CONTENTS

SECTION IV Methods

SECTION V Case Studies

CHAPTER 22

Linguistics and Inscriptions

FRANCESCA DELL'ORO

Introduction

It is genuinely encouraging to see how much progress has been made in recent years in integrating the linguistic analysis and enhancing the linguistic investigation of Ancient Greek and Latin inscriptions. Not long ago, it was still possible to write that "[e]ven though databases containing Greek and Latin inscriptions do exist, the syntactician who wishes to work on inscriptions has to face the problem that . . . annotated corpora of Greek and Latin inscriptions are still not available" (Dell'Oro 2015, 274; see, already, Adamik 2009, 13). The situation is changing considerably across all levels of linguistic analysis, and the main aim of this chapter is to illustrate how inscriptional resources are becoming increasingly responsive to the needs of linguistic research, what has already been achieved, and which future directions might be pursued (see Barmpoutis and Bozia, this volume).

This chapter stands at the intersection of historical linguistics, corpus linguistics—with its recent developments in natural language processing (henceforth NLP)—epigraphy, and the digital humanities. It begins by outlining certain features of inscriptions that distinguish them from the literary texts with which linguists—and, of course, philologists and literati—are typically more familiar. The second part presents a selection of openly accessible projects, offering an overview of the current state of the art. The chapter concludes with some reflections on possible future developments. Ultimately, it seeks to underscore the value of inscriptions not merely as historical-archaeological artifacts, but also as dynamic linguistic data, capable of enriching, complementing, and reshaping our understanding of language use within ancient life and culture.

Engaging with Inscriptions as a Linguist

Before outlining some characteristic features of inscriptional records, it is worthwhile to clarify the use of the term "corpus." In epigraphic studies, "corpus" refers to printed, digitized, or born-digital collections of texts that often aim to be *as complete as possible*—for example, the *Corpus Inscriptionum Latinarum* or the I.Sicily corpus. In linguistics, it refers to any collection of texts. However, in the field of corpus linguistics, "corpus" typically denotes a *representative selection* of texts (e.g., Stefanowitsch 2020, 22–23). Some scholars refer to ancient languages as "corpus languages," with reference to the (predominantly literary) texts preserved through the indirect tradition and their (tendentially) limited and finite nature (e.g., Langslow 2002, 23–24; Clackson 2011, 2; Dell'Oro 2015, passim; Cuzzolin 2019, 6). This usage contrasts with the concept of a corpus in corpus linguistics, where the selection is not the result of chance but rather a deliberate and informed choice of texts. There is therefore an inherent ambiguity in the use of the term "corpus" in many linguistic studies on ancient languages. Moreover, inscriptional material continues to grow (Adamik 2009, 13–14;

Prag 2021, 181–183), making the definition of Latin and Ancient Greek as "corpus languages" particularly ill-suited; "text language" would be a much better choice. In the context of digital resources, "corpus" as *any collection of texts* must be clearly distinguished from databases that catalogue forms or entries (e.g., Prag 2021, 185).

Inscriptions—together with, for instance, papyri—constitute direct attestations from antiquity (i.e., sources not mediated through the manuscript tradition). Methodologically, inscriptional texts cannot be "detached" from their physical support and treated solely as "texts" without forfeiting important information on various levels, including their linguistic aspects (e.g., Prag 2021, 188). Because of this, their linguistic analysis presents certain inherent challenges. Some of these issues are briefly outlined and discussed here (see also Dell'Oro 2015).

One issue is the lack of authorship in the traditional sense, or the difficulty in recovering the identity of the author of the textual part of the inscription. Behind the setting-up of an inscription, there is often a commissioner—a person or some collective entity (e.g., the senate or the *boulḗ*). This person or group may be the actual author of the text. However, the task of composing the text can also be given to another individual—for instance, a poet, in the case of a funerary epigram. Equally, the execution of the inscription depends on the type of material. Stone inscriptions presuppose the preparation of the stone to be inscribed, sometimes with lines drawn to guide the stonecutter in producing orderly writing (see, e.g., https://stonemasters.uw.edu.pl/workshop-identification). When a linguist encounters a form that deviates from the expected norm, a key issue may be whether it reflects a genuine linguistic variation or rather a mechanical mistake by the stonecutter (e.g., Adamik 2009, 18). Often, it is the recurrence of the same deviation in various inscriptions that helps to clarify the matter.

Another issue is text length. Some inscriptions consist of a complete text comprising just one word, often a personal name. For example, inscriptions from the necropolis of Praeneste are excluded from the CLaSSES corpus (discussed below) because they contain only personal names in the nominative case (Marotta et al. 2020, 49n16). On the other hand, it should be noted that much depends on the particular research questions in view. A project in Ancient Greek anthroponomastics such as LGPN-Ling (also discussed below) shows that diatopic, semantic, and even micro-syntactic analysis may be possible—and indeed meaningful—even with isolated forms (see the section "Corpora and Databases for Linguistic Investigation").

Some inscriptions may be fragmentary. This feature does not depend on the inscription itself but on its history and state of preservation. However, the issue of what is readable and what is supplemented remains a significant problem for linguistic analysis. In particular, NLP tools allow for the parsing of raw text but are unable to distinguish editorial signs. The I.Sicily Sketch Engine Corpus, for instance, provides annotated inscriptions without distinguishing between extant and supplemented letters, or between abbreviated and unabbreviated words (Fendel 2024), whereas the outputs of the automatic parsing needed to be manually corrected in the case of the inscriptions from the WoPoss corpus (see the section "Corpora and Databases for Linguistic Investigation").

