[redacted] 13(2-3):265-283, 2006
2006 [redacted] Ltd
ISSN: [redacted] Print/1744-5035

A not so anonymous review on:

About labbe's "intertextual distance"*

[redacted]* and C. N. Ledoux

French architect
(1776 - 1802)

Université de Franche-Comte, Besançon.

Mr Jourdain

é

Maison de Sciences de l'Homme
Claude Nicolas Ledoux

*Address correspondence to: Université de Franche-Comte, 32 rue Megevand, Besançon, Cedex, 25030, France. Email: jean-marie.viprey@univ-fcom.frte.

Where you will be afraid by:

- Unexpected ***ctrl-C ctrl-V*** side effects
- Mathematical ***Monsters (MM)***
- Graphical ***Monsters (GM)***
- Three level of exercises, ***high school* (*)**, ***primary school* (**)**, **Open problems (***)**

Please quote precisely: Cyril and Dominique

In the 2001, Volume 8, Number 3, issue of the *Journal of Quantitative Linguistics* (pp. 213 – 231) M. M. Dominique and Cyril Labbé published a paper entitled "Inter-Textual Distance and Authorship Attribution. Corneille and Molière". Dominique and Cyril Labbé (hereafter referred to as DCL) propose a new formula for the computation of dissimilarity between texts, as well as a distances scale. They intend to apply it in the field of authorship attribution, and especially in one particular case: a controversy about the authorship of several plays signed by Molière. The object of this paper is to discuss their rationale and conclusions in the light of several simple experiments. Though we will quote DCL's paper throughout, some previous knowledge of its content is recommended. The reader will be also referred to a book written by Dominique Labbé (hereafter DL) in 2003 (Labbé, 2003). In that book, DL recapitulates for a larger audience most of the contents of the paper.

But (simply) wrong

Don't forget to provide your own def.

1. Conceptual frameworks; text theory

[1.1] DCL refer to three papers dealing with authorship attribution, as well as the works of four researchers in the field of lexical statistics. However, they do not refer to any particular theory, either of language, or of texts. This leads them to invoke several concepts, such as the concept of *actual distance* between texts, without providing the reader with an actual definition. Most researchers whose object of investigation is *text* will consider the term *inter—textual distance* itself to be inadequate to name their method, which consists of a single measurement, valued by a single scale. More modestly, Muller (1992a, b) talks of *lexical connexion*. This latter term makes it clear that massive vocabulary is just one component of textuality among others.

[1.2] Adam (1999), referring to Harris (1969) and overall to Bakhtine (1977) together with several works on discourse analysis, defines text as a combination of a *structure* and a *texture*[1] (for us: *microstructure*). The *vocabulary* of a text can no more be reduced to a list of items, even if that list includes information on their frequency. It basically consists of a rhythm of occurrence

See the special issue of the French review *Corpus* entitled « La distance intertextuelle ».

Irrelevant: self-referencing footnote

To improve your knowledge on this subject see CDL : "How to measure the meanings of words ?".

(*macrodistribution*) which corresponds to thematic constitution and progression (variety, breaks, increasing), and of a network of *collocations*.

Don't forget to mention the ones you are referring to!

[1.3] Here is the first and most basic objection we put forward to DCL: by no mean between lexical inventories of two texts be sufficient to draw conclusions about thei the end point of meticulous, interdisciplinary work, utilising a variety of approache statistical measurements, could we contemplate conclusions of general import.

[1.4] Furthermore, DCL use the terms *genre* and *theme* without referring to any literary and/or linguistic theory. The same can be said about the concept of *author* and generally about their way of dealing with literary history, which is also deserving of criticism. Nevertheless, we will move on to a critical analysis of DCL's proposition.

Unknown operator: perhaps a wrong ctrl C/ctrl V?

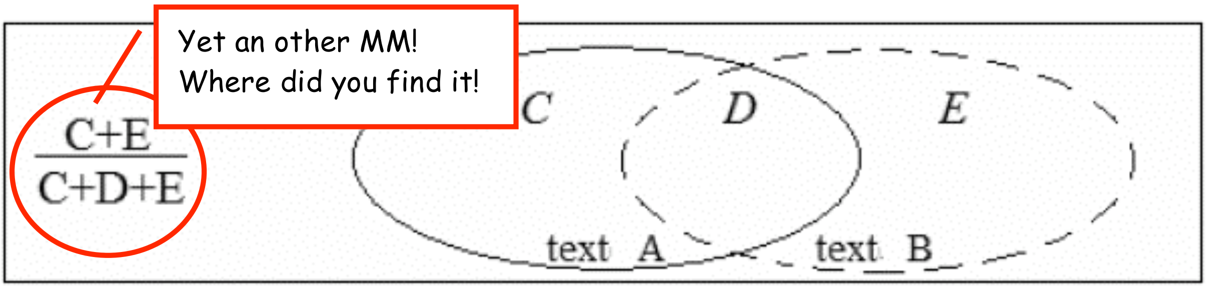
L's formula

[2.1] DCL propose a *distance* calculation *D*(*a*,*b*) (*a* referring to text *A*, *b* to text *B*). This is the quotient of the symmetrical difference between the vocabularies of the two texts, by a number describing the size of the global vocabulary. In set notation:

$$\frac{VF_a \cdot \Delta : VF_b}{VF_a \cdot \cap VF_b .}, \quad where\ \mathrm{VF} = vocabularyfrequency.$$

This is a Mathematical Monster (MM)...
You can't intersect numbers and/or you can't divide sets.

Or, as presented in the figure and notations of DCL:



Yet an other MM!
Where did you find it!

[2.2] This *distance* is positive (it takes the value 0 when both texts share exactly the same lexical material, in the same proportions) and less than 1 (the value 1 means that they do not share any lexical item).

Congratulations! : no mistake here!

[2.3] If, exceptionally, both texts have the same length (*N*, overall number of occ *absolute distance* is the sum of the absolute values of differences between occurrences of items in *Va* and *Vb*. The *relative distance* is then the quotient of the *absolute distance* by the quantity :

$$(Na + Nb)\text{-}Na \cup b.$$

Yet another MM!

[2.4] In the general case (where the texts have different lengths), DCL proceed in four steps:

1. They modify the actual occurrence of the items of the longer text (B), by the coefficient *U*(*a*,*b*) = *Na*/*Nb*. They thus obtain values that they denote (for each item *i*) *Eia*(*u*). At that stage, the sum of all these values is equal to *Na* (by construction).

2. For each item of A, they calculate |*Fia* - *Eia*(*u*)|. The sum of the absolute values of these differences falls into the calculation of the *absolute difference*.

3. The items of B which are missing in A (B\A) are taken into account only if *Eia*(*u*) ≥ 1. In that case, *Eia*(*u*) is added to the *absolute difference* (since *Fia* = 0, |*Fia*-*Eia*(*u*)| = *Eia*(*u*)) and also to *N'b*[2]. If *Eia*(*u*) < 1, *Eia*(*u*) is not counted into *N'b*.

(R1) Remember this over the next few pages: it's important!

4. The *intertextual distance* (further: *DI*) is then the quotient of the *absolute distance* by *Na* + *N'b*.

[2.5] Furthermore (2001, p. 218) DCL add a precision: they exclude from the summing of the numerator (in step 1) any individual absolute differences smaller than 0.5. They do not mention whether *Fia* must be then deducted from *Na*.

It is not mentioned, because it's obvious! **

[2.6] The authors do, however, add several restrictions. They suggest that it is invalid to apply the formula to texts smaller than 1000 tokens, as well as to pairs of texts for which *Nb*/*Na* > 10 (i.e. pairs where B is more than 10 times longer than A). They also suggest that texts must be *normalized* and all tokens lemmatized, i.e., attached to their dictionary entries. This point will be discussed further below.

Did you really understand this one? This means that there is a known length effect on ID.

3. DCL's "standardized scale"

[3.1] The value thus obtained must then be interpreted. DCL therefore present what they coin the *Inter-textual Distance Standardized Scale*, which is reproduced here as Table 1. They claim (2001, p. 218) that this scale has been established via tests on several corpora, representing about 10 million tokens, from various genres and periods, including several novels from the last three centuries. On the other hand, in Labbé (2003, p. 14), DL reports that these tests have examined several thousand texts.

Science: *noun.* The intellectual & practical activity encompassing the systematic study of the structure & behaviour of the physical & natural world through observation & experiments. (Compact Oxford English Dictionary).

Empirical: *adjective.* Based on observation or experience rather than theory or pure logic: *they provided considerable empirical evidence to support their argument* (Compact Oxford English Dictionary).

[3.3] Neither theatre plays, nor poetry, essays, etc. are mentioned among the texts submitted to testing. This being the case, it seems difficult to consider DCL's "scale" as a scientific process, nor even an *empirical* one in the usual meaning of that term.

