Inter-textual Distance and Authorship Attribution

J. Savoy
Université de Neuchâtel


Distance between Two Texts

- Easy to understand, difficult to define a measure
- Select the units
  - letters, words (only function words or part of them), lemmas, grammatical categories, syntactic structures, themes
- Based only on the vocabulary (binary feature) or lexical connection (frequency of occurrence)
- Measure?
- The intersection of two texts (A and B) vocabularies ($V_A \cap V_B$).
- But in such case we ignore the frequencies!
Distance between Two Texts

- Properties (wished) of the distance $\delta(A,B)$
  - not sensitive to length difference
  - applicable to several texts
  - varying smoothly from 0 (same vocabulary and similar frequencies) to 1 (no common type)
  - symmetric, for two texts A and B, we have $\delta(A,B) = \delta(B,A)$
  - as transitive as possible
    if we have $\delta(A,B) < \delta(A,C) < \delta(B,C)$
    then $\delta(A,B) < \delta(A,(B \cup C))$
  - robust (marginal changes must produce small variations)
Distance between Two Texts

- Previous experiments tend to show that textual distance depends on
  - the author
  - the epoch (chronology, e.g., texts written within a 20 years difference)
  - the subject (context, themes with its own and specific vocabulary)
  - the genre (written vs. spoken, prose vs. verse, poetry, novel, theatre, fiction)
- Authorship attribution: comparing a doubtful work with undisputed works within the same epoch, subject, and genre
Define a Distance

- **Absolute** distance between A and B is the size of both text (A & B) less the size they have in common. Having
  \[ N_A = \text{size of the A text} \]
  \[ N_B = \text{size of the B text} \]
- The *absolute* distance (Muller) is defined as:
  \[
  D_{abs}(A, B) = (N_A \cup N_B) - (N_A \cap N_B)
  \]
- Nothing in common, the distance reaches a maximum
- If A = B, then the distance is null
- Useful? Difficult to interpret…
Jaccard Distance

- A relative distance (between 0 and 1)

\[ D(A, B) = 1 - \text{Jaccard}(A, B) = \frac{(V_A \cap V_B)}{(V_A \cup V_B)} \]

where \( V_A \) = the vocabulary of text A and \( V_B \) = the vocabulary of text B

- Simple interpretation:
  Ratio between the size of the vocabulary common to the two texts and the total vocabulary

- But …
  If the two texts have very different sizes! Does not take the frequency into account
Jaccard-Based Distance

- Correction is needed (Brunnet, 1988)

\[ D(A, B) = \frac{|V_A| - |V_A \cap V_B|}{|V_A|} + \frac{|V_B| - |V_A \cap V_B|}{|V_B|} \]

- where we count
  the size of the vocabulary exclusive to A
  +
  the size of the vocabulary exclusive to B

Jaccard-Based Distance

- \( D(A, B) \) varies between 0 and 2
- The smaller the distance, the greater the similarity
- After defining the distance, we can use PCA (principal component analysis) / clustering
- Some examples …
Jaccard-Based Distance

Examples based on vocabularies

\[ D(A, B) = \frac{1000 - 300}{1000} + \frac{1000 - 300}{1000} = 0.7 + 0.7 = 1.4 \]

\[ D(A, B) = \frac{1000 - 600}{1000} + \frac{1000 - 600}{1000} = 0.4 + 0.4 = 0.8 \]
Jaccard-Based Distance

- From the examples
  - increasing the vocabulary in common (in fact we double it), the distance $D(A,B)$ is reduced from 1.4 to 0.8
  - Not so easy to explain the concept of distance

- But
  - Emphasis on *hapax* (words occurring once in the corpus), wrong spelling, names
  - The frequencies are ignored
  - Use with texts having the same (vocabulary) size
Jaccard-Based Distance

Examples based on vocabularies

\[
D(A, B) = \frac{5000 - 800}{5000} + \frac{1000 - 800}{1000} = 0.84 + 0.2 = 1.04
\]

\[
D(A, B) = \frac{5000 - 400}{5000} + \frac{1000 - 400}{1000} = 0.92 + 0.6 = 1.52
\]
Jaccard-Based Distance

- From the examples
  - facing with texts having different vocabulary size, it becomes more difficult to interpret the distance values

- But
  - we need to be able to consider texts having different sizes
  - How to include the frequency information? Is it useful?
Other Measures

