Classicists, Digital Humanists and Computational Linguists. Allies, not Enemies

Marco Passarotti
Università Cattolica del Sacro Cuore, Milan, Italy
marco.passarotti@unicatt.it

With a corpus of more than a billion words of texts written in Ancient Greek and Latin, the collaboration between Classics and corpus linguistics seems to be nothing but mandatory.

Ancient Greek and Latin linguistics is, indeed, corpus linguistics. The absence of native speakers makes the texts that are extant from the past the only source of empirical evidence that we have to support our understanding of Classical languages. Over the last decades, corpus linguistics has faced a methodological turn driven by the always growing size of (easily) accessible data and the availability of increasingly powerful computational tools to process them at different linguistic levels (ranging from morphology to semantics). This has made computational linguistics gain a progressively more central role in corpus linguistics, at the point that it is nowadays almost impossible to be a corpus linguist without having any expertise in computational linguistics.

Ancient Greek and Latin being dead languages, their textual corpus, although huge, is finished; no more new texts in Ancient Greek and Latin are produced (apart from isolated cases, like the official documents of the Roman Catholic Church). In principle, this would make it possible to draw definitive conclusions about Classical languages; although such a result is still far to be achieved, it is reasonable to foresee that in the (hopefully near) future almost all the texts written in Ancient Greek and Latin will be finally available in digital format.

Despite this fortunate situation, most classicists still ignore even basic methods, resources and tools of corpus and computational linguistics, like statistical analysis, annotated corpora and NLP tools. This is due to a number of reasons, among which is the obsolescence of many Faculties of Classics (which do not offer courses in corpus and computational linguistics, and sometimes not even in theoretical linguistics) and a certain degree of reluctance by classicists to learn and apply computational methods for textual analysis (and sometimes to use computers at all).

Granted, there are cultural barriers to collaborative work between classicists and computational linguists. These barriers stem from the well-known rift between the “two cultures” of the Humanities and the Sciences. The phrase “two cultures” was first introduced by C.P. Snow in 1959 (the numbers of pages reported here

Snow writes of “two groups, comparable in intelligence, identical in race, not grossly different in social origin, earning about the same incomes, who have almost ceased to communicate at all” (p. 169). These two groups “have a curious distorted image of each other” (p. 169). “The non-scientists have a rooted impression that the scientists are shallowly optimistic, unaware of man’s condition. On the other hand, the scientists believe that the literary intellectuals are totally lacking in foresight” (p. 170). Nevertheless, “the clashing point of two subjects, two disciplines, two cultures […] ought to produce creative chances” (p. 172). Snow concludes that “there is only one way out of all this: it is, of course, by rethinking our education”.

This was written in 1959. Sure, things have moved on since then, but not as much as it could (and should) be. If we consider computational linguists and classicists as representatives of scientists and humanists respectively, digital humanists should place themselves somewhere in the middle, playing the role of those who are provided with enough expertise to fill the gap between the two groups. Unfortunately, their relation with both “traditional” humanists/classicists and computational linguists is far from being perfect. Actually, sometimes classicists look at digital humanists as lost children who keep spending their precious time playing with computers; sometimes digital humanists look at traditional humanists as conservative scholars who do not want to admit that their time is gone; sometimes computational linguists look at digital humanists as unripe computational linguists; and sometimes digital humanists just do not look at computational linguists at all, reinventing the wheel for the nth time.

In the talk, I will discuss my view on the relation between computational linguistics and digital humanities, by focussing particularly on their different approaches to (linguistic) data. Despite many commonalities, the two disciplines are often separate worlds. For instance, most of the workshops co-located with computational linguistics conferences that deal with topics that should attract digital humanists (like LaTeCH, ACRH and Computational Linguistics for Literature) are attended mostly (or solely) by computational linguists. On the other side, you do not see many computational linguists (and traditional classicists, as well) attending events of digital humanities, like the annual DH Conference.

Digital humanists who deal with Classical languages must look at computational linguistics more than they have done so far. Ancient Greek and Latin still lack advanced NLP tools and state-of-the-art language resources. Several problems that “digital classicists” have to deal with while building new resources and tools for Classical languages have already been solved by computational linguistics in similar projects for modern languages, and solutions can be ported thanks to their (usually) language-independent nature.

On its side, digital humanities poses a number of specific problems that are simply unknown to computational linguistics, like philological issues, which imply different versions of the same text (something computational linguists are not used
to deal with), the absence of native speakers, and stylistic questions: most of the
texts written in Classical languages are of high register, raising problems that are
just different from those of *The Wall Street Journal*.

As a computational linguist who has to do with Classical languages, I will
discuss a number of pros and cons of the different relations with data pursued by
the disciplines here concerned. I will also bring arguments to show that the very
name ‘digital humanities’ is indeed self-defeating and not strategic to fully reach
the world of classicists, since it creates two (linked, but still separate) groups in the
humanities: the digital and the non-digital ones. In 2014, a humanist who refuses
to “be digital” in his everyday work is simply a bad professional: paradoxically,
the name ‘digital humanities’ warrants her/his existence.

Following the status of position paper of these words, my aim is not to suggest
one conclusive solution, but to pose the problem as the first step to overcome the
present state of affairs and finally make classicists, digital humanists and computa-
tional linguists what they deserve to be: allies, not enemies.