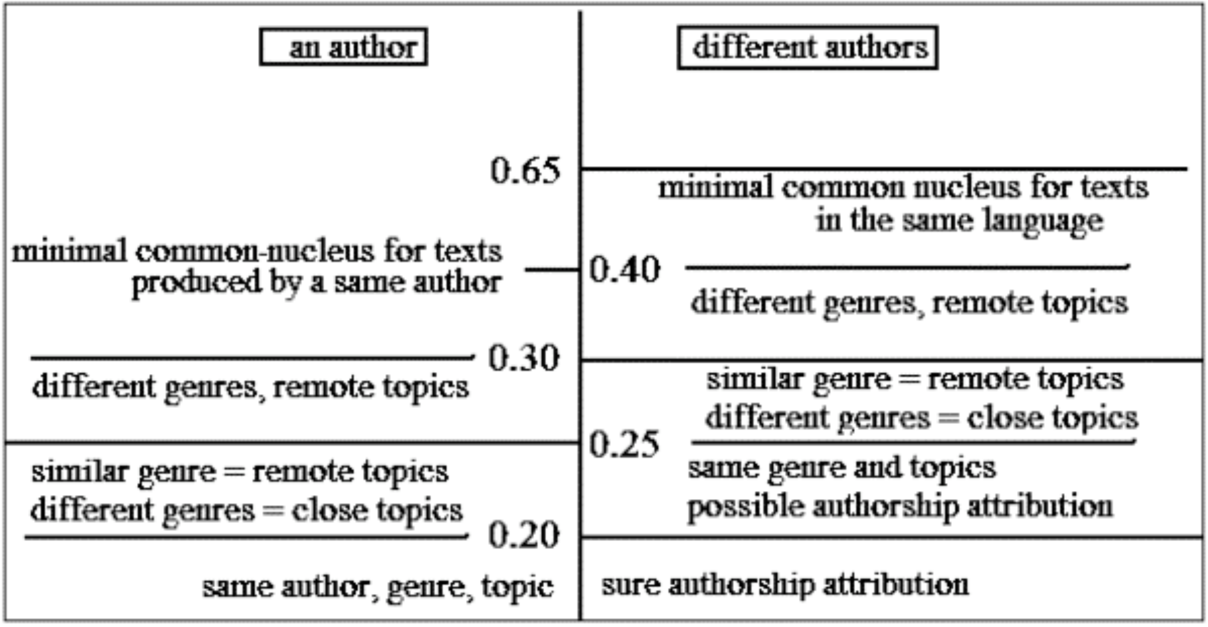

an author
different authors
0.65
minimal common-nucleus for texts produced by a same author
minimal common nucleus for texts in the same language
0.40
different genres, remote topics
0.30
different genres, remote topics
similar genre = remote topics
different genres = close topics
0.25
same genre and topics
possible authorship attribution
similar genre = remote topics
different genres = close topics
0.20
same author, genre, topic
sure authorship attribution

Fig. 1. Reproduction of the inter-textual standardised scale (Labbé, 2001).

[3.4] Two major subtleties introduced by DCL's subsequent commentary must also be noted. Where the scheme indicates *same author*, the commentary asserts that "distances smaller than 0.20 *usually* do not exist between two different authors". And when the scheme states "sure authorship attribution", the commentary adds the adverb *quite* ("quite sure"), which is equivocal in English and may be equivalent to either *entirely* or *somewhat*). But what DL keeps from this in all subsequent publications is what the scheme specifically says. In Labbé (2003, p. 14) he writes: "*Une distance inférieure ou égale à 0.20 désigne avec certitude un auteur unique. Même quand un écrivain en 'pastiche' un autre, la distance entre le pastiche et les originaux est toujours supérieure à ce seuil.*" (A distance lower than or equal to 0.20 indicates with certainty a single author. Even when a writer makes a "pastiche" of another, the distance […] is always greater than that threshold.) We will show below (Section 6) that this is wrong.

Section 6 shows that this is right

[3.5] DCL claim (2001, p. 219) that "for the same author we a[illegible]ller than those existing between two different and contemporary authors (w[illegible] the same topic)". This statement is again made unverifiable by the parenthesis. It is indeed thus far impossible to know when, and to what extent, two texts deal with the same topic, *unless* we examine their lexical kinship. Similarly, in the following paragraph, DCL explain the problem of two texts, of known different authors, possibly having a *ID* inferior to 0.20. They write that "one of them was 'inspired' by the other". But what is the possible measurement of t[illegible] another (and in general by many others)? Is this not a [illegible] the face of very basic experiments such as we demonst[illegible]

Do you understand?
this word?

Do you mean they are not good scientists?
What about yourself: Scientist or Inquisitor?
Nevertheless, test your technical competence:
try to solve exercice n°1 (next page)!

[3.6] Furthermore, is it acceptable, from the *scientific* point of view, to write on one side (at the right of the scheme) "sure authorship attribution" and, on the other side, at the same level, to nuance "same author" by the clause we have just quoted? Why should the *inspiration* of one author by another not perturb *attribution* certainty as well?

[3.7] To demonstrate the ultimate consequences of DCL's statements, we must again quote Labbé (2003, p. 15): "*Pour trouver l'auteur d'un texte douteux ou anonyme, il n'est pas nécessaire de rechercher tous les écrivains susceptibles de l'avoir écrit, il suffit d'en trouver un pour lequel la distance, entre une partie de son oeuvre et le texte analysé, sera inférieure aux seuils indiqués ci-dessus*" (To find the author of an uncertain or anonymous text, it is not necessary to search all the writers suspected of having written it; it is enough to find one writer for whom the distance between a part of his work and the analysed text is lower than the thresholds indicated above). Should not scientific cautiousness demand, on the contrary, that we foresee the case in which a third candidate might occur? By which empirical means did DL make sure that a single text could not be attributed, by his method, to two or more authors? What solution does he propose in that case, whose probability is by no means nil?

[3.8] More generally, a validation scale such as that proposed by DCL requires two properties which are obviously lacking here: it must be non-discrete (setting rigid and regularly spaced thresholds such that 0.20, 0.25, 0.30 is arbitrary), and formulated in probabilistic terms.

You're right. But overall, it's
so simple, seems to be a joke?

4. The biases of the "intertextual distance"

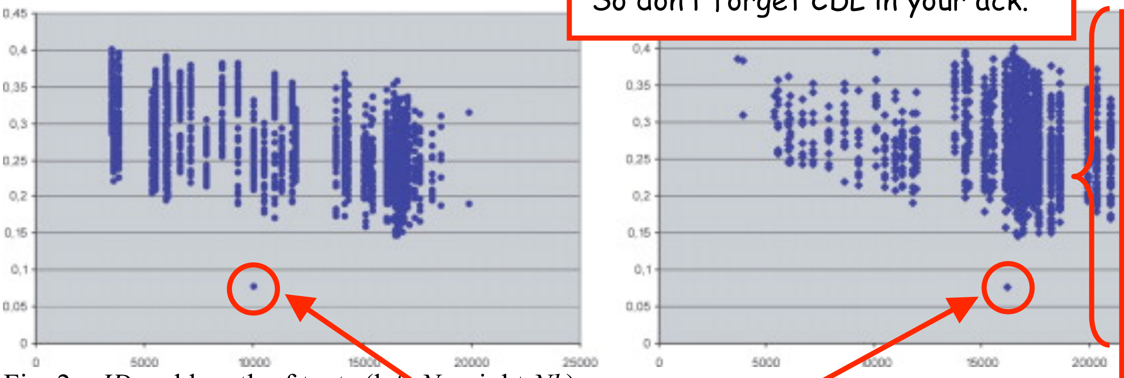
Mistake: see below under 6.6

[4.1] For anyone who has dealt for some time with questions of lexical connexion, it is not difficult to suppose that DCL's formula will bring two orders of biases, even if some artifices allow these to be limited within certain zones of application.

Exercise n°1 (*): Find n, the number of texts such that n(n-1)/2 = 2114

[4.2] First, it is clear that the *ID* are inversely dependent from the length (*N*) of the studied texts. Indeed, the longer the texts are, the more the chancy part of the distributional differences muffles whatever is the ratio *Na/Nb*. We can show this phenomena with the help of two scatter diagrams (Fig. 2) showing ordinate *DI*, on abscissa, *Na* (left graph), and *Nb* (right graph).
There are 2114 points, which represent all the possible pairs of texts in the lemmatised corpus Corneille-Molière provided by DCL.

So don't forget CDL in your ack.



Here is your first Graphical Monster (GM). I tried very hard to reproduce it! See the next inserted page for details.

Fig. 2. *ID* and length of texts (left *Na*, right *Nb*).

Where did you find this 0.07 distance? Again a wrong ctrlC-ctrlV side effect!

[4.3] This test can be applied to all sort ... nd gives this result. This bias is not a bias ... O discredits the *distance standardized scale*. DCL do not explicitly mention this bias related to the quotient *Na/Nb*. They allude to it (as well as to the bias related to *Na* and *Nb*) with the following offhand comment (2001, p. 218): "It is convenient not to apply the calculation to too small texts […] and to avoid too large a scale of sizes (around 1/10)".

[4.4] Indeed, except in the case where *Na* = *Nb*, the calculation conjures away all the hapaxes of B (for which *Eia*(*u*) < 1). If *Na/Nb* > 2, then it conjures away all items whose frequency is 1 or 2. And so on, until items whose frequency is 10 are discarded when *Na/Nb* = 10 (if we respect DCL's limit for *Na/Nb*).

NO! As said in R1 (above p. 2): only the ones which are not in the other text.

[4.5] The second bias is very underhand, since it causes no visible dependence between *ID* and *Na/Nb*. In such conditions, *ID* may have no reliable worth. The achievement of its symmetrical property (the requirement that *D*(*a*,*b*) = D(*b*,*a*)) is questionable.

