



STATE

Assisted Text Transcription System

David Llorens, Andrés Marzal, Federico Prat, Juan Miguel Vilar (UJI, Spain)

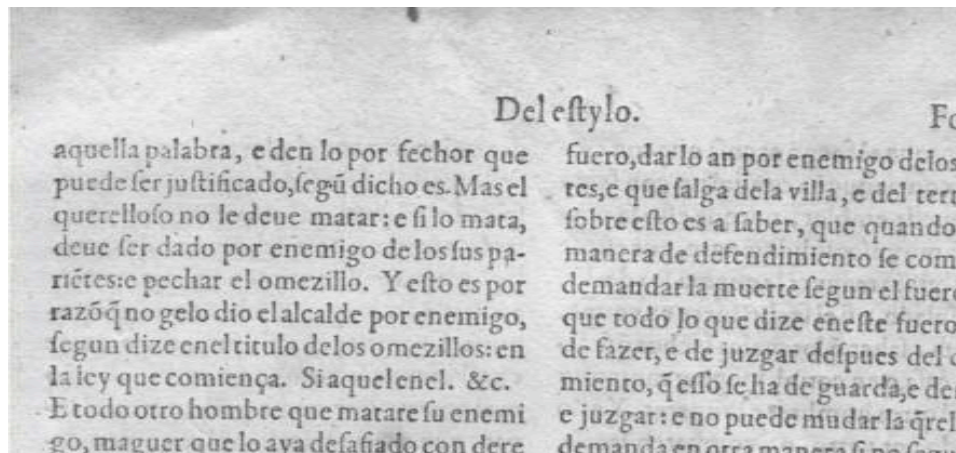
María José Castro, Salvador España, Francisco Zamora (UPV, Spain)

The people



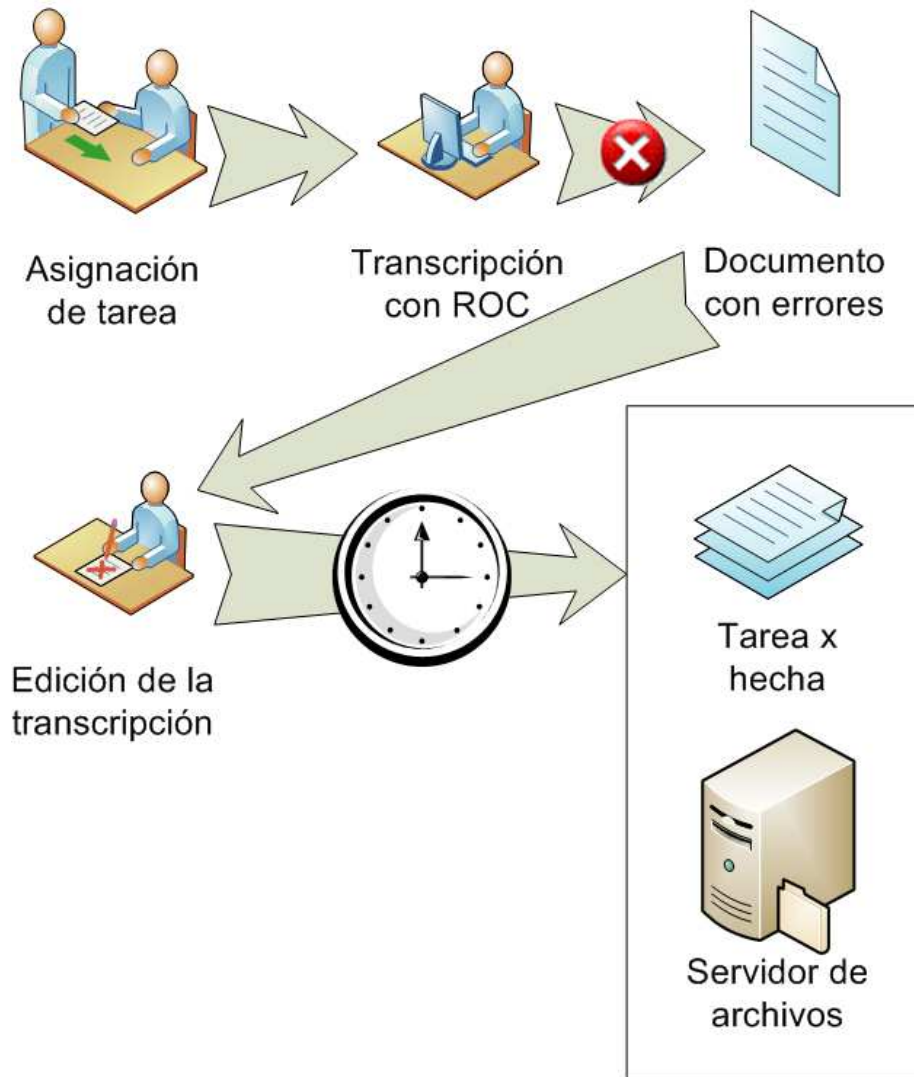
Automatic? Transcription

- ▶ **Optical Character Recognition (OCR)** systems perform **poorly**:
 - ▶ On damaged documents.
 - ▶ On ancient documents.
 - ▶ On handwritten documents.
 - ▶ On text with non-estandar characters or fonts.
- ▶ It is always needed to **supervise and correct** the automatic transcription.



