STATE

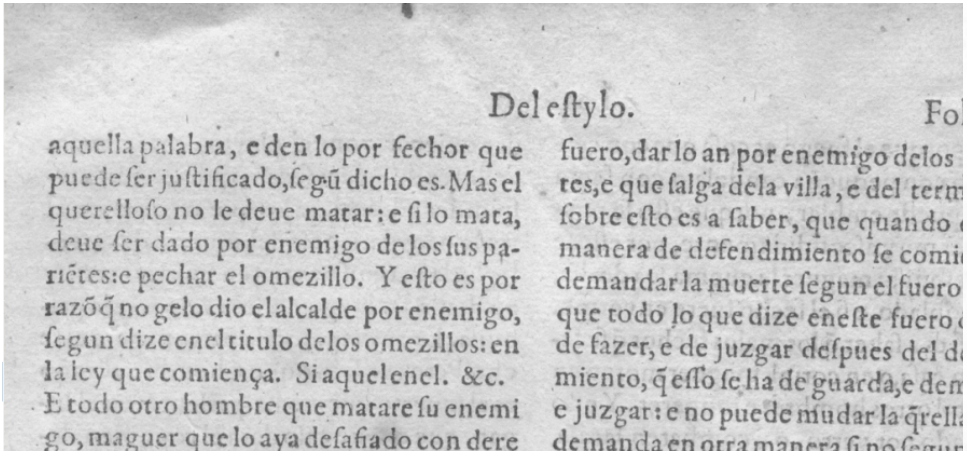
Assisted Text Transcription System

David Llorens, Andrés Marzal, Vicente Palazón, Federico Prat, Juan Miguel Vilar (UJI, Spain)
María José Castro, Salvador España, Joan Pastor, Francisco Zamora (UPV, Spain)

The people

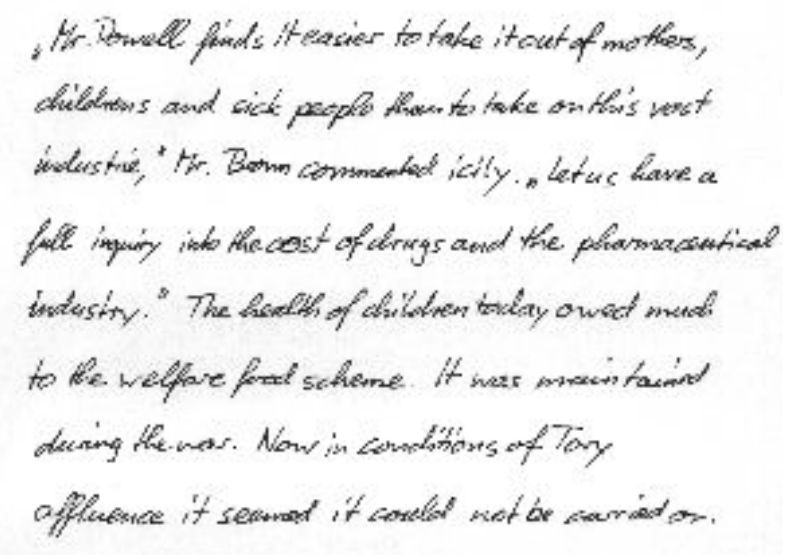
Automatic? Transcription

- **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.

