CLaSSES: a new digital resource for Latin epigraphy

Irene De Felice
Department of Philology,
Literature and Linguistics
University of Pisa
irene_def@yahoo.it
ILC (CNR, Pisa)
irene.defelice
@ilc.cnr.it

Margherita Donati
Department of Philology,
Literature and Linguistics
University of Pisa
margherita.donati
@for.unipi.it

Giovanna Marotta
Department of Philology,
Literature and Linguistics
University of Pisa
gmarotta@ling.unipi.it

Abstract

English. CLaSSES (Corpus for Latin Socio-linguistic Studies on Epigraphic textS) is an annotated corpus for quantitative and qualitative sociolinguistic analyses on Latin inscriptions. It allows specific researches on phonological and morphophonological phenomena of non-standard Latin forms with crucial reference to the typology of the text, its origin and chronological collocation. This paper presents the first macrosection of CLaSSES, focused on the inscriptions from the archaic-early period.

Italiano. *CLaSSES (Corpus for Latin Socio-linguistic Studies on Epigraphic textS) è un corpus annotato finalizzato all'analisi socio-linguistica quantitativa e qualitativa delle e-pigrafi latine. Permette di analizzare i fenomeni fonologici e morfofonologici che caratterizzano le forme latine non standard, in relazione alla tipologia testuale, all'area geografica di provenienza e alla datazione delle iscrizioni. L'articolo presenta la prima macrosezione di CLaSSES, incentrata sulle iscrizioni risalenti al periodo preletterario e arcaico.*

1 Digital resources for Latin inscriptions

Available digital resources for Latin epigraphy include some important databases. The Clauss-Slaby database (*http://www.manfredclauss.de/gb/index.html*) records almost all Latin inscriptions (by now 696.313 sets of data for 463.566 inscriptions from over 2.480 publications), including also some pictures. It can be searched by records, province, place and specific terms, thus providing users with quantitative information. The Epigraphic Database Roma EDR (*http://www.edr-edr.it/English/index_en.php*) is part of the international federation of Epigraphic Databases called Electronic Archive of Greek and Latin Epigraphy (EAGLE). It is possible to look through EDR both as a single database or together with its partner databases accessing EAGLE's portal (*www.eagle-eagle.it*).[1]

Although they collect a large amount of data, these resources cannot provide linguists with rich qualitative and quantitative linguistic information focused on specific phenomena. The need for a different kind of information automatically extracted from epigraphic texts is particularly pressing when dealing with sociolinguistic issues.

There is a current debate on whether inscriptions can provide direct evidence on actual linguistic variations occurring in Latin society or they cannot. As Herman (1985) points out, the debate on the linguistic representativity of inscriptions alternates between totally skeptical and too optimistic approaches. Following Herman (1970, 1978a, 1978b, 1982, 1985, 1987, 1990, 2000), we believe that epigraphic texts can be regarded as a fundamental source for studying variation phenomena, provided that one adopts a critical approach. Therefore, we cannot entirely agree with the skeptical view adopted by Adams (2013: 33-34), who denies the role of inscriptions as a source for sociolinguistic variation in the absence of evidence also from metalinguistic comments by grammarians and literary authors.

That said, the current state-of-the-art digital resources for Latin epigraphic texts does not allow researchers to evaluate the relevance of inscriptions for a sociolinguistic study that would

---

[1] As regards the representation of epigraphic texts in digital form, the international project *EpiDoc* provides guidelines for encoding scholarly and educational editions in XML (*http://sourceforge.net/p/epidoc/wiki/Home/*).

like to rely on direct evidence. Furthermore, it is worth noting that within the huge amount of epigraphic texts available for the Latin language not every inscription is equally significant for linguistic studies: e.g., many inscriptions are very short or fragmentary, others are manipulated or intentionally archaising. Obviously, a (socio)linguistic approach to epigraphic texts should take into account only linguistically significant texts.

2   Aims of the corpus

The resource we present is part of a research project devoted to the sociolinguistic variation in the Latin language (see Donati et al., in press, for further details on this project). Sociolinguistic variation of Latin in Rome and the Empire is a promising research area (Rochette, 1997; Adams et al., 2002; Adams, 2003; Adams, 2007; Biville et al., 2008; Dickey and Chahoud, 2010; Clackson, 2011; Adams, 2013). Since the seminal work by Campanile (1971), many scholars have underlined that sociolinguistic categories and methods can be usefully applied to ancient languages (Lazzeroni, 1984; Vineis, 1984, 1993; Giacalone Ramat, 2000; Molinelli, 2006), even if cautiously.