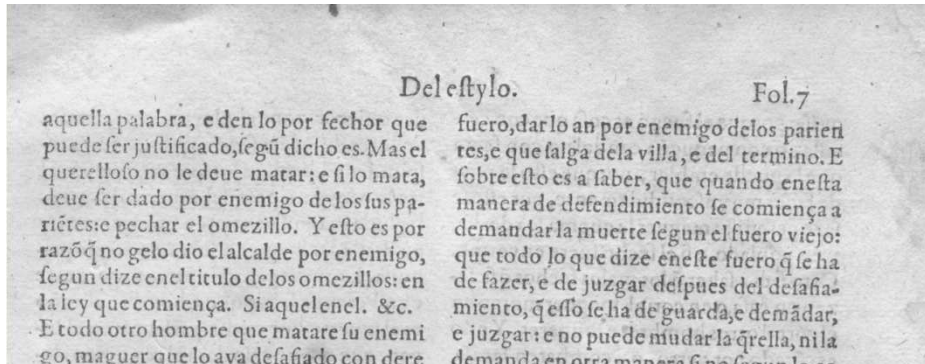
Mr. Dowell finds it easier to take it out of mothers, childrens and sick people than to take on this vast industrie," Mr. Bonn commented icily. "Let us have a full inquiry into the cost of drugs and the pharmaceutical industry." The health of children today owes much to the welfare food scheme. It was maintained during the war. Now in conditions of Tory affluence it seemed it could not be carried on.

Documentary resources digitization



- ▶ Each document is **transcribed with an OCR system**.
- ▶ The resulting text contains **many errors**.
- ▶ Text editing: the transcriber must **manually correct** the text.
- ▶ It is the most **time consuming stage**, where it is easy to introduce **errors**.
- ▶ Finally, the correct transcription is uploaded to the server.
- ▶ Now, the digital information can be exploited in many ways.

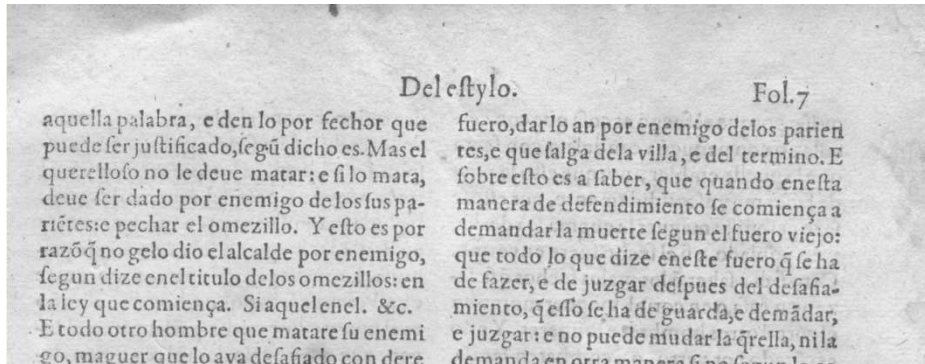
How is text editing performed?



- ▶ Comparing the page with its transcription is **very uncomfortable and difficult.**

aquella palabra , e den lo por fechor que
.puede ser j u [st]i[si]cado,seg[un] dicho es.Mas el
quer.elloso no le deue matar: e si lo mata,
deue ser dado por enemigo de los sus pa-
ri[en]tes:e pechar el omezillo. [ss] e[st]o es por
raz[on][que] no gelo dio el alcalde por enemigo,
segun dize enel titulo delos omezillos: en
.l.a ley que comienca. bi aquel enel. ac.
.E todo otro hombre que matare su enemi
go, m.uguer que lo aya desa[si]ado con dere

How is text editing performed?



aquella palabra , e den lo por fechor que
.puede ser j u [st]i [si] cado, seg[un] dicho es. Mas el
quer.eloso no le deve matar: e si lo mata,
deve ser dado por enemigo de los sus pa-
ri[en]tes: e pechar el omezillo. [ss] e[st]o es por
raz[on][que] no gelo dio el alcalde por enemigo,
segun dize en el titulo de los omezillos: en
.l.a ley que comiença. bi aquel enel. ac.
.E todo otro hombre que matare su enemi-
go, m.uguer que lo aya desa[si]ado con dere

- ▶ Comparing the page with its transcription is **very uncomfortable and difficult**.
- ▶ The human expert must
 - ▶ **jump** continuously from the image to the text,
 - ▶ **find** each error,
 - ▶ **move** with the mouse or keyboard,
 - ▶ and **type** the correction.

Ancient documents

Problems get worse with ancient documents

- ▶ Stains, handwritten annotations on the margins, bad conditions, cracks, patches, disappearing ink in some regions, and so on:

Text extraction is difficult for OCR systems.