(R1) Again! Not all! Only the ones which are not in A

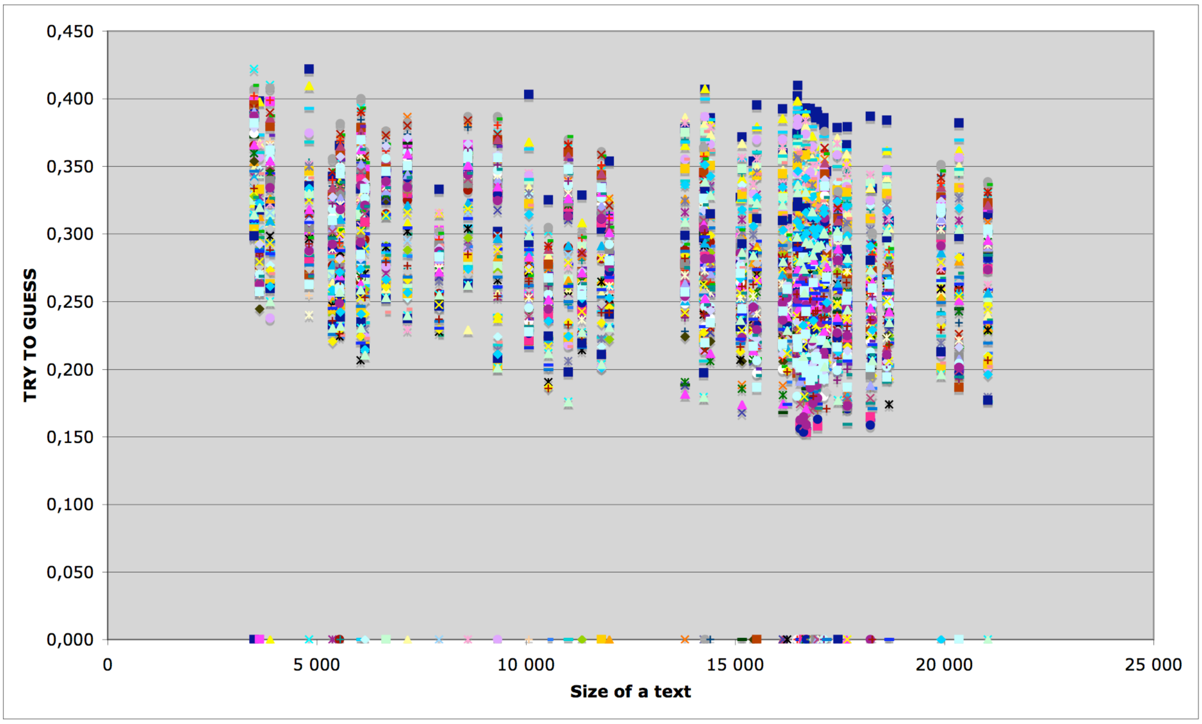
[4.6] In addition, conjuring away lots of items introduces a destructive threshold eff ... special case where the pair of texts presents the identity *Na* = *Nb*, their *ID* will then b ... the whole of both their frequency vocabularies. If, for any reason (or for th ... experiment), one token (one single token) is deleted from A, then all the hapaxes of B suddenly get left out of the calculation, which inevitably provokes a collapse of *ID*. Let us demonstrate this experimentally on a selection of 101 *Contes* of Maupassant, as published by Conard in 1929 (table of contents in appendix).

Hum... Did you read the definition of ID? But you are right: for a good scientist 1+1=2 is questionable!

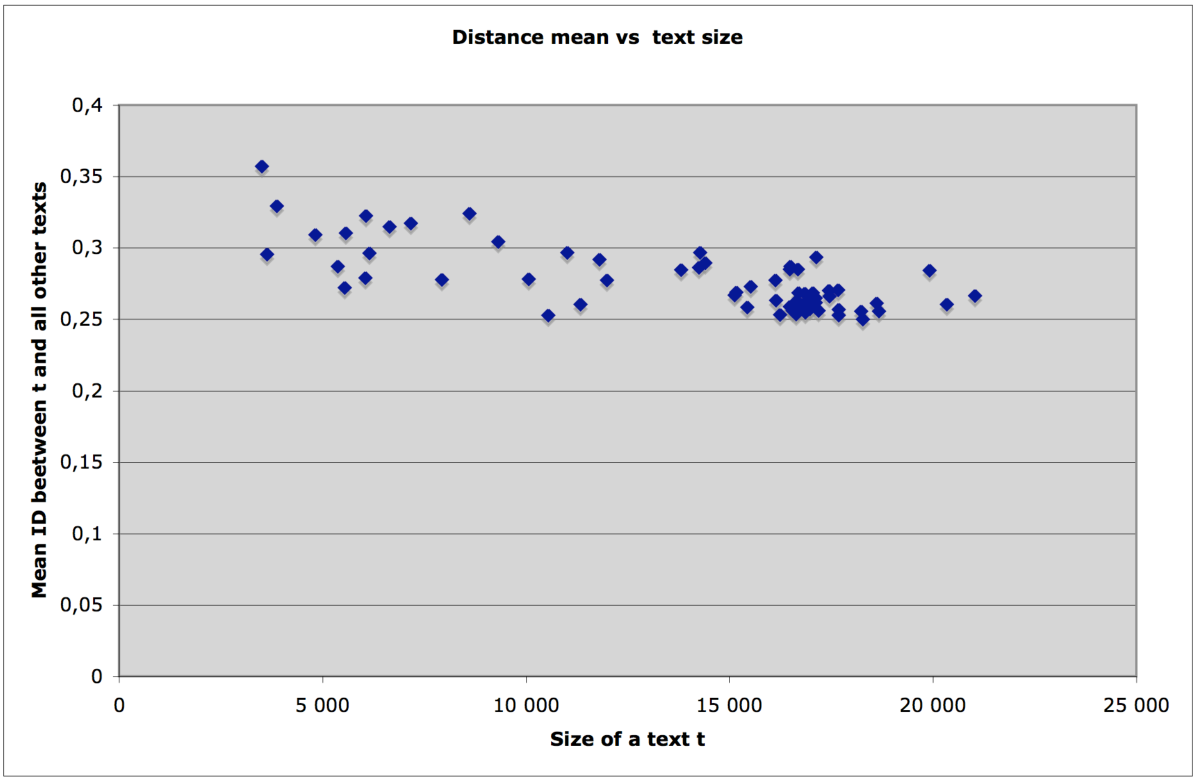
This edition does not exist!

Inserted Page n°1

I tried very hard to reproduce your graph…Firstly, I produced this one:



Then I understood what you did… And it became clear: you were looking to produce this graph, which is supposed to demonstrate the slight effect of text lengths on ID:



I am sorry, but you are counting punctuation marks as words!

[4.7] In the non-lemmatised corpus, *Na* is strictly equal to *Nb* for 3 pairs of texts, for instance *Le Remplaçant* and *La Tombe* (1647 tokens). *ID* between these two texts is 0.507. If in the vocabulary of *Le Remplaçant*, we reduce the frequency of a single type, by a si [illegible] *ID* falls to 0.451. We are dealing here with an outstanding example of a threshold effect.

You've got a real problem of symmetry!

[4.8] The discovery of this perturbation also reveals the actual asymmetry of DCL's calculation. This asymmetry is only hidden by the implicit assumption that *Na* should always be different from *Nb*. Indeed, if we conversely now modify *Vb* by the same tiny quantity that we did previously for *Va* (*me*, from 17 to 16 tokens), *ID* again falls drastically, but this time, to 0.465. As stated above, the symmetry claimed here could be illusory.

Oh my goodness! Symmetry Again! You will have to work very hard on this notion…

[4.9] In authorship attribution c [illegible] plied, one is often in possession of fragments of texts; cutting such fragments may be highly hazardous. It cannot be allowable that a distance calculation, otherwise so global, might be so sensitive to whether the calculation is made from A towards B, or from B towards A (once we have noticed that the calculation can only be done in one of these two ways).

Exercise 2 (**):

Find "Le remplaçant" and "La Tombe".

Count the number of "tokens" (with or without punctuation marks).

Exercise 3 (***):

Try to reproduce an ID=0.507 between "Le Remplaçant" and "La Tombe".

Try to reproduce the effect observed here.

Try to conclude.

Bingo! I found the 0.07 distance that occurred on Fig 2 after a wrong ctrlC-ctrlV

5. Lemmatization

[5.1] DCL then stipulate that texts must be *normalized* (without any reference to give any precise meaning to that term) and, "from [their] point of view [...] tagged". We see here that DCL consider *tagging* to be the same as *lemmatization*, when they are distinct operations (Habert et al., 2000) and the English actually use the verb *lemmatize*. And indeed, the results of *ID* are noticeably different depending on whether we utilize, for the corpus Corneille-Molière, raw or lemmatised text (DL provided us with his lemmatized corpus). This is shown by Figure 3, where the smooth curve represents the *ID* between lemmatized texts (sorted by increasing order), and the bumpy curve, the *ID* between rough texts. The maximum difference between the two results is 0.07, the minimum 0.025, the mean difference 0.042. This difference is not proportional to the values.

You are confusing word spelling standardization and lemmatization

What is the meaning of this term?

Yet another Graphical Monster! By the way, for all your graphs, don't forget to add labels on axes before publication.
This shows that ID-lemma and ID-token don't measure the same things.