Assuming this methodological perspective, our empirical analysis of Latin texts is focused on identifying and classifying specific sociolinguistic variants, mostly at the phonological and the morphophonological level. Being aware of the debate on the reliability of inscriptions currently ongoing (§ 1), we intend to investigate whether it is possible to find out relevant evidence for sociolinguistic variation in Latin *via* integration of the modern quantitative and correlative sociolinguistics with a corpus-based approach. Since digital resources devoted to this particular kind of research are actually lacking, our first step was the creation of an original resource for sociolinguistic research on Latin epigraphic texts.

First of all, we collected a corpus including a quite large amount of linguistic and metalinguistic data, to allow grounded quantitative analyses. Our hypothesis is that sociolinguistic aspects eventually emerging from the inscriptions can be detected first identifying the occurrence of non-standard forms in terms of frequency, with crucial reference to the typology of text, its origin

and chronological collocation (§ 3), and then also comparing them with their standard variants.[2]

In our analysis of the inscriptions from the archaic and the early period, we considered as non-standard those forms which deviate from the standard as it will be established between the 3rd and the 1st century BCE. For this reason we prefer here the more neutral term "non-standard" (instead of "substandard", used e.g. in Cuzzolin and Haverling, 2009), in the sense of "non-classical", i.e. not present in standard/classical Latin (for a more detailed discussion of this terms see Donati et al., in press).[3] So, e.g. in CIL I², inscription 8 (*L Cornelio L f Scipio aidiles cosol cesor*, ca. 250-200 BCE), *Cornelio* can be identified as a non-standard nominative form for the standard *Cornelius*.

3   Methods

3.1   The Corpus CLaSSES

As a first step, we collected the texts of the inscriptions we were interested in and built a corpus. Inscriptions are from the *Corpus Inscriptionum Latinarum* (CIL), the main and most comprehensive source for Latin epigraphy research. Here we present the work carried out during the first phase of our project, corresponding to one macrosection of CLaSSES.

As for the chronology, inscriptions selected are dated from 350 to 150 BCE with most of them falling into the 3rd century BCE (i.e. Archaic-Early Latin). The volumes of CIL covering this chronological segments that were systematically examined are the following: CIL I² *Pars II, fasc. I*, section *Inscriptiones vetustissimae*; CIL

---

[2] It is worth noting that assuming non-standard forms is not a trivial epistemic operation for every phase of Latin, in particular for the archaic (7th century BCE-ca. 240 BCE) and the early period (ca. 240 BCE-ca. 90 BCE). A Latin linguistic and literary standard gradually emerges between the second half of the 3rd century BCE and the 1st century BCE, culminating in the Classical period (Mancini, 2005, 2006; Clackson and Horrocks, 2007; Cuzzolin and Haverling, 2009).

[3] Even if assuming non-standard forms in archaic and early Latin may seem anachronistic in some way, this choice is based on two fundamental aspects: a) many phenomena occurring in these "deviant" 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 the Early to the Late Latin (this point is not uncontroversial, see e.g. Adams, 2013: 8); b) in any case, they provide evidence for phonological and morphophonological variation within archaic epigraphs, thus presumably indicating different levels in the diasystem.

I² *Pars II, fasc. II, Addenda Nummi Indices*, section *Addenda ad inscriptiones vetustissimas;* CIL I² *Pars II, fasc. III, Addenda altera Indices*, section *Addenda ad inscriptiones vetustissimas;* CIL I² *Pars II, fasc. IV, Addenda tertia*, section *Addenda ad inscriptiones vetustissimas.*

It is worth noting that the texts offered by CIL were also revised and checked by means of the available philological resources for Latin epigraphy of this period (Warmington, 1940; Degrassi, 1957, 1963; Wachter, 1987), in order to guarantee the most reliable and updated philological accuracy.

Since inscriptions are not all equally relevant for (socio)linguistic studies, the following texts have been excluded: 1) legal texts, since generally prone to be intentionally archaising; 2) too short (single letters, initials) or fragmentary inscriptions; 3) inscriptions from the necropolis of Praeneste, since containing only an-throponyms in nominative form.

To sum up, the final number of inscriptions in the archaic-early section of CLaSSES is 379 (1804 words). These 379 inscriptions are classified into four textual typologies:

1. *tituli sepulcrales* (n. 27), i.e. epitaphs;
2. *tituli honorarii* (n. 18), i.e. inscriptions celebrating public people;
3. *tituli sacri* (n. 96), i.e. votive inscriptions;
4. *instrumenta domestica* (n. 238), i.e. inscriptions on domestic tools.

