ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 1, Number 1
december 2015

Emerging Topics at the First Italian Conference on Computational Linguistics

aA Accademia university press

IJCoL
Italian Journal of Computational Linguistics

1-1 | 2015

Emerging Topics at the First Italian Conference on Computational Linguistics

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/ijcol/308
DOI: 10.4000/ijcol.308
ISSN: 2499-4553

Publisher
Accademia University Press

Electronic reference
IJCoL, 1-1 | 2015, "Emerging Topics at the First Italian Conference on Computational Linguistics" [Online], Online since 01 December 2015, connection on 12 April 2023. URL: https://journals.openedition.org/ijcol/308; DOI: https://doi.org/10.4000/ijcol.308

registrazione in corso presso il Tribunale di Trento

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

ISSN 2499-4553
ISBN 978-88-99200-63-3

www.aAccademia.it/IJCoL_01

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it

aAccademia university press

IJCoL

Volume 1, Number 1
december 2015

Emerging Topics at the First Italian Conference on Computational Linguistics

a cura di
Roberto Basili, Alessandro Lenci,
Bernardo Magnini, Simonetta Montemagni

CONTENTS

CLaSSES: a New Digital Resource for Latin Epigraphy

Irene De Felice[*]
Università di Pisa

Margherita Donati[§]
Università di Pisa

Giovanna Marotta[†]
Università di Pisa

CLaSSES (Corpus for Latin Sociolinguistic Studies on Epigraphic textS) is an annotated corpus aimed at (socio)linguistic research on Latin inscriptions. Provided with linguistic, extra- and meta-linguistic features, it can be used to perform quantitative and qualitative variationist analyses on Latin epigraphic texts. In particular, it allows the user to analyze spelling (and possibly phonetic-phonological) variants and to interpret them with reference to the dating, the provenance place, and the type of the texts. This paper presents the first macro-section of CLaSSES, focused on inscriptions of the archaic and early periods (CLaSSES I).

1. Introduction[1]

This paper presents CLaSSES I, the first macro-section of CLaSSES (Corpus for Latin Sociolinguistic Studies on Epigraphic textS), an epigraphic corpus built for variationist studies on Latin inscriptions. This resource was developed within a research project devoted to sociolinguistic variation and identity dynamics in the Latin language (for further details on the project, see Donati et al. in press; Marotta in press).

In the first section of the paper, some of the digital resources available for Latin epigraphy will be briefly introduced, then the most important aspects of innovation of CLaSSES will be highlighted (§ 2). The following section will address the current debate about the role played by epigraphic texts as a source of evidence for linguistic variation within dead languages, as well as the theoretical grounds for variationist research on epigraphic Latin (§ 3). The core part of the paper describes the sources of our corpus and the linguistic, meta- and extra-linguistic annotation conducted (§ 4); some results of such annotation are also reported (§ 5). Finally, the last section will draw some conclusions and will sketch the future directions of our work (§ 6).

[*] Department of Philology, Literature and Linguistics, University of Pisa.
E-mail: irene_def@yahoo.it
[§] Department of Philology, Literature and Linguistics, University of Pisa.
E-mail: margherita.donati@for.un
[†] Department of Philology, Literature and Linguistics, University of Pisa.
E-mail: gmarotta@ling.unipi.it

[1] This research was developed at the Laboratory of Phonetics and Phonology of Pisa University within the PRIN project *Linguistic representations of identity. Sociolinguistic models and historical linguistics* (PRIN2010, prot. 2010HXPFF2_001). The results related to the project are available online at http://www.mediling.eu/. The paper was conceived by the three authors together. For academic reasons only, the scientific responsibility is attributed as follows: § 1 is common; § 2, § 4.5, § 4.6, § 5 to I. De Felice; § 3, § 4.2, § 4.3, § 4.4 to M. Donati; § 4.1, § 6 to G. Marotta.

2. Digital resources for Latin inscriptions

The available open-access digital resources for Latin epigraphy include, at present, some important databases (cf. Feraudi-Gruénais 2010; Elliott 2015). The Epigraphic Database Clauss-Slaby (EDCS)[2] is the most extensive online resource and records almost all Latin inscriptions (to date, 735.664 sets of data for 491.190 inscriptions from 3.500 publications), together with a very large number of pictures (so far, 98.897). It allows simple as well as combined queries, by publication, Roman province, place, and specific terms (possibly by using boolean operators and simple regular expressions); in addition, users can search also for misspelled words. The text of the inscriptions is presented without abbreviations and, when possible, in its complete form.

Another very useful online resource is the Epigraphic Database Roma (EDR);[3] it is part of the Electronic Archive for Greek and Latin Epigraphy (EAGLE),[4] an international network of epigraphic databases aiming to provide an open-access digital version of all published Greek and Latin inscriptions up to the 7th century AD. The main purpose of EDR is to collect all inscriptions from Rome and Italy, including Sardinia and Sicily (with the exception of Christian inscriptions of Rome). Besides the information about the content of the inscriptions, EDR also provides information about the writing support (e.g. typology, material, dimension) and a wide-ranging bibliography; often, also images and photographs are supplied (Panciera 2013; Caldelli et al. 2014). To date, EDR material includes 70294 inscriptions and 42022 photographs. Through the online query interface, the user can perform a number of simple or combined searches, through the following sections: text (words or groups of letters, possibly with boolean operators AND/OR), place of provenance, date, type of object, material, size, preservation condition (intact or fragmentary texts), writing technique, language (e.g. Greek, Latin, Greek - Latin bilingual), type of inscription, social role of people mentioned, edition (Evangelisti 2010).