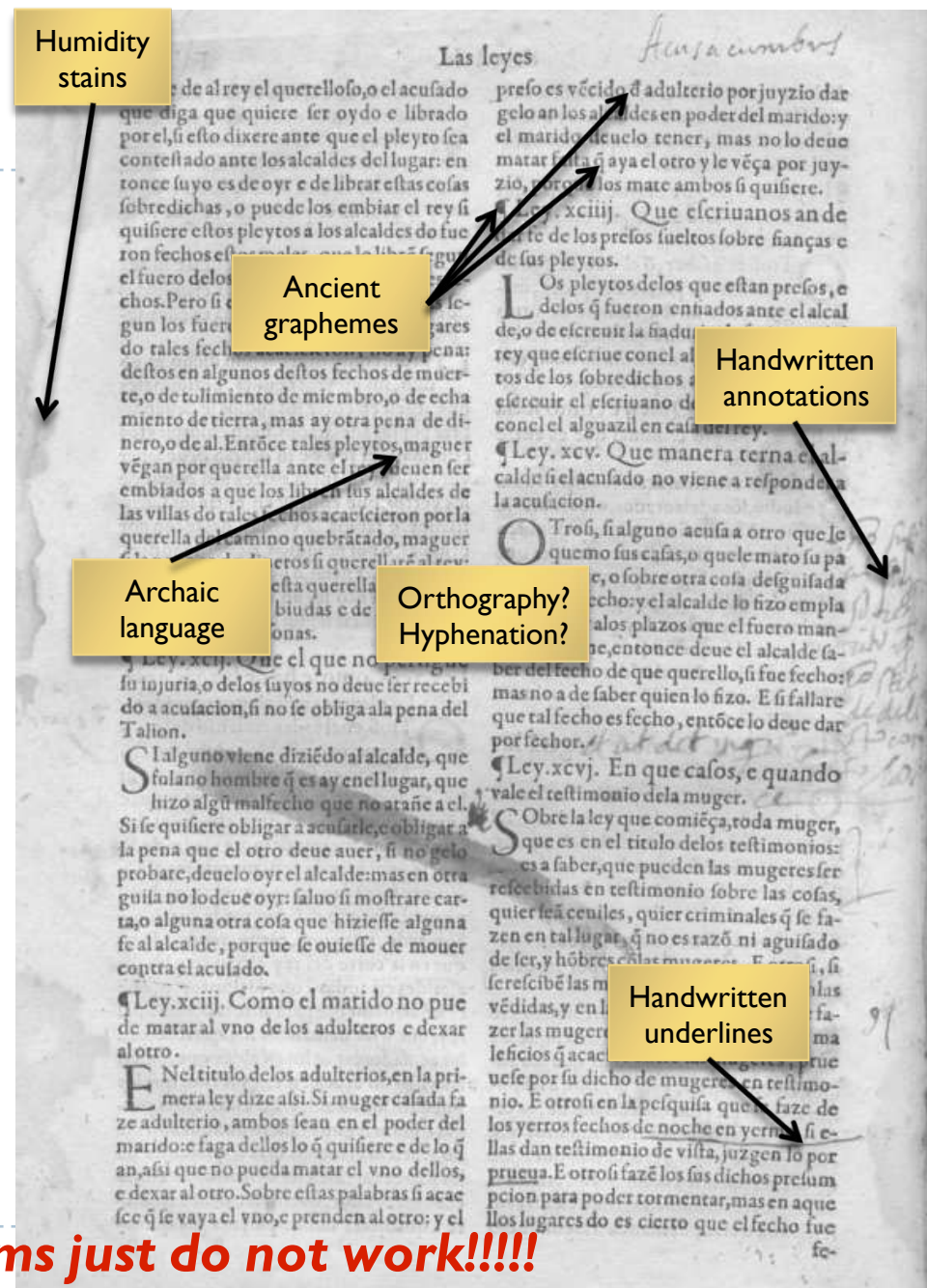
- ▶ Special fonts or ancient graphemes:

OCR systems do not recognize them.

- ▶ Non-modern orthography or syntax and archaic language:

Language models of OCR systems are not suitable.

Standard OCR systems just do not work!!!!

