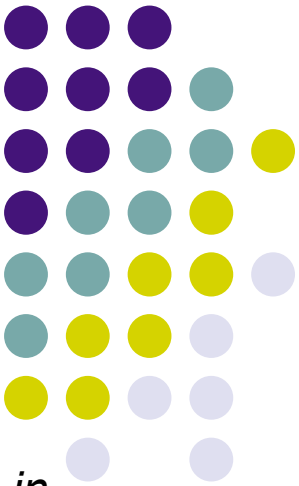
Inter-textual Distance and Authorship Attribution

J. Savoy
Université de Neuchâtel

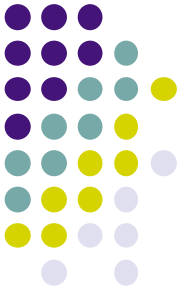


D. Labbé : *Experiments on Authorship Attribution by Intertextual Distance in English*. Journal of Quantitative Lingustics, 14(1), 2007, 33-80.

C. Labbé, D. Labbé : *Inter-textual Distance and Authorship Attribution Corneille and Molière*. Journal of Quantitative Lingustics, 8(3), 2001, 213-231.

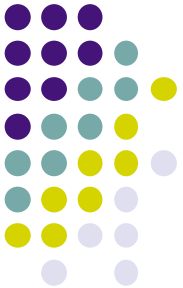
D. Labbé : *Corneille in the Shadow of Molière*. Seminar French Department, Trinity College, Dublin, 2004.

Distance between Two Texts



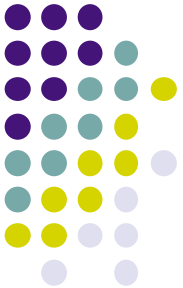
- Easy to understand, difficult to define a measure
- Select the units
  letters, words (only function words or part of them), lemmas, grammatical categories, syntactic structures, themes
- Based only on the vocabulary (binary feature) or lexical connection (frequency of occurrence)
- Measure?
- The 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



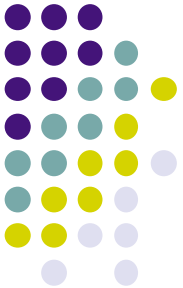
- Properties (wished) of the distance $\delta(A,B)$
  - not sensitive to length difference
  - applicable to several texts
  - varying smoothly from 0 (same vocabulary and similar frequencies) to 1 (no common type)
  - symmetric, for two texts A and B,
    we have $\delta(A,B) = \delta(B,A)$
  - as transitive 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

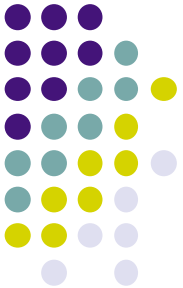


- *Absolute* distance between A and B is the size of both text (A & B) less the size they have in common. Having
  N_A = size of the A text
  N_B = size of the B text
- The *absolute* distance (Muller) is defined as:

$$D_{abs}(A, B) = (N_A \cup N_B) - (N_A \cap N_B)$$

- Nothing in common, the distance reaches a maximum
- If A = B, then the distance is null
- Useful? Difficult to interpret…

Jaccard Distance

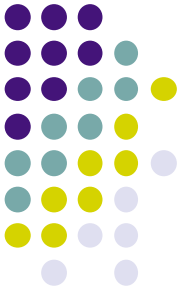


- A relative distance (between 0 and 1)

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

  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



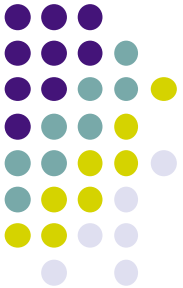
- Correction is needed (Brunnet, 1988)

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

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

E. Brunnet (1988). Une mesure de la distance intertextuelle: la connexion lexicale. Revue informatique et statistique dans les sciences humaines. Université de Liège.

