Neutralising the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels

José Calvo Tello
jose.calvo@uni-wuerzburg.de
University of Würzburg

Daniel Schlör
daniel.schloer@informatik.uni-wuerzburg.de
University of Würzburg

Ulrike Henny
ulrike.henny@uni-wuerzburg.de
University of Würzburg

Christof Schöch
christof.schoech@uni-wuerzburg.de
University of Würzburg

Summary

We propose a way to work with the stylometric distance measure Delta to analyse the subgenre of texts written by different authors. For that, we neutralize the author signal by penalizing the texts from the same writer, allowing the texts to have their shortest distances to other authors' works. We test this method with several subcorpora of Spanish prose and a corpus of French theatre.

Stylometry and Delta beyond Authorship

Since John Burrows proposed it in 2002, Delta has been one of the most used and researched methods in stylometry and authorship attribution. Burrows explained it as "expression of difference, pure difference" (2002: 269) and is based on basic statistical concepts like most frequent words, z-scores and the Manhattan distance between each pair of texts.[2] Burrows closes his paper with an unanswered question about why Delta works so well.

Other researchers such as Hoover (2004b: 454), Argamon (2008), Plasek (2014) or Evert et al (2015: 79) have confirmed that we are still far from being able to answer this question. This lack of understanding has not stopped the stylometric community of trying to improve Delta (Hoover 2004a; Argamon 2008; Eder 2013). Smith and Aldridge (2011) have proposed Cosine Delta which gives the best results in different languages (Jannidis et al. 2015).

Since Delta is sensitive to aspects or *signals* like genre or period (Burrows 2002), the corpora for authorship attribution tend to be homogenous in those aspects. Research has been conducted to try to separate signals (Schöch 2013 and 2014) or selecting the words that contribute to them using Recursive Feature Elimination (Büttner and Proisl 2016). Jannidis and Lauer (2014) and Hoover (2014) show how Delta can be used to distinguish genre and periods within the works of a single author. Other researchers have used other methods such as classification (Hettinger et al., 2016; Underwood 2014) or logistic regression (Jockers 2013; Riddell and Schöch 2014) to similar ends.

Neutralizing Author Signal in Delta

Our proposal is to neutralize the author signal directly on the Delta matrix. We use a testing corpus of texts from three Spanish authors and three subgenres. Detailed information about the corpora, files, parameters and scripts is in our GitHub repository. We applied Cosine Delta (5000 MFW) with Stylo (Eder, Rybicki and Kestemont 2016) and visualized the resulting distance matrix with Python:

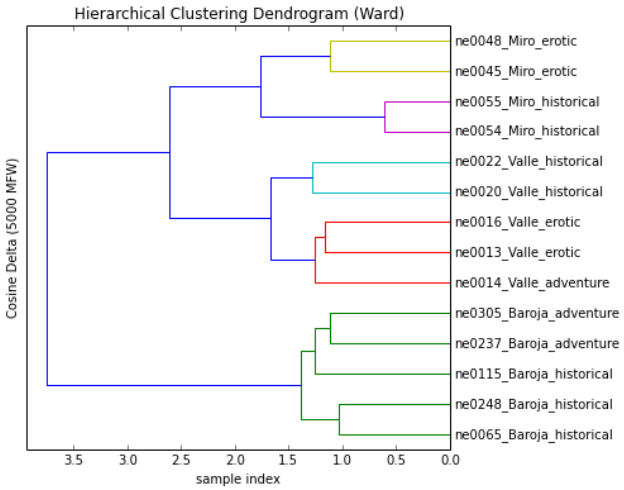

Figure 1: Dendrogram from Cosine Delta

As expected, the texts are clustered by author, with sub-clusters of subgenres. The underlying Delta Matrix contains distances between all texts:

