pisa (Italy)
alessandro.bondielli@unipi.it