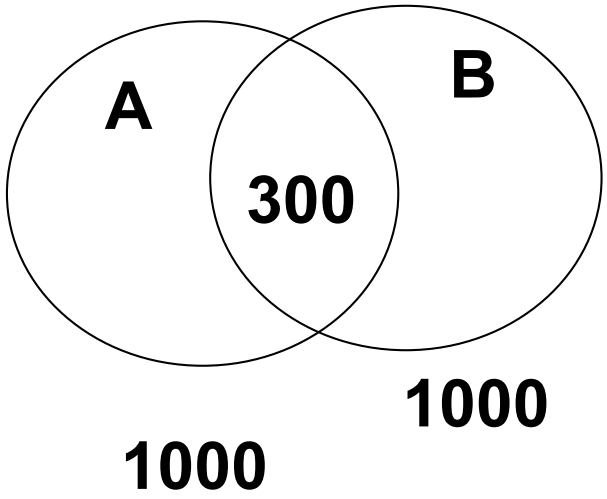
Jaccard-Based Distance

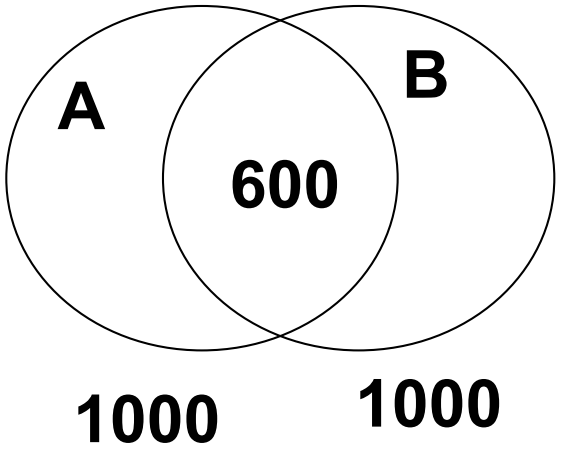


- D(A, B) varies between 0 and 2
- The smaller the distance, the greater the similarity
- After defining the distance, we can use PCA (principal component analysis) / clustering
- Some examples …

Jaccard-Based Distance

Examples based on vocabularies

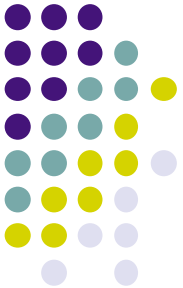




$$D(A,B) = \frac{1000-300}{1000} + \frac{1000-300}{1000} = 0.7 + 0.7 = 1.4$$

$$D(A,B) = \frac{1000-600}{1000} + \frac{1000-600}{1000} = 0.4 + 0.4 = 0.8$$

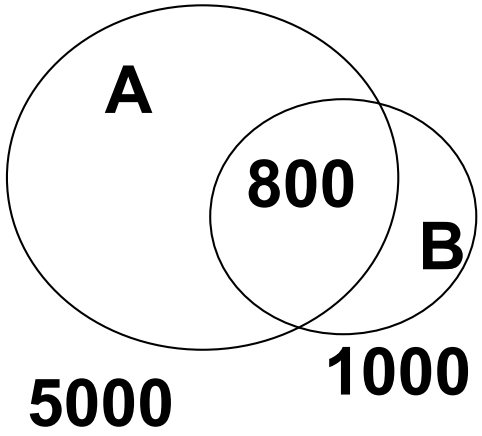
Jaccard-Based Distance

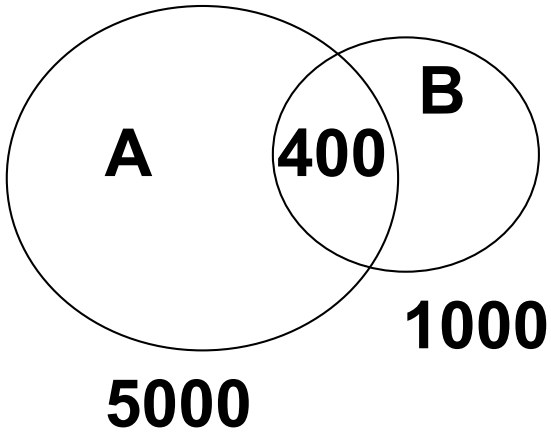


- From the examples
  - increasing the vocabulary in common (in fact we double it), the distance D(A,B) is reduced from 1.4 to 0.8
  - Not so easy to explain the concept of distance
- But
  - Emphasis on *hapax* (words occurring once in the corpus), wrong spelling, names
  - The frequencies are ignored
  - Use with texts having the same (vocabulary) size

Jaccard-Based Distance

Examples based on vocabularies

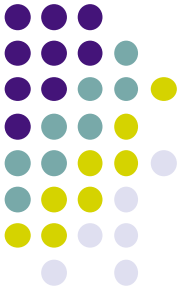




$$D(A,B) = \frac{5000 - 800}{5000} + \frac{1000 - 800}{1000} = 0.84 + 0.2 = 1.04$$