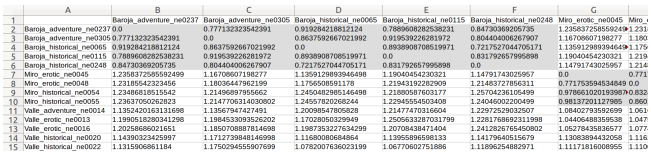

Figure 2: Cosine Delta Matrix

We see a tendency of lower Delta values for documents of the same author (below 1.0) in comparison to documents of different authors (above 1.0). But what about the closest texts written by a different author? For the historical novel in column E, they are in the rows 14 and 15 and are historical novels, as well. This pattern is found for the majority of the texts. How could we cluster the texts preferring the closest text from other authors? And if we are able to neutralize the author signal, will we see noise or subgenre clusters?

Our proposal is to penalize the distances between the texts of the same author (cf. Lu and Leen 2007 for penalization in image clustering), making them closer to the average distance of texts of different authors, then cluster the neutralized distance matrix and measure the cluster homogeneity by author and subgenre.

We define the set of all documents by an author *a* as A_a, the collection containing all documents by all authors as *C* and total number of documents in the collection is defined as *c*:

$$A_a := \{d_1, \cdots, d_{m_a}\}$$
$$C = \{A_1, \cdots, A_n\}$$
$$c := |\bigcup C|$$

Note that each document is in exactly one author-document set A_i.

First, we calculate the average distance of texts of all pairwise different authors (in fig. 2, all the distances in black). We call this value the **mean of different authors or *M(C)*** and for this collection its value is 1.16.

$$M(C) := \frac{\sum_{\substack{A_a, A_b \in C, a \neq b \\ d_i \in A_a, d_j \in A_b}} \Delta(d_i, d_j)}{\sum_{A} |A| \cdot (c - |A|)}$$

Second, we calculate the **mean of the texts of each author a *M(A_a)*** (in fig. 2, the distances in grey).

$$M(A_a) := \frac{\sum_{\substack{d_i, d_j \in A_a \\ i \neq j}} \Delta(d_i, d_j)}{|A_a| \cdot |A_a - 1|}$$

For each author, we subtract his/her mean value from the mean of different authors *M(C) - M(A_a)* resulting in the **difference of the author**. This value represents how far the texts of a specific author are to the mean of different authors:[4]

author	mean	difference
Miro	0.607	0.552
Baroja	0.669	0.490
Valle	0.752	0.407

Figure 3: Means and differences of author

Third, we add the difference of the author *M(C) - M(A_a)* to the Delta values of text of the same author. This gives a Neutralized Delta-function as follows:

$$\forall d_i \in A_a, d_j \in A_b$$
$$\tilde{\Delta}(d_i, d_j) := \begin{cases} \Delta(d_i, d_j) & \text{for } a \neq b \\ \Delta(d_i, d_j) + (M(C) - M(A_a)) & \text{for } a = b \end{cases}$$

This converts the table from Figure 1 into a Neutralized Delta matrix:

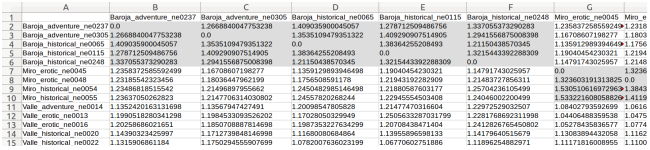

Figure 4: Author-Neutralized Delta matrix

The values in grey are now in general above 1.0: the texts of the same author have been separated, showing relations between texts independently of authorship. Now the adventure and historical novels of Baroja in columns C and D have their closest text in works of different authors but belonging to the same subgenre.

