Studying formulaic language in Latin inscriptions through Named Entity Recognition: two new Trismegistos databases

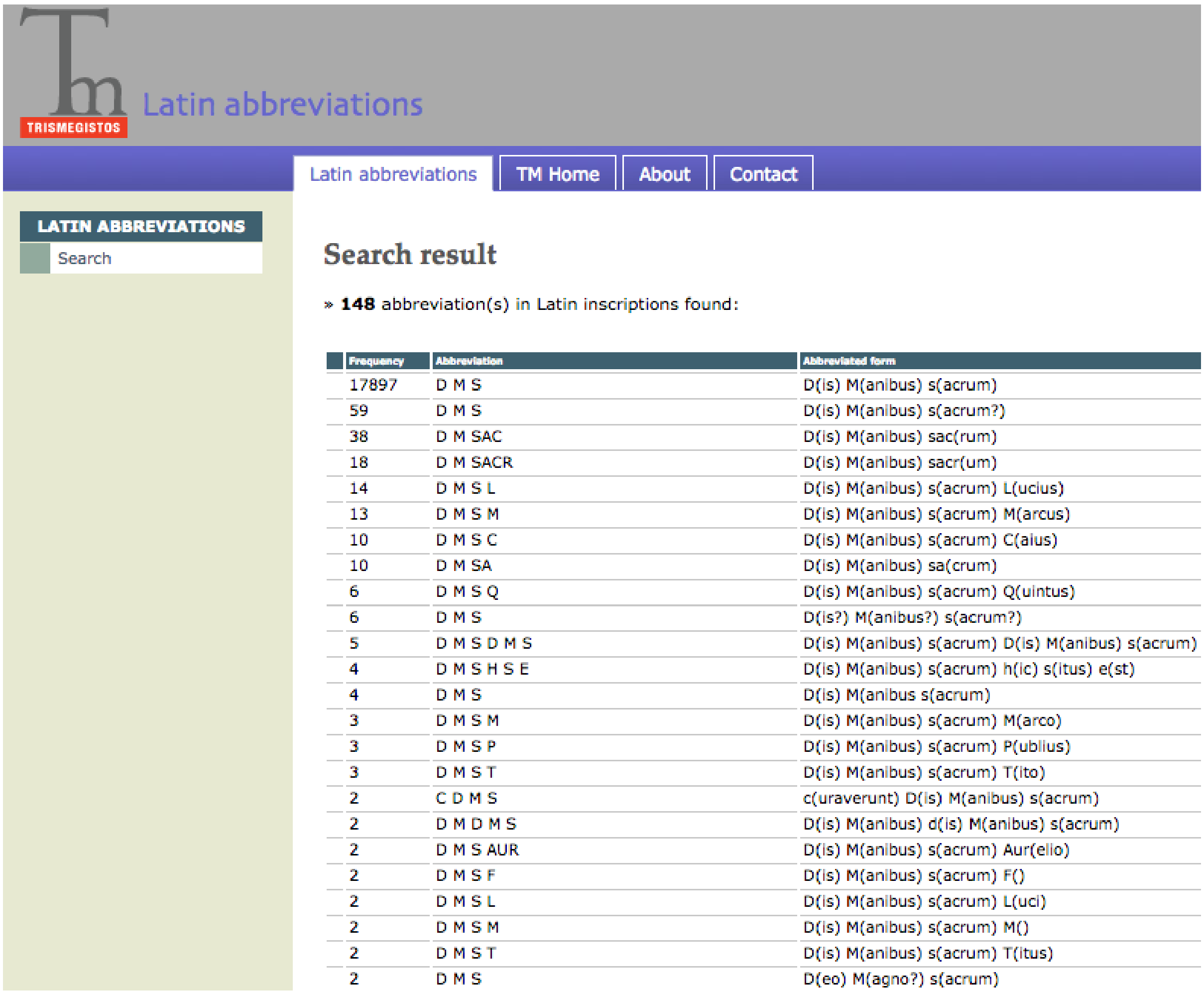
Mark Depauw
KU Leuven
mark.depauw@kuleuven.be

Tom Gheldof
KU Leuven
tom.gheldof@kuleuven.be

The availability of the full text of Latin inscriptions through projects such as EAGLE (the Europeana portal of Ancient Greek and Latin Epigraphy) opens up many exciting possibilities for new research. The application of Named Entity Recognition [NER] allows extracting all kinds of information from the full text, in an exponentially faster way than used to be possible.