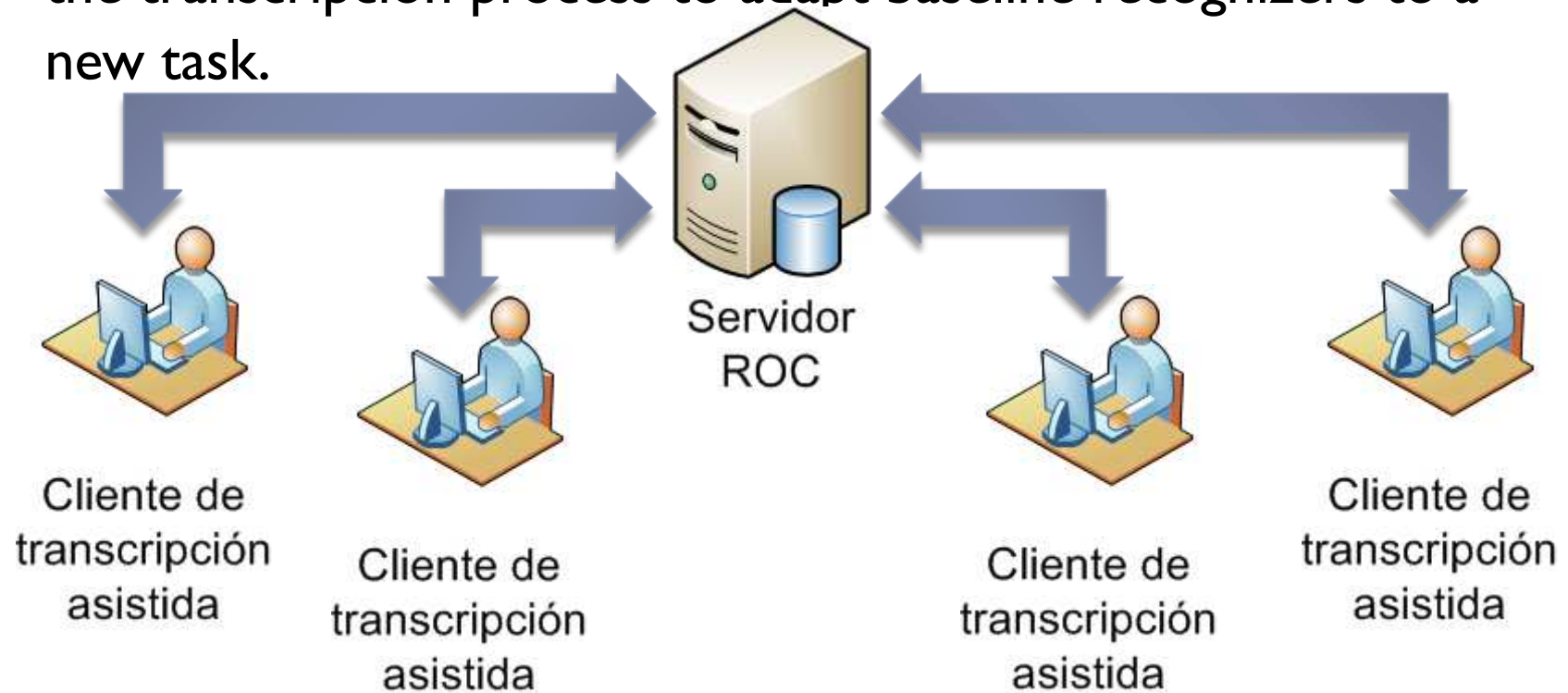


Our aims

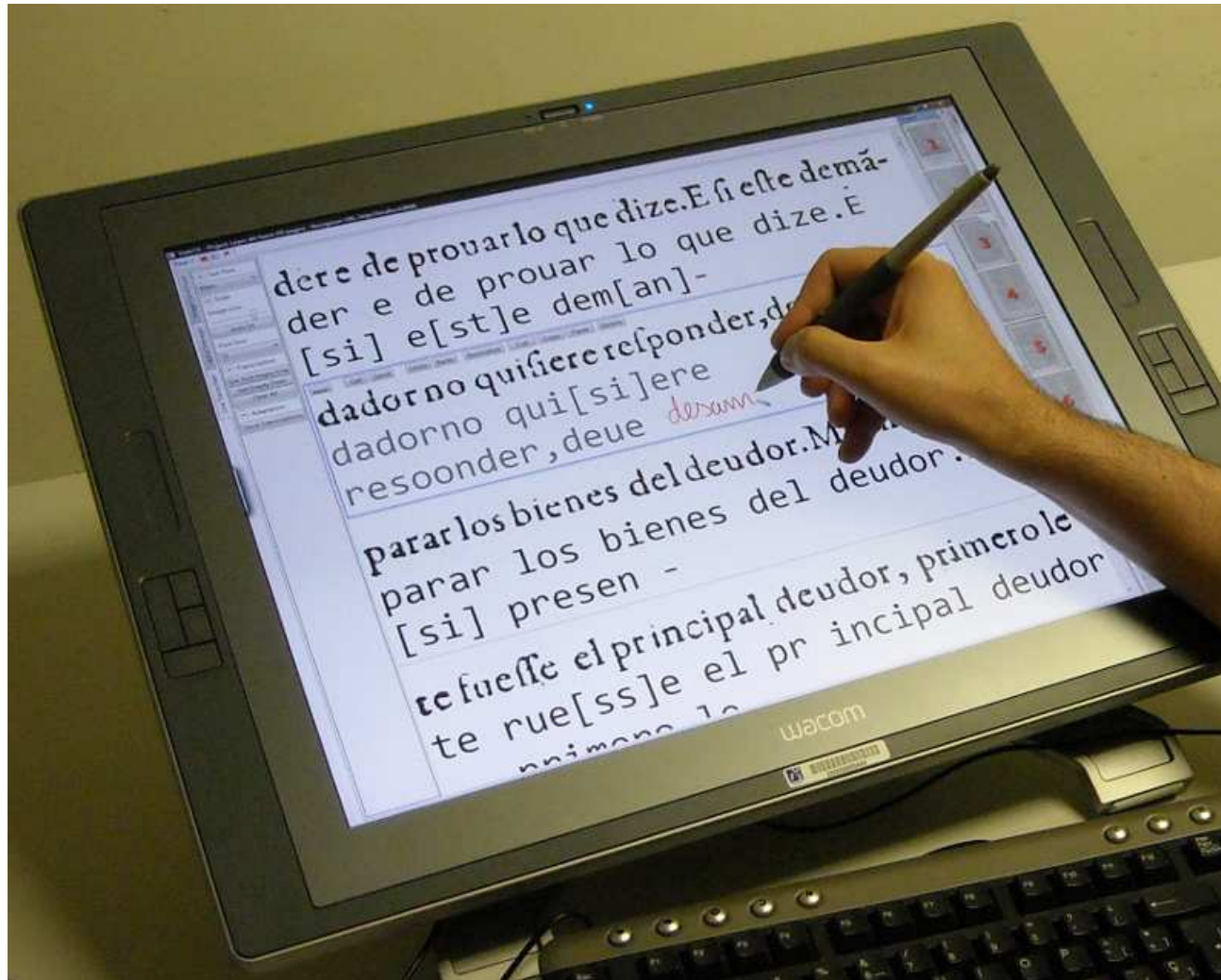
- ▶ Our transcription system **integrates advanced and accurate tools**
 - ▶ for *image processing*,
 - ▶ to *detect page layout*,
 - ▶ for *text recognition*.
- ▶ **But also... it assists the human expert** to
 - ▶ *partially automatize* the text editing process, and
 - ▶ attend to usability in order to ease the *interactive text editing process*.
- ▶ And the recognition system includes **adaptive learning**: it learns from samples of each new task.
- ▶ With all of this: we aim to **drastically reduce the time devoted to text editing**.

Adaptive learning recognition system

- ▶ We have designed a **client/server** architecture: what the system learns from a user, all users benefit immediately.
- ▶ New **glyphs, lexicon entries** or **LMs** can be added during the transcription process to adapt baseline recognizers to a new task.