- Using TLE (Table Lexical Entries) meaning the word types with their frequency
  - follows a Zipf's law or a power law with $f_r$ the frequency of occurrence of the $r^{th}$ ranked item
  - Power law:
    $$f_r \approx \frac{c}{z^\alpha} = c \cdot z^{-\alpha}$$
  - could be rather large for a given corpus
  - remove low frequency forms (less than 5 occurrences)?
  - used absolute or relative frequencies?
Interesting Starting Point

- Another definition
  \( f_{iA} = \) frequency of word type \( i \) in text \( A \)
  \( N_A = \) size (number of tokens) of text \( A \)

\[
N_A = \sum_{i \in V_A} f_{iA}
\]

\( V_A = \) vocabulary of text \( A \)

\[
D(A, B) = \frac{1}{2} \cdot \left( \frac{\sum_{i \in V_A} |f_{iA} - f_{iB}|}{N_A} + \frac{\sum_{i \in V_B} |f_{iB} - f_{iA}|}{N_B} \right)
\]

- We use the vocabulary and the frequency information
Interesting Starting Point

- Works well when the size of the two texts A and B are very similar ($N_A = N_B$)
- The minimum is 0 (only if the two texts have the same length)
- Maximum is 1 (nothing in common and whatever the text length)
- The difference in size is still a problem…
- We need to be precise when defining the elements used in the comparison. Surface words (with inflection)? Lemmas? Stems?
Intertextual Distance

If we have two documents (A and B) with different sizes ($N_A$ and $N_B$), we can reshape the largest (say B) to the size of the smallest (to obtain a size $B' = A$)

we can reuse previous formula with $N_{B'}$ instead of $N_B$
Intertextual Distance

- We assume that text $B$ is larger than $A$
  - $f_{iB} =$ frequency of word type $i$ in text $B$
  - $V_A =$ vocabulary of text $A$
  - $N_A =$ size (number of tokens) of text $A$ ($= N_B$)
- We define the expected frequency value in $B'$ as

$$e_{iB} = \frac{N_A}{N_B} \cdot f_{iB}$$

- Our final distance measure:

$$D(A, B) = \frac{\sum_{i \in V_A, V_B'} |f_{iA} - e_{iB}|}{N_A + N_B'}$$
Intertextual Distance

- Interpretation
- If \( f_{iA} = e_{iB} \) then the distance \( D(A,B) = 0 \)
- If \( f_{iA} > 0 \rightarrow e_{iB} = 0 \) and if \( e_{iB} > 0 \rightarrow f_{iA} = 0 \)
  then the distance \( D(A,B) = (N_A + N_{B'}) / (N_A + N_{B'}) = 1 \)
- When \( D(A,B) = 0.5 \), the two texts tend to share 50% of their whole extent
- When \( D(A,B) = 0.25 \), the two texts have three quarters in common
- Inside a work, the author may use many expressions (e.g., in Latin, jargon, colloquial language, abbreviations (e.g., in letters))
Intertextual Distance

- But …
- If $f_{iA}$ is always an integer, $e_{iB'}$ could be a fraction
- Remove all values $e_{iB} < 0.5$ (or $e_{iB} < \theta$) in the numerator
- The difference in size between A and B must be lower than 1/10.
- With a lot a low frequency types, we may have many fractions in defining the distance. Thus we need at least 1,000 tokens per text.
- It is important to apply the same word normalization procedure for both texts (e.g., not to using poem (with uppercase in each line) with prose)…
Interpreting Intertextual Distance

We need to distinguish between two cases, the same author or different authors (contemporary texts)

<table>
<thead>
<tr>
<th>Minimal common nucleus for texts produced by the same author</th>
<th>0.65</th>
<th>Minimal common nucleus for texts in the same language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different genres, remote topics</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>Different genres, remote topics</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>Similar genre = remote topics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Different genres = close topics</td>
</tr>
<tr>
<td>Similar genre = remote topics</td>
<td>0.2</td>
<td>Same genre and topics</td>
</tr>
<tr>
<td>Different genres = close topics</td>
<td></td>
<td>Possible authorship attribution</td>
</tr>
<tr>
<td>Same author, genre, topic</td>
<td>0.1</td>
<td>Sure authorship attribution</td>
</tr>
</tbody>
</table>
Interpreting Intertextual Distance

- Labbé used lemmas (French language)
- Labbé’s findings
  - For the same author, the distance is always smaller than those existing between two different authors
  - Smaller than 0.2: usually do not exist between different authors (plagiarism, one "inspired" the other (same topic, genre, vocabulary)
  - Between 0.2 and 0.25: texts are very similar. One author: change in theme and genre. If one author is unknown: possible attribution (but other proofs are useful (stylistic))
Interpreting Intertextual Distance