Two other components of EAGLE well worth mentioning are the Epigraphische Datenbank Heidelberg (EDH),[5] which mostly includes Latin or bilingual (Greek - Latin) inscriptions of provinces of the Roman empire, and the Epigraphic Database Bari (EDB),[6] which collects Christian inscriptions of Rome from the 3rd to the 8th century AD.

Some electronic resources of utility are also made freely available by the Corpus Inscriptionum Latinarum (CIL) research centre, in particular the Archivium Corporis Electronicum database (a collection of bibliographical references, squeezes, and photographs), the word indices to a few CIL volumes, and the concordances (that link inscription numbers adopted in early editions to those adopted in the CIL volumes).[7]

For what regards the representation of epigraphic or papyrological texts in digital form, the international and collaborative project EpiDoc (Epigraphic Documents),[8] which involves a large community of scholars working on Greek and Latin inscriptions (cf. Bodard 2010), provides tools and guidelines for the encoding of editions of ancient documents in XML, the Extensible Markup Language. EpiDoc adopts a subset of the XML defined by the Text Encoding

[2] http://www.manfredclauss.de/gb/index.html.
[3] http://www.edr-edr.it/English/index_en.php.
[4] http://www.eagle-network.eu.
[5] http://www.uni-heidelberg.de/institute/sonst/adw/edh.
[6] http://www.edb.uniba.it.
[7] All these resources are accessible from the website http://cil.bbaw.de.
[8] http://sourceforge.net/p/epidoc/wiki/Home/.

Initiative's (TEI) standard for the digital representation of texts, which is now widely used in the humanities. This flexible system allows not only to transcribe a Greek or Latin text, but also, for instance, to encode its translation, description, and other pieces of information such as dating, history of the inscription, bibliography, and the object on which the text is written. At the moment, we decided not to follow the EpiDoc guidelines, due to the current aims of the project. However, we do not exclude a conversion of our existing corpus in the XML interchange format in the future.

Although the current state-of-the-art digital resources for Latin inscriptions briefly presented here collect a copious number of epigraphic texts and often provide useful extra-linguistic data, such as provenance place, dating, material, etc., they do not allow researchers to directly access specific information about relevant linguistic variation phenomena. They do not satisfactorily meet the needs of the linguist to study Latin epigraphic texts from a variationist perspective. In order to systematically address the massive graphic and linguistic variation observable in Latin inscriptions, a specific tool is necessary. We argue that the corpus CLaSSES is a new and useful resource, since it consists not only of raw epigraphic texts, but also of linguistic information about specific spelling variants that can be regarded as clues for phonetic-phonological (and morpho-phonological) variation (cf. § 4).

3. Studying variation in Latin through inscriptions

There is a current debate[9] on whether inscriptions can provide direct evidence for actual linguistic variation in Latin. In other words, can epigraphic texts be regarded as primary and reliable sources for reconstructing variation dynamics related to social strata, different language registers, and geographic variability? It is obviously true that inscriptions are the only direct evidence left by antiquity (although they can be influenced by literary uses, writers' education, and many other factors), since every other kind of written text, even comedy or the so-called "vulgar" texts, is necessarily mediated by philological and manuscript tradition. In this sense, inscriptions are likely to keep record of linguistic variation. However, the story is not that simple.

As Herman (1985) points out, the debate on the evaluation of late or "vulgar" inscriptions as linguistically representative texts is ancient and alternates between approaches that are either totally skeptical or too optimistic. Herman argues for a critical approach (1978b, 1985): epigraphic texts are fundamental sources for studying variation phenomena, provided that scholars take into account the issues related to their philological, paleographic, archaeological and historical interpretation, as well as the complex relationship between speech and writing. He states "mon article [...] veut sans doute constituer une mise en garde à l'adresse de ceux qui espèrent entrevoir grâce aux inscriptions [...] de nettes différences dialectales dans le latin des provinces de l'Empire, il tend cependant à prouver, en même temps, que les données épigraphiques, analysées avec critique et soin, correspondent bien à la réalité d'un état de langue déterminé et permettent par conséquent de suivre, de province en province, le cheminement inégal des innovations" (1985: 207). However, Herman's fundamental studies on Latin demonstrate that epigraphic texts are actually fruitful for studying linguistic variation (Herman 1970, 1978a, 1978b, 1982, 1987, 2000, among others; see also Loporcaro 2011a, 2011b).

[9] We just touch on this topic; for further discussion see Donati et al. in press; Marotta 2015, in press.