STATE: An Assisted Text Transcription System



STATE: An Assisted Text Transcription System

- ▶ A **Transcription Assistant** that helps to
 - ▶ Clean images
 - ▶ Find page layout
 - ▶ Correct the given transcription

StateTA

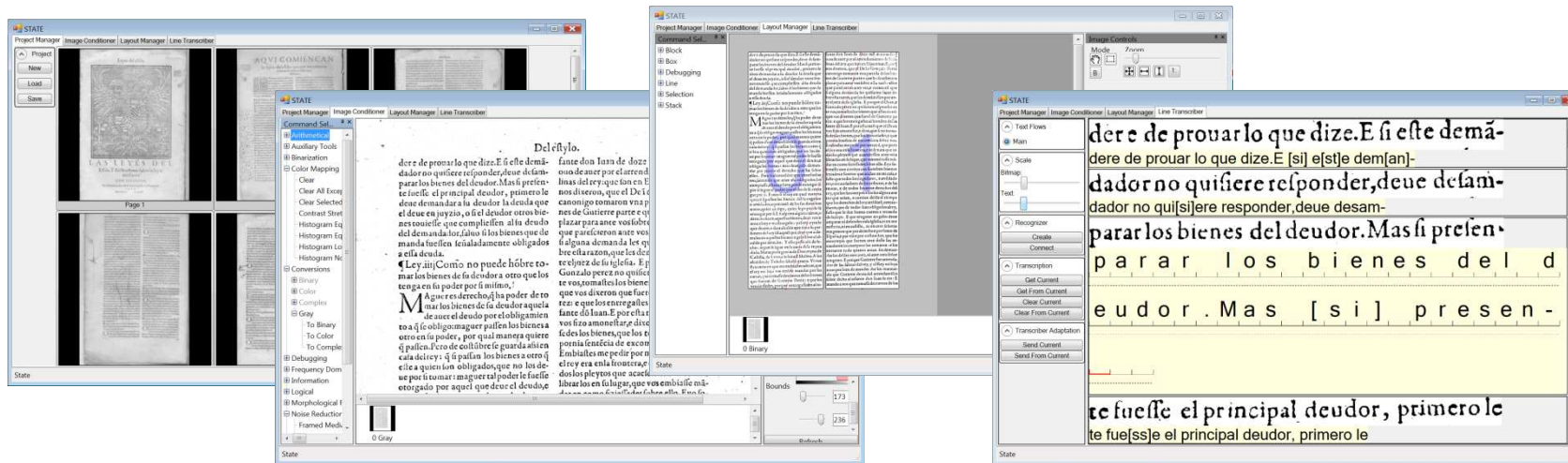
- ▶ A **text Recognition Engine**
 - ▶ For ancient printed documents
 - ▶ For handwritten documents

StateRE_NN

StateRE_HMM/ANN

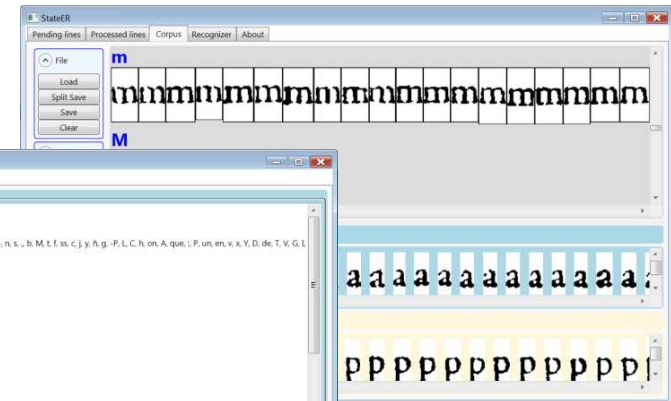
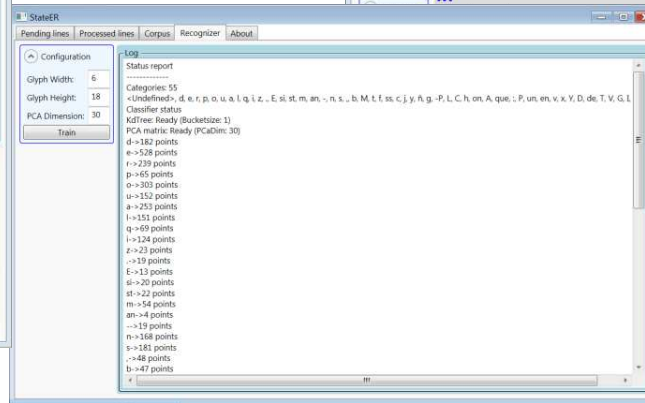
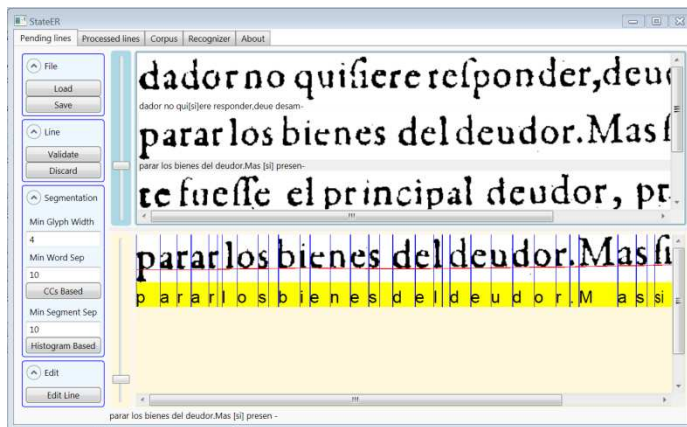
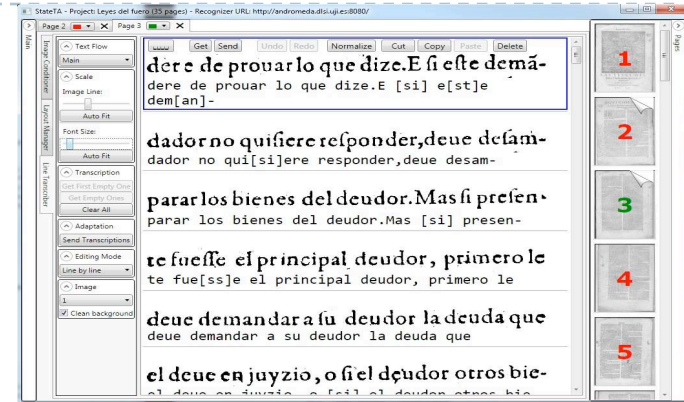
StateTA: State Transcription Assistant