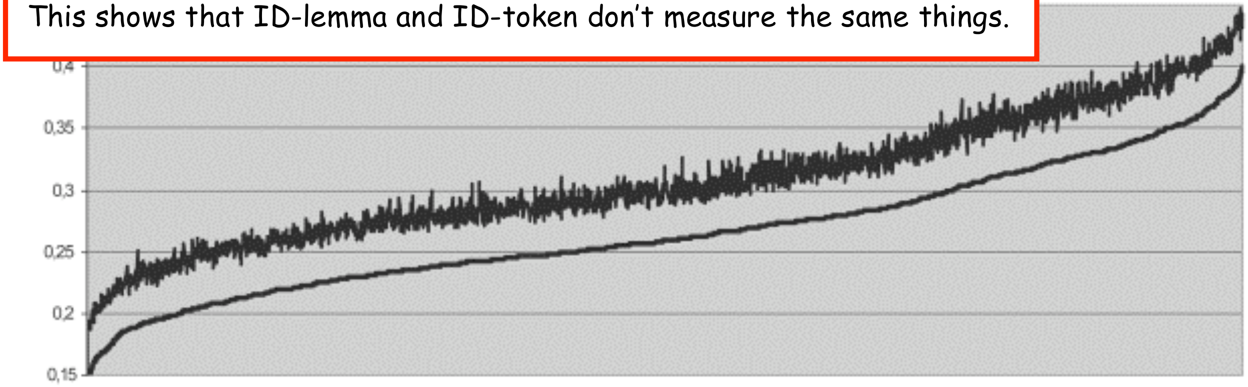


Fig. 3. Rough parallelism of *ID* depending on whether texts are lemmatised (smooth curve) or not.

[5.2] It is not our intention to tease apart, in these gap variations, what comes fr themselves, and what comes from the lemmatizing process. Crucially, we woul agree with this statement: "One can see that the distance calculation implies standards" (2001, p. 218). Such an agreement being difficult to achieve, the unlikely to be generalized or submitted to experiment.

Of course! "cesse" is a word and "sans cesse" is not one (Ask Messrs. Robert or Larousse)

Well… At least they are not counting*, ; »'. !? as types…

[5.3] In any case, such an agreement could not be reached with regard to the lemmatization standards used by DL. For instance, for the whole corpus (928,000 tokens, 9947 different lemmas found by DL), only 16 lemmas are compounds (plus 12 named entities). Compounds as frequent as *afin que*, *afin de*, *bien que*, *sans cesse*, are systematically counted as two lemmas; *cesse* is considered to be a noun, etc. Admittedly most French corpora, submitted to diverse lexical statistic operations, have suffered from insufficient reference to modern lexicological and semantic theories and to computational linguistics. Admittedly, these mistakes and naivetics are widespread. Nevertheless, that should not conceal the main point: some inaccuracy can be compensated for using probabilistic statistics, but nothing is probabilistic in the DL approach.

Sooo… That's why you are counting punctuation marks… You should tell us a little bit more on these probalistic statistics that can be used to compensate mistakes and inaccuracy. But be careful, it seems to be inefficient for ctrlV-ctrlC side effects!

[5.4] At any rate, if necessary for an experiment, the calculation of *ID* may be performed on non-lemmatized texts. This is worth noting, because if scientists intend to test the method at the desired scale (thousands of texts, hundreds of millions of tokens), demanding that the text should be lemmatized (and, what is more, following a specific norm), is exorbitant.

But you can do incredible things… You're so brilliant!

6. Experiments

When we tried to verify DCL's assertion that *ID* below 0.20 indicates a single author, we first began work on non-lemmatized texts. We thus established that, for instance, bert's *Madame Bovary* shows an *inter-textual distance* of 0.223 with Maupassant's *Une* and also 0.223 with the same author's *Fort comme la mort*. Mean *ID* between *Madame Bovary* and Maupassant's eight novels is 0.241. It is far higher in the case of *Salammbô*: 0.348. See Table 1.

Again: you count punctuation marks as words…

To be continued: next page

From previous page!

Infinitesimal precision is also coming with probalistic statistics?

Table 1. *ID between three novels of Flaubert and Maupassant's eight novels (1. Une Vie; 2. Bel-Ami; 3. Mont-Oriol; 4. Pierre et Jean; 5. Fort comme la mort; 6. Notre cœur; 7. L'Âme étrangère; 8. L'Angelus).*

	1	2	3	4	5	6	7	8	**Moyenne**
Madame Bovary	0.223	0.239	0.231	0.242	0.223	0.238	0.280	0.250	0.240750
Salammbô	0.321	0.351	0.340	0.356	0.346	0.358	0.358	0.350	0.347500
L'Education sentimentale	0.245	0.234	0.245	0.244	0.237	0.246	0.288	0.268	0.250875

Did you use CDL software? Quite nice isn't it? I promise you, we will work on a nice GUI for you

[6.2] In a second step, we then set about lemmatizing *Madame Bovary* and *Une Vie*, adhering most closely to DCL's "word for word" technique. The result confirmed our expectation: between the lemmatized texts, *ID* decreases to 0.197. In other words, below the threshold of 0.20 under which DCL rule out the existence of two different authors.

I'm so disappointed… You didn't use CDL's software, after all. Again you are counting punctuation marks as words… Nevertheless, did you have a look at fourth and fifth decimals? Because… you know… it's important: 0.19701 and 0.19699 are not the same!

[6.3] Among numerous possible tests, we also noticed an exceptionally low *ID* between the non-lemmatized texts of Balzac's *Père Goriot* and Dumas' *Comte de Monte-Cristo*, as v diverse other novels of those two writers in particular. See Table 2.

Disappointing: the fifth and sixth decimals have disappeared…

Table 2. *ID between three novels of Balzac and five of Dumas (1. Fernande; 2. Le Cor Cristo; 3. Joseph Balsamo; 4. Le Collier de la Reine; 5. Les mille et un fantômes).*

	1	**2**	**3**	**4**	**5**	**Moyenne**
Histoire des Treize	0.221	0.240	0.241	0.241	0.232	0.2350
Le Père Goriot	0.239	0.218	0.221	0.224	0.230	0.2264
Le Médecin de campagne	0.250	0.238	0.251	0.256	0.220	0.2430

You have convinced me: your mixing-all-dialectic-scientific-probabilistic-statistic-approach is clearly the right one!

[6.4] At least some of those pairs would necessarily fall under 0.2 if lemmatized. This simply demonstrates that DCL's tests were too incomplete to claim scientific status.

[6.5] If we apply *DI* inside the whole of the *Comédie humaine* (CH), we see that many pairings give a *DI* far greater than that obtaining between several of Balzac's novels and several of Dumas'. This empirical evidence contradicts the claims of DCL about their own scale (see Section 3 above).

Well! I think it's time for me to tell you something. As you know, there is a length effect on ID. You have already studied Na and Nb and now comes N=Na+Nb. If I guess correctly, the next ones will be: Na/Nb, N/Na, N/Nb, 1/N, 1/Na, 1/Nb, but also Na*Nb and Nb*Na (as symmetry is questionable!).

[6.6] Moreover, this observation throws light on the bias related to *N*, by a commonly admitted and very robust statistical test: Spearman's rank correlation (Kendall, 1962). As CH consists of 86 texts, 3655 pairs are possible. Only 3148 are permitted since 507 have a quotient *Nb*/*Na* > 10. Of these 3148, 888 have a *DI* greater than 0.3. Those 888 pairs involve 77 of the 86 texts. All of remaining nine texts are among the 20 longest texts, and have more than 100,000 tokens. We therefore have counted the number of times when each of the 86 texts of CH is involved in a *DI* superior to 0.3: from 0 times

Including punctuation marks?

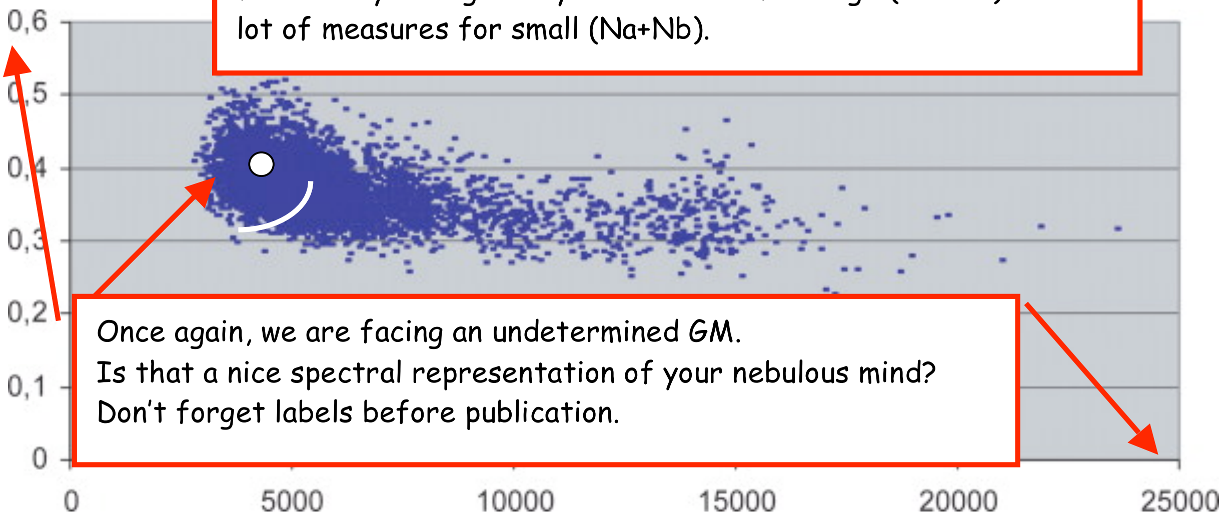
(nine texts) to 58 times (*Jésus-Christ en Flandre*, 7838 tokens). Then we established the Spearman correlation index between the ranking of the 86 texts, by decreasing length on the one hand, and by increasing frequency of involvement in a *DI* > 0.3 on the other. The result is clear: 0.842 for 86 items, which suggests a very probable inverse correlation between *N* and *DI*.

