# Classification of Manuscripts: potential computational insights

Sophie Robert-Hayek

University of Lorraine
Écritures - *Maison des Sciences de l'Homme*

UNIVERSITÉ DE LORRAINE
ÉCRITURES EA 3943

16/01/2025 - Halle

# Introduction

# Outline

# Classification of manuscripts

Classification of manuscripts in inevitable:

- To reduce the number of witnesses (exclude certain types of text, *i.e.* Byzantine);
- To compute preliminary proximity for further stemmatology work (*i.e.* detect families);
- To understand more generally the proximity between texts and understand their relationship;

# Existing approaches

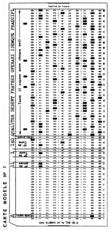
Many existing approaches existing for **grouping Greek NT manuscripts**:

- Shared errors in classical philology;
- Study of *TestTellen* (INTF and Alands);
- Quantitative Analysis (Colwell and Tune 1969): Study of proximity between known text-types for some selected readings (Colwell and Tune approach);
- Claremont Profiling Method (CPM, Wisse 1982): compute the absence/presence of set readings against the TR and consider that co-occurence of readings show a dynamic;
- **Index de variabilité** (Amphoux 1989): compute a method to measure distance between texts using the type of readings that is considered.

# What can computational insights offer us?

Potentiality of computational approaches have been understood since the advent of the computer itself!

# What can computational insights offer us?



But how can one sift through all the created data?

# What can computational insights offer us?

What can computational approaches offer us for the question of textual proximity?

- **Improve visualization**: existing classification scheme can be hard to read and hard to exploit because of the richness of material;
- **Speed-up computation and extend scope**: Existing methods have to be limited to a set of readings or even a selected chapter because of the quantity of readings to consider;
- **Provide new methods**: systematize classification approaches (called clustering in the ML place) is a core task of Machine Learning.

# Improving visualization

# Outline

# Example of results

## Example from Text Und Textwert:

```
F. KORREKTUREN AN   1 TESTSTELLE

   TST. 55:   ACTA 16,33
          C : LA 1/2    οι  αυτου  παντες
============================================================================

  ■ ■ HS.-NR.: 1360        TESTSTELLEN:  98

A. LA   2 :  78                                              SUMME:  1 TST

B. LA 1/2 :  10, 11, 18, 28, 29, 35, 36, 41, 42, 44, 45, 48, 52, 53, 55, 56,
             76, 84, 87, 88, 91, 97,100,102
      1/2B:  20                                              SUMME: 25 TST

C. LA   1 :   2,  7-  9, 12, 13, 15- 17, 19, 21- 27, 30, 31, 34, 37- 40, 43,
```

# Example of results visualization

**Example of Wisse profiling:**

GROUP PROFILES IN LUKE 1

| | B | $K^r$ | $K^x$ | M27 | M106 | Λ | $\Pi^a$ | $\Pi^b$ | 1 | 13 | 16 | $22^a$ | $22^b$ | 291 | 1167 | 1216 | 1519 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | X | | | | | | | | | | |
| 2 | X | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | X | | X | | | | | | | |
| 4 | | | | | | | X | X | | X | | | | X | | | X |
| 5 | | | | | | | | | | | | | | | | | |
| 6 | X | | X | X | | X | | | | | | | X | X | X | | X |
| 7 | | | | | | | | | | X | | | | | | | |
| 8 | X | | | | | X | ● | | | | X | | | ● | | ● | |

# Existing approaches relying on statistical analysis

Several tentative visualisation in 2D of existing approaches, using *Principal Component Analysis*:

- O. M. Kvalheim, D. Apollon, and R. H. Pierce (1988). "A Data-Analytical Examination of the Claremont Profile Method for Classifying and Evaluating Manuscript Evidence". In: *Symbolae Osloenses* 63.1, pp. 133–144;
- Jean Duplacy and Éric Huret (1977). "Classification des états d'un texte, mathématiques et informatique : repères historiques et recherches méthodologiques". In: *Revue d'Histoire des Textes* 5-1975, pp. 249–309

# Improvement of displays

- The development of new software architecture and new Web framework;
- Display of items has become easy to implement into Web applications;
- Lots of improvement and research regarding efficiency of information display.