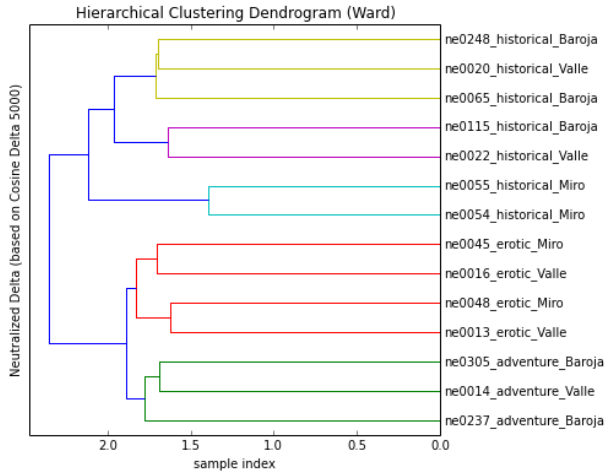

Fig. 5: Author-neutralized Delta dendrogram

In comparison with Figure 1, this dendrogram allows us to see new text relations beyond authorship but within subgenre, showing clusters with different authors but the same subgenre: for example, the cluster of historical novels by Baroja and Valle or the two very close subclusters of erotic novels by Miró and Valle.

Tests and Evaluation

For the evaluation, the homogeneity of the clusters (Rosenberg and Hirschberg, 2007) was measured. This measure yields values between 0 and 1. As ground truth, the metadata about author and subgenre have been used. The results for the dummy corpus:

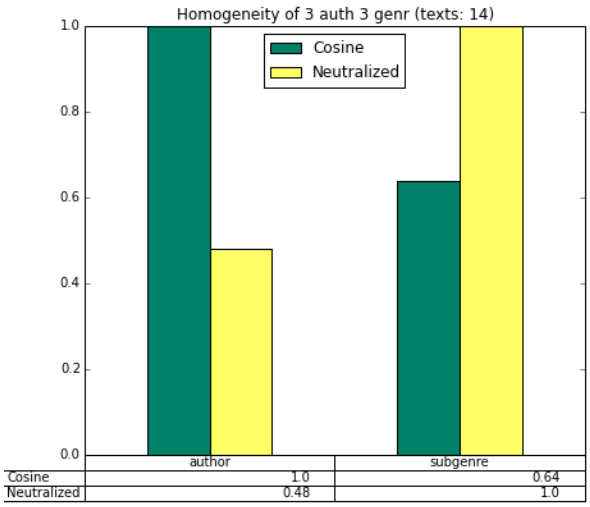

Figure 6: Homogeneity of Cosine and Neutralized Delta for author and subgenre

The homogeneity of the clusters of Cosine Delta (see fig. 1) are perfect for authors and much lower for subgenre, because the author clusters contain subgenre subclusters. The homogeneity of the clusters of Neutralized Delta (see fig. 5) is lower for authorship (as expected), but not for subgenre. In this case the neutralization of the author signal only deteriorates the homogeneity for authorship but improves the homogeneity for subgenre.

We have analysed different subgenres present in the whole corpus for test the method. We created subcorpora of historical, bildungsroman, erotic and adventure novels:[5]