Do you know that bias is a mathematical notion: the bias of an estimator is the difference between an estimator's expectation and the value of the parameter being estimated. As ID is not an estimator, the phenomenon you are studying here is a property; in any case it is a bias!

[6.7] Those first tests were conducted on large texts, i.e., "favoured" by one of the two biases inherent in DCL's *ID*. We then tested *ID* on very short texts, Maupassant's 101 *Contes* listed in the appendix. We entirely lemmatized that corpus (more than 300,000 tokens), following again DL's "word for word" technique (except for some compounds DL identifies in his own Corneille-Molière corpus). Except for *La Bécasse*, which we excluded on DCL's recommendation (it has only 859 tokens), lengths (*N*) of the 100 others go from 1322 (*La Folle*) to 12,212 (*La Maison Tellier*), with a mean of 2938. So, all of them have more than the 1000 token-threshold beneath which DCL find it *convenient* not to calculate *ID*.

[6.8] Results are in accordance with our predictions, i.e., irremediably affected by the bias described. The mean for the 4950 pairs is established at 0.371, i.e. very close to the fateful limit of 0.4, which DCL consider the *minimal common nucleus for texts produced by a same author*. The mean for the couples with *Na* + *Nb* < 4000 is 0.401. See the scatter diagram in Figure 4, showing dependence of *ID* towards *Na* + *Nb*.

Come-on, be serious! The only thing shown on this diagram is the fact that you've got only one measure for large (Na+Nb) and a lot of measures for small (Na+Nb).



Once again, we are facing an undetermined GM.
Is that a nice spectral representation of your nebulous mind?
Don't forget labels before publication.

Fig. 4. *ID* and cumulative length of texts – Maupassant's *Contes*.

[6.9] If we were to trust *ID* and its *standardized scale*, we should conclude that those 100 texts have been written by several different writers. Furthermore, we would remain unable to determine any more precise attribution unless some substantial amendment were made to the interpretation process. The huge number of incompatibilities (couples whose *ID* > 0.4), the frequency of which increases as fast as *Na* + *Nb* falls, discourages any reasonable clustering.

I have reread many times. It seems not to be very clear. Are you working here with ID-lemma or ID-token?
But I'm sure you are still counting punctuation marks as words!

Well… Of course CDL are not using your standards where "*" and ";" are words…

7. First conclusions

[7.1] Before examining how first DCL, and then DL, applied their proposition to the case of Corneille and Molière, let us summarize our prior observations.

You forgot to mention yours!

[7.2] We have shown

* in Section 1, that when they advance their thesis, DCL do not lean on any theoretical reference concerning the key notions they deal with;

* in Section 3, that the proposed Inter-textual Distance Standardized Scale corresponds to no scientific standard, either in its making, or in its presentation;

* in Section 4, that the formula for ID includes two major biases, one patent – dependence towards Na, Nb, and Na + Nb – the other one underhand: uncontrolled conjuring away of distance factors (the least frequent items of the longer text). That second bias moreover involves a threshold effect, strong enough to discredit the whole method;

* in Section 5, that the suggestion by DCL that texts should be lemmatised (which became a strong demand in Labbé, 2003) is formulated in such a way that it throws doubt upon any generalisation of this method, and upon any verification done by third parties;

* in Section 6, that if we apply ID to actual cases, we are led to unacceptable and/or absurd conclusions. This is due to its biases on the one hand, to the naively discrete character of its interpretation scale, on the other hand.

Don't forget to add Na*Nb, Nb+Na and Nb*Na here!

Be careful, the only thing you are discrediting here is yourself!

An absurd person always leads you to absurds conclusions.

You are right… Good scientists know that 0.197999 and 0.20001 are very close.

You should say here that CDL are good scientists as they gave you the opportunity to work with their own data and software.

8. Application to corneille a

[8.1] Given all this, we may suspect that the application of DCL's method to the case of Corneille et Molière, which is mentioned in the very title of their paper, should be questionable. That application is [illegible] successive chapters: *Molière's plays, then Corneille and Molière*.

You are right. So what? What exactly is your point?

Hum… Is that a joke or are you really completely off your head?

Genre? To which theory are you referring?

[8.2] In *Molière's plays*, DCL begin with a selection of eight plays, which they present as Molière's best-known plays (further "main masterpieces"). This cannot fail to astonish a scientific mind, since no notoriety criterion is made explicit. Moreover, in the table presented, DCL fail to mention that, of those eight plays, four are written in verse and four are not; they do not use that distinction for their results. Mentioning this fact would have made apparent the strong influence of versification upon *ID*. This influence is linked to the lexical restriction which occurs in the rhyme position, even more particularly when constrained by a genre and a century. If we want a clear idea of that effect, we just have to sort the plays *according to that criterion*. Then we establish the mean values of every block. This works even within the limits of DCL's restrictive selection.

Table 3. *Rearrangement of DCL's Table 2, putting the versified plays together (MI: Le Malade imaginaire).*

	T	M	FS	DJ	A	BG	MI
Ecole des Femmes (EF)	0.183	0.194	0.198	0.205	0.200	0.231	0.223
Tartuffe (T)		0.167	0.170	0.199	0.199	0.230	0.219
Le Misanthrope (M)			0.173	0.204	0.210	0.239	0.239
Les Femmes savantes (FS)				0.219	0.214	0.234	0.226
Dom Juan (DJ)					0.170	0.207	0.205
L'Avare (A)						0.194	0.187
Le Bourg. gentilhomme (BG)							0.196

Biased? You should say that length has a slight influence on ID.

So, why don't you simply call it "mean distances"?

[8.3] Mean *ID* is, between the versified plays, 0.181; between those in prose, 0.193; for the other pairs, 0.218. That observation seems important enough to be noted.

[8.4] Then DCL present another table, containing the *overall distances* (which would be more correctly named *mean distances* as it is done in the paper itself). They quickly draw the following summary conclusion: "except for these few plays [those presenting the highest mean ID], it is quite certain that all the work is from a single author". 8

[8.5] Then, in *Corneille and Molière*, DCL do essentially two sorts of things. On the one hand, they apply to the *ID* matrix for the 67 plays of the joint corpus (33 of Corneille's, 32 of Molière's, and both versions of *Psyché*) two synthetic analysis methods: *cluster analysis* and *tree classification*. We may remind the reader that the data submitted to those analyses are biased.

Could you find only one author who has written this before?

You are right, everybody was aware of it. But did nobody write about it, as it was too obvious?

[8.6] Let us carry on regardless of those biases, and co
century theatre precisely what they already know

1. that Molière's play *Dom Garcie de Navarre*, because of its genre (it is his only *heroic comedy*), is related to Corneille's texts of the same genre;
2. that Corneille's last two comedies (the *Menteurs*) are more related to Moliere's comedies (especially to his versified ones) than to Corneille's early comedies;
3. that both versions of *Psyché*, which strongly intersect, are rather eccentric (being written by several hands, a well recognised fact).

Rather eccentric? Just like you!

These "several hands", or "pens" are two: Corneille and Molière. Why did you hide this fact?

"Menteurs
quite in the centre of Molière's works", they overvalue a pure graphical artifice. Actually, even with their biased data, DCL should limit themselves to the statement that 15 and 16 are attached to the *Molière's verses* cluster. Indeed, both arrangements (Fig. 5) are strictly equivalent as graphs from the same tree-analysis: the left one is the one published by D variant from exactly the same results. One can see that the right o are "in the centre" of Molière's works.