- ▶ 4 components:
 - ▶ Project manager
 - ▶ Image conditioner
 - ▶ Layout manager
 - ▶ Line transcriber



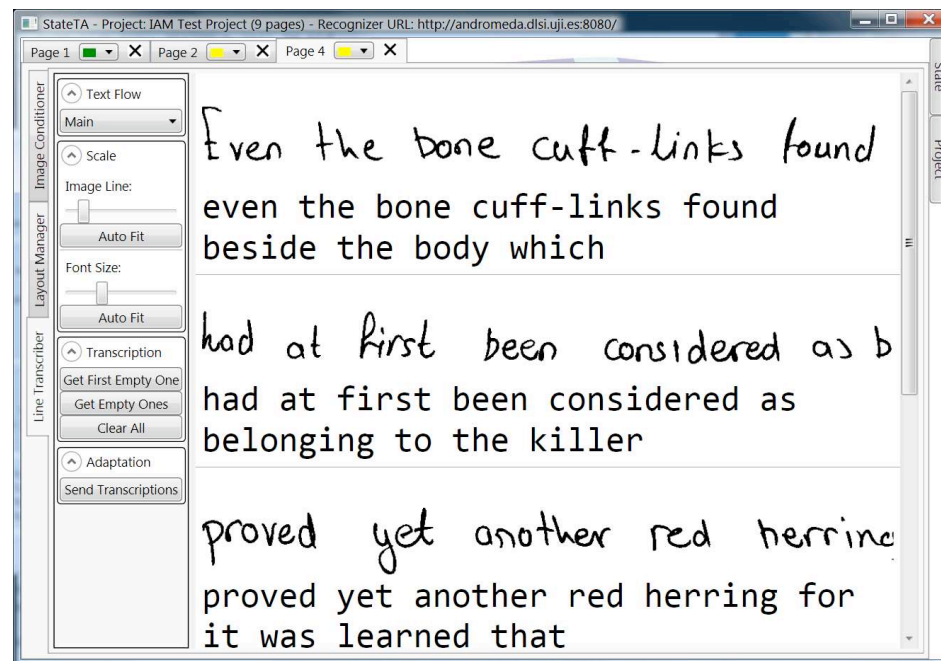
StateRE_NN: State Recognition Engine

- ▶ It is a line-by-line **OCR** based on a **Nearest Neighbors** ...
- ▶ ...and an application **manager** to
 - ▶ **Enlarge the corpus** to learn.
 - ▶ **Re-Train** the recognition engines.



StateRE_HMM/ANN: State Recognition Engine

- ▶ It is a line-by-line **Handwritten Text Recognizer**.
- ▶ The recognition engine is based on **HMM/ANN**: Hidden Markov Models hybridized with Artificial Neural Networks to estimate the emission probabilities.



But... Does STATE increase the productivity?

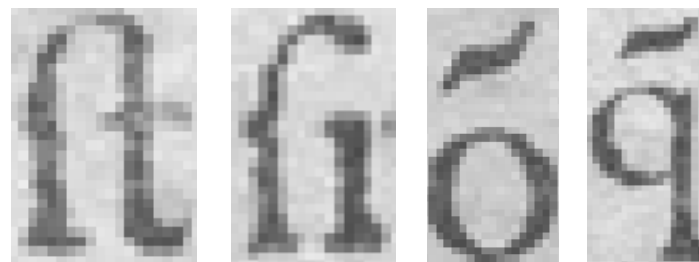
System evaluation

dere de provar lo que dize. E si este demãdador no quisiere responder, deue defãmpar los bienes del deudor. Mas si presente fuesse el principal deudor, primero le deue demandar a su deudor la deuda que el deue en juyzio, o si el deudor otros bienes touiesse que cumplieren al su deudo del demandador, salvo si los bienes que de manda fuesen señaladamente obligados a essa deuda.

¶ Ley. iiii. Como no puede hõbre tomar los bienes de su deudor a otro que los tenga en su poder por si mismo.

Muger es derecho, q̄ ha poder de tomar los bienes de su deudor a quella de auer el deudo por el obligamien to a q̄ se obligo: maguer passen los bienes a otro en su poder, por qual manera quiere q̄ passen. Pero de costũbre se guarda asy en casa del rey: q̄ si passan los bienes a otro q̄

- ▶ **Task:** a Spanish book dated in 1569, printed in ancient font, with archaic lexicon and syntax, hundreds of abbreviations, and no consistent rule about word separation.