The entire collected corpus was then manually 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. Each epigraphic text of CLaSSES was also enriched with metalinguistic information, regarding its geographic origin, its textual typology and its dating. For example, in CIL I², inscription 45 (*Diana mereto noutrix Paperia*), *mereto* is identified by the string CIL-I²-45/2, while CIL-I²-45 is associated to the following data: loc.: *Gabii*, text. typ.: *tit. sacr.*, dat.: 250-200 BCE.

3.2 Annotation of non-standard forms

In a second step, CLaSSES has been linguistically analysed (for textual interpretation of inscriptions, we mainly referred to the rich information included within CIL, as well as to Warmington, 1940; Degrassi, 1957, 1963; Wachter, 1987). This is the core part of the anno-

tation phase, that provides the corpus with a rich set of qualitative data.

Each non-standard form (already identified by its token-ID) was manually retrieved by two annotators, then also associated to both its corresponding standard form and its lemma, e.g. *cosulibus* (non-standard dat. pl.) - *consulibus* (standard dat. pl.) - *consul* (lemma). Uncertain cases were discussed by the annotators to achieve consensus.

5	Publ_ID	Token_ID	Text. Typ.	Origin	Dating	Non-standard Form	Standard Form	Lemma
6	CIL-I2-7	31	*Tit. Sep.*	Roma	200-150	*abdoucit*	*abduxit*	*abduco*
7	CIL-I2-8	2	*Tit. Sep.*	Roma	250-200	*Cornelio*	*Cornelius*	*Cornelius*
8	CIL-I2-8	6	*Tit. Sep.*	Roma	250-200	*aidiles*	*aedilis*	*aedilis*
9	CIL-I2-8	7	*Tit. Sep.*	Roma	250-200	*cosol*	*consul*	*consul*
10	CIL-I2-9	1	*Tit. Sep.*	Roma	200	*honc*	*hunc*	*hic*
11	CIL-I2-9	2	*Tit. Sep.*	Roma	200	*oino*	*unum*	*unus*

Table 1. Sample excerpt from the Excel sheet containing the annotation of CLaSSES non-standard forms.

Then, all non-standard forms were classified into three classes: vocalism, consonantism and morphophonology, according to the level in which they deviate from the standard form. For example, the nominative *consol* shows a vocalic phenomenon, because it deviates from the standard *consul* for the vowel lowering *u>o*.

A finer-grained analysis of non-standard forms led to a sub-classification of the phenomena investigated. Relevant categories adopted for this classification are the following:

1. for vowels, timbric alterations (lowering, raising), length (*apex*, I *longa*, gemination), syncope, deletion, insertion, monophthongization and archaic spellings of diphthongs;
2. for consonants, final consonant deletion (*-s, -m, -t, -r*), nasal deletion (*-ns->-s-, -nf->-f-*), insertion, assimilation, dissimilation, length (gemination, degemination), voice (voiceless *pro* voiced and voiced *pro* voiceless stops), deaspiration.

Some of these phenomena are especially relevant in the current discussion about social stratification of Latin, namely vowel lowering (*i>e, u>o*), monophthongization (*ae>e, au>o*), synchope, final *-s* and *-m* deletion (cf. among others Campanile, 1971; Pulgram, 1975; Leumann, 1977; Vineis, 1984; Herman, 1987; Weiss, 2009; Loporcaro, 2011a, 2011b; Adams, 2013; Benedetti and Marotta, 2014). Data related to vocalism and consonantism were also classified according to morphophonology: for example, the non-standard nominative *Cornelio* for *Cornelius* is annotated for the lowering *u>o, for*

the final -*s* deletion and for the non-standard -*o* ending nominative of the second declension.

This fine-grained annotation allows researchers to evaluate the statistical incidence of these discussed non-standard phenomena with respect to the corresponding standard forms, also with reference to textual typology, period, geographical origin. Thus, the linguistic annotation of CLaSSES is original and innovative, because it provides not only a list of non-standard occurrences, but especially a collection of data well-suited for a systematically grounded quantitative and qualitative analysis.

4   Possible applications

The data collected so far, together with those deriving from our future work (§ 5), will be the input for the creation of a database that will allow users to make different queries through a web interface.

