Syllabic Quantity: Rhythmic Clues for Latin Authorship

Description

**Syllabic Quantity: Rhythmic Clues for Latin Authorship**

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani

In today’s digital landscape, it’s extremely easy for people to communicate while hiding their true identities, if they so wish. Indeed, since many digital communication platforms do not perform proper identity checks, users can create and freely use anonymous or false accounts. This has incentivized researchers to develop techniques for uncovering who’s behind these anonymous identities, and unmasking forged ones. Such techniques are known as “Authorship Identification” (AID, for short), and are part of the field of “Authorship Analysis”, whose goal is to infer characteristics (such as the gender, the age, or the native language) of the authors of written documents.
Generally speaking, AID techniques try to spot the “hand” of a given author, by analyzing the peculiarities of writing style that can “give an author away.” The core of this practice, also known as “stylometry” (i.e., the quantitative measurement of writing style), does not rely on the investigation of the artistic value of a written work, or of its meaning, or of the events and facts that the text describes, but on a quantifiable characterization of its style. This characterization is typically achieved through an analysis of the frequencies of linguistic events (“style markers”). Examples of such events can be the use of a certain punctuation symbol, of a certain adverb, or of a peculiar sequence of a given conjunction followed by an adverb. Actually, it has been demonstrated that it is precisely the elements of apparently minimal significance, such as the use of punctuation symbols, of personal pronouns, or of words of a certain length, that are the best style markers, since they tend to be out of the conscious control of the author, and hence harder to modify or imitate.

---

Modern computational approaches to AID employ such linguistic “features” in order to train a machine-learned classification algorithm to recognize the writing style of the author(s) of interest. AID is thus employed in any task in which the paternity of a document is unknown, disputed, or possibly forged, and thus must be uncovered. It was indeed linguistic analysis that allowed to uncover the identity of the “Unabomber” (see https://en.wikipedia.org/wiki/Ted_Kaczynski), one of the most famous serial killers of recent history. Moreover, countless manuscripts in our cultural heritage had their authorship lost in the centuries, or purposefully hidden or mystified. We focus on the application of AID in these ancient settings (and, specifically, on the Latin language).
Why can syllabic quantity be relevant for authorship?

As mentioned above, the basis of an AID method are the style markers, a.k.a. “stylistic features.” Since the very inception of stylometry in the late 19th century, researchers have proposed many stylistic features, some of which have become standard nowadays. However, the quest for new, potentially better features is still ongoing.

The Latin language is based on syllables, that is, oscillations of sound in the pronunciation of a word, that are characterized by their “quantity” (“long” or “short”, depending on the amount of time required to actually pronounce it). In fact, well-chosen patterns of these syllables were the basis of Latin prosody: Latin authors gave a “rhythm” to the discourse by combining long and short syllables in rhythmic schemes, not only in poetry, but in prose as well. Orators such as Cicero were particularly aware of the effects of such metric devices, and many scholars have pointed out how certain Latin authors consistently show a certain (more or less conscious) preference for specific rhythmic patterns. The “musicality” of an author’s production, i.e., the rhythmic patterns s/he used, might thus play an important role in the identification of that author’s style; this is the intuition where our work starts.

Our contribution

In our research, we extract the syllabic quantity patterns from Latin texts and use them as stylistic features for a machine learning algorithm (Support Vector Machine). This algorithm trains a classifier that guesses the author of a Latin document from among a set of possible candidates. Tests that we have run on different datasets of Latin texts show that these rhythmic features improve the accuracy of the authorship identification, thereby confirming the idea that rhythm is indeed a stylistic marker for Latin prose.

This opens up possible directions for future work, such as investigating the importance of rhythm on AID tasks for other languages, starting from ones linguistically close to Latin such as Italian or Spanish. In
particular, it would be fascinating to study whether also modern-day prose writers unconsciously opt for specific rhythmic patterns, to the point of being somehow recognizable thanks to them.

If you are interested in this subject and want more details on the research hypothesis and experimentation setting of this work, the full article by Corbara, Moreo, and Sebastiani is available at http://doi.org/10.1002/asi.24660.