# Improvement of displays

- The development of new software architecture and new Web framework;
- Display of items has become easy to implement into Web applications;
- Lots of improvement and research regarding efficiency of information display.

Lots of potential of new Web development approaches for dynamic visualization **of manuscripts relationship**:

- Interact in an interactive way with distance between manuscripts;
- Evaluate visually existing classifications in an easier way.

# Example of results visualization

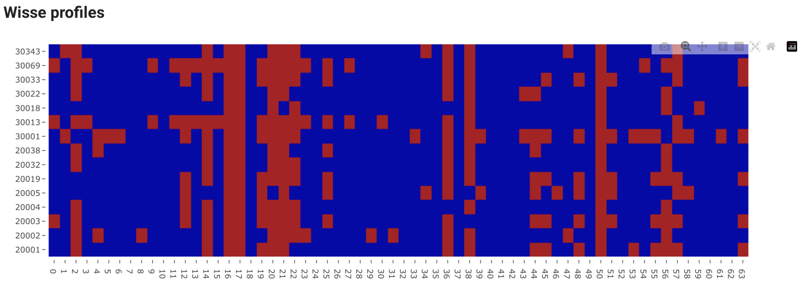
**Example of Wisse profiling:**

GROUP PROFILES IN LUKE 1

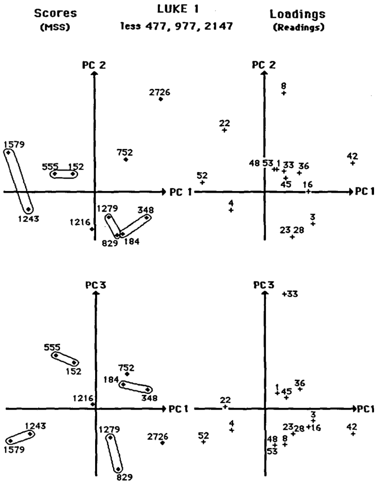
| | B | $K^r$ | $K^x$ | M27 | M106 | Λ | $\Pi^a$ | $\Pi^b$ | 1 | 13 | 16 | $22^a$ | $22^b$ | 291 | 1167 | 1216 | 1519 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | X | | | | | | | | | | |
| 2 | X | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | X | | X | | | | | | | |
| 4 | | | | | | | X | X | | X | | | | X | | | X |
| 5 | | | | | | | | | | | | | | | | | |
| 6 | X | | X | X | | X | | | | | | | X | X | X | | X |
| 7 | | | | | | | | | | X | | | | | | | |
| 8 | X | | | | | X | • | | | | X | | | • | | • | |

# Example of results visualization

**Example of web based visualization of Wisse profile:**

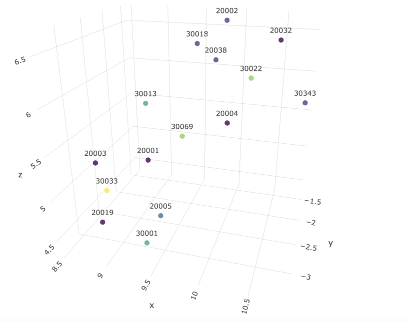


# Existing approaches relying on statistical analysis




Kvalheim, Apollon, and Pierce 1988

# Example of results visualization




But does not prevent one for interpretation so hard to **interpret regardless**!

# Example of results visualization

Dynamic computation and visualization of **collation results**:

| | | | | | |
|---|---|---|---|---|---|
| 20001 | ελεγεν | δε | προς αυτους ο μεν θερισμος πολυς οι δε εργαται ολιγοι | δεηθηται | ουν του κυριου του θερισμου οπως εκβαλη εργατας εις τον θερισμον αυτου |
| 20002 | ελεγεν | ουν | προς αυτους ο μεν θερισμος πολυς οι δε εργαται ολιγοι | δεηθητε | ουν του κυριου του θερισμου οπως εκβαλη εργατας εις τον θερισμον αυτου |

# Speeding-up computations

# Outline

# Speeding up computations: profile reading

- Development of a Web application for automatic detection of **Wisse profile**;

# Speeding up computations: profile reading

- Development of a Web application for automatic detection of **Wisse profile**;
- New manuscripts can be analyzed **through the profile methods** in a few seconds.