On the other hand, Adams (2003, 2007, 2013) limits the role of the inscriptions as a source for direct evidence of the spoken language and linguistic varieties of Latin. He argues that one can never be sure whether the variants found in inscriptions reflect the actual pronunciation, or are just misspellings or archaisms: only the critical evaluation of deviant spellings together with metalinguistic data, such as those provided by grammarians and authors, can ensure that these spellings actually reflect a phonetic reality. Moreover, even if deviant spellings can be recognized as reflecting speech, ascribing it to a given social class or level is a further step that needs to be confirmed, again, by grammarians, rhetors, and literary authors. Adams states that "certain misspellings are so frequent that there can be no doubt that they reflect the state of the language. Cases in point are the omission of *-m* and the writing of *ae* as *e*. But the state of what varieties of the language? Those spoken by a restricted educational/social class, or those spoken by the majority of the population? This is a question that cannot be answered merely from an examination of texts and their misspellings or absence thereof, because good spellers will stick to traditional spellings whether they are an accurate reflection of their own speech or not. If, roughly speaking, we are to place the pronunciation lying behind a misspelling in a particular social class, we need additional evidence, such as remarks by grammarians or other speakers" (2013: 33-34). So, in Adams' approach to Latin sociolects, grammarians and their remarks occupy a very prominent place.

In our opinion, epigraphic texts can be regarded as a fundamental source for studying variation in Latin, provided that one adopts a critical approach. This position is shared by several scholars, who in recent works highlight the relevance of the epigraphic data (Consani in press; De Angelis in press; Kruschwitz 2015; Marotta 2015, in press; Rovai 2015). Nevertheless, the critical points raised by Adams cannot be ignored.

Furthermore, sociolinguistic variation of Latin in Rome and the Empire is a promising research area (Adams et al. 2002; Adams 2003, 2007, 2013; Biville et al. 2008; Dickey and Chahoud 2010; Rochette 1997). From the seminal work by Campanile (1971), many scholars highlight that sociolinguistic categories and methods can be usefully applied to ancient and dead languages (Giacalone Ramat 2000; Lazzeroni 1984; Molinelli 2006; Vineis 1984, 1993), even if cautiously, since ancient languages are corpus languages[10] and we are forced to rely on written sources only (Cuzzolin and Haverling 2009; Giacalone Ramat 2000; Winter 1998).

Assuming this methodological perspective, our empirical analysis of Latin epigraphic texts is focused on identifying and classifying specific spelling variants, which can be regarded as clues for variation also at the phonetic-phonological, and consequently morpho-phonological level. Being aware of the debate on the reliability of inscriptions currently ongoing, we intend to investigate whether it is possible to find out relevant evidence for sociolinguistic variation in epigraphic Latin *via* the integration of the modern quantitative and correlative sociolinguistics with a corpus-based approach. Since, at present, there is a lack of digital resources devoted to this particular kind of research (cf. § 2), our first step was the creation of an original resource for studying Latin epigraphic texts, which will be described in what follows.

[10] A corpus language can be defined as a language "known only through written documents" (Clackson 2011: 2).

4. Building CLaSSES I

4.1. Materials

As a matter of fact, Latin inscriptions of the archaic and early periods are characterized by a wide array of variation in spelling that may well correspond to a variation at the linguistic level as well. In order to analyze epigraphic texts from a variationist perspective, it is methodologically necessary to compare the attested forms with a fixed point of reference, which can be identified in Classical Latin. In our analysis of the inscriptions of the archaic and early periods (macro-section CLaSSES I), we classified as "non-classical" those forms, attested mainly in the archaic and early periods, that do not belong to the tradition of Classical Latin.[11] Therefore, in CLaSSES I we avoid terms such as "non-standard" or "substandard", currently in use in the scientific literature. For example, in CIL I^2 8 (L CORNELIO L F SCIPIO AIDILES COSOL CESOR), CORNELIO is identified as a non-classical nominative form for the classical CORNELIUS. Indeed, identifying non-classical forms is not a trivial operation for every chronological phase of Latin, in particular for the archaic (7th century BC - ca. 240 BC) and the early (ca. 240 BC - ca. 90 BC) periods. A Latin linguistic and literary standard gradually emerges between the second half of the 3rd century BC, when literature traditionally begins, and the 1st century BC, when Cicero makes explicit the Latin linguistic norm in his rhetorical works (Clackson and Horrocks 2007; Cuzzolin and Haverling 2009; Mancini 2005, 2006).[12]

CLaSSES I includes inscriptions of the archaic and early periods. Inscriptions are from the *Corpus Inscriptionum Latinarum* (CIL), the main and most comprehensive source for Latin epigraphy research. Inscriptions selected for this macro-section of our corpus are dated from 350 to ca. 150 BC, with most of them falling into the 3rd century BC. The volumes of the CIL that cover this chronological segment were systematically examined: CIL I² *Pars II, fasc. I,* section *Inscriptiones vetustissimae* (Lommatzsch 1918); CIL I² *Pars II, fasc. II, Addenda Nummi Indices,* section *Addenda ad inscriptiones vetustissimas* (Lommatzsch 1931); CIL I² *Pars II, fasc. III, Addenda altera Indices,* section *Addenda ad inscriptiones vetustissimas* (Lommatzsch 1943); CIL I² *Pars II, fasc. IV, Addenda tertia,* section *Addenda ad inscriptiones vetustissimas* (Degrassi and Krummrey 1986). It is worth noting that the texts offered by the CIL were also revised and checked by means of the available philological resources for Archaic Latin epigraphy (Warmington 1940; Degrassi 1957-1963; Wachter 1987), in order to guarantee the most reliable and updated philological accuracy.