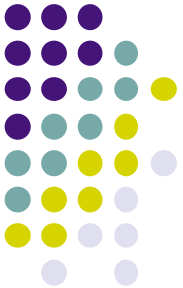
$$D(A,B) = \frac{5000 - 400}{5000} + \frac{1000 - 400}{1000} = 0.92 + 0.6 = 1.52$$

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)

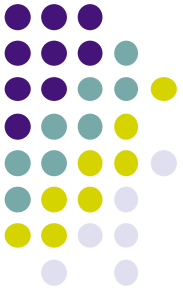
  meaning the word types with their frequency

  - 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

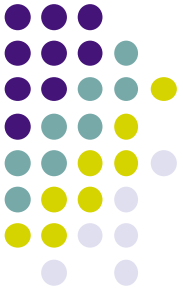
$$N_A = \sum_{i \in V_A} f_{iA}$$

  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

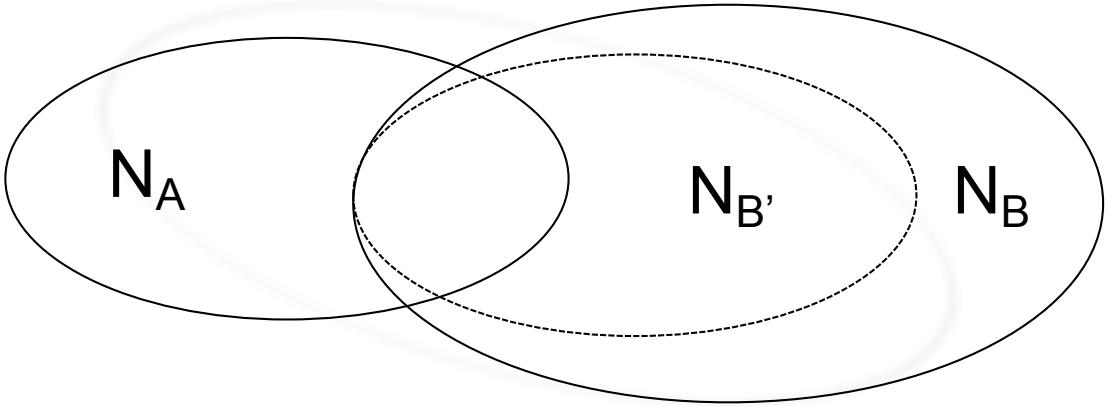
Interesting Starting Point



- 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?

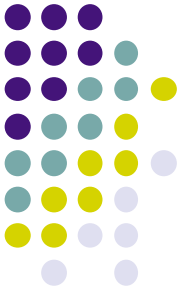
Intertextual Distance

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



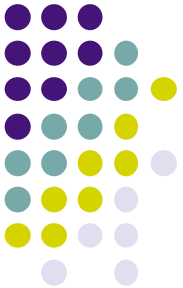
- We assume that text B is larger than A
  f_{iB} = frequency of word type *i* in text B
  V_A = vocabulary of text A
  N_A = size (number of tokens) of text A (= $N_{B'}$)
- We define the expected frequency value in B' as

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

- Our final distance measure:

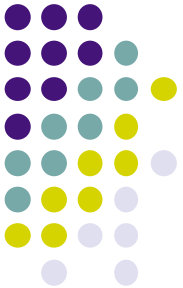
$$D(A,B) = \frac{\sum_{i \in V_A, V_{B'}} |f_{iA} - e_{iB}|}{N_A + N_{B'}}$$

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))