- Above 0.25: authors are probably different or genres and topics too far
- Above 0.4: Authors are different
- Above 0.65: Texts are written with different languages
**Example**

Between two texts written by two different authors but within the same period, topic and genre (tragedy in verses with a distance = 0.256)

<table>
<thead>
<tr>
<th></th>
<th>Tite et Bérénice (Corneille, 1670)</th>
<th>Bérénice (Racine, 1670)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CORNEILLE:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agésilas (1666)</td>
<td>0.159</td>
<td>0.278</td>
</tr>
<tr>
<td>Attila (1667)</td>
<td>0.180</td>
<td>0.289</td>
</tr>
<tr>
<td>Tite et Bérénice (1670)</td>
<td>0</td>
<td><strong>0.256</strong></td>
</tr>
<tr>
<td>Pulchérie (1672)</td>
<td>0.155</td>
<td>0.271</td>
</tr>
<tr>
<td>Suréna (1672)</td>
<td>0.156</td>
<td>0.264</td>
</tr>
<tr>
<td><strong>RACINE:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andromaque (1667)</td>
<td>0.259</td>
<td>0.225</td>
</tr>
<tr>
<td>Britannicus (1669)</td>
<td>0.251</td>
<td>0.209</td>
</tr>
<tr>
<td>Bérénice (1670)</td>
<td><strong>0.256</strong></td>
<td>-</td>
</tr>
<tr>
<td>Bazajet (1672)</td>
<td>0.262</td>
<td>0.220</td>
</tr>
<tr>
<td>Mithridate (1673)</td>
<td>0.248</td>
<td>0.206</td>
</tr>
</tbody>
</table>
Example

- Distances are larger between works written in prose (P) and in verse (V)
- The smallest 0.167 (*Tarfuffe, Misanthrope*)
- The largest 0.239 (*Misanthrope, {Bourgeois… or Malade imaginaire}*)
- D (*Tarfuffe, Dom Juan*) = 0.199, this the same author
- For the others, relatively small distances, thus the same author for all
Application

Distance between one play and all the others (mean) in Moliere’s works

<table>
<thead>
<tr>
<th>Title</th>
<th>Year of création</th>
<th>Nature</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>L'Avare</td>
<td>1668</td>
<td>Prose</td>
<td>.216</td>
</tr>
<tr>
<td>Dom Juan</td>
<td>1665</td>
<td>Prose</td>
<td>.220</td>
</tr>
<tr>
<td>L'Ecole des femmes</td>
<td>1662</td>
<td>Verse</td>
<td>.220</td>
</tr>
<tr>
<td>Le Tartuffe</td>
<td>1664</td>
<td>Verse</td>
<td>.224</td>
</tr>
<tr>
<td>Le Misanthrope</td>
<td>1666</td>
<td>Verse</td>
<td>.229</td>
</tr>
<tr>
<td>L'Ecole des maris</td>
<td>1661</td>
<td>Verse</td>
<td>.230</td>
</tr>
<tr>
<td>Femmes savantes</td>
<td>1672</td>
<td>Verse</td>
<td>.232</td>
</tr>
<tr>
<td>Dépit amoureux</td>
<td>1658</td>
<td>Verse</td>
<td>.235</td>
</tr>
<tr>
<td>Malade imaginaire</td>
<td>1673</td>
<td>Prose</td>
<td>.235</td>
</tr>
</tbody>
</table>

Application

“What troubles me in Moliere’s works is this high frequency of “e”!”
“Not the frequency of “o” and “u”?"
“Not at all. Only this high frequency of “e”.”
Application

What if … the distance between a Moliere’s play and a Corneille’s play is too small?

By the way, *Psyché* (1671) was written by both!
Using clustering, the farther the points, the greater the distance.