Sometimes inscriptions exhibit archaisms, that is, features that do not reflect the actual evolution of the language but are introduced artificially—for instance, to lend the text a more solemn tone. For this reason, legal texts were excluded from the CLaSSES corpus (Marotta et al. 2020, 49n16). Archaicizing features have also been employed to showcase the skills of stonecutter workshops, as is probably the case with some Latin spellings in the well-known Ancient Greek–Latin bilingual inscription from Palermo (*IG* XIV 297 = *CIL* X 7296; see Alföldy 1989). Formulaic language is also typically regarded as an obstacle to linguistic investigation. Nonetheless, it is worth noting that formulae may reflect diaphasic variation and

are subject to diachronic change, thus offering valuable insights into linguistic evolution and cultural change (for an example, see Dell'Oro 2015, 275).

As is the case with literary attestations, text types and genres are not evenly distributed either chronologically or geographically in the case of inscriptions. This is clearly illustrated by the sub-corpora of the corpus CLaSSES (see the section "Corpora and Databases for Linguistic Investigation"). However, the availability of digitized material and/or inter-project cooperation also affects the material that can ultimately be included in a digital corpus.

Finally, with specific reference to the sociolinguistic (variationist) investigation of misspellings in Latin (imperial) inscriptions, a skeptical stance toward the usefulness of inscriptional evidence has been noted (see, e.g., Adamik 2009, 12; 2012, with the references therein). Indeed, Adams (2007, 6–8, with the references therein) describes inscriptions as the weakest form of evidence for diatopic variation across the Roman Empire. Adamik traces an initial wave of skepticism to the early decades of the twentieth century, when inscriptional material from certain provinces (Gaul, Spain, and Dalmatia), along with texts from some late antique authors such as Gregory of Tours, was thought to exhibit the same vulgarisms across all regions of the Roman Empire. From this, it was inferred that linguistic differentiation in Latin did not begin until after 600 CE. On the debate, see also Marotta et al. (2020, 40). It remains important to clarify the extent to which this debate can be extended to other levels of the linguistic analysis.

Corpora and Databases for Linguistic Investigation

Any digital epigraphic project holds considerable potential for linguistic research. Projects originally developed for historical purposes already yield valuable data for linguistics; for instance, annotations of provenance and dating enable the reconstruction of diatopic and diachronic variation, respectively. By contrast, the aim of this section is to highlight the few projects that explicitly annotate linguistic information using linguistic methodologies and tools with the goal of illustrating the methodologies and challenges involved in the linguistic annotation of inscriptions. Without claiming to be exhaustive, this chapter briefly outlines two types of openly accessible corpora and databases focused on Ancient Greek and/or Latin: those providing linguistically annotated editions of inscriptions, and those annotating specific linguistic phenomena. In describing these resources, the focus is placed on linguistic aspects, although they may encompass many other types of features.

Linguistic information may pertain to various levels of linguistic analysis—writing,[1] phonetics/phonology,[2] morphology,[3] syntax,[4] lexicon,[5] semantics,[6] pragmatics,[7] textual analysis,[8] and discourse analysis[9]—or to axes of (sociolinguistic) variation within the language as a diasystem—diatopic,[10] diastratic,[11] diaphasic,[12] diamesic,[13] and

[1] The study of writing systems, spelling, and script features.
[2] The study of speech sounds and sound patterns.
[3] The study of word structure and formation.
[4] The study of sentence structure and word order.
[5] The study of the vocabulary of a language, including word meanings and relations.
[6] The study of meaning in language.
[7] The study of meaning in context and language use.
[8] The study of how meaning is built across texts.
[9] The study of language in use across contexts and interactions.
[10] Variation across geographical space.
[11] Variation across social strata or groups.
[12] Variation according to communicative situation or register.
[13] Variation according to the communication channel (e.g., spoken vs. written).

diachronic.[14] Additional aspects—such as metrics in verse inscriptions or translation practices—may also be encompassed within linguistic investigation (see also Mambrini, this volume). In each subsection, the projects are arranged according to their date of inception, with multilingual endeavors preceding those focused solely on Ancient Greek or Latin. (See also Vierros and Yordanova on papyri, this volume.) Alongside a general overview of the annotation criteria and entries for each project, particular features or strategies are emphasized—for example, how new editions are handled, or how editorial updates are accommodated within a digital framework.

Epigraphic Corpora and Databases Providing Linguistic Annotation and Tools

The project Crossreads: Text, Materiality and Multiculturalism at the Crossroads of the Ancient Mediterranean (Prag 2021) is developing tools for investigating Sicilian epigraphy by expanding the I.Sicily corpus and focusing on the study of the following aspects: linguistic history, the materiality of text (in particular through petrography), and palaeography. The I.Sicily corpus aims to provide a newly edited digital collection of all the inscriptional records in the languages of ancient Sicily (Ancient Greek, Latin, Oscan, Messapian, Sikel, Elymian, Hebrew, (Neo)Punic, and Phoenician), covering the period from the seventh century BCE to the seventh century CE. Editions are prepared anew, based on autopsy wherever possible, and encoded using the EpiDoc schema.

Each inscription is documented according to the following categories: a unique "ISicXXXXXX" identifier; "Editorial Status"; an "Image" (when possible, high-resolution); "Previous Editions" (with link to a Zotero reference list); "Discussions"; "Digital Editions" and "Identifiers," including the Trismegistos identifier; "Location"—comprising "Provenance," "Original Location" (with a map), "Last Recorded Location" (with a map), and "Autopsy Date"; "Physical Description"—covering "Material," "Object Type," and "Dimensions"; "Epigraphic Description"—encompassing "Epigraphic Type," "Execution Type," "Letter Heights," "Interlinear Heights," "Letter Forms," and "Date"; "Text"—presented in an interpreted version alongside the diplomatic and EpiDoc editions; "Apparatus"; "Translation"; "Commentary"; "Contributions"—including "Editor," "Principal Contributor," "Contributors," and "Date of Last Revision"; downloadable "XML Source File"; and "Citation Information"—detailing how the edition should be cited and the URI (uniform resource identifier) of the inscription. Users can also share descriptions via social networks such as Facebook and Twitter. The EpiDoc-XML files are archived on Zenodo. Notably, when an edition is updated, the previous version remains preserved in the Zenodo repository, enabling users to access earlier versions that are no longer available in the online corpus. A search interface allows inscriptions—or groups of inscriptions—to be retrieved according to various criteria. They may also be consulted via an interactive map on the same webpage.