Nice handmade ctrlC-ctrlV

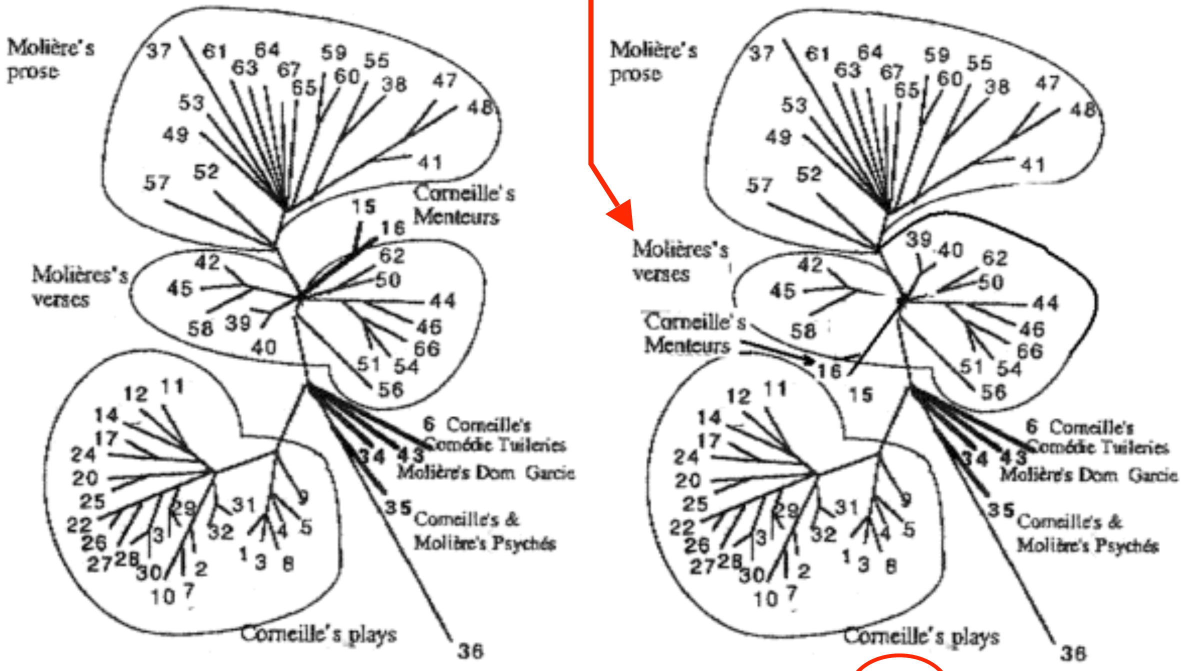


Fig. 5. Two equivalent graphs (but one more suggestive than the other) of DCL's tree-analysis.

Please quote exactly: "Luong's tree-analysis" (not CDL)!

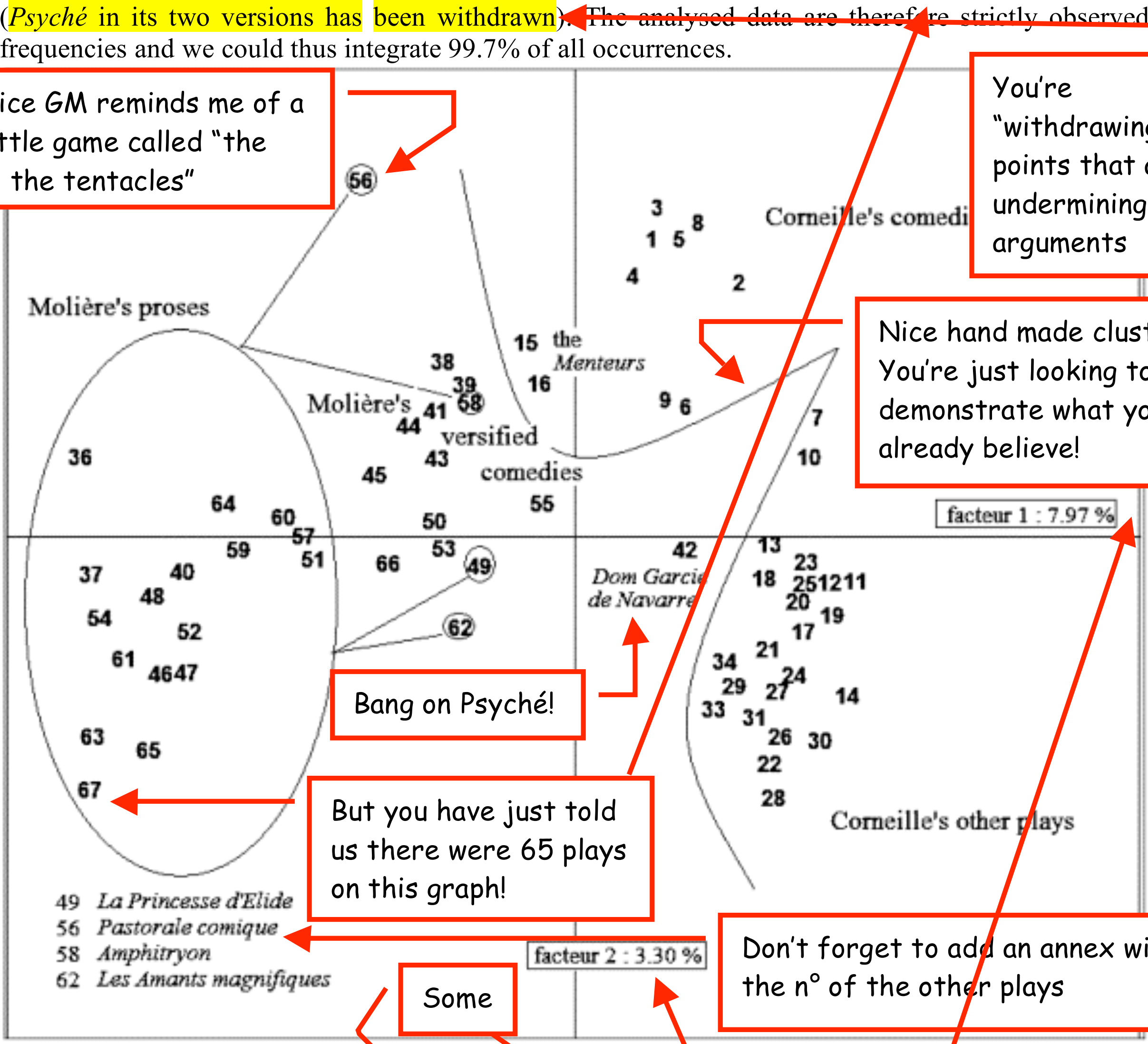
Have you read this book? It states that CA is an interesting exploration tool that has to be used in conjunction with automatic clustering…

[8.8] Only one serious conclusion could be drawn from this analysis, if the matrix data were not biased. The *Menteurs* are indeed significantly nearer, following only that criterion, to Molière's versified comedies than to his prose comedies and than to Corneille's interpreting that graph as DCL do: "in other words, the *Menteurs* authorsl most of Molière's masterpieces" – this is akin to a conjuring trick.

Yes it has been proved and you know that!

[8.9] Where are the control contrasting analyses? Why did DCL not contra and Maupassant's or Dumas' and Balzac's works? Did they clearly prove that such a phenomenon cannot be met elsewhere, among distinct but notoriously related authors, to various extents?

[8.10] We will compare the result obtained so indirectly and hazardously by DCL with one from a classical Correspondence Analysis. For the CA principles in the context of text analysis, see for example Lebart, Salem and Berry (1998). Here, the submitted matrix is a very large table (6200 lines, 65 columns). It contains the distribution of all lemmas (hapaxes excepted) in the 65 plays in question (*Psyché* in its two versions has been withdrawn). The analysed data are therefore strictly observed frequencies and we could thus integrate 99.7% of all occurrences.



This nice GM reminds me of a nice little game called "the day of the tentacles"

You're "withdrawing" the points that are undermining your arguments

Nice hand made clustering! You're just looking to demonstrate what you already believe!

Bang on Psyché!

But you have just told us there were 65 plays on this graph!

Don't forget to add an annex with the n° of the other plays

Some

Fig. 6. CA graph of distribution of ~~all~~ lemmas in ~~all the~~ plays of the corpus (columns only shown).

[8.11] This graph indicates very well the medium position of the *Menteurs*, kinship of *Dom Garcie* with Corneille's tragedies and tragi-comedies, and even an interesting position of *Mélicerte* (55).

[8.12] It is worth noting the locations of *Dom Juan* (51) and *L'Avare* (60), which DCL attribute with certainty to Corneille. This graph presents, just more clearly and without any bias, the data which are *grosso modo* on DCL's tree-analysis graph. Who would interpret this as a proof of Corneille's authorship of 16 plays of Molière?

You should investigate these percentages…

Inserted Pages n°2

The graph you should have shown according to your key (caption) (using R-software):