There are many possible operations that can be done on what we already have. For example, as a conclusion of the annotation work conducted on texts from CIL I^2, we automatically created a *Lexicon* (that will be shortly published) of non-standard forms that contains 340 lemmas. For each lemma, all inflected non-standard forms are reported, with their corresponding inflected standard form, the indication of the inscription they belong to, the indication of their position within the inscription, e.g. *curo*: *coiraveront* (*curaverunt*), CIL I^2, 364(20); *coraveron* (*curaverunt*), CIL I^2, 59(5).

Comparing the total number of non-standard tokens with the *Index* resulting from tokenization, we are also allowed to highlight the proportion of standard and non-standard forms for a given lemma. We registered a 38,4% presence of non-standard forms in the overall corpus:
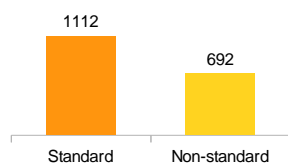

Figure 1. Non-standard vs. standard forms in the corpus (tot. 1804 words).

Similarly, for those interested in particular linguistic issues (such as vowel raising or lowering, monophthongization, etc.), a frequency count of the occurrences of a given phenomenon can be easily done, with or without considering the position of the word within the inscription.

Finally, cross-researches that take into account not only linguistic information (lemma, morphological form, phenomena) but also metalinguistic information (origin, dating, textual typology) are supported. This is one of the strongest points of our resource, because it allows to find correlations among categories. For instance, the following graph shows the percentages of non-standard forms over the total number of forms with respect to the different typologies of text:
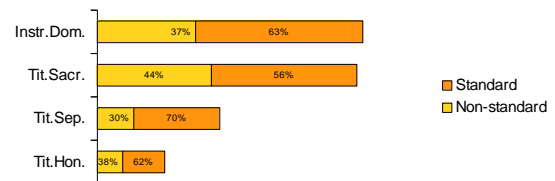

Figure 2. Percentages of non-standard and standard forms with respect to the different typologies of inscriptions.

Moreover, it is also possible to analyze the correlation between a particular phenomenon and the dating of an inscription, or its typology (whether it is classified among *instrumenta domestica*, *tituli sacri*, etc.).

This is exactly the kind of evidence we need to foster a sociolinguistic approach to epigraphic texts. These examples of possible queries follow the belief that quantitative evidence is a necessary requirement for a grounded, systematic linguistic study, even in the case of a corpus language.

5   Conclusion

CLaSSES is an epigraphic Latin corpus for quantitative and qualitative sociolinguistic analyses on Latin inscriptions, that can be useful for both historical linguists and philologists. It is annotated with linguistic and metalinguistic features which allow specific queries on different levels of non-standard Latin forms.

We have here presented the first macrosection of CLaSSES, containing inscriptions from the archaic-early period. In the next future we will collect comparable sub-corpora for the Classical and the Imperial period. Moreover, data will be organized in a database available on the web.

Aknowledgments

els and historical linguistics, coordinator Piera Molinelli (PRIN2010, prot. 2010HXPFF2_001). The research and the results related to the project are presented at *http://www.mediling.eu/.*

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, 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. Actes du colloque international*. Maison de l'Orient et de la Mediterranée-J. Pouilloux, Lyon.

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, 1918 ed.).

CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. II, Addenda Nummi Indices* (Lommatzsch, 1931 ed.).

CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. III, Addenda altera Indices* (Lommatzsch, 1943 ed.).

CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. IV, Addenda tertia* (Degrassi and Krummrey, 1986 eds.).

Clackson, James (ed.). 2011. *A Companion to the Latin Language*. Blackwell Publishing, Malden, Mass.

Clackson, James and Geoffrey Horrocks. 2007. *The Blackwell History of the Latin Language*. Oxford and Carlton, Malden, Mass.

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.

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, Giovanna Marotta, and Francesco Rovai. in press. Prospettive sociolinguistiche sul latino: un corpus per l'analisi dei testi epigrafici. In *Proceedings of the Eleventh International Conference on Latin vulgaire - Latin tardif (LVLT 11)*, Oviedo 2014.

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.

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. *Mèlanges Skok*. Jugoslavenska Akademija Znanosti i Umjetnosti, Zagreb:207-216 [also in Herman 1990:183–194].

Herman, József. 1987. La disparation 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, Pécs, 2–5 septembre 1985*. 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.

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, Etudes 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.

Molinelli, Piera. 2006. Per una sociolinguistica del latino. In Arias Abellán, Carmen (ed.). *Latin vulgaire - Latin tardif VII*. Secretariado de Publicaciones Univ. de Sevilla, Sevilla:463-474.

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

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

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.