Two Methods of Author Identification: the Gary/Ajar Case

Vina Tirvengadum

138 Kingsway avenue, Winnipeg, Manitoba, R3M OH1, Canada

KEYWORDS: author style statistics

AFFLIATION: University of Manitoba

E-MAIL: thngfish@alpha.remcan.ca
FAX NUMBER: (204) 452 0481
PHONE NUMBER: (204) 452 0481

1. The Romain Gary/Émile Ajar controversy

In 1974 the well-established French author Romain Gary, a recipient of the *Prix Goncourt* (France's highest literary award) wanting to escape from "le parisianisme" and the context in which critics and readers alike had pegged him, published *Gros-Câlin* under the pseudonym of Émile Ajar. By adopting this new name, he wanted to start over and have his work judged on its own merits and not on his established reputation. His stratagem payed off: *Gros-Câlin* immediately attracted the attention of critics and readers alike, became a best-seller while new novels published under the Gary name were not as successful. When a few astute critics noticed similarities between Gary and Ajar, Gary vehemently denied having any connection with Ajar. But when these rumours persisted, Gary fearing that he might be found out, coaxed Paul Pavlowich, his nephew, to impersonate Ajar. To quash any further rumour that he was Ajar, he even accused Ajar of plagiarising him. With increasing paranoia, he wrote a second Ajar novel, *La Vie devant soi* which became an immediate success and was awarded *the Prix Goncourt*. Romain Gary thus became the first author (and presumably the last one) to receive this award twice, which is strictly forbidden by the Goncourt academy.

It was after his suicide in 1980, that two books *Vie et Mort d'Émile Ajar* by Romain Gary (published posthumously in 1981) and *L'Homme que l'on croyait* (1981) by Paul Pavlowich enabled readers to demystify the double disguise: Ajar was the pseudonym of Gary and not of Paul Pavlowich. Critics immediately saw similarities between the Gary and the Ajar novels in terms of ideas, characters, images, recurring motifs and phrasings. But so far, no one has undertaken a comparative analysis of the Gary and Ajar style using statistical methods. If Buffon's assertion that "le style c'est l'homme même" is true, the Émile Ajar corpus should prove to be statistically similar to the Romain Gary corpus. Romain Gary thus provides an excellent example for authorship attribution study, or stylometric study in it's broad sense. Authorship attribution study is the analysis of stylistic idiosyncracies of an author as an index of authenticity, or an attempt to capture quantitatively the essence of an author's use of language. However, the purpose of this paper is not to attribute unknown works to Romain Gary (we now know that Gary and Ajar are the same author) but rather to compare his style in two novels written under the pseudonym of Émile Ajar to two novels written under the Romain Gary name.

2. Hypothesis

Nearly all experts from Buffon to Barthes postulate that style, which is dictated by the subconscious, forms the genetic fingerprint of a writer's work. It is therefore impossible to disguise one's style. It would then follow that works written under a pseudonym should contain the genetic fingerprint of the original writer. A stylistic analysis of the pseudonymous corpus would reveal that it is statistically similar to the work of the original writer. Therefore, there should be no significant difference between the style of Romain Gary and that of Émile Ajar.

The problem with this assumption, however, is that there is no hard evidence to support the idea that authors have an unconscious as well as a conscious aspect to their style. The two applications of author recognition, namely (1) authorship attribution and (2) chronological studies, have made contradictory claims. The attribution method claims that the unconscious aspect of a writer's style remains constant throughout his/her life, in other words, an author will leave his/her stylistic fingerprint on each and every one of his/her works. However the chronological studies claim that the unconscious features change throughout the author's life and develop rectilinearly which then allow a work to be dated.

3. Methodology

Since the findings of Mosteller and Wallace on the *Federalist papers* in 1964, linguists, statisticians and literary critics alike use the method of stylistic fingerprinting to attribute authorship to disputed works. Like many experts, they assert that style which comes from the unconscious mind, forms the genetic fingerprint of any writing and helps distinguish one author from the next. While the methods used are not the same, statistical models play an important role for their findings.

The various methods used for the determination of authors and the measurement of style have been:

word-lengths (Mendenhall, 1887), (Brinegar, 1963), (Mosteller and Wallace, 1964); number of syllables per word (W. Fucks, 1952); sentence-length (Yule, 1938), (Williams, 1940) and Kjetsså (1979) etc. But this paper focuses on vocabulary distribution as a general style discriminant. The analysis deals mostly with (a) high frequency words and (b) synonyms as discriminants of style. In their analysis of *The Federalist Papers* Mosteller and Wallace focus their research mainly on the use of synonyms such as "while" and "whilst" as style discriminants to make their conclusions, while J.F. Burrows in *Computation into criticism: A Study of Jane Austen's novels and an experiment in Method* concentrates on the use of high frequency words as an essential element of an author's style. Burrow's assertion (and the one shared in this paper) rests on the premise that the essential element of an author's style is not confined in the rare lexical words likely to evoke love, hate or war, but in the forty or fifty unambiguous and most common word types in the entire corpus. Although this method has been put into question by F.J. Damereau in 1975 it is still widely used and is the one used in this paper.