Moreover, it is noteworthy that within the vast quantity of epigraphic texts available for this phase of Latin not every inscription is significant for linguistic studies. As a consequence, the following texts have been excluded: 1) legal texts, since they are generally prone to archaisms; 2) too short (single letters, initials) or fragmentary inscriptions; 3) inscriptions from the necropolis of Praeneste, as they contain only anthroponyms in nominative form.

[11] For a more detailed discussion of this term, see Donati et al. in press.

[12] The standard is based on the Roman variety of Latin (Clackson and Horrocks 2007), first developed in texts written by a few authors of high repute and later transmitted by grammarians (Cuzzolin and Haverling 2009); however, standardization is not only a literary operation, but it is also developed in connection with (linguistic) politics and the process of codification of the right (Poccetti et al. 1999). Once standardized, these forms of written Latin changed very little throughout antiquity and the Middle Ages.

4.2. Tokenization and lemmatization

CLaSSES I includes 386 inscriptions, for a total number of 1869 words. The entire collected corpus was tokenized and an index was created, so that each token of the corpus is univocally associated to a token-ID containing the CIL volume, the number of the inscription and the position in which the token occurs within the inscription. We intend tokens as character sequences without spaces. We count among tokens lacunae as well (i.e. gaps in the inscription identified by the string "[…]"), since they occupy a specific position within the text, and they actually exist in its critical edition.

Each token has also been manually lemmatized, when possible. For this operation, we mainly relied upon the Oxford Latin Dictionary.

4.3. Extra- and meta-linguistic data

Each epigraphic text of CLaSSES I was enriched with extra-linguistic information, i.e. related to its place of provenance and dating, and meta-linguistic information, i.e. related to the text type. In particular, we identified five text types, largely following the traditional classification by CIL and Warmington (1940); however, we decided to further distinguish, within the group of the inscriptions traditionally classified as *tituli sacri*, between *tituli sacri privati* and *tituli sacri publici* (for details, see Donati 2015):

a. *tituli honorarii* (n. 18), i.e. inscriptions celebrating public people and inscriptions on public monuments (e.g. CIL I^2 363 L RAHIO L F C[...] AIDILES [D]E[DERE]);

b. *tituli sepulcrales* (n. 26), i.e. epitaphs and memorial texts (e.g. CIL I^2 52 C FOURI M F);

c. *instrumenta domestica* (n. 246), i.e. inscriptions on domestic tools (e.g. CIL I^2 441 BELOLAI POCOLOM);

d. *tituli sacri privati* (n. 82), i.e. votive inscriptions offered by private individuals or brotherhoods (e.g. CIL I^2 384 L OPIO C L APOLENE DONO DED MERETO);

e. *tituli sacri publici* (n. 14), i.e. votive inscriptions offered by people holding public offices or whole communities (e.g. CIL I^2 395 A CERVIO A F COSOL DEDICAVIT).

As an example of the extra- and meta-linguistic information included in CLaSSES I, in CIL I^2 45 DIANA MERETO NOUTRIX PAPERIA the word MERETO is identified by the token-ID CIL-I^2-45/2, while the inscription CIL-I^2-45 is associated to the following data: place of provenance *Gabii*, dating 250 - 200 BC, text type *tituli sacri privati*.

In order to account for the rich and manifold linguistic material of the inscriptions included in CLaSSES I, each word of the corpus is also classified according to different parameters, as the next sections illustrate. The criteria adopted for the annotation were jointly discussed and the manual annotation was performed by two annotators, who constantly worked in parallel. Moreover, each one of them also checked a sample of the annotation made by the other one.

4.4. Graphic form annotation

The graphic forms occurring in epigraphic texts are of different kinds, mainly due to the conservation status of the writing support. Therefore, we make a distinction between the following types:

a. complete words (e.g. CIL I^2 45 DIANA);
b. abbreviations, i.e. every kind of shortening, including personal name initials (e.g. CIL I^2 46 DON for DONUM);
c. incomplete words, i.e. words partly integrated by editors (e.g. CIL I^2 448 ME[NERVAE);
d. words completely integrated by editors (e.g. CIL I^2 2875c [LAPIS]);
e. misspellings (e.g. CIL I^2 550 CUDIDO for CUPIDO);[13]
f. uncertain words, i.e. words that cannot be interpreted, not even in their graphical form (e.g. CIL I^2 59 STRIANDO);
g. numbers;
h. lacunae.

4.5. Language annotation

Since Latin archaic inscriptions sometimes include foreign words, we distinguish Latin words, which constitute the largest part of the corpus, from words belonging to other languages:[14]

a. Greek (e.g. CIL I^2 565 DOXA);
b. Oscan (e.g. CIL I^2 394 BRAT);
c. Umbrian (e.g. CIL I^2 2873 NUMESIER);
d. Etruscan (e.g. CIL I^2 554 MELERPANTA);
e. hybrid, for mixed forms (e.g. CIL I^2 553 ALIXENTROM);
f. unknown, for words of uncertain origin (e.g. CIL I^2 576 VIET).