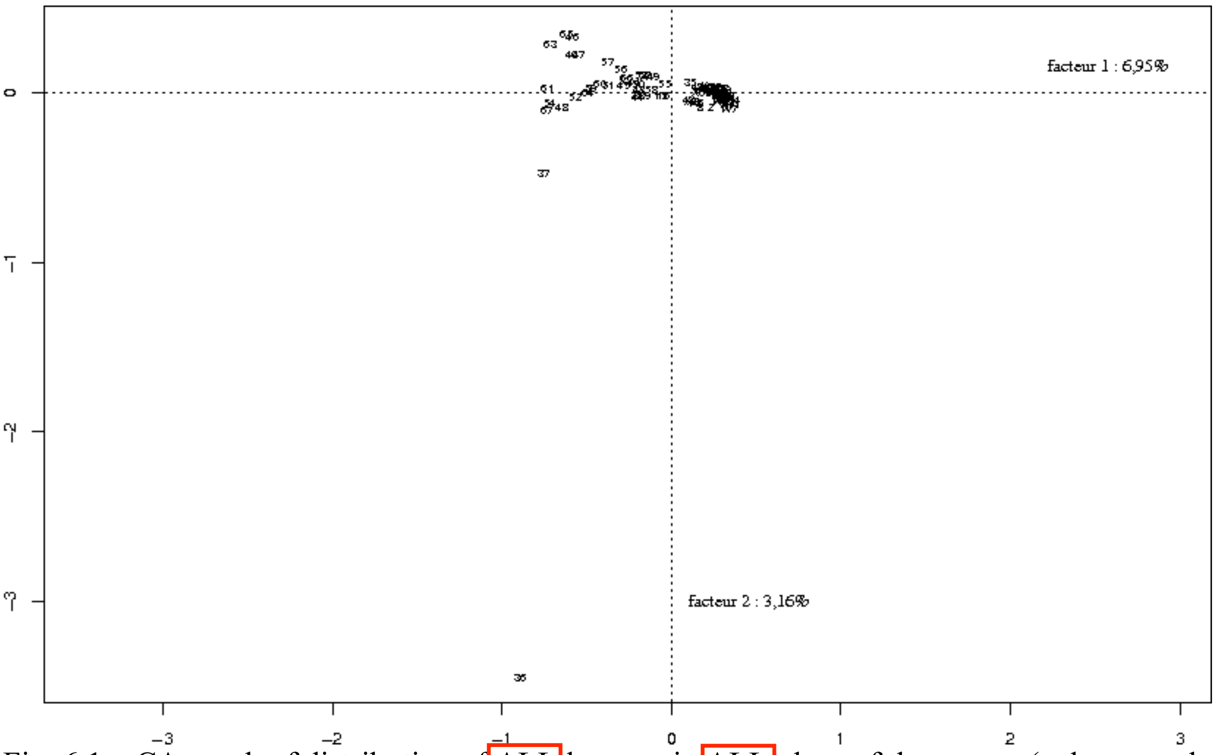
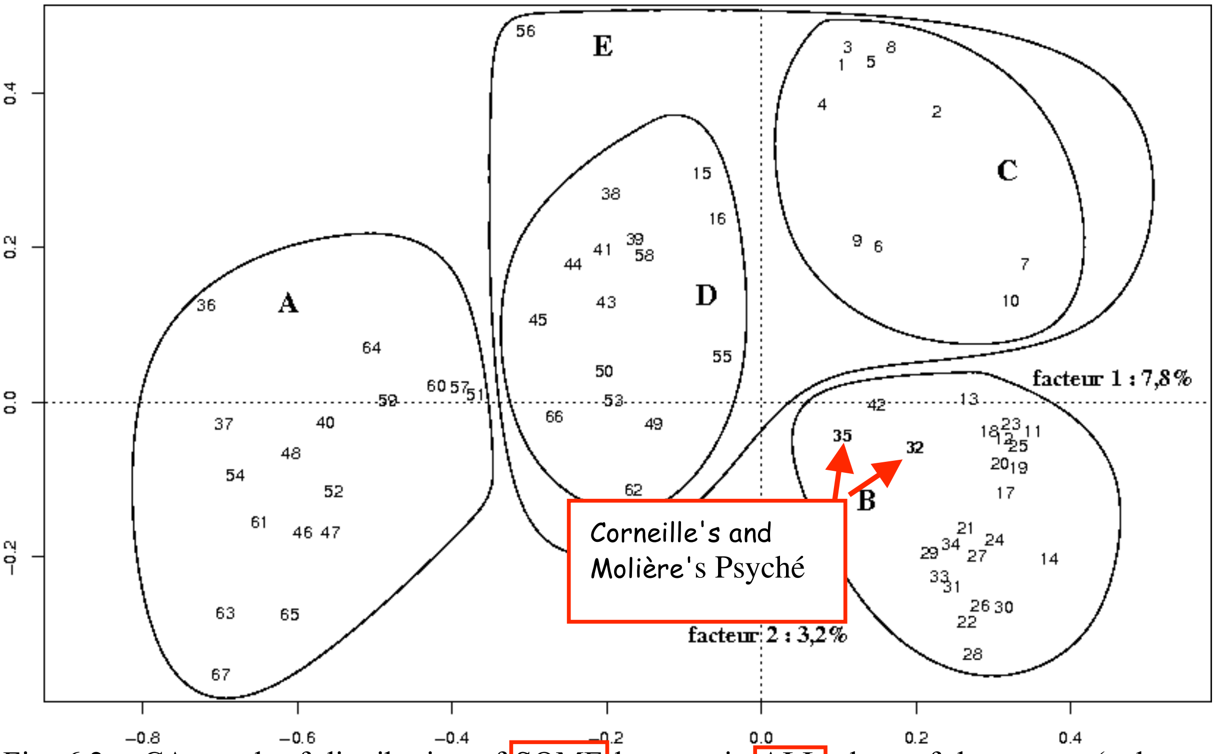


Fig. 6.1. CA graph of distribution of ALL lemmas in ALL plays of the corpus (columns only shown).

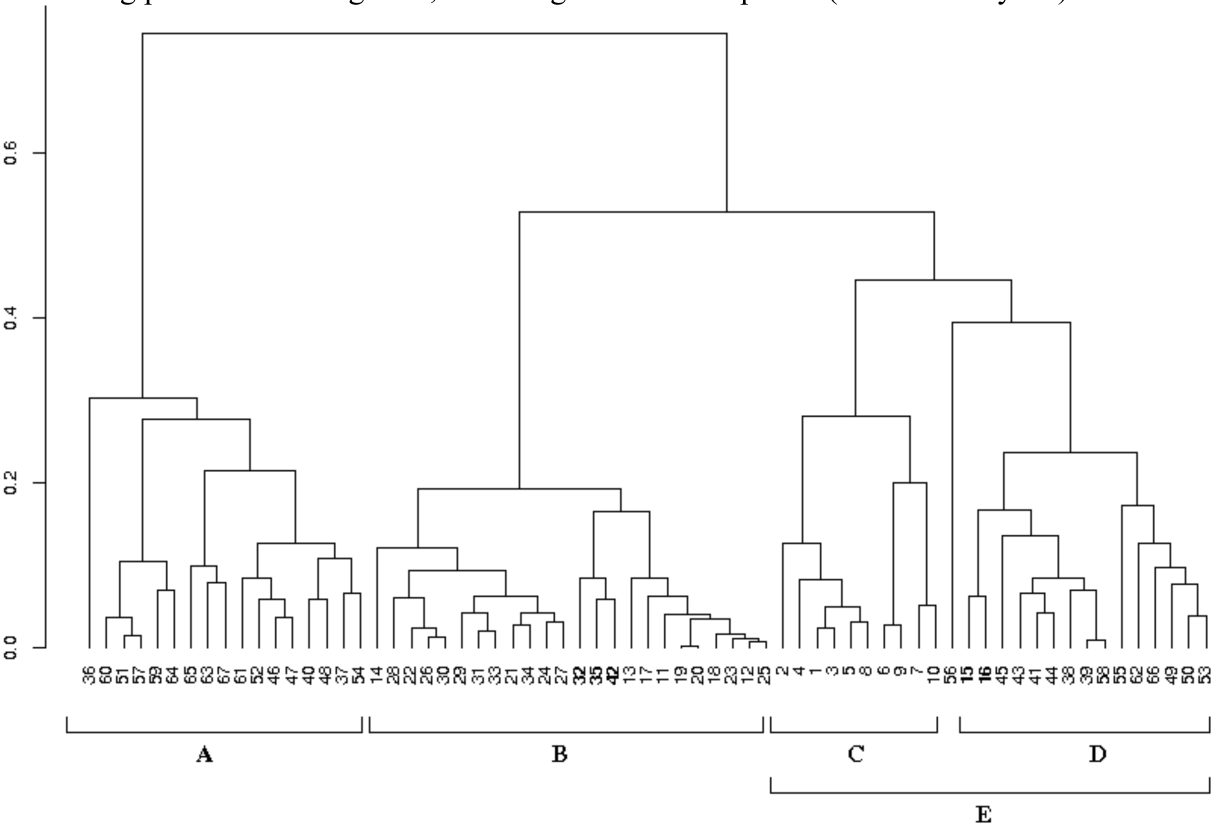
With the help of an automatic clustering (next page) you should have got this graph:



Corneille's and Molière's Psyché

Fig. 6.2. CA graph of distribution of SOME lemmas in ALL plays of the corpus (columns only shown).

Clustering points of own figure 6, including the 2 hidden points (32 & 35 : Psyché)



Groupe A

36. Jalousie du B.
60. Avare
51. Dom Juan
57. Sicilien
59. Georges Dandin
64. Fourb. de Scapin
65. Escarbagnas
63. Bourgeois gentil.
67. Malade imag.
61. Pourceaugnac
52. Amour médecin
46. Crit. de l'Ecole
47. Impromptu de V.
40. Préc. ridicules
48. Mariage forcé
37. Médecin volant
54. Méd. malgré lui

Groupe B

14. Pompée
28. Othon
22. Nicomède
26. Sertorius
30. Atilla
29. Agésilas
31. Tite et Bérénice
33. Pulchérie
21. Don Sanche
34. Suréna
24. Oedipe
27. Sophonisbe
32. Psyché (Corneille)
35. Psyché (Molière)
42. DomGarcie(Molière)
13. Polyeucte
17. Rodogune
11. Cinna
19. Héraclius
20. Andromède
18. Théodore
23. Pertharite
12. Horace
25. Toison d'Or

Groupe C

2. Clitandre
4. Galerie du Palais
1. Mélite
3. Veuve
5. Suivante
8. Place royale
6. Comédie des T.
9. Illusion comique
7. Médée
10. Cid

*56. **Com. pastorale** (Molière)*

Groupe D

15. Menteur (Corneille)
16. Suite du Menteur (Corneille)
45. Ecole des femmes
43. Ecole des maris
41. Sganarelle
44. Fâcheux
38. Dépit amoureux
39. l'Etourdi
58. Amphytrion
55. Mélicerte
62. Amants magnifiques
66. Femmes savantes
49. Princesse d'Elide
50. Tartuffe
53. Misanthrope

Your Conclusions?
Cluster D. The author of n° 15 & 16 (Corneille) also wrote n°: 38, 39, 41, 43, 44, 45, 49, 50, 53, 55, 58, 62 66; Cluster B; Corneille also wrote: 32, 35, 42

Did you look at the sixth decimal to see if it was 0.20499 or 0.20501?