A sub-corpus of 723 linguistically annotated Greek and Latin inscriptions is available for upload in the Sketch Engine platform (Fendel 2024), a tool for corpus analysis. To annotate the sub-corpus, a tokenizer for EpiDoc texts (Crellin 2024) was developed. As the tokenizer removes the Leiden editorial conventions, users must consult the corresponding edition of the inscription in the I.Sicily corpus to determine the condition of the text (see the section "Engaging with Inscriptions as a Linguist"). Lemmatization and part-of-speech (henceforth

[14] Variation over time.

POS) tagging were performed using the PROIEL model via the UDPipe.[15] The results of the automatic annotation were then manually corrected, which led to the creation of a gold standard. The reference lemmas are in Attic-Ionic Greek and Classical Latin. Abbreviations have been expanded, raising issues comparable to those encountered with tokenization. The project also provides a syntax treebank annotator that is agnostic as to any annotation formalism (https://github.com/rsdc2/syntax-treebank-annotator), though it remains compatible with the annotation schemas available in Arethusa.

The multifaceted project Digital Marmor Parium (Berti 2016) includes a treebank annotated by G. G. A. Celano (see https://www.digitalmarmorparium.org/linguistics.html) according to the Ancient Greek and Latin Dependency Treebank formalism (Celano 2014), which also includes an "Advanced Syntax Layer / Semantic Layer." Importantly, this was the first instance in which a digital edition of an inscription was manually treebanked—that is, analyzed in terms of lemmatization, POS-tagging, morphosyntax, and syntactic dependencies. Treebanked inscriptional corpora are still extremely rare. A small treebank corpus of archaic and classical Greek inscriptions of the Euboean colonies of Sicily, based on the first volume of *Inscriptions grecques dialectales de Sicile* (*IGDS* I treebank; Dubois 1989), was recently released as open access. The manual annotation of this corpus served as a test case to identify several challenges in treebanking inscriptions—ranging from the selection of lemmas for dialectal forms (ultimately aligned with Attic-Ionic to ensure interoperability with other corpora) to the treatment of the peculiarities of the local Ancient Greek alphabets (preserved), the interpretation of inscriptional punctuation marks, the handling of ellipsis (with elliptical nodes introduced only when strictly necessary), and the management of certain editorial interventions—as detailed in Dell'Oro and Celano (2019). The treebank annotation of both the *Marmor Parium* and the *IGDS* I inscriptions was carried out using the Arethusa platform.

The project DĀMOS Database of Mycenaean at Oslo (Aurora 2025) makes all published Mycenaean documents available online for reading, browsing, and pedagogical use (Aurora 2017), and it endeavors to follow the text of the most recent reference editions. Discrepancies between editions are planned to be included in the "Notes" section. The search interface enables users to query archaeological, epigraphic, and content-related features of the texts, alongside linguistic information.

The section "Classification, Hands and Seals" allows users to search by selecting and combining features across the following categories: site, document, hand, stylus, series, subseries, set, and seal. Documents may also be retrieved via their reference sigla. The section "Chronological and Spatial Data" allows users to search by selecting and combining features across the following categories: chronology, find area, findspot, provenance (for vases), find year, location, and inventory number. Each text is also provided with a link to other relevant resources such as the LiBER (Linear B Electronic Resources) project (Del Freo, Di Filippo, and Rougemont 2024) for images and to the Trismegistos identifier. Additional metadata includes joins (in case of broken documents) and a basic bibliography.

It is possible to query the texts for individual words or multiple words with a wide range of selectable and combinable features. For example, under "word type" users may target categories such as common word (i.e., words spelled according to syllable-based phonetic principles), abbreviation, logogram, and measure unit. Search results can be downloaded

[15] Interested users can test the outputs of the models available for Ancient Greek and Latin here: https://github.com/WoPoss-project/automatic_annotation (Bermúdez Sabel and Dell'Oro 2020).

as CSV files. Morphosyntactic analysis of the forms will be available shortly. Competing readings and linguistic interpretations have been annotated and will also be released. The annotated files are currently being converted to the EpiDoc XML schema to ensure optimal interoperability (Aurora et al. 2025).

The project Iscrizioni Latine Arcaiche "Latin Archaic Inscriptions" (ILA; Rocca, Sarullo, and Muscariello 2017) will provide a digital edition of the archaic Latin inscriptions (seventh–fifth centuries BCE) considered particularly significant for the study of the Latin language, whether already published or hitherto unpublished. It is worth specifying for the non-specialist reader that the earliest non-fragmentary literary attestations of Latin date only from the comedies of Plautus in the third century BCE. Moreover, during the second half of the twentieth century, the corpus of Latin archaic inscriptions expanded considerably, and several inscriptions remain unpublished. Particular attention is paid to the relationship between the text and the object, combining epigraphic and linguistic approaches. Among the project's principal areas of focus is writing, with particular attention to the chronological and geographical distribution of signs and their variants. While a comprehensive description of the linguistic phenomena is still in progress, the linguistically relevant elements within the epigraphic analysis include "Position of the inscription on the object," which refers to the spatial relationship between the text and the inscribed item; "Scriptura," which details the technique employed in the physical realization of the inscription; "Direction of writing," indicating the orientation of the script; "Dividing signs," which concerns the presence of punctuation and its potential functions; "Dimensions of the letters," which are significant for understanding the visibility of the inscription in relation to both the object and its observer; "Epigraphic commentary," which provides a detailed description and analysis of each letter, addressing both formal characteristics (i.e., the shape and model of each character) and actual features (such as any distinctive traits in the execution), alongside more general observations.