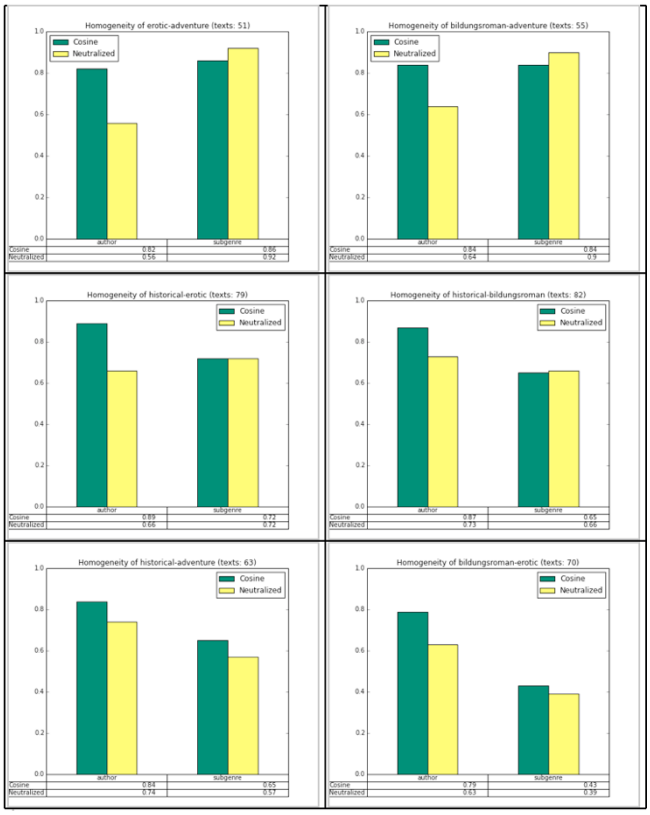

Figure 7: Homogeneities for Spanish prose subcorpora

As expected, the neutralization consistently deteriorates the homogeneity for author (between -0.26 and -0.1) while the homogeneity for subgenre is not deteriorated (between -0.08 and 0.06). The homogeneity for subgenre of adventure compared to erotic and bildungsroman get the best results (over 0.9) and they even improved on results with Cosine. Adventure novels are also best recognized in classification tasks (Hettinger et al. 2016). Subgenres which are very difficult to differentiate like historical and adventure (Pedraza

Jiménez and Rodríguez Cáceres 1983: 672 and 1987: 459) get one of the worst results.

The results are similar when testing other corpora, such as a corpus of French drama (Schöch et al. 2015) and a corpus of Spanish American novels:

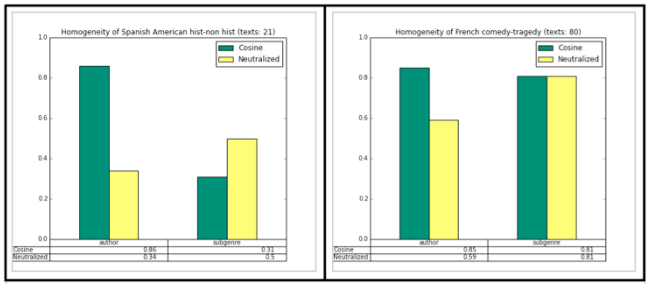

Figure 8: Homogeneity values for French drama and Spanish American novels

Conclusion and future work

Our main goal was to present a method to neutralize the Delta distances of the same author using the difference between the mean of the author and the mean of different authors. Tested on eight subcorpora, this procedure, as we expected, deteriorates the homogeneity of authorship clusters but maintains the subgenre homogeneity, improving it for some cases. That discovers relations between texts (see fig. 5) that were hidden by authorship. This procedure brings a new way of working with Delta beyond authorship attribution.

Both Cosine and Neutralized Delta show very different results for the comparison of different subgenres, something which points to the different internal structure of the subgenres. The comparison of very different subgenres (like adventure against erotic or bildungsroman) gets higher subgenre cluster homogeneity. Neutralized Delta could be used for comparing different corpora of specific subgenres and test the significance of the results to better characterize these subgenres. In an ideal scenario, we would like to test on a perfect balanced corpus where a set of authors are represented in all subgenres of the same period.

For future work, we will analyse how different parameters like versions of Delta or number of MFW affect the results. We also plan to transfer the approach to an earlier step in the Delta procedure and penalize the word z-score vectors.

We look forward to the feedback of the international DH community about this new use of the very effective "expression of difference, pure difference" which is Delta.

Acknowledgements

To avoid confusion regarding intellectual property, we would like to make it clear that the main idea and implementation are the work of the first author. Other authors have brought important remarks, feedback, some of the corpora and have helped with the redaction and the formalisations.

Bibliography

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, **17**(3): 267–87 http://revistacaracteres.net/revista/vol5n1mayo2016/entendiendo-delta.

Büttner, A. and Proisl, T. (2016). Stilometrie interdisziplinär: Merkmalsselektion zur Differenzierung zwischen Übersetzer- und Fachvokabular. *DHd 2016: Modellierung, Vernetzung, Visualisierung*. Leipzig: Universität Leipzig, pp. 66–69 http://www.dhd2016.de/abstracts/sektionen-002.html.