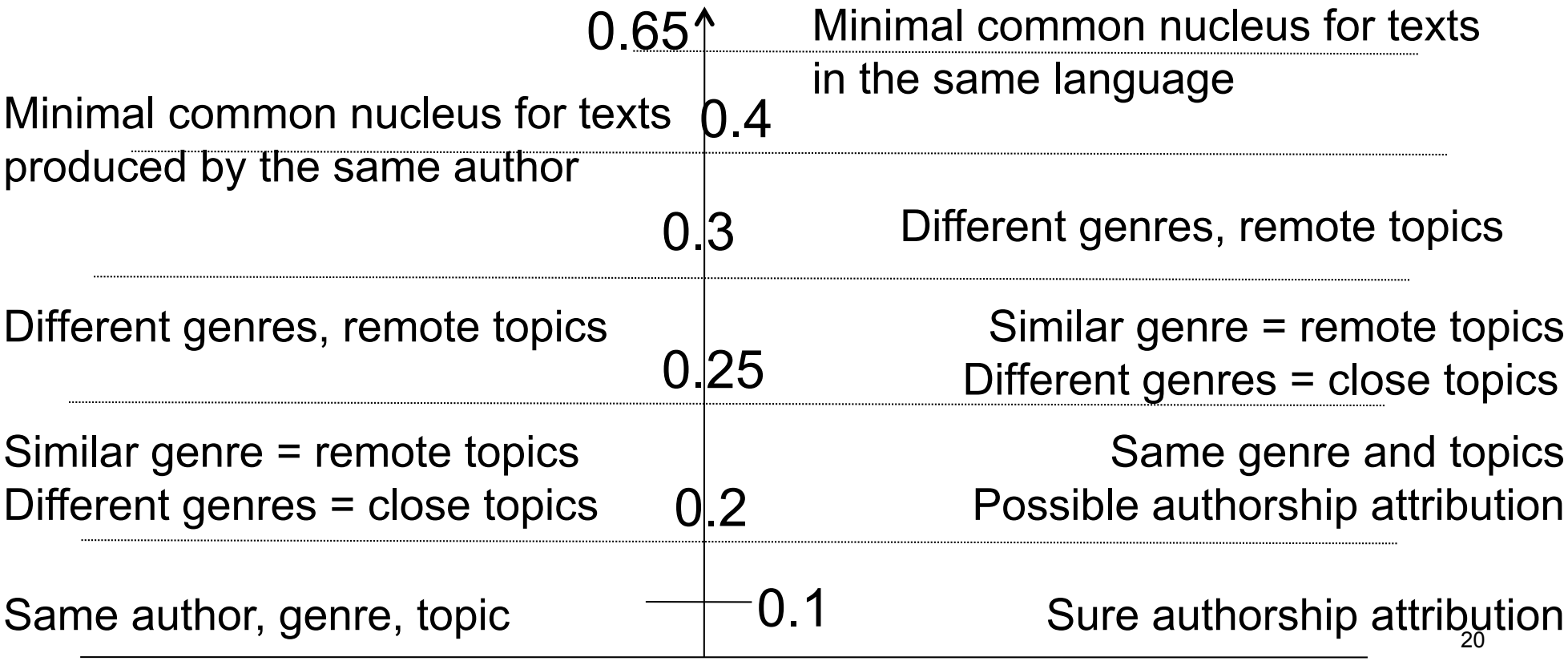
Intertextual Distance



- 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)



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)

	Tite et Bérénice (Corneille, 1670)	Bérénice (Racine, 1670)
CORNEILLE :		
Agésilas (1666)	0.159	0.278
Attila (1667)	0.180	0.289
Tite et Bérénice (1670)	0	**0.256**
Pulchérie (1672)	0.155	0.271
Suréna (1672)	0.156	0.264
RACINE :		
Andromaque (1667)	0.259	0.225
Britannicus (1669)	0.251	0.209
Bérénice (1670)	**0.256**	-
Bazajet (1672)	0.262	0.220
Mithridate (1673)	0.248	0.206

Example

	V	V	P	V	P	P	V	P
	Ecole des femmes	Tartuffe	Dom Juan	Le Misanthrope	L'Avare	Bourgeois gentilh.	Femmes savantes	Malade imaginaire
Ecole des femmes	0	.183	.205	0.194	0.200	.231	.198	.223
Le Tartuffe		0	.199	.167	.199	.230	.170	.219
Dom Juan			0	.204	.170	.207	.219	.205
Le Misanthrope				0	.210	.239	.173	.239
L'Avare					0	.194	.214	.187
Bourgeois gentilh.						0	.234	.196
Femmes savantes							0	.226
Malade imaginaire								0