These projects, though primarily epigraphic, incorporate linguistic tools and annotations of great value for linguistic research. The next section turns to corpora and databases specifically designed with linguistic analysis as their main objective.

Corpora and Databases Designed for Linguistic Research

With one exception, all the projects outlined in this section focus on Latin. This highlights an implicit imbalance in the digital development of Latin as compared to Ancient Greek linguistics.

The project Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE (CEIPoM; Pitts 2022) provides a linguistically annotated database of the languages of the Italian peninsula during the first millennium BCE. At the time of writing (CEIPoM 1.3), it includes almost all Messapic, Venetic, and Osco-Umbrian inscriptions, alongside Latin inscriptions up to approximately 100 BCE. The dataset is organized across four levels—Texts, Sentences, Tokens, and Analysis—each corresponding to a CSV table. The annotation of CEIPoM has been mostly compiled or corrected manually, though automatic annotation has also been implemented.

Texts are identified by "Text_ID"—the identifier within the corpus; "Reference"—the identifier in the reference edition; and "Name"—their informal designation, such as Fibula Praenestina. Language information includes "Language, Language_Family," "Language_Variety," and "Script." Contextual information encompasses chronological—"Date_after" and "Date_before"—along with geographical fields—"Provenance, GeoID" according to Trismegistos, "Latitude," and "Longitude." The fields "Finite_Verb," "Analysable_Token,"

and "Text_Length" facilitate the filtering of the data, for example, to exclude inscriptions that lack a finite verb or analyzable content.

Each sentence is identified by its "Text-ID," "Sentence-ID," "Sentence-position," and its strings of characters ("Sentence Field"). Where appropriate, some inscriptions are also subdivided into sections ("Section"). The first three fields also serve to identify the token and its analysis. Each token is further described by the fields "Token_ID," "Token_position," "Token," "Token_clean," "Relation," and "Head." Usefully, "Token-clean" provides the character sequence of the token stripped of special or non-alphabetic characters or in transliteration (in cases where the original characters are non-Latin). Syntactic analysis is provided at the token level, drawing on AGDT (Ancient Greek Dependency Treebank) 2.0, via the fields "Relation" (indicating the token's syntactic relation to its head) and "Head"—that is, the "Token_ID" of the syntactic head. Sentence punctuation—whether original or editorially supplied—is not annotated. Missing syntactic elements are represented by the sign "-" in the "Token" field.

Analysis at additional linguistic levels is identified by a dedicated "Analysis_ID." This includes information on lemma, morphology, semantics, phonology, and orthography. Owing to space constraints, a detailed account of each field cannot be provided here (for which, see the "Vademecum" at CEIPoM). It is worth noting that the terminology has been selected to facilitate cross-linguistic comparability within the corpus—see, for example, the use of A-stem in place of "first declension." These descriptions are also mapped onto a conventional set of nine-slot POS tags ("POS_code").

Semantic annotation, being the least frequently represented level of linguistic analysis, warrants particular attention. It includes the fields "Meaning" (an English translation), "Meaning_category" (e.g., proper nouns, body parts, relations), and "Meaning_subcategory" (e.g., "Personal," "Toponym," "Theonym," under the category of proper nouns). Further fields include "Classical_Latin_equivalent" (a close cognate or semantic counterpart) and "Classical_Latin_form" (its inflected Latin form). Variant forms may be retrieved by combining the fields "Standard_aligned" and "Form_aligned," as in the case of the classical form *curaverunt* and the archaic or archaicizing form *coeraverunt* ("they took care of").

The Computerized Historical Linguistic Database of the Latin Inscriptions of the Imperial Age (LLDB; Adamik 2009, 2016) provides access to Vulgar Latin as attested in Latin inscriptions, papyri, and parchment documents of the Roman Empire, contributing to a deeper insight into the dynamics that gave rise to the Romance languages and influenced the linguistic, ethnic, and cultural configuration of medieval and modern Europe. The acronym LLDB—standing for Late Latin Data Base—refers to the title of the project of József Herman that constitutes the original foundation of the Computerized Historical Linguistic Database. See Herman (1991) for his guidelines on data collection, which continue to underpin the methodology of the current project. Their revised and extended version is available at http://lldb.elte.hu/admin/doc_guidelines.php.

LLDB collects misspellings, defined as "deviations from the classical norm" (Adamik 2016, 15). It is important to emphasize that only misspellings are recorded in the database. This decision mainly reflects Herman's methodological approach, which focused on the distributional analysis of misspellings based on calculating the relative frequency of various types of spelling deviation per region. Nevertheless, the complete text is accessible via links to external digital resources listed in the section "Bibliography."

Each misspelling is assigned an identifier within the database (LLDB-number) and described according to the following categories: "Code" (and alternative code, when a second interpretation is possible)—the descriptor of the relevant linguistic phenomenon, referring to deviations from the classical norm as well as simple technical spelling errors (up to the

time of writing, the project has identified up to 562 codes); "Relevant text"—the word or the string of text that deviates from the classical "norm"; "Classical text"—the classical counterpart of the relevant text; "Bibliography"—references to the edition; "equivalent bibliography"—used to avoid duplicate recordings; "Localization"—province, city, or other place where the inscription was found or erected; "Date"—the date of the inscription; "Type of inscription"—this includes various features such as whether the text is Christian or not, whether it is in prose or verse, whether it is official or private, along with the material used; "Evaluability factors"—conditions that affect the reliability of the data (e.g., legibility, state of preservation, or whether the individuals mentioned were native Latin speakers); "Remark"—interpretive or noteworthy comments in one of six languages of the European Union; and "Collector"—the collaborator who entered the data.