4.6. Annotation of non-classical variants

The core part of the annotation phase, which provides the corpus with a rich set of qualitative data, consists of a linguistic analysis of CLaSSES I.[15] The two annotators manually retrieved all the non-classical forms in the corpus (tot. 690), then they also associated them to their corresponding classical form, e.g. nom. sg. CORNELIO

[13] Misspellings are mistyped words, i.e. words that are written in a different way with respect to their Classical form for an error of the stone-cutter.
[14] Obviously, lacunae are excluded from this classification.
[15] For textual interpretation of inscriptions, we mainly referred to the information included within CIL, as well as to Warmington 1940; Degrassi 1957-1963; Wachter 1987.

(non-classical) - CORNELIUS (classical). Uncertain cases were discussed by the annotators to achieve consensus.

All non-classical forms were then classified according to the type of variation phenomena that distinguish them from the corresponding classical equivalents. Variation phenomena may regard vowels, consonants, as well as morpho-phonology (i.e. when vocalic and consonantal phenomena occur in morphological endings). For instance, the nominative CONSOL (CIL I^2 17) shows a vocalic phenomenon, because it deviates from the standard CONSUL for the vowel alternation <o>-<u>.

a. *Vowels*. Among the phenomena related to vowels, we distinguish the followings: alternations (CIL I^2 2909 MENERVA for MINERVAE; CIL I^2 560a PISCIM for PISCEM); gemination (CIL I^2 365 VOOTUM for VOTUM); syncope (CIL I^2 37 VICESMA for VICESIMA); epenthesis (CIL I^2 59 MAGISTERE for MAGISTRI); monophthongization (CIL I^2 376 DIANE for DIANAE); archaic spellings of diphthongs (CIL I^2 397 FORTUNAI for FORTUNAE).

b. *Consonants*. Among the phenomena related to consonants, we distinguish the followings: final consonant deletion (CIL I^2 8 CORNELIO for CORNELIUS); nasal deletion within consonant clusters (CIL I^2 8 COSOL for CONSUL; CIL I^2 560c COFECI for CONFECI); assimilation (CIL I^2 7 OPSIDESQUE for OBSIDESQUE); gemination (CIL I^2 16 [P]AULLA for PAULA); degemination (CIL I^2 563 APOLO for APOLLO); voice alternations (CIL I^2 462a ECO for EGO; CIL I^2 389 PAGIO for PACIUS); deaspiration (CIL I^2 555 TASEOS for THASIUS). Some of these phenomena are especially relevant in the current discussion about sociolinguistic variation in Latin, namely vowel alternations, monophthongization, synchope, final *-s* and *-m* deletion (as already discussed in a body of works; cf. among others Adams 2013; Benedetti and Marotta 2014; Campanile 1971; Herman 1987; Leumann 1977; Loporcaro 2011a, 2011b; Marotta 2015, in press; Pulgram 1975; Vineis 1984; Weiss 2009).

c. *Morpho-phonology*. If a given variant occurs in a morpho-phonological position (typically, in the word ending), then an additional level of annotation is added, which keeps track of the particular ending attested. For instance, among the most frequent phenomena annotated, we highlight the *–a* ending of the dative singular of the first declension (CIL I^2 43 DIANA for DIANAE); the *–os* and *-o* endings of the nominative singular of the second declension (CIL I^2 406b CANOLEIOS and CIL I^2 408 CANOLEIO for CANOLEIUS); the *–om* ending of the accusative singular of the second declension (CIL I^2 2486a DONOM for DONUM); and the *–et* ending of the 3rd person of the perfect (CIL I^2 2867 DEDET for DEDIT).

This fine-grained annotation creates the prerequisites for the evaluation of the statistical incidence of each kind of non-classical variant, as well as to perform cross-queries taking into account text type, dating, and place of provenance.

5. Results

We can now present the results of the annotation conducted on CLaSSES I. As **Table 1** shows, the text type most represented in the corpus is the *instrumentum domesticum*, with 246 epigraphic texts (726 words), followed by 82 inscriptions classified as *tituli sacri privati* (523 words), 26 inscriptions classified as *tituli*

sepulcrales (310 words), 18 inscriptions classified as *tituli honorarii* (182 words), and finally 14 texts pertaining to the *tituli sacri publici* category (128 words).

Table 1
Classification of the 1869 words constituting CLaSSES I according to which text type they pertain.

Text type				
instr. domestica	*tit. sacri privati*	*tit. sepulcrales*	*tit. honorarii*	*tit. sacri publici*
726	523	310	182	128
38.9%	28%	16.6%	9.7%	6.8%

For what regards the annotation of a word's graphic form (**Table 2**), only 54.4% of the words constituting the corpus are complete, whereas 30% are abbreviated (most of these forms stand for proper nouns, such as C for GAIUS or L for LUCIUS), and 8.2% are incomplete. Moreover, 3.3% of the words are missing, either because the editors classified them as lacunae, or because they totally integrated them; 3% are uncertain and cannot be interpreted. Misspellings and numbers constitute the minor part of the corpus.