15, 16 *Menteur* of Corneille
34, 35 *Psyché*
43 *Dom Gracie*
06 *Comédie des tuileries* (Corneille)
44 *Ecole des maris*
46 *Ecole des femmes*
66 *Femmes savantes*
51 *Tarfuffe*
54 *Misanthrope*
36 *Psyché* (Quinault, 1671)
According to the Distance

- One Moliere's play (*Dom Gracie*, #43) appears near of Corneille's works
- Large distance between Psyche written by Quinault, (1671) and other works
- Not all Moliere's works are questionable!
Other Elements

- No manuscripts by Moliere (only around 20 signatures)
- No single description of him at work, no explanation of his creative methods, books read
- No indication of how Moliere has conceived his works
- *Psyché* was written by both (1/3 Molière, 2/3 Corneille) and recognized as it
- Molière and Corneille were in Rouen together during around sixth months
- After Molière moved to Paris (1662) and produced many master works. The distance between them is small
- The tragedy was the most important genre at that time
- A first doubt raised by Pierre Louÿs (1919) (stylistic)
Written by Molière
(at least not by Corneille!)

<table>
<thead>
<tr>
<th>Title</th>
<th>Acts</th>
<th>Genre</th>
<th>Date</th>
<th>Size (Tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>La jalousie du barbouillé</td>
<td>1</td>
<td>Prose</td>
<td>1659</td>
<td>3 501</td>
</tr>
<tr>
<td>Le médecin volant</td>
<td>1</td>
<td>Prose</td>
<td>1659</td>
<td>3 876</td>
</tr>
<tr>
<td>Les précieuses ridicules</td>
<td>1</td>
<td>Prose</td>
<td>1660</td>
<td>6 651</td>
</tr>
<tr>
<td>Critique de l’école des femmes</td>
<td>1</td>
<td>Prose</td>
<td>1663</td>
<td>8 613</td>
</tr>
<tr>
<td>Impromptu de Versailles</td>
<td>1</td>
<td>Prose</td>
<td>1663</td>
<td>7 170</td>
</tr>
<tr>
<td>Le mariage forcé</td>
<td>1</td>
<td>Prose</td>
<td>1664</td>
<td>6 059</td>
</tr>
<tr>
<td>L'amour médecin</td>
<td>3</td>
<td>Prose</td>
<td>1665</td>
<td>6 148</td>
</tr>
<tr>
<td>Le médecin malgré lui</td>
<td>3</td>
<td>Prose</td>
<td>1666</td>
<td>9 319</td>
</tr>
<tr>
<td>La comtesse d'Escarbagnas</td>
<td>1</td>
<td>Prose</td>
<td>1671</td>
<td>5 565</td>
</tr>
</tbody>
</table>

### Written by Corneille?

<table>
<thead>
<tr>
<th>Titles</th>
<th>Acts</th>
<th>Genre</th>
<th>Date</th>
<th>Size (tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L'étourdi</td>
<td>5</td>
<td>Vers</td>
<td>1658?</td>
<td>18 674</td>
</tr>
<tr>
<td>Le Dépit amoureux</td>
<td>5</td>
<td>Vers</td>
<td>1656?</td>
<td>16 243</td>
</tr>
<tr>
<td>Sganarelle ou le cocu imaginaire</td>
<td>1</td>
<td>Vers</td>
<td>1660</td>
<td>6 042</td>
</tr>
<tr>
<td>Dom Garcie de Navarre</td>
<td>5</td>
<td>Vers</td>
<td>1661</td>
<td>17 049</td>
</tr>
<tr>
<td>L'Ecole des maris</td>
<td>3</td>
<td>Vers</td>
<td>1661</td>
<td>10 536</td>
</tr>
<tr>
<td>Les fâcheux</td>
<td>3</td>
<td>Vers</td>
<td>1661</td>
<td>7 922</td>
</tr>
<tr>
<td>L'Ecole des femmes</td>
<td>5</td>
<td>Vers</td>
<td>1662</td>
<td>16 625</td>
</tr>
<tr>
<td>La princesse d'Elide</td>
<td>5</td>
<td>Vers et prose</td>
<td>1664</td>
<td>11 333</td>
</tr>
<tr>
<td>Le Tartuffe</td>
<td>5</td>
<td>Vers</td>
<td>1664</td>
<td>18 272</td>
</tr>
<tr>
<td>Dom Juan</td>
<td>5</td>
<td>Prose</td>
<td>1665</td>
<td>17 454</td>
</tr>
<tr>
<td>Le Misanthrope</td>
<td>5</td>
<td>Vers</td>
<td>1666</td>
<td>17 182</td>
</tr>
<tr>
<td>Mélicerte</td>
<td>2</td>
<td>Vers</td>
<td>1666</td>
<td>5 540</td>
</tr>
<tr>
<td>Amphytrion</td>
<td>3</td>
<td>Vers libres</td>
<td>1668</td>
<td>15 117</td>
</tr>
<tr>
<td>L'Avare</td>
<td>5</td>
<td>Prose</td>
<td>1668</td>
<td>21 033</td>
</tr>
<tr>
<td>Psyché</td>
<td>5</td>
<td>Vers</td>
<td>1671</td>
<td>16 182</td>
</tr>
<tr>
<td>Les Femmes savantes</td>
<td>5</td>
<td>Vers</td>
<td>1672</td>
<td>16 865</td>
</tr>
</tbody>
</table>
Moliere's Plays Written by?