Two search modules enable the combination of multiple search criteria. The results can be exported in Word and Excel formats. A particularly useful and relatively uncommon feature is the ability to generate charts illustrating the distribution of the selected features according to various parameters. Currently, the database is being further developed within the project DiLaDi (Digital Latin Dialectology: Tracing Linguistic Variation in the Light of Ancient and Early Medieval Sources), which aims to incorporate parchment charters, primarily from the domain of private law.

The Corpus for Latin Sociolinguistic Studies on Epigraphic textS (CLaSSES; Marotta et al. 2020) is designed to facilitate variationist phonetic-phonological and, with regard to morphological endings, morphophonological investigations. The linguistic annotation focuses on (ortho)graphic variants and, in particular, on misspellings, defined as "those spellings that are not congruent with the 'standard' language as exhibited in the literary texts of the Classical period" (Marotta et al. 2020, 40). The interpretation of such misspellings is intentionally left open to the researcher, who is invited to consider where they fall on a continuum between a mechanical miswriting and sociolinguistic variants.

Despite the adjective "epigraphic," the corpus—which initially comprised only inscriptional non-literary records—has expanded to include ink tablets, documentary papyri, and ostraca. The texts are drawn from various collections. As mentioned in the section "Engaging with Inscriptions as a Linguist," the record type is often strictly related to the provenance region. Of the four geographical sections, "Rome and Italy" contains 1,250 Latin inscriptions from the time period from the sixth century BCE to the first century CE; "Roman Britain" comprises 762 ink-written tablets from the auxiliary fort of Vindolanda, dating from the first century to the third century CE; "Egypt and the Eastern Mediterranean" includes 220 documentary letters written on papyri and ostraca from the period between the first century and the sixth century CE; and "Sardinia" features 1,184 inscriptions from the time period from the first century BCE and the beginning of the seventh century CE. (See also discussion of text type in the following paragraph.) It is worth noting that certain texts were excluded on principle, broadly reflecting issues raised in the section "Engaging with Inscriptions as a Linguist" such as the brevity and fragmentary condition of some documents. Examples include the inscriptions from the necropolis of Praeneste (brevity) and legal texts (archaisms).

The corpus can be searched by combining various criteria. The extralinguistic data includes the place of provenance and dating. The metalinguistic data covers several categories: "Text type"—whose categories vary by geographic section (for example, the section "Rome and Italy" includes *tituli honorarii*, *tituli sepulcrales*, *instrumenta domestica*, *tituli sacri privati*, and *tituli sacri publici*; the section "Sardinia" adds military diplomas to the mentioned types, while for the texts in the section "Egypt and Eastern Mediterranean" the texts are just classified as formal or informal); "Graphic form"—which indicates whether the word is complete, incomplete, abbreviated, editorially supplied in full, presumably

misspelled, graphically uncertain, or whether the token corresponds to a number or symbol (the latter referring to non-alphabetic signs occurring in the sections "Roman Britain" and "Sardinia"), and, finally, whether there are lacunae; and "Language"—which, in addition to Latin, encompasses Ancient Greek, Oscan, Umbrian, Persian, Etruscan, Iberian, Semitic, Hebrew, (Neo-)Punic, Egyptian, and Coptic, along with the categories "Hybrid" and "Unknown" (i.e., words of uncertain origin). For the letters from Vindolanda, metadata also includes the author and the addressee. The linguistic annotation is based on the identification and classification of "non-classical" (ortho)graphic spellings, each of which is consistently linked to its corresponding classical form. These variants are categorized according to the linguistic phenomenon involved: vowel alternations, marking of vowel length, vowel omission, diphthong-related phenomena, omission of final consonant, omission of nasal before consonant, assimilation, doubling or reduction of consonants, and <B>/<V> confusion. The annotation has been further refined to enable users to search for specific phenomena. Linguistic phenomena were subdivided according to whether they affect vowels, consonants, or morphophonology. More fine-grained distinctions have been introduced among the types of linguistic phenomena, but these cannot be outlined in detail here.

The sections, texts, and annotated features can be accessed through a dedicated search interface. Texts are also accessible via an interactive map on the website's home page. The searchable annotated features can be combined and the results exported in various formats. The corpus thus enables the evaluation of statistical evidence for specific linguistic phenomena. The most recent addition is the linking of CLaSSES's lemmas to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin (De Felice et al. 2023).

The project LGPN-Ling: Etymology and Semantics of Ancient Greek Personal Names (Minon 2020) comprises a digital database and a printed lexicon, the *Lexonyme*, which is in the process of being published in three volumes (for the first of these, see Minon 2023). LGPN-Ling's relevance to this chapter lies in the fact that, although personal names are not attested exclusively in inscriptional documents, many are attested exclusively or primarily in inscriptions. Moreover, inscriptional records provide access to diatopic variation—that is, the dialectal forms of the personal names—whereas in the manuscript tradition their spelling has often been normalized.

In the interface, the section "Name" presents the lemma in the Greek alphabet alongside its transliteration into Latin characters, as well as the genitive form (and its variants), which indicates the declensional class. A link to the Lexicon of Greek Personal Names database provides information on the number of attestations of the personal name (as a lemma) and its gender, along with the dates of its earliest and latest occurrence. An additional link directs users to the Trismegistos identification system. Where applicable, dialectal and gender variants are also listed in this section. It should be noted that the earliest attested variant is adopted as the lemma. The "Geolinguistic" section provides information on the region of first attestation and/or the regions where the personal name is most frequently attested. The following sections—"Bases" and "P(ersonal)N(ame) Suffixes"—are dedicated to the morphological analysis of the personal name. Each name contains at least one base (*root1*) and may also include a "prefix" and a second base (*root2*). A name may also include up to four suffixes, which are counted from the outermost one inward. The section "Bases Morphosyntax" provides information about the POS to which the roots belong, along with the hierarchal relationships between them (when more than one is present) in the form of X (head) and Y (dependent element). The section "Semantics" provides information about the interpretation of the personal name—that is, how to interpret it. The section "Lexis" provides sources and references—including links to the LOGEION dictionary (logeion.uchicago.edu/λόγος)—relevant to the semantic and lexical interpretation. It also offers

details about the relevant lemmas in the Ancient Greek (non-onomastic) lexicon. A final section, "References," presents linked sources and onomastic references—including Bechtel's *Historische Personennamen des Griechischen* (1964)—and various notes concerning, for example, the interpretation of the personal name. For instance, the name *Polyterpon* came to be associated with a general of Alexander the Great, following the latter's exploits.