System evaluation: Experiment

- ▶ Time saving when using STATE with respect to using only a text editor.

| System | Time for each page |
|-------------|----------------------------------|
| Editor only | between 27 and 35 minutes |
| STATE | around 12 minutes |

System evaluation

- ▶ **Task:** unrestricted handwritten task from the IAM database: collection of forms written by different people.

He said there concerned Mr. Weaver's alleged association with organizations blacklisted by the Government. Immediately Mr. Kennedy pushed a letter to Senator Robertson saying the Federal Bureau of Investigation had reported on Mr. Weaver.

"Of course you must count about two hundred for legal charges and stamp duties, maybe less, depending on the price of the house, and whether it has been registered. I take it you have a mortgage lined

"Mr. Donnell finds it easier to take it out of mothers, childrens and sick people than to take on this vast industry," Mr. Bonn commented icily. "Let us have a full inquiry into the cost of drugs and the pharmaceutical industry." The health of children today owed much to the welfare food scheme. It was maintained during the war. Now in conditions of Tory affluence it seemed it could not be carried on.

System evaluation: Experiment

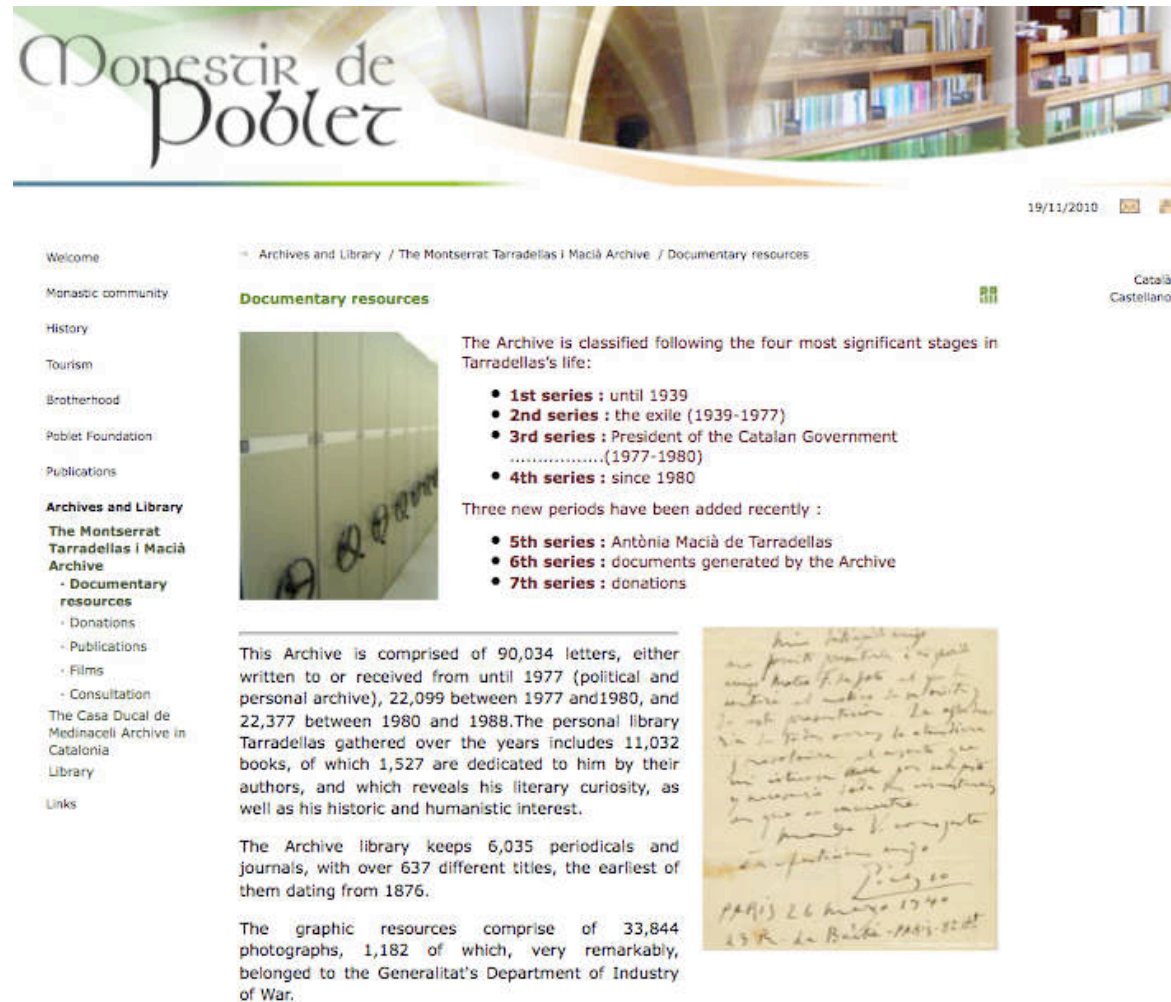
- ▶ Time saving when using STATE with respect to using only a text editor.

| CER (Character Error Rate) | Editor only | STATE |
|----------------------------|-------------|-------|
| $CER < 15$ | 100 | 49 |
| $15 \leq CER < 30$ | 100 | 69 |
| $30 \leq CER$ | 100 | 95 |

To summarize

- ▶ **STATE** is a **practical solution** to the assisted transcription problem for difficult ancient printed or handwritten documents.
- ▶ It offers a **multimodal interface**: mouse, keyboard and stylus.
- ▶ It can be connected to **different recognition engines** (at the moment one based on NN and other based on HMM/ANNs).
- ▶ It can be easily **adapted to new documents**: it learns from samples obtained from the documents to be transcribed.