The same mechanism can be (is?) applicable to the *TestTellen* of the Alands.

# Speeding up computations

- **Collation**: **DNA based** algorithms allow for very fast automatic collation: a few minutes to collate the Vetus Latina in John;
- **Morphological analysis**: RNN based approach for automatic morphological analysis.

Collation algorithms are for now relatively mechanical, but will be tweaked to account for the subtility of semantic similarity.

# Link to various projects

- Collatex: `https://collatex.net`;
- Software for computation of the profiles: `https://github.com/metz-theolab/manuscript-clusterer`.

# Systematizing existing algorithms

# Outline

# Possibility to rely on quantitative and statistical approaches for clustering

Ever since the invention of computational systems, **computational approaches have been used to better understand collected data**.

## Digital Humanities

**Computational Humanities** are:

- an interdisciplinary field;
- combining research in traditional humanities;
- with **tools from computer science and mathematics**;
- to bring new knowledge to humanities related problems.

# Clustering in Machine Learning approaches

Inventor of clustering (Benzecri 1969):

> The help of a computer is needed to apply to the data previously collected a set of quasi-universal computations or rather transformations which give them such a shape that the man of the field may unarbitrarily read on the output what was undecipherable in the input.

The purpose of clustering is to find pattern in data that **is otherwise impossible to find due to the multivariate nature of the data**.

# Clustering in Machine Learning Approaches

An algorithm designed to group together a set of items without any *a priori*.

Given a set of items group them together according to **how much they look like each other**.

These approaches have a strong potential for text/manuscript classification **as they share a common goal**!

# Clustering in Machine Learning Approaches

Two requirements to use these methods:

- The **choice of an algorithm**;
- The definition of a **distance between two manuscripts/texts**.

# Clustering in Machine Learning Approaches

Two requirements to use these methods:

- The **choice of an algorithm**;
- The definition of a **distance between two manuscripts/texts**.

A **distance** numerical measure of how far apart two individuals are.

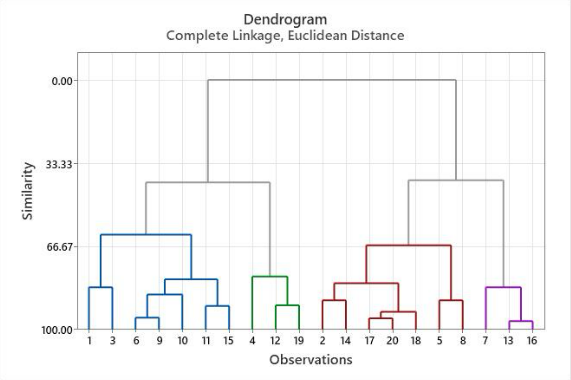
# Agglomerative clustering

**Agglomerative clustering** consists in iteratively grouping together individuals **that look the most like each other**, according to the defined **distance**.

# Agglomerative clustering

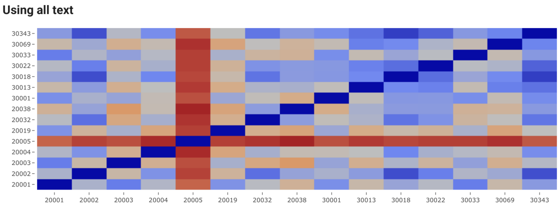
**Agglomerative clustering** consists in iteratively grouping together individuals **that look the most like each other**, according to the defined **distance**.