The project A World of Possibilities. Modal Pathways over an Extra-Long Period of Time: The Diachrony of Modality in the Latin Language (WoPoss; Dell'Oro 2023) aims to reconstruct the evolution of modal forms throughout the history of Latin. Within the project theoretical framework, the notion of modality encompasses the domains of necessity, possibility, and volition. Both lexical items—such as *possum*, "I can"—and morphological markers—such as *–bilis*, "that can be X (e.g., done)"—are annotated. Modal passages are initially annotated via automatic processing using the Stanza NLP tool and the Perseus model for Latin (for tokenization, lemmatization, POS-tagging, morphosyntactic analysis, and dependency parsing). This phase is followed by manual refinement, including the correction of automatic annotations and the semantic analysis of modal passages. To ensure the open access of the corpus, the texts are drawn from open sources. However, to maintain philological quality, the text of the annotated passages has been modified in accordance with the modern reference editions.

Adopting a comprehensive view of Latin as a diasystem and seeking to maximize linguistic variation, the WoPoss corpus has begun to incorporate inscriptions. This process started with republican-era inscriptions containing modal passages (based on Warmington 1940 and the EDR EpiDoc dataset: see Panciera et al. 2020). Their integration into WoPoss marks the first annotation of inscriptions at the syntax–semantics–pragmatics interface, where semantics is understood in its linguistic rather than computational sense. WoPoss annotators conduct a detailed analysis encompassing the "Modal marker" (e.g., polarity), the "Modal scope" (e.g., ± dynamicity and ± control, in relation to the description of the modalized event), the "Modal relation" (e.g., dynamic, deontic, epistemic, among others), and the "participant" in the modalized event (e.g., ± animate). With respect to editorial conventions, the raw text parsed by Stanza retained the Leiden Conventions. Original spellings have been preserved alongside their normalized forms. Manual corrections were subsequently applied. This approach has enabled the provision of a linguistically annotated corpus while preserving essential editorial information.

Before concluding, it is worthwhile to briefly mention some other recently developed tools. One such example is the LatinNow WebGIS, created as part of the project Latinization of the North-Western Roman Provinces: Sociolinguistics, Epigraphy and Archaeology (LatinNow). This tool enables users to locate inscriptions from the (Roman) West, dated between 800 BCE and 800 CE, according to various criteria such as language zones, writing equipment, and types of settlements (civilian by type, civilian by status, or military). Within the same project, an annotation schema for bilingual and multilingual texts was also developed (Mullen and Willi 2024). A promising tool worth mentioning here is AGILe (Ancient Greek Inscriptions Lemmatizer), a lemmatizer specifically designed for Greek inscriptions (Graaf et al. 2022).

Owing to space constraints, this section does not address advances made in the relevant fields for other ancient languages. Notable examples include DigItAnt—the Platform for the Languages of Ancient Italy (Murano et al. 2023; see also Mallia et al. 2023 for the web application EpiLexO), which, at the time of writing, provides annotated inscriptions in Oscan, Faliscan, and Venetic; the list of lemmas and associated forms for Latin, Ancient Greek, Hebrew, and Aramaic available in the Inscriptions of the Israel/Palestine corpus; the Corpus of Gāndhārī Texts, which offers linguistic analysis of documents in Middle Indic varieties,

a dictionary linking forms to their source texts, and an accompanying grammar; and the linguistically annotated corpus of Hittite texts accessible via the Hethitologie Portal Mainz (portal of Hittitology).

Directions for the Future

This chapter has discussed some of the challenges associated with using inscriptional records in linguistic research and has offered an overview of several projects that either integrate linguistic analysis or focus primarily on linguistic investigation based on inscriptions. The challenges identified include: ensuring philological accuracy in both automatic parsing and the documentation of editorial interventions; establishing standards that promote maximum interoperability and reusability among projects based on the same inscriptional material but with differing emphases—historical, epigraphic, or linguistic; developing the technical and methodological expertise required to address the uneven availability of inscriptional sources across space and time; and enhancing the interaction between manual and automatic annotation by supporting the creation of gold standard corpora and refining automatically annotated texts through systematic manual correction.

This overview also allows for two key observations. First, as noted at the outset of this chapter, developments in the field are progressing rapidly. Second—and perhaps most refreshingly and encouragingly—NLP and digital humanities tools have drawn attention to the historical neglect of inscriptions and are helping to address that gap within the broader field of linguistics.

Looking ahead to the needs of linguists, the next generation of corpora, databases, and tools should, at a minimum, provide basic linguistic annotation—namely, tokenization and lemmatization—to enable lemma-based research and reduce reliance on keyword searches (Marotta et al. 2020, 42). This will, in turn, facilitate further lines of inquiry, such as the application of corpus linguistic methodologies (e.g., co-occurrence likelihood measures) to inscriptional corpora, alongside the integrated use of both documentary and non-documentary text collections.

Acknowledgments

I am deeply indebted to many colleagues who have provided information and data about their projects. I list them in alphabetic order: Béla Adamik, Federico Aurora, Giovanna Marotta, Sophie Minon, Alex Mullen, Giovanna Rocca, Jonathan Prag, and Giulia Sarullo.

Reference List

Adamik, Béla. 2009. "In Memoriam József Herman: Von der *Late Latin Data Base* bis zur *Computerized Historical Linguistic Database of Latin Inscriptions of the Imperial Age*." *Acta antiqua Academiae scientiarum Hungaricae* 49 (1): 11–22.