STATE in use



Monestir de Poblet

19/11/2010

Wellcome

Monastic community

History

Tourism

Brotherhood

Poblet Foundation

Publications

Archives and Library

The Montserrat Tarradellas i Macià Archive

- Documentary resources

- Donations

- Publications

- Films

- Consultation

The Casa Ducal de Medinaceli Archive in Catalonia

Library

Links

Archives and Library / The Montserrat Tarradellas i Macià Archive / Documentary resources

Documentary resources

The Archive is classified following the four most significant stages in Tarradellas's life:

- **1st series** : until 1939
- **2nd series** : the exile (1939-1977)
- **3rd series** : President of the Catalan Government(1977-1980)
- **4th series** : since 1980

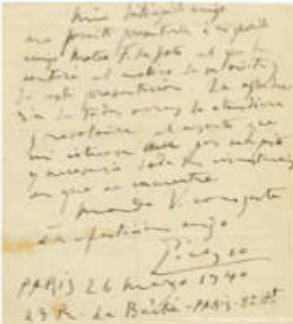
Three new periods have been added recently :

- **5th series** : Antònia Macià de Tarradellas
- **6th series** : documents generated by the Archive
- **7th series** : donations

This Archive is comprised of 90,034 letters, either written to or received from until 1977 (political and personal archive), 22,099 between 1977 and 1980, and 22,377 between 1980 and 1988. The personal library Tarradellas gathered over the years includes 11,032 books, of which 1,527 are dedicated to him by their authors, and which reveals his literary curiosity, as well as his historic and humanistic interest.

The Archive library keeps 6,035 periodicals and journals, with over 637 different titles, the earliest of them dating from 1876.

The graphic resources comprise of 33,844 photographs, 1,182 of which, very remarkably, belonged to the Generalitat's Department of Industry of War.



STATE in use



ARXIU VIRTUAL JAUME I

Documents d'època medieval relatius a la Corona d'Aragó

1237, juliol 9. El Puig

Jaume I dona a l'abat de Lacrassa l'església de Sant Vicent i unes terres, a la ciutat de València

Arxiu de la Corona d'Aragó. Barcelona. Cancelleria Reial. Registre 5, f. 1r. Original. És la nota escrita al Llibre del Repartiment del Regne de València, on tan sols figura la breu regesta feta pels notaris reials

B. abbas de Crassa, ecclesiam Sancti Vincentii et XXX jovatas terre in termino de Valentia. VII idus julii.

Universitat Jaume I. Castelló. Arxiu Virtual Jaume I - <http://www.jaumeprimer.uji.es> - Document nº 001397

[castellano]

Pàgina principal

Presentació

Notícies

Comentaris a documents

Obres de referència

Documents extensos



Contactar

STATE in use

The screenshot displays the website 'La Biblioteca Virtual del Español' with the following layout:

- Header:** 'El Bibliotecario' on the right, and 'La Biblioteca Virtual del Español' and 'Biblioteca Virtual Miguel de Cervantes' in the center.
- Left Sidebar:**
 - Logo: BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES
 - Navigation menu:
 - Catálogo general
 - Literatura
 - Lengua
 - Historia
 - Biblioteca Americana
 - Biblioteca de Signos
 - Biblioteca Joan Lluís Vives
 - Biblioteca Letras Galegas
 - Literatura Infantil y Juvenil
 - Últimos contenidos incorporados
 - Obras más consultadas
 - Efemérides
 - Noticias
 - Suscripción al Boletín
 - Revista de Novedades Editoriales
 - La biblioteca accesible en L.S.E.
- Main Content Area:**
 - Novedades:** A grid of book covers including 'José Enrique Rodó', 'Joanot Martorell, el Tirant lo Blanc', 'LETRAS MEXICANAS', 'Pedro Antonio de Alarcón', and 'Antonio Rodríguez Blinodóum'.
 - Recomendaciones:** A section titled '-El Boomeran(g)-' featuring 'Revista Etiqueta Negra' with the text: 'Una especie de The New Yorker en castellano, pero mejor diseñada. ¿Se podía hacer una revista así en el Perú?' and a small image of the magazine cover.
 - Colaboradores:** Logos for 'Santander' and 'Instituto Cervantes'.
- Right Sidebar:**
 - Actualidad:** A box titled 'EL BLOG' with the logo and text: 'VISITA NUESTRO BLOG Y...'. It lists three points: 'Estarás al día de todas nuestras noticias.', 'Te enterarás de los últimos contenidos incorporados.', and 'Suscríbete al blog por correo o RSS.'
 - publicidad:** A placeholder for an advertisement.

Perspectives and future work

- ▶ STATE: Sistema de Transcripción Asistida para Texto Escrito
(TIN2006-12767)
2006 – 2009
- ▶ HITITA: Herramienta Interactiva para la Transcripción de Imágenes de Textos Antiguos
(TIN2010-1958)
2011 – 2013

Perspectives and future work

▶ Immediate plans:

- ✓ to include **recognition confidence** information for each character, so that the user can go quickly to places where corrections may be needed
- ✓ to improve the recognition engine for ancient printed/handwritten text using **language models** and **HMM-based** decoders
- ✓ to enrich the obtained transcription with **scholarly descriptions in XML format** (TEI standard, for example)

A live demo!!!

dadorno quifiere responder,deue defam-
dadorno qui[si]ere
resoonder,deue desam-
parar los bienes del deudor.Mas si prefen-
parar los bienes del
deudor.Mas [si] preien -
te fuesse el principal deudor, primero le
te rue[ss]e el pr incipal
deudor , primero le
deue demandar a su deudor la deu
deue deman dar a iu deu