The author is not clearly either Moliere of Corneille

<table>
<thead>
<tr>
<th>Titles</th>
<th>Acts</th>
<th>Genre</th>
<th>Date</th>
<th>Size (tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le sicilien ou l'amour peintre</td>
<td>1</td>
<td>Prose</td>
<td>1667</td>
<td>5 375</td>
</tr>
<tr>
<td>Georges Dandin</td>
<td>3</td>
<td>Prose</td>
<td>1668</td>
<td>11 009</td>
</tr>
<tr>
<td>Monsieur de Pourceaugnac</td>
<td>2</td>
<td>Prose</td>
<td>1669</td>
<td>11 803</td>
</tr>
<tr>
<td>Les amants magnifiques</td>
<td>5</td>
<td>Prose</td>
<td>1670</td>
<td>11 983</td>
</tr>
<tr>
<td>Le bourgeois gentilhomme</td>
<td>5</td>
<td>Prose</td>
<td>1670</td>
<td>17 136</td>
</tr>
<tr>
<td>Les fourberies de Scapin</td>
<td>3</td>
<td>Prose</td>
<td>1671</td>
<td>14 245</td>
</tr>
<tr>
<td>Le malade imaginaire</td>
<td>3</td>
<td>Prose</td>
<td>1673</td>
<td>19 920</td>
</tr>
</tbody>
</table>

French Presidential Discourse

• Which ones are the most similar / the most dissimilar?
• Which president is closer to de Gaulle?
Another Application

- Two main trends during the V\textsuperscript{th} republic
  - De Gaulle and Mitterand
  - The centre for Giscard and Chirac

<table>
<thead>
<tr>
<th></th>
<th>de Gaulle</th>
<th>Pompidou</th>
<th>Giscard</th>
<th>Mitterrand1</th>
<th>Mitterrand2</th>
<th>Chirac</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Gaulle</td>
<td>0</td>
<td>0,158</td>
<td>0,215</td>
<td>0,220</td>
<td>0,229</td>
<td>0,218</td>
</tr>
<tr>
<td>Pompidou</td>
<td>0,158</td>
<td>0</td>
<td>0,170</td>
<td>0,184</td>
<td>0,184</td>
<td>0,168</td>
</tr>
<tr>
<td>Giscard</td>
<td>0,215</td>
<td>0,170</td>
<td>0</td>
<td>0,184</td>
<td>0,178</td>
<td>0,159</td>
</tr>
<tr>
<td>Mitterrand1</td>
<td>0,220</td>
<td>0,184</td>
<td>0,184</td>
<td>0</td>
<td>0,106</td>
<td>0,164</td>
</tr>
<tr>
<td>Mitterrand2</td>
<td>0,229</td>
<td>0,184</td>
<td>0,178</td>
<td>0,106</td>
<td>0</td>
<td>0,151</td>
</tr>
<tr>
<td>Chirac</td>
<td>0,218</td>
<td>0,168</td>
<td>0,159</td>
<td>0,164</td>
<td>0,151</td>
<td>0</td>
</tr>
</tbody>
</table>
Another Application

• Two main trends during the Vth republic
  • De Gaulle and Mitterand
  • The centre by Giscard and Chirac
• The distances (according to the vocabulary and frequencies)
  • Distance (De Gaulle - Mitterand2) = 0.229
  • Distance (Mitterand1 - Mitterand2) = 0.106
  • Distance (De Gaulle - Pompidou) = 0.158
  • Distance (De Gaulle - Chirac) = 0.218
• Does not respect the chronology
• Difference in the terms used
  "Immigration" by Chirac, "Immigrants" by Mitterand
Another Application

A graphical view for the French presidential speeches during the year (1958-2002)
Conclusion

- Various metrics
  - based on most frequent words
  - based on function words or part of them
  - on suffix productivity
  - on the vocabulary
  - on both types and their frequency
- Assume texts with similar length (using the entire text)
- Labbé’s method (word types and their frequency)
- Authorship attribution is a difficult question!