Adamik, Béla. 2012. "In Search of the Regional Diversification of Latin: Some Methodological Considerations in Employing the Inscriptional Evidence." In *Latin vulgaire—Latin tardif IX: Actes du IXe colloque international sur le latin vulgaire et tardif, Lyon, 6–9 septembre 2009*, edited by Frédérique Biville, Marie-Karine Lhommé, and Daniel Vallat, 123–139. Lyon: Maison de l'Orient et de la Méditerranée "Jean Pouilloux."

Adamik, Béla. 2016. "*Computerized Historical Linguistic Database of the Latin Inscriptions of the Imperial Age*: Search and Charting Modules." In *From* Polites *to* Magos: *Studia György Németh*

sexagenario dedicata, edited by Szabó Ádám, 13–27. Budapest: University of Debrecen Department of Ancient History.

Adams, James N. 2007. *The Regional Diversification of Latin 200 BC–AD 600*. Cambridge, UK: Cambridge University Press.

Alföldy, Géza. 1989. "Epigraphische Notizen aus Italien III: Inschriften aus Nursia (Norcia)." *Zeitschrift für Papyrologie und Epigraphik* 77: 155–180.

Aurora, Federico. 2017. "*pa-ro*, *da-mo*: Studying the Mycenaean Case System through *DĀMOS* (*Database of Mycenaean at Oslo*)." In *Aegean Scripts: Proceedings of the 14th International Colloquium on Mycenaean Studies Copenhagen, 2–5 September 2015*. Vol. 1, edited by Marie-Louise Nosch and Hedvig Landenius Enegren, 83–98. Rome: Istituto di Studi sul Mediterraneo Antico.

Aurora, Federico. 2025. "*DĀMOS, Database of Mycenaean at Oslo*—Developments and Perspectives." In ko-ro-no-we-sa*: Proceedings of the 15th International Colloquium on Mycenaean Studies September 2021*, edited by John Bennet, Artemis Karnava, and Torsten Meißner, 143–160. Athens: Rethymnon. https://doi.org/10.26248/ariadne.vi.1847.

Aurora, Federico, Damir Nedić, Asgeir Nesøen, and Dag Haug. 2025. "Exporting Mycenaean: From a Relational Database to EpiDoc XML Files (and Back Again?)." *Digital Humanities in the Nordic and Baltic Countries Publications* 7 (2). https://doi.org/10.5617/dhnbpub.12295.

Bechtel, Friedrich. 1964. *Die historischen Personennamen des Griechischen bis zur Kaiserzeit*. Hildesheim: Olms.

Bermúdez Sabel, Helena, and Francesca Dell'Oro. 2020. "Automatic Annotation of Latin and Greek Texts." GitHub. WoPoss project. https://github.com/WoPoss-project/automatic_annotation.

Berti, Monica. 2016. "The Digital Marmor Parium." In *Epigraphy Edit-a-thon: Editing Chronological and Geographic Data in Ancient Inscriptions. Leipzig, April 20–22, 2016*, edited by Monica Berti. Leipzig: Universitätsbibliothek.

Celano, Giuseppe G. A. 2014. "Guidelines for the Annotation of the Ancient Greek Dependency Treebank 2.0." GitHub. Perseus DL. https://github.com/PerseusDL/treebank_data/edit/master/AGDT2/guidelines.

Clackson, James. 2011. "Introduction." In *A Companion to the Latin Language*, edited by James Clackson, 1–6. Malden, UK: Wiley-Blackwell.

Crellin, Robert. 2024. PyEpiDoc tokenizer, version May 2024. GitHub. PyEpiDoc. https://github.com/rsdc2/PyEpiDoc.

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

De Felice, Irene, Lucia Tamponi, Federica Iurescia, and Marco Passarotti. 2023. "Linking the Corpus *CLaSSES* to the *LiLa Knowledge Base of Interoperable Linguistic Resources for Latin*." In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30–Dec 02, 2023, Venice, Italy*, 173–179. Turin: Accademia University Press. https://ceur-ws.org/Vol-3596/paper20.pdf.

Del Freo, Maurizio, Francesco Di Filippo, and Françoise Rougemont. 2024. *Linear B Electronic Resources (LiBER)*. Rome: CNR Edizioni. https://liber.cnr.it.

Dell'Oro, Francesca. 2015. "What Role for Inscriptions in the Study of Syntax and Syntactic Change in the Old Indo-European Languages? The *Pros* and *Cons* of an Integration of Epigraphic Corpora." In *Perspectives on Historical Syntax*, edited by Carlotta Viti, 271–290. Amsterdam: Benjamins.

Dell'Oro, Francesca. 2023. "WoPoss Guidelines for the Annotation of Modality: Revised Version." University of Neuchâtel–Swiss National Science Foundation. Zenodo. https://doi.org/10.5281/zenodo.3560950.

Dell'Oro, Francesca, and Giuseppe G. A. Celano. 2019. "Epigraphic Treebanks: Some Considerations from a Work in Progress." FirstDrafts@Classics. Harvard CHS. https://chs.harvard.edu/wp-content/uploads/2020/11/DellOroCelano_4.pdf.

Dubois, Laurent, ed. 1989. *Inscriptions grecques dialectales de Sicile: Contribution à l'étude du vocabulaire grec colonial*. Rome: École française de Rome.

Fendel, Victoria B. 2024. "The *I.Sicily* Sketch Engine Corpus." *Journal of Open Humanities Data* 10: 56. https://doi.org/10.5334/johd.258.

Graaf, Evelien de, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. "*AGILe*: The First Lemmatizer for Ancient Greek Inscriptions." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck et al., 5334–5344. Marseille: European Language Resources Association. https://aclanthology.org/2022.lrec-1.571/.

Herman, József. 1991. *Late Latin Data Base: Guidelines for Data Collection*. Budapest: Research Institute for Linguistics of the Hungarian Academy of Sciences.