In order to test if indeed the Gary corpus is statistically similar to the Ajar corpus, four books were scanned, two by Romain Gary: *Au-delà de cette limite votre ticket n'est plus valable* (1975) and *Clair de femme* (1977), plus two Émile Ajar books: *Gros-Câlin* (1974) and *La Vie devant soi* (1975). As Romain Gary's literary career spanned nearly thirty years, these four books, all written within a four year period, were chosen in order to avoid the problem of chronology. After the scanning process, alphabetical concordances and word counts (using the O.C.P.) were established. From these another programme sorted out the words in descending order yielding a list of highest frequency words in the Ajar and Gary novels. But as testing Ajar against Gary would not have been conclusive, other French twentieth century novels were included in the tests: Camus' *L'Étranger* (1942); Gide's *L'Immoraliste* (1902) and *La Porte Étroite* (1909); Mauriac's *Le Noeud de vipères* (1932). For these novels, a series of texts held in the ARTFL database were used. These five books were chosen because they are of similar lengths to the Gary and Ajar books, they belong to the same genre (the novel) and in all of them, as in the Gary and Ajar novels, the narrator is the first person singular. Furthermore a list of high frequency words compiled by Engwall in 1974 and made up of the most common words found in twentyfive best sellers in France from 1962 to 1968 was included in this study. All work was done on keywords, not lemmas.

4. Results

After a graph on occurrences of words per thousand in each text was plotted, a definite pattern was seen to emerge: the second Ajar novel *La Vie devant soi* deviated considerably from the other Gary novels as well as all the other novels. In order to determine the significance of this difference, three statistical tests were done: the t-test, the Pearson correlation and the chi-squared test. All these tests showed that *La Vie devant soi* was statistically different from the Gary novels as well as all the other novels. After a confidence interval of 99% for the t-test was constructed, it was observed that among the high frequency words *La Vie devant soi* had 55.5 percent of occurrences that fell outside the expected range, while the other novels ranged between 11.6 and 28.3 percent of the expected value. The Pearson correlation yielded similar results. While correlation between all the books were high (as expected) they were highest among nearly all the other texts ranging between 0.94 and 0.96. However *La Vie devant soi* showed a lower correlation with the other books, ranging between 0.66 with Engwall (lowest correlation) and 0.82 with *Gros-Câlin* (highest correlation). It's correlations with the two Gary novels were 0.78 and 0.80, while the correlations between the two Gary novels were 0.95.

The Chi-squared test also showed that among the high frequency words *La Vie devant soi* had eleven observations that fell above 3.84 (a significant chi-squared value at one degree of freedom) while the other novels had between one and six observations falling above 3.84. When the high frequency list was condensed to a context-free list, *La Vie devant soi* still gave the highest chi-squared value of 43.36, while chi-squared values for the other Gary novels ranged between 9.08 and 18.23. A further chi-squared test on pairs of synonyms was done. As expected it showed very high chi-squared values, but the highest value was again found in *La Vie devant soi* at 6,198 followed by *Gros-Câlin* at 1,746.

5. Conclusions

The statistical results found in this paper prove beyond any doubt that high frequency words and pairs of synonyms, which are considered, by many, to be the unconscious elements of an author's style, can indeed be consciously manipulated by the author. The notion that function words (and synonyms) constitute a genetic fingerprint of an author's style has therefore been disputed in the Romain Gary – Émile Ajar case. While *Gros-Câlin*, the first Ajar novel closely resembles the two Gary novels, *La Vie devant soi* is so significantly different from the two Gary novels that it could have been written by another author. It would appear that Gary did not feel the need to drastically

change his style in *Gros-Câlin* his first Ajar novel, feeling confident that nobody would make the connection between him and Ajar. But when critics saw similarities between *Gros-Câlin* and the Gary novels, he became increasingly paranoid, set out to prove that he was not Ajar and wrote *La Vie devant soi*. In so doing, he consciously or unconsciously changed the genetic fingerprint of the Gary style in that novel.

References

Brinegar, C.S. "Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship." *Journal of the American Statistical Association*, 58 (1963), 85-96.

Burrows, J.F. *Computation into Criticism: A Study of Jane Austen's Novels and an experiment in Method*. Oxford: Clarendon Press, 1987.

Damerau, F.J. "The Use of Function Word Frequencies as Indicators of Style." *Computers and the Humanities*, 9 (1975): 271-280.

Ellegard, A. *A Statistical Method for determining Authorship: The Junius Letters, 1769-1772*. Gothenburg: University of Gothenburg, 1962.

Fucks, W. "On the Mathematical Analysis of Style." *Biometrika*, 39 (1952): 122-129.

Gary, R. *Vie et mort d'Émile Ajar*. Paris: Gallimard, 1981?

Kjetsså, G. "*And Quiet Flows the Don* Through the Computer." *Association for Literary and Linguistic Computing Bulletin*, 7 (1979): 248-256.

Mendenhall, T.C. The Characteristic Curves of Composition." *Science*, IX (1887): 237-249.

Mosteller, F. and D.L. Wallace. "Inference and Disputed Authorship: *The Federalist*." Reading, M.A: Addison-Wesley, 1964.

Pavlowich, P. *L'Homme que l'on croyait*. Paris: Fayard 1981.

Williams, C.B. "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style." *Biometrika*, 31 (1940), 356-361.

Yule, G.U. "On Sentence-Length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship." *Biometrika*, 30 (1938): 363-390.