- 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

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

D. Labbé : *Corneille in the Shadow of Molière*. Seminar French Department, Trinity College, Dublin, 2004.

Application

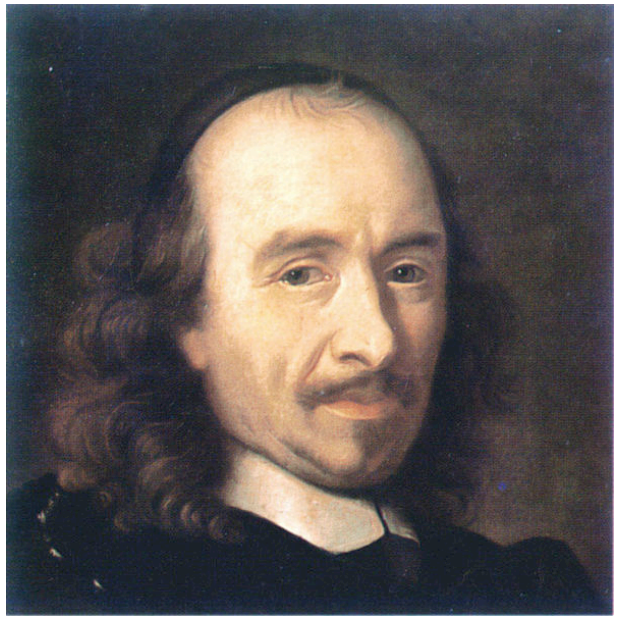
“What troubles me in Moliere’s works is this high frequency of “e”!”

“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!

Synthetic Viev

Using clustering, the farst 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*