Langslow, David R. 2002. "Approaching Bilingualism in Corpus Languages." In *Bilingualism in Ancient Society: Language Contact and the Written Text*, edited by James N. Adams, Mark Janse, and Simon Swain, 23–51. Oxford: Oxford University Press.

Mallia, Michele, Andrea Bellandi, Alessandro Tommasi, Cesare Zavattari, Michela Bandini, Valeria Quochi. 2023. "'EpiLexO.' ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics 'A. Zampolli,'" National Research Council, in Pisa. http://hdl.handle.net/20.500.11752/ILC-1005.

Marotta, Giovanna, Francesco Rovai, Irene De Felice, and Lucia Tamponi. 2020. "*CLaSSES*: Orthographic Variation in Non-Literary Latin." *Studi e saggi linguistici* 58 (1): 39–65.

Minon, Sophie. 2020. "Le projet LGPN-Ling: Analyse étymologique et sémantique des anthroponymes grecs antiques." *Bulletin de la Société de linguistique de Paris* 115 (1): 253–297.

Minon, Sophie, ed. 2023. *Lexonyme: Lexique étymologique et sémantique des anthroponymes grecs antiques*. Vol. 1. Geneva: Droz.

Mullen, Alex, and Anna Willi. 2024. "Appendix 1." In *Latinization, Local Languages, and Literacies in the Roman West*, edited by Alex Mullen and Anna Willi, 413–415. Oxford: Oxford University Press.

Murano, Francesca, Valeria Quochi, Angelo Mario Del Grosso, Luca Rigobianco, and Mariarosaria Zinzi. 2023. "Describing Inscriptions of Ancient Italy: The ItAnt Project and Its Information Encoding Process." *Journal on Computing and Cultural Heritage* 16 (3): 1–14.

Panciera, Silvio, Giuseppe Camodeca, Giovanni A. Cecconi, Silvia Orlandi, Lanfranco Fabriani, and Silvia Evangelisti. 2020. "EDR—Epigraphic Database Roma Epidoc Files." Zenodo. https://doi.org/10.5281/zenodo.7561922.

Pitts, Reuben J. 2022. "Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE (CEIPoM)." *Journal of Open Humanities Data* 8 (1). http://doi.org/10.5334/johd.65.

Prag, Jonathan R. W. 2021. "*I.Sicily* and *Crossreads*: A Digital Epigraphic Corpus for Ancient Sicily." In *Trinacria, "an Island outside Time": International Archaeology in Sicily*, edited by Christopher Prescott, Arja Karivieri, Peter Campbell, Kristian Göransson, and Sebastiano Tusa, 181–192. Oxford: Oxbow. https://doi.org/10.5281/zenodo.5567727.

Rocca, Giovanna, Giulia Sarullo, and Marta Muscariello. 2017. "The Digital Edition of the Archaic Latin Inscriptions (7th–5th Century B.C.)." In *Digital and Traditional Epigraphy in Context: Proceedings of the Eagle 2016 International Conference*, edited by Silvia Orlandi, Raffaella Santucci, Francesco Mambrini, and Pietro Maria Liuzzo, 67–72. Rome: Università La Sapienza. https://www.eagle-network.eu/wp-content/uploads/2017/07/419-736-1-SM.pdf.

Stefanowitsch, Anatol. 2020. *Corpus Linguistics: A Guide to the Methodology*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3735822.

Warmington, Eric H. 1940. *Archaic Inscriptions*. Cambridge, MA: Harvard University Press.

Online Resources

AGDT 2.0. https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/Greek_guidelines.md.

Arethusa. https://www.perseids.org/tools/arethusa/app/#/.

CEIPoM Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE. https://github.com/ReubenJPitts/Corpus-of-the-Epigraphy-of-the-Italian-Peninsula-in-the-1st-Mil

lennium-BCE/tree/main; Vademecum. https://reubenjpitts.github.io/Corpus-of-the-Epigraphy-of-the-Italian-Peninsula-in-the-1st-Millennium-BCE/vademecum.
CLaSSES Corpus for Latin Sociolinguistic Studies on Epigraphic textS. http://classes-latin-linguistics.fileli.unipi.it.
Computerized Historical Linguistic Database of the Latin Inscriptions of the Imperial Age. https://lldb.elte.hu/.
Corpus of Gāndhārī Texts. https://gandhari.org/corpus.
DĀMOS Database of Mycenaean at Oslo. https://damos.hf.uio.no/1.
DiLaDi. Digital Latin Dialectology: Tracing Linguistic Variation in the Light of Ancient and Early Medieval Sources. https://pric.unive.it/projects/diladi/home.
EpiDoc. https://epidoc.stoa.org/.
Epigraphic Database Roma (EDR). http://www.edr-edr.it/default/index.php.
Epigraphik Datenbank Clauss-Slaby. http://www.manfredclauss.de/.
Hethitologie Portal Mainz. https://www.hethport.uni-wuerzburg.de/HPM/index.php.
IGDS I treebank. https://github.com/PerseusDL/treebank_data/tree/master/IGDS.
Inscriptions of Israel/Palestine corpus. https://search.inscriptionsisraelpalestine.org/inscriptions/map. (For the wordlists, see, e.g., https://search.inscriptionsisraelpalestine.org/wordlist/greek.)
I.Sicily corpus. http://sicily.classics.ox.ac.uk/.
LatinNow. https://latinnow.eu/about/.
Lexicon of Greek Personal Names (online version). https://www.lgpn.ox.ac.uk/.
LGPN-Ling. https://lgpn-ling.huma-num.fr/exist/apps/lgpn-ling7/about.html.
Sketch Engine. https://www.sketchengine.eu/.
UDPipe. https://lindat.mff.cuni.cz/services/udpipe/run.php.
WoPoss corpus. https://woposs.unine.ch/form.html.