[8.13] On the other hand, DCL produce a table (5) of *ID* between the *Menteurs* and each of Molière's plays. What essential would those data show if they were not biased? That the *ID* are regularly spaced from 0.205 to 0.341 with *Le Menteur*, from 0.206 to 0.331 with *La Suite du Menteur*. We particularly notice that no *ID* is lower than 0.2. That does not prevent DCL from concluding in favour of a sure attribution to Corneille of all Molière's versified plays, as well as of *Dom Juan*, and *L'Avare*.

Exercise (*): find the first and last date of Corneille's plays. Find Molière's date of birth and death. Proceed to a dialectic-scientific-probabilistic-statistic analysis of these facts and finally conclude.

[8.14] In order to justify their spectacular intervention into the field of literary studies, DCL claim (2001, p. 220) that "From the very beginning, it was rumoured that Molière was not the writer of his plays." Overall, they claim that "Since then, the problem has been discussed many times." Indeed, it was raised three times in total: at the beginning of the 20th century by Pierre Louÿs, a French poet; in 1957 by Henry Poulaille, a French writer; in 1990, by two lawyers, Hippolyte Wouters and Christine de Ville de Goyet. Their theses are, moreover, fairly different one from another. Overall, DCL omit this key point: so far not a single specialist scholar of French classical theatre, or even of the 17th century or of theatre in general, in France or in the whole world, ever validated those hypotheses. Labbé (2003) evokes a silent plot, organized by Molierists and/or Corneillists. That suspicion would perhaps be more justified if the "problem" was more recent and if relevant specialists were not counted by hundreds, all around the world.

Correct!

9. Conclusions

[9.1] 1. The original objective of DCL was to submit a new measurement of distance between texts. The result is disappointing since the *inter-textual distance* has two biases which made it unusable, even to compare contrastive pairings (text A is nearer to B than to C…). Moreover, the idea of a rigid

1) You're fired (for incompetence).
2) Don't forget to send me your next publication. I will check it over carefully for you so you won't make such big mistakes again.
3) Even better still! Forget about publication.
4) Add an acknowledgement to CDL as they provided you with data and software.
5) Stay at home.
6) What a waste of time and effort! What a Pity!

research. But it would be even less serious to settle, by a single measurement, a question of such an importance.

[9.4] 4. DCL's paper and its widespread repercussions are likely to seriously weaken the credibility of statistical methods in the humanities, and particularly in literature. It is worth noting, furthermore, that all the authors referred to by DCL's paper in the field of lexical statistics have expressed themselves against DCL's proposition: Etienne Brunet (Brunet, 2004), Charles Muller (*Le Point* 11.04.2003), Jean-Pierre Barthélémy (*Le Monde* 11.06.2003). Meanwhile DCL have not received any significant approbation during the last three years.

The only thing you have destroyed is your own credibility

Appendix

Table of the 101 Maupassant's Contes selected for study mentioned in Sections 4 and 6.

1	Sur l'Eau	36	Normand (Un)	71	Bonheur (Le)
2	Maison Tellier (La)	37	Parricide (Un)	72	Aveu (L')
3	Aventure parisienne (Une)	38	Réveillon (Un)	73	Coco
4	Partie de campagne (Une)	39	Ruse (Une)	74	Crime au Père Boniface (Le)
5	Aux Champs	40	Veillée (La)		
6	Aveugle (L')	41	Vieux Objets	75	Gueux (Le)
7	Bécasse (La)	42	Voleur (Le)	76	Ivrogne (L')
8	Bûche (La)	43	Yveline Samoris	77	Lettre trouvée

9	Ce cochon de Morin
10	Clair de Lune
11	Confessions d'une femme
12	Correspondance
13	Farce normande
14	Folle (La)
15	Fou ?
16	Gâteau (Le)
17	Histoire vraie
18	Lit (Le)
19	Loup (Le)
20	Madame Baptiste
21	Mademoiselle Fifi
22	Marroca
23	Menuet
24	Mots d'amour
25	Nuit de Noël
26	Peur (La)
27	Pierrot
28	Relique (La)
29	Rempailleuse (La)
30	Roche aux Guillemots (La)
31	Rouerie
32	Saut du Berger (Le)
33	Testament (Le)
34	Coq chanta (Un)
35	Fils (Un)
44	A cheval
45	Ami Joseph (L')
46	Auprès d'un Mort
47	Aventure de Walter Schnaffs (L')
48	Confession (La)
49	Denis
50	Deux Amis
51	En Mer
52	Farce (La)
53	Ficelle (La)
54	Humble Drame
55	Main (La)
56	Mon oncle Jules
57	M. Jocaste
58	Orphelin (L')
59	Père Milon (Le)
60	Petit (Le)
61	Première Neige
62	Remplaçant (Le)
63	Réveil
64	Sabots (Les)
65	Saint-Antoine
66	Serre (La)
67	Tombouctou
68	Duel (Un)
69	Vendetta (Une)
70	Vengeur (Le)
	sur un Noyé
78	Mère Sauvage (La)
79	Notes d'un voyageur
80	Parure (La)
81	Petit Fût (Le)
82	Rose
83	Souvenir
84	Tombe (La)
85	Lâche (Un)
86	Vieux (Le)
87	Bête à Maît' Belhomme (La)
88	Mes Vingt-cinq jours
89	Cri d'alarme
90	Epave (L')
91	Fermier (Le)
92	Mademoiselle Perle
93	Etrennes
94	Allouma
95	Hautot père et fils
96	Soir (Un)
97	Champ d'oliviers (Le)
98	Mouche
99	Après
100	Colporteur (Le)
101	Père (Le)

Notes

1. We prefer the contrasting terms *macrostructure* and *microstructure* (Viprey, 1997, 2002).
2. represents the subset *E* of DCL's scheme, but by construction it is smaller than it. It will be used to establish the denominator of the final value.
3. It is necessary to make so many suppositions *precisely because* DCL have not published any more precise details than those they give in their paper (and in Labbé, 2003).
4. Our emphasis.
5. Bias related to *N*.
6. Bias related to *Na/Nb*.
7. Flaubert's and Maupassant's w ditions, Dumas' works from the Calmann-Levy editions.

Remember: N=Na+Nb. You can add 1/(Na+Nb), 1/Na, 1/Nb,...

8. Once again, we may note the highly equivocal adverb (see Section 3).
9. Ter relaps, see note 8.

References

[1] Adam, J. -M. (1999) *Linguistique textuelle: des genres de d*

[2] Bakhtine, M. (1977) *Marxisme et philosophie du langage*, P

[3] Barthélémy, J. -P. and Guénoche, A. (1991) *Trees and Prox* Sons.

[4] Brunet, E. (2004) Où l'on mesure la distance entre les distances, *Texto!*, [en ligne], mars 2004. Rubrique Dits et inédits (http://www.revuetexto.net/Inedits/Brunet/Brunet_Distance.html).

[5] Habert, B., Nazarenko, A. and Salem, A. (1997) *Les linguistiques de corpus*, Paris: Colin.

[6] Harris, Z. S. (1969) Analyse du discours, *Langages*, 13, pp. 11–65.

[7] Kendall, M. G. (1962) *Rank Correlation Methods*, London: Griffin.

[8] Lebart, L, Salem, A. and Berry, L. (1998) *Exploring Textual Data*, Boston: Kluwer Academic Publisher.

[9] Labbé, D. (2003) *Corneille dans l'ombre de Molière*, Paris, Bruxelles: Les Impressions nouvelles.

[10] Labbé, D. and Labbé, C. (2001) Inter-textual distance an authorship attribution, *Journal of Quantitative Linguistics*, 8(3), pp. 213–228.

[11] Luong, X. (1988) Using a tree-model in textual analysis, *Computers and the Humanities*, 23, pp. 397–402.

[12] Muller, C. (1992a) *Initiation aux méthodes de la statistique linguistique*, Paris: Champion.

[13] Muller, C. (1992b) *Principes et méthodes de statistique lexicale*, Paris: Champion.

[14] Muller, C. (1993) *Langue française: débats et bilans*, Paris: Champion.

[15] Viprey, J. -M. (1997) *Dynamique du vocabulaire des Fleurs du mal*, Paris: Champion.

[16] Viprey, J. -M. (1998) Une norme endogène pour le calcul stylistique du vocabulaire, *JADT 1998, 4èmes Journées internationales d'Analyse statistique des Données Textuelles*. Nice: CNRS-UNSA.

[17] Viprey, J. -M. (2002) *Analyses textuelles et hypertextuelles des Fleurs du mal*, Paris: Champion.

"Ter Relaps"… So… You are not a scientist after all! Just a stupid Torquemada