Table 2
Classification of the 1869 words constituting CLaSSES I according to their graphic form.

Graphic form							
complete	*abbreviat.*	*incomplete*	*integrated*	*misspelling*	*uncertain*	*number*	*(lacunae)*
1017	560	153	28	12	56	9	34
54.4%	30%	8.2%	1.5%	0.6%	3%	0.5%	1.8%

As **Table 3** shows, Latin is the language most represented in the corpus (93.5% of the words), whereas only 4.7% of the words have a different origin.

Table 3
Classification of the 1869 words constituting CLaSSES I with regard to their language.

Language							
Latin	*Greek*	*Oscan*	*Umbrian*	*Etruscan*	*hybrid*	*unknown*	*(lacunae)*
1748	11	12	3	9	17	35	34
93.5%	0.6%	0.6%	0.2%	0.5%	0.9%	1.9%	1.8%

6. Conclusions and future directions

CLaSSES I is a corpus that allows quantitative and qualitative analysis on graphemic variation occurring in Latin inscriptions, satisfying basic requirements for grounded and systematic linguistic studies. It is annotated with linguistic, extra- and meta-linguistic features, which permit specific cross-queries on the text, also considering the dating, the geographic origin, and the type of the inscription.

As we have illustrated in the previous sections, the initial hypothesis in our project is that, given the wide array of variation detectable in archaic and early Latin inscriptions, sociolinguistic aspects possibly emerging may be highlighted by identifying and classifying the occurrences of non-classical variants. Even if the search for non-classical forms in Archaic and Early Latin might seem anachronistic in some way, this choice is based on two fundamental aspects. First, many phenomena occurring in these forms seem to represent the basis for diachronic developments occurring from Late Latin to the Romance languages, thus revealing some continuity at least at some (sociolinguistic?) level from Early to Late Latin (this point is not uncontroversial, see e.g. Adams 2013: 8). Second, different spellings in any case provide evidence for orthographic - and possibly phonological - variation within archaic inscriptions, thus presumably pointing to different levels in the diasystem.

There are a number of case studies that have already been conducted on CLaSSES I. For instance, the analysis of the distribution of non-classical and classical forms, presented in Donati et al. (in press), confirms in quantitative terms that the linguistic standard is not yet established in the chronological period considered in CLaSSES I. Marotta (2015) analyzes vowel alternations: the spellings <e> and <o>, alternating with <i> and <u>, are interpreted as possible clues for the existence of a phonological opposition grounded on vowel quality rather than vowel quantity, at least at some level of the Latin diasystem. In Donati (2015), the possible correlation between the distribution of non-classical variants and diaphasic factors related to the type of text are analyzed, as well as the distribution of non-classical variation phenomena in vowels and consonants.

Our primary current aim is to build and develop other sections of CLaSSES, by using the same annotation criteria already adopted for CLaSSES I and described above (cf. § 4.2 - § 4.6). In particular, two macro-sections are now in progress, CLaSSES II and CLaSSES III. CLaSSES II includes inscriptions of the period 150 - 50 BC, whereas CLaSSES III is focused on Classical Latin, i.e. 50 BC - 50 AD. Moreover, we plan to add a morphological layer of annotation to the lemmatized corpus. This operation will provide the word tokens with information related to morphological properties, such as the part of speech (PoS), and possibly the morphological categories (case, number, tense, person, etc.). Furthermore, given the high frequency of proper names in epigraphic texts, we also intend to annotate the named entities.

Finally, all the data collected will be the input for the creation of a database available through a web interface in the near future.

References

Adams, James N. 2003. *Bilingualism and the Latin Language*. Cambridge University Press, Cambridge.

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

Adams, James N. 2013. *Social Variation and the Latin Language*. Cambridge University Press, Cambridge.

Adams, James N., Mark Janse, and Simon Swain (eds.). 2002. *Bilingualism in Ancient Society. Language Contact and the Written Word*. Oxford University Press, Oxford.

Benedetti, Marina and Giovanna Marotta. 2014. Monottongazione e geminazione in latino: nuovi elementi a favore dell'isocronismo sillabico. In Molinelli, Piera, Pierluigi Cuzzolin, and Chiara Fedriani (eds.). *Latin vulgaire - Latin tardif X. Actes du Xe colloque international sur le latin vulgaire et tardif*. Sestante Edizioni, Bergamo: 25-43.

Biville, Frédérique, Jean-Claude Decourt, and Georges Rougemont (eds.). 2008. *Bilinguisme gréco-latin et épigraphie*. Maison de l'Orient et de la Méditerranée-J. Pouilloux, Lyon.