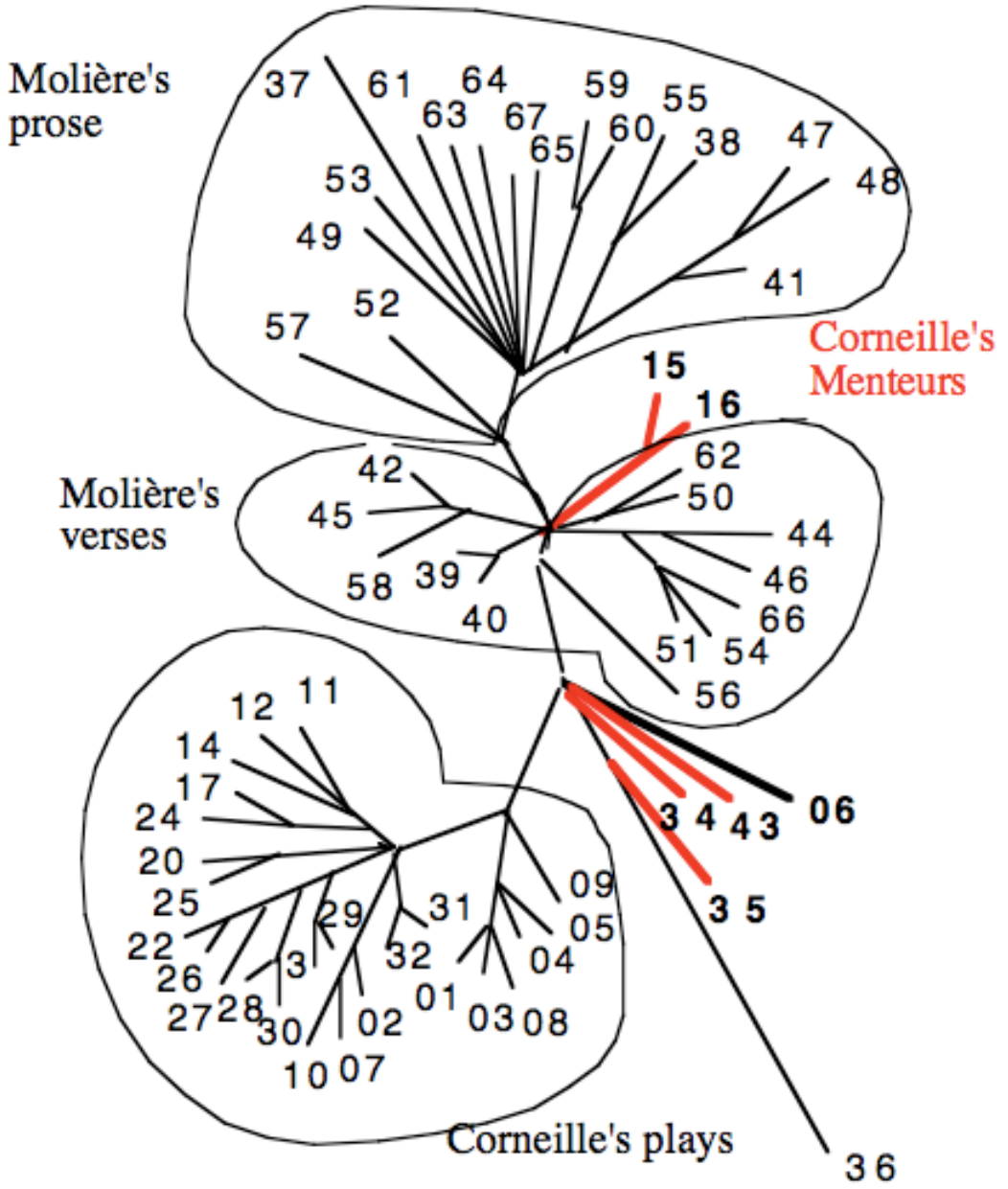
46 *Ecole des femmes*

66 *Femmes savantes*

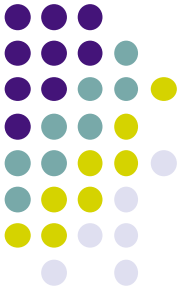
51 *Tarfuffe*

54 *Misanthrope*

36 *Psyché* (Quinault, 1671)

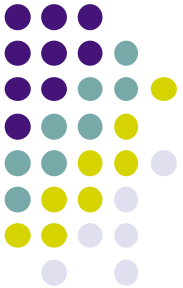


According to the Distance



- One Moliere's play (*Dom Gracie*, #43) appears near of Corneille's works
- Two pieces of Corneille (*Le Menteu*r, #15, *La Suite du Menteur,* #16) appear in the middle of Moliere's verse plays.
- 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)

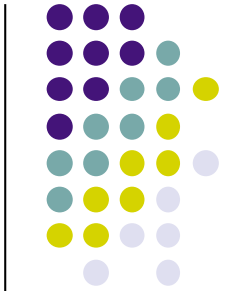
Written by Moliere

(at least not by Corneille!)

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

D. Labbé : *Corneille in the Shadow of Molière*. Seminar French Department, Trinity College, Dublin, 2004.

Written by Corneille?



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

Moliere's Plays Written by?

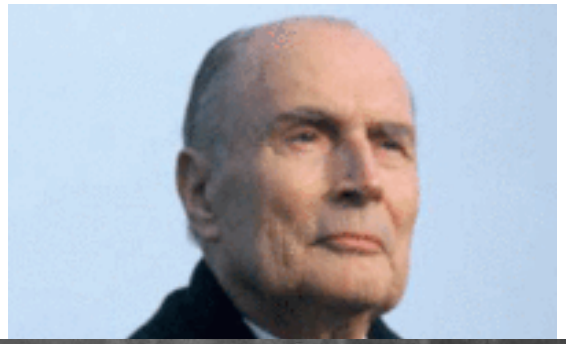
The author is not clearly either Moliere of Corneille

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

D. Labbé : *Corneille in the Shadow of Molière*. Seminar French Department, Trinity College, Dublin, 2004.

French Presidential Discourse

- Which ones are the most similar / the most dissimilar?
- Which president is closer to de Gaulle?

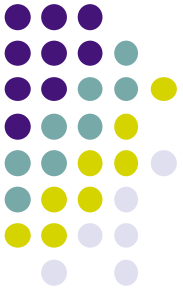


Another Application

- Two main trends during the Vth republic
  - De Gaulle and Mitterand
  - The centre for Giscard and Chirac

	de Gaulle	*Pompidou*	*Giscard*	*Mitterrand1*	*Mitterrand2*	*Chirac*
de Gaulle	0	0,158	0,215	0,220	0,229	0,218
Pompidou	0,158	0	0,170	0,184	0,184	0,168
Giscard	0,215	0,170	0	0,184	0,178	0,159
Mitterrand1	0,220	0,184	0,184	0	0,106	0,164
Mitterrand2	0,229	0,184	0,178	0,106	0	0,151
Chirac	0,218	0,168	0,159	0,164	0,151	0

