description of the style of each author is in general subjective, and therefore hard to derive in natural language; it is even harder to find a description which enables a machine to automatically tell one author from the other. A literature review on modern authorship attribution methods, usually coming from the fields of machine learning and statistical analysis, is reported in Stamatatos (2009); Jockers and Witten (2010); Koppel et al. (2009); Grieve (2007); Juola (2006). Among these, algorithms based on similarity measures such as Benedetto et al. (2002) and Koppel et al. (2011) are widely employed and usually assign an anonymous text to the author of the most similar document in the training data.

During the last decade, compression-based distance measures have been effectively applied to cluster texts written by different authors (Cilibrasi and Vitányi, 2005) and to perform plagiarism detection (Chen et al., 2004). Such universal similarity measures, of which the most well-known is the Normalized Compression Distance (NCD), employ general compressors to estimate the amount of shared information between two objects. Similar concepts are also used by methods using runlength histograms to retrieve and classify documents (Gordo et al., 2013). Experiments carried out in Oliveira et al. (2013) conclude that NCD-based methods for authorship analysis outperform state-of-the-art classification methodologies such as Support Vector Machines. A study on larger and more statistically meaningful datasets shows NCD-methods to be competitive with respect to the state of the art (de Graaff, 2012), while Stamatatos (2009) reports that compression-based