Calvo Tello, J. (2016). Entendiendo Delta desde las Humanidades. *Caracteres. Estudios culturales y críticos de la esfera digital*, **5**(1): 140–76.

Eder, M. (2013). Bootstrapping Delta: a safety-net in open-set authorship attribution. *DH2013*. Lincoln: UNL https://sites.google.com/site/computationalstylistics/preprints/m-eder_bootstrapping_delta.pdf?attredirects=0.

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: A package for computational text analysis. *The R Journal*, **16**(1): 1–15 https://journal.r-project.org/archive/accepted/eder-rybicki-kestemont.pdf.

Evert, S., Proisl, T., Jannidis, F., Pielström, S., Schöch, C. and Vitt, T. (2015). Towards a better understanding of Burrows's Delta in literary authorship attribution. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver CO: Association for Computational Linguistics, pp. 79–88 .

Hettinger, L., Reger, I., Jannidis, F. and Hotho, A. (2016). Classification of Literary Subgenres. *DHd 2016*. Leipzig: Universität Leipzig, pp. 154–58 http://dhd2016.de/boa.pdf.

Hoover, D. L. (2004a). Testing Burrows's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Hoover, D. L. (2004b). Delta Prime?. *Literary and Linguistic Computing*, **19**(4): 477–95.

Hoover, D. L. (2014). A Conversation Among Himselves: Change and the Styles of Henry James. In Hoover, D. L., Culpeper, J. and O'Halloran, K. (eds), *Digital Literary Studies*. New York & London: Routledge, pp. 90–119.

Jannidis, F. and Lauer, G. (2014). Burrows's Delta and Its Use in German Literary History. In Erlin, M. and Tatlock, L. (eds), *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*. Rochester: Camden House, pp. 29–54 gerhardlauer.de/index.php/download_file/view/335/1/.

Jannidis, F., Pielström, S., Schöch, C. and Vitt, T. (2015). Improving Burrows' Delta – An empirical evaluation of text distance measures. *DH 2015*. Sydney: ADHO http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html.

Jockers, M. L. (2013). *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Lu, Z. and Leen, T. K. (2007). Penalized Probabilistic Clustering. *Neural Computation*, **19**(6): 1528–67

Pedraza Jiménez, F. B. and Rodríguez Cáceres, M. (1983). *Manual de literatura española. 7: Época del realismo*. Pamplona: Cénlit.

Pedraza Jiménez, F. B. and Rodríguez Cáceres, M. (1987). *Manual de Literatura Española. 9: Generación de Fin de Siglo: Prosistas*. Pamplona: Cénlit.

Plasek, A. (2014). Incommensurability? Authorship, Style, and the Need for Theory. DH2014: Lausanne: ADHO http://dharchive.org/paper/DH2014/Paper-755.xml.

Riddell, A. and Schöch, C. (2014). Progress through Regression. *Digital Humanities DH2014:*. Lausanne: ADHO http://dharchive.org/paper/DH2014/Paper-60.xml.

Rosenberg, A. and Hirschberg, J. (2007). V-Measure: A conditional entropy-based external cluster evaluation measure. Prague: Association for Computational Linguistics, pp. 410–20 https://aclweb.org/anthology/D/D07/D07-1043.pdf.

Schöch, C. (2013). Fine-tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater. *DH2013*. Lincoln: UNL http://dh2013.unl.edu/abstracts/ab-270.html.

Schöch, C. (2014). Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In Schöch, C. and Schneider, L. (eds), *Literaturwissenschaft im digitalen Medienwandel*. pp. 130–57 http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf.

Schöch, C., Henny, U., Calvo Tello, J. and Popp, S. (2015). *The CLiGS Textbox*. Würzburg: University of Würzburg https://github.com/cligs/textbox.

Smith, P. W. H. and Aldridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method. *Journal of Quantitative Linguistics*, **18**(1): 63–88

Underwood, T. (2014). Understanding Genre in a Collection of a Million Volumes, Interim Report. https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251