Another Application

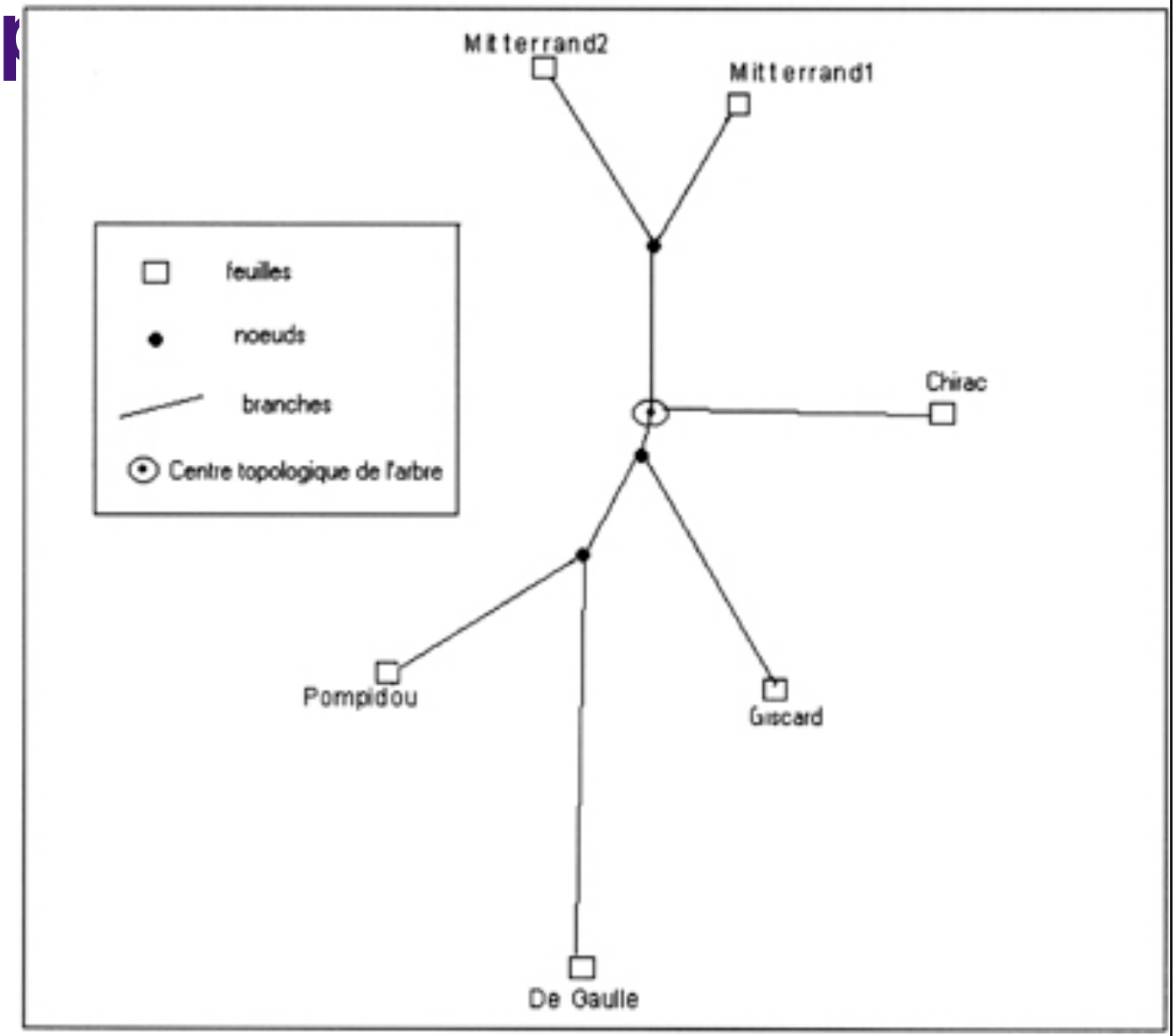


- Two main trends during the Vth republic
  - De Gaulle and Mitterand
  - The centre by Giscard and Chirac
- The distances (according to the vocabulary and frequencies)
  - Distance (De Gaulle - Mitterand2) = 0.229
  - Distance (Mitterand1 - Mitterand2) = 0.106
  - Distance (De Gaulle - Pompidou) = 0.158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

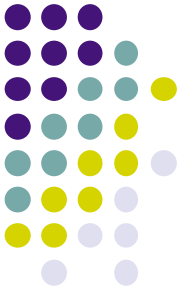
  "Immigration" by Chirac, "Immigrants" by Mitterand

Another Ap

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!