

# TWO DEMONSTRATIONS OF THE MACHINE TRANSLATION APPLICATIONS TO HISTORICAL DOCUMENTS

**Authors:** Miguel Domingo, Francisco Casacuberta  
{midobal, fcn}@prhlt.upv.es

Pattern Recognition and Human Language Technology Research Center  
Universitat Politècnica de València



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA



## INTRODUCTION

- Historical documents are an important part of our cultural heritage.
- The linguistic variation in historical documents has always been a concern for scholars in humanities.
- Human language evolves with the passage of time.
- Orthography changes depending on the author and time period.
- e.g., the data in *LALME*<sup>a</sup> indicate 45 different forms recorded for the pronoun *it*, 64 for the pronoun *she* and more than 500 for the preposition *through*.
- This language barrier limits historical documents accessibility mostly to scholars.

<sup>a</sup>Linguistic Atlas of Late Medieval English.

## APPLICATION 1: LANGUAGE MODERNIZATION

**Goal:** Make historical documents accessible to a broader audience by generating a new version of the document written in the modern version of its original language.

Example<sup>a</sup>:

As thou art to thyself.

Such was the very armour he had on

When he the ambitious Norway combated.

So frowned he once when, in an angry parle,

He smote the sledded Polacks on the ice.

'Tis strange.

Yes, as much as you look like yourself.

The king was wearing exactly this armor

when he fought the king of Norway.

And the ghost frowned just like the king did once

when he attacked the Poles, traveling on the ice in sleds.

It's weird.

<sup>a</sup>John Crowther (2003). No Fear Shakespeare: Hamlet. SparkNotes.

## APPLICATION 2: SPELLING NORMALIZATION

**Goal:** Achieve an orthography consistency by adapting the document's spelling to modern standards.

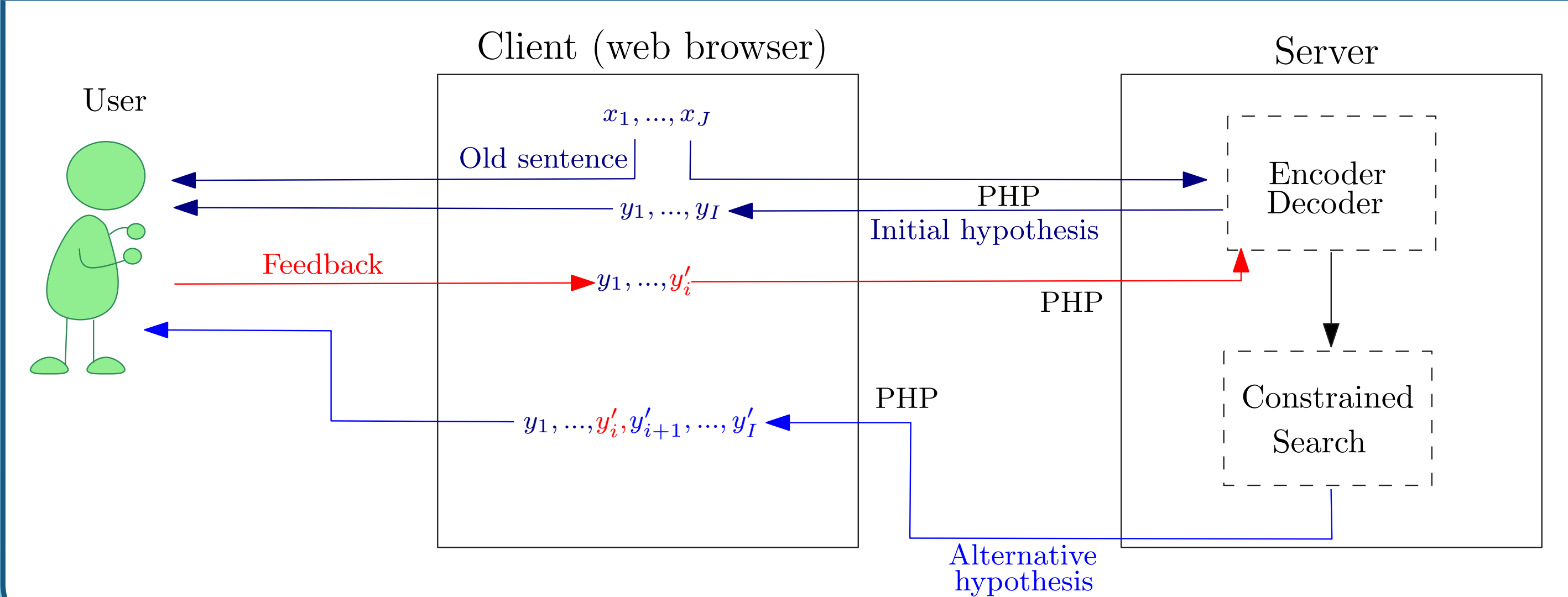
Example<sup>a</sup>:

Bien responde la esperan<sup>ç</sup>a  
en que enga<sup>ñ</sup>ado he vi<sup>u</sup>ido  
al cu<sup>y</sup>dado que he tenido  
de tu estudio y tu crian<sup>ç</sup>a!

Bien responde la esperan<sup>z</sup>a  
en que enga<sup>ñ</sup>ado he vivi<sup>d</sup>o  
al cuida<sup>d</sup>o que he tenido  
de tu estudio y tu crian<sup>z</sup>a!

<sup>a</sup>Fred F. Jehle (2001). Works of Miguel de Cervantes in Old- and Modern-spelling. Indiana University Purdue University Fort Wayne.

## SYSTEM ARCHITECTURE



## RESULTS

The screenshot shows the web application interface. At the top, there is a navigation bar with the PRHLT logo, the title 'MACHINE TRANSLATION APPLICATIONS TO HISTORICAL DOCUMENTS', and links for 'MORE INFO' and 'CONTACT'. The main content area is titled 'Sentence #1' and displays two versions of a sentence. On the left, the 'Old sentence' is shown: 'Señor, debaxo destes lienços estan vnas imagines de reliebe y entabladura, que han de seruir en vn retablo que hazemos en nuestra aldea.' Below it is a 'MODERNIZE' button. On the right, the 'System hypothesis and User feedback' area shows the modernized sentence: 'Señor, debajo de estos lienzos están unas imagines de reliebe y entabladura, que han de servir en un retablo, lo que pasaran en nuestra aldea.' Below this is a 'VALIDATE' button. A 'Learn from sample' button is also present. The interface includes labels for 'Generate initial hypothesis' and 'Incremental adaptation'.

## CONCLUSIONS

- We presented a demonstration of two MT applications to historical documents.
- We described its client-server architecture and developed a website to facilitate the use of the system.

## FUTURE WORK

- Improve our website's front end.
  - Create a better visualization of the hypothesis.
  - Visualize relevant attributes of the neural system to understand better the model predictions and behavior.

## ACKNOWLEDGEMENTS

The research leading to these results has received funding from Generalitat Valenciana (GVA) under project PROMETEO/2019/121. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for part of this research, and Andrés Trapiello and Ediciones Destino for granting us permission to use their book in our research.