Results can then be organized into a dendrogram:



# Agglomerative clustering

Agglomerative clustering starts from **a distance matrix** between manuscripts:



# Defining a distance between texts

**Possible distances** include:

1. The distance as the number of shared words between two texts;

# Defining a distance between texts

**Possible distances** include:

1. The distance as the number of shared words between two texts;
2. The distance between the profiles in Wisse profiling method;

# Defining a distance between texts

**Possible distances** include:

1. The distance as the number of shared words between two texts;
2. The distance between the profiles in Wisse profiling method;
3. The distance in terms of shared readings;

# Defining a distance between texts

**Possible distances** include:

1. The distance as the number of shared words between two texts;
2. The distance between the profiles in Wisse profiling method;
3. The distance in terms of shared readings;
4. The distances by measuring a variability score between two readings.

# Defining a distance between texts

**Naive approach** as a distance between words:
$d(Rehdigeranus, Corbeiensis^2) = 1$

| chapter | verse | VL11 | VL8 |
|---|---|---|---|
| 16 | 8 | et | et |
| 16 | 8 | cum | cum |
| 16 | 8 | uenerit | aduenerit |
| 16 | 8 | ille | ille |
| 16 | 8 | arguet | arguet |
| 16 | 8 | mundum | mundum |
| 16 | 8 | de | de |
| 16 | 8 | peccato | peccato |
| 16 | 8 | et | et |
| 16 | 8 | de | de |
| 16 | 8 | iustitia | iustitia |
| 16 | 8 | et | et |
| 16 | 8 | de | de |
| 16 | 8 | iudicio | iudicio |

# Defining a distance between texts

Defining the distance as the difference between two Wisse profiles as the **sum of readings in common**: $d(B, 13) = 0$

GROUP PROFILES IN LUKE 1

| | B | $K^r$ | $K^x$ | M27 | M106 | Λ | $\Pi^a$ | $\Pi^b$ | 1 | 13 | 16 | $22^a$ | $22^b$ | 291 | 1167 | 1216 | 1519 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | × | | | | | | | | | | |
| 2 | × | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | × | | × | | | | | | | |
| 4 | | | | | | | × | × | | × | | | | × | | | × |
| 5 | | | | | | | | | | | | | | | | | |
| 6 | × | | × | × | | × | | | | | | | × | × | × | | × |
| 7 | | | | | | | | | | × | | | | | | | |
| 8 | × | | | | | × | • | | | | × | | | • | | • | |

# Defining a distance between texts

Amphoux (Amphoux 1989) **variability index**:
Each selected variant reading has a difference score:

- 2 for words with a lexeme (verbs, nouns, adjectives);
- 1 for all other (pronouns, conjunctions, prepositions, adverbs, particles)

Additionally, the type of difference counts different:

- Presence/Absence: 1 unit;
- Replacement: 2, unless synonym, then 1;
- Displacement: 1.

# Defining a distance between texts

***Applications to Vetus Latina: Pastorelli*** (David Pastorelli (2017). "A Classification of Manuscripts Based on a New Quantitative Method: The Old Latin Witnesses of John's Gospel as Test Case". In: *Journal of Data Mining and Digital Humanities*).

In John 14, the clusters revealed by the algorithm show clearly that the **text-types synthesized by Bonifatius Fischer are not to be doubted**.

# Defining a distance between texts

For all approaches, the readings must be selected using human insight:

> "*The correct orientation could only be determined by* ***evaluating the quality of the variants, which no machine is capable of doing***." (West 1973, p. 72)

# Defining a distance between texts

For all approaches, the readings must be selected using human insight:

"*The correct orientation could only be determined by* ***evaluating the quality of the variants, which no machine is capable of doing***." (West 1973, p. 72)

*This may explain the failure of one of the first attempts at textual taxonomy in the United States. If the rumor I heard is accurate, nothing emerged from the computer that could be called a classification: by* ***taking into account all the variations, even the slightest ones, every version of the text ended up being more or less atypical***.
Duplacy and Huret 1977, p. 280

# Defining a distance between texts

Fully automated approach seem to be limited by **the lack of selection of variants**, as **large scale collation is now possible**.