Although NER originally stems from the domain of Computational Linguistics, it is currently more and more being used in the field of Digital Humanities as well. The Trismegistos project (http://www.trismegistos.org) successfully used a gazetteer-based version of this technique to distill personal and place names from the papyrological full text database DDBDP (Duke Databank of Documentary Papyri), now fully integrated in the Papyrological Navigator (papyri.info). By using the capitalization of proper names in Latin and Greek texts, a system was set up to extract clusters of capitalized words and lower-case linking words. These clusters were then compared with Greek and Latin naming patterns and analyzed accordingly. Especially the complex Roman naming system and the multiplication of genealogical identifiers proved technical hurdles. The names themselves were matched with different Trismegistos gazetteers (TM Places for toponyms, TM People for personal names). Given that not every cluster was properly recognized and some clusters were still unmatched, a second phase of manually checking these clusters remained necessary. However, most of the work had been done by computer-aided NER and the success ratio was about 85%. The procedure is now also applied to new full-text databases, again to extract personal names and place names. Trismegistos hopes soon to include not only the attestations from Egypt, but also the much larger set of data from the ancient Mediterranean in general, preserved through Latin and Greek inscriptions (Depauw and Gheldof, 2014)

This presentation, however, focuses on two new, narrowly linked projects applying NER in a different way. The first, already finished and to be launched soon, is a database of abbreviations in Latin inscriptions, the second, under construction, is one of formulaic language in the same texts, e.g. funerary inscriptions. An example is the Dis Manibus formula, meaning 'to the Manes ['gods of the underworld]', often attested at the beginning of texts, and sometimes expanded with sacrum 'devoted to'. The group can be abbreviated as D M S, but some words can be written in full, e.g. the final sacrum.



Search result

» **148** abbreviation(s) in Latin inscriptions found:

	Frequency	Abbreviation	Abbreviated form
	17897	D M S	D(is) M(anibus) s(acrum)
	59	D M S	D(is) M(anibus) s(acrum?)
	38	D M SAC	D(is) M(anibus) sac(rum)
	18	D M SACR	D(is) M(anibus) sacr(um)
	14	D M S L	D(is) M(anibus) s(acrum) L(ucius)
	13	D M S M	D(is) M(anibus) s(acrum) M(arcus)
	10	D M S C	D(is) M(anibus) s(acrum) C(aius)
	10	D M SA	D(is) M(anibus) sa(crum)
	6	D M S Q	D(is) M(anibus) s(acrum) Q(uintus)
	6	D M S	D(is?) M(anibus?) s(acrum?)
	5	D M S D M S	D(is) M(anibus) s(acrum) D(is) M(anibus) s(acrum)
	4	D M S H S E	D(is) M(anibus) s(acrum) h(ic) s(itus) e(st)
	4	D M S	D(is) M(anibus s(acrum)
	3	D M S M	D(is) M(anibus) s(acrum) M(arco)
	3	D M S P	D(is) M(anibus) s(acrum) P(ublius)
	3	D M S T	D(is) M(anibus) s(acrum) T(ito)
	2	C D M S	c(uraverunt) D(is) M(anibus) s(acrum)
	2	D M D M S	D(is) M(anibus) d(is) M(anibus) s(acrum)
	2	D M S AUR	D(is) M(anibus) s(acrum) Aur(elio)
	2	D M S F	D(is) M(anibus) s(acrum) F()
	2	D M S L	D(is) M(anibus) s(acrum) L(uci)
	2	D M S M	D(is) M(anibus) s(acrum) M()
	2	D M S T	D(is) M(anibus) s(acrum) T(itus)
	2	D M S	D(eo) M(agno?) s(acrum)

Figure 1: Different abbreviated forms of D M S in database TM Latin Abbreviations

At the end of the text H S E, in full hic situs (or sita) est or 'here lies', shows similar variation, with sepultus 'buried' as alternative. The regional and chronological variation of these and other formulae has hitherto only been investigated through samples, and the patterns of combination of formulae in a single text are virtually unexplored. We show how NER extracts abbreviations from the HTML; how the full form of the abbreviations is then used to distill formulae; how the information gathered can immediately be processed chronologically and geographically, through the link with Trismegistos; and how the combination of formulae and abbreviations can be graphically presented and analyzed through network analysis. This allows for new insights into chronological and regional differentiation of formulaic language in Latin inscriptions.

References

Depauw, M. and Gheldof, T. (2014). 'Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information', *Communications in Computer and Information Science*, vol. 416, pp. 40-52.