Bodard, Gabriel. 2010. EpiDoc: Epigraphic Documents in XML for Publication and Interchange. In Feraudi-Gruénais, Francisca (ed.). *Latin on Stone: Epigraphic Research and Electronic Archives*. Lexington Books, Lanham: 101-118.
Caldelli, Maria Letizia, Silvia Orlandi, Valentina Blandino, Valerio Chiaraluce, Luca Pulcinelli, and Alessandro Vella. 2014. EDR – Effetti collaterali. *Scienze dell'Antichità*, 20 (1): 267-289.
Campanile, Enrico. 1971. Due studi sul latino volgare. *L'Italia Dialettale*, 34: 1-64.
CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. I, Inscriptiones Latinae antiquissimae* (Lommatzsch, E. 1918 ed.).
CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. II, Addenda Nummi Indices* (Lommatzsch, E. 1931 ed.).
CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. III, Addenda altera Indices* (Lommatzsch, E. 1943 ed.).
CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. IV, Addenda tertia* (Degrassi, A. and J. Krummrey 1986 eds.).
Clackson, James and Geoffrey Horrocks. 2007. *The Blackwell History of the Latin Language*. Blackwell, Malden, Mass.
Clackson, James. 2011. Introduction. In Clackson, James (ed.). *A Companion to the Latin Language.* Wiley/Blackwell, Chichester/Malden: 1-6.
Consani, Carlo. in press. Fenomeni di contatto a livello di discorso e di sistema nella Cipro ellenistica (Kafizin) e le tendenze di "lunga durata". In Di Giovine, Paolo (ed.). *Atti del Convegno "Dinamiche sociolinguistiche in aree di influenza greca: mutamento, variazione e contatto" (Roma, 22-24 settembre 2014), Linguarum Varietas*, 5.
Cuzzolin, Pierluigi and Gerd Haverling. 2009. Syntax, sociolinguistics, and literary genres. In Baldi, Philip and Pierluigi Cuzzolin (eds.). *New Perspectives on Historical Latin Syntax*: *Syntax of the Sentence*. De Gruyter, Berlin-New York: 19-64.
De Angelis, Alessandro. in press. Un esito palatale nel latino di Sicilia: a proposito del bilinguismo greco-latino. In Di Giovine, Paolo (ed.). *Atti del Convegno "Dinamiche sociolinguistiche in aree di influenza greca: mutamento, variazione e contatto" (Roma, 22-24 settembre 2014), Linguarum Varietas*, 5.
Degrassi, Attilio. 1957-1963. *Inscriptiones latinae liberae rei publicae*. La Nuova Italia, Firenze.
Dickey, Eleonor and Anna Chahoud (eds.). 2010. *Colloquial and Literary Latin*. Cambridge University Press, Cambridge.
Donati, Margherita. in press. Variazione e tipologia testuale nel corpus epigrafico *CLaSSES I. Studi e Saggi Linguistici*, 53 (2).
Donati, Margherita, Francesco Rovai, and Giovanna Marotta. in press. Prospettive sociolinguistiche sul latino: un corpus per l'analisi dei testi epigrafici. In *Latin vulgaire - Latin tardif XI*.
Elliott, Tom. 2015. Epigraphy and Digital Resources. In Bruun, Christer and Jonathan Edmondson (eds.). *The Oxford Handbook of Roman Epigraphy*. Oxford University Press, Oxford-New York: 78-85.
Evangelisti, Silvia. 2010. EDR: History, Purpose, and Structure. In Feraudi-Gruénais, Francisca (ed.). *Latin on Stone. Epigraphic Research and Electronic Archives*. Lexington Books, Lanham: 119-134.
Feraudi-Gruénais, Francisca. 2010. An inventory of the Main Archives of Latin Inscriptions. In Feraudi-Gruénais, Francisca (ed.). *Latin on Stone: Epigraphic Research and Electronic Archives*. Lexington Books, Lanham: 157-160.
Giacalone Ramat, Anna. 2000. Mutamento linguistico e fattori sociali: riflessioni tra presente e passato. In Cipriano, Palmira, Rita D'Avino, and Paolo Di Giovine (eds.). *Linguistica Storica e Sociolinguistica*. Il Calamo, Roma: 45-78.
Glare, Peter G. W. (ed.) 1968-1982. *Oxford Latin Dictionary*. Oxford University Press, Oxford.
Herman, József. 1970. *Le latin vulgaire*. Press Universitaires de France, Paris.
Herman, József. 1978a. Évolution a>e en latin tardif? Essai sur les liens entre la phonétique historique et la phonologie diachronique. *Acta Antiquae Academiae Scientiarum Hungariae*, 26: 37-48 [also in Herman 1990: 204-216].
Herman, József. 1978b. Du latin épigraphique au latin provincial. Essai de sociologie linguistique sur la langue des inscriptions. In *Étrennes de septantaine: Travaux de linguistique et de grammaire comparée offerts à Michel Lejeune*. Éditions Klincksieck, Paris: 99-114 [also in Herman 1990: 35-49].
Herman, József. 1982. Un vieux dossier réouvert: les transformations du système latin des quantités vocaliques. *Bulletin de la Société de Linguistique de Paris*, 77: 285-302 [also in Herman 1990: 217-231].

Herman, József. 1985. Témoignage des inscriptions latines et préhistoire des langues romanes: le cas de la Sardaigne. In Deanović, Mirko (ed.). *Mélanges de linguistique dédiés à la mémoire de Petar Skok (1881–1956)*. Jugoslavenska Akademija Znanosti i Umjetnosti, Zagreb: 207-216 [also in Herman 1990: 183-194].

Herman, József. 1987. La disparition de *-s* et la morphologie dialectale du latin parlé. In Herman, József (ed.). *Latin vulgaire-Latin tardif. Actes du Ier colloque international sur le latin vulgaire et tardif*. Niemeyer, Tübingen: 97-108.

Herman, József. 1990. *Du latin aux langues romanes. Études de linguistique historique*. Niemeyer, Tübingen.

Herman, József. 2000. Differenze territoriali nel latino parlato dell'Italia: un contributo preliminare. In Herman, József and Anna Marinetti (eds.). *La preistoria dell'italiano. Atti della Tavola Rotonda di Linguistica Storica. Università Ca' Foscari di Venezia 11-13 giugno 1998*. Niemeyer, Tübingen: 123-135.

Kruschwitz, Peter. 2015. Linguistic Variation, Language Change, and Latin Inscriptions. In Bruun, Christer and Jonathan Edmondson (eds.). *The Oxford Handbook of Roman Epigraphy*. Oxford University Press, Oxford-New York: 721-743.

Lazzeroni, Romano. 1984. Lingua e società in Atene antica. *Studi classici e orientali*, 34: 16-26.

Leumann, Manu. 1977. *Lateinische Laut- und Formenlehre*. Beck, München.

Loporcaro, Michele. 2011a. Syllable, segment and prosody. In Maiden, Martin, John Charles Smith, and Adam Ledgeway (eds.). *The Cambridge History of the Romance Languages. I: Structures*. Cambridge University Press, Cambridge: 50-108.

Loporcaro, Michele. 2011b. Phonological Processes. In Maiden, Martin, John Charles Smith, and Adam Ledgeway (eds.). *The Cambridge History of the Romance Languages. I: Structures*. Cambridge University Press, Cambridge: 109-154.

Mancini, Marco. 2005. La formazione del neostandard latino: il caso delle *differentiae uerborum*. In Kiss, Sándor, Luca Mondin, and Giampaolo Salvi (eds.). *Latin et langues romanes, Études linguistiques offertes à J. Herman à l'occasion de son 80ème anniversaire*. Niemeyer, Tübingen: 137-155.

Mancini, Marco. 2006. *Dilatandis litteris*: uno studio su Cicerone e la pronunzia 'rustica'. In Bombi, Raffaella, Guido Cifoletti, Fabiana Fusco, Lucia Innocente, and Vincenzo Orioles (eds.). *Studi linguistici in onore di Roberto Gusmani*. Ed. dell'Orso, Alessandria: 1023-1046.

Marotta, Giovanna. in press. Talking stones. Phonology in Latin inscriptions. *Studi e Saggi Linguistici*, 53 (2).

Marotta, Giovanna. in press. Sociolinguistica storica ed epigrafia latina. Il corpus *CLaSSES I*. In Di Giovine, Paolo (ed.). *Atti del Convegno "Dinamiche sociolinguistiche in aree di influenza greca: mutamento, variazione e contatto" (Roma, 22-24 settembre 2014)*, *Linguarum Varietas*, 5.

Molinelli, Piera. 2006. Per una sociolinguistica del latino. In Arias Abellán, Carmen (ed.). *Latin vulgaire - Latin tardif VII. Actes du VIIe colloque international sur le latin vulgaire et tardif*. Secretariado de Publicaciones Univ. de Sevilla, Sevilla: 463-474.

Panciera, Silvio. 2013. Notizie da EAGLE. *Epigraphica*, 75: 502-506.

Poccetti, Paolo, Diego Poli and Carlo Santini. 1999. *Una storia della lingua latina*, Carocci, Roma.

Pulgram, Ernst. 1975. *Latin-Romance Phonology: Prosodics and Metrics*. Fink Verlag, Munich.

Rochette, Bruno. 1997. *Le latin dans le monde grec*. Latomus, Bruxelles.

Rovai, Francesco. in press. Notes on the inscriptions of Delos. The Greek transliteration of Latin names. *Studi e Saggi Linguistici*, 53 (2).

Vineis, Edoardo. 1984. Problemi di ricostruzione della fonologia del latino volgare. In Vineis, Edoardo (ed.). *Latino volgare, latino medioevale, lingue romanze*. Giardini, Pisa: 45-62.

Vineis, Edoardo. 1993. Preliminari per una storia (e una grammatica) del latino parlato. In Stolz, Friedrich, Albert Debrunner, and Wolfgang P. Schmidt (eds.). *Storia della lingua latina*. Pàtron, Bologna: xxxvii-lviii.

Wachter, Rudolf. 1987. *Altlateinische Inschriften. Sprachliche und epigraphische Untersuchungen zu den Dokumenten bis etwa 150 v. Chr*. Peter Lang, Bern-Frankfurt am Main-New York-Paris.

Warmington, Eric Herbert. 1940. *Remains of Old Latin. Vol. 4, Archaic inscriptions*. Harvard University Press-Heinemann, Cambridge MA-London.

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

Winter, Werner. 1998. Sociolinguistics and Dead Languages. In Jahr, Ernst Håkon (ed.). *Language Change. Advances in Historical Sociolinguistics*. Mouton de Gruyter, Berlin: 67-84.