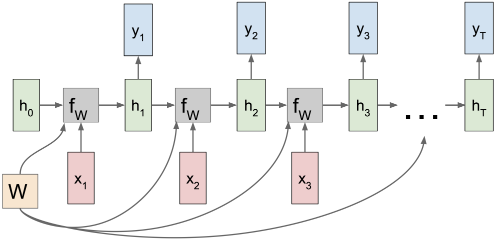
| chapter | verse | VL11 | VL8 |
|---|---|---|---|
| 5 | 41 | honorem | gloriam |
| 5 | 41 | ab | ab |
| 5 | 41 | hominibus | hominibus |
| 5 | 41 | non | non |
| 5 | 41 | accipio | adcipio |

Variants should be analyzed and sorting them through the automatic collations is **not a gain of time compared to manual approaches** (181 pdf pages for John in 9 manuscripts!)

# Recent improvements in Natural Language Processing

**But can this be overcome?**

Natural Language Processing (NLP) aims to enable computers to **comprehend, interpret, and interact with human language in a manner that is both meaningful and contextually appropriate.**



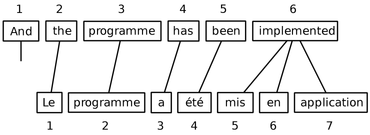
# Towards the possibility of automatic variant selection ?

Neural networks offer the possibility to **automatically compute a score based on similarity and reading type**:

| chapter | verse | VL11 | VL8 |
|---|---|---|---|
| 5 | 41 | honorem | gloriam |
| 5 | 41 | ab | ab |
| 5 | 41 | hominibus | hominibus |
| 5 | 41 | non | non |
| 5 | 41 | accipio | adcipio |

# Going further: towards the possibility of automatic detection of rendering?

Whenever working with translations, a recent field using NLP based approaches is automatic alignment of translations:



# Going further: towards the possibility of automatic detection of rendering?

For example, automatize and systematize group detection in VL manuscripts by performing alignment (Burton 2000):

| **Greek word** | **Group 1 rendering** | **Group 2 rendering** |
|---|---|---|
| μικρόν | *pusillum* | *modicum* |
| ἐντολή | *mandatum* | *praeceptum* |
| ἀγάπη | *caritas* | *dilectio* |
| φῶς | *lumen* | *lux* |
| λόγος | *verbum* | *sermo* |

# Going further: towards the possibility of automatic detection of rendering?

Will this give us the possibility to perform groups using renderings?

# Current works on the Vetus Latina for John

# Outline

# Current research perspective

***Research questions*:**

- Can one select automatically interesting variants and compute a variability score for classification of?
- How do the automatic classes confirm/infirm hypothesis regarding VL transmission (Pastorelli? Hougthon? Fischer?)?

Current advances using pipelines trained on Latin and Greek:

| witnesses | chapter_number | verse_number | witness_0_tokens | witness_0_lemmatized | witness_0_tense | witness_0_number | witness_0_pos | witness_1_tokens | witness_1_lemmatized | witness_1_tense | witness_1_number | witness_1_pos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **VL11-VL8** | 8 | 29 | relinquit | relinquo | Pres | - | VERB | reliquit | relinquo | Perf | - | VERB |
| **VL11-VL8** | 8 | 29 | me | ego | - | Sing | PRON | me | ego | - | Sing | PRON |
| **VL11-VL8** | 8 | 29 | solum | solus | - | - | ADV | solum | solus | - | - | ADV |
| **VL11-VL8** | 8 | 29 | quia | quia | - | - | SCONJ | quia | quia | - | - | SCONJ |

# Current research perspectives

Distance score will then be computed to perform a clustering using **agglomerative clustering**:

| chapter | verse | manuscript 1 | manuscript 2 | manuscript 1 (value) | manuscript 2 (value) | variability score |
|---|---|---|---|---|---|---|
| 16 | 6 | VL11 | VL8 | quia | quia | 0 |
| 16 | 6 | VL11 | VL14 | quia | quoniam | 1 |
| 16 | 6 | VL11 | VL15 | quia | quia | 0 |
| 16 | 6 | VL11 | VL2 | quia | quoniam | 1 |
| 16 | 6 | VL11 | VL3 | quia | quia | 0 |
| 16 | 6 | VL11 | VL4 | quia | quia | 0 |
| 16 | 6 | VL11 | VL5 | quia | quoniam | 1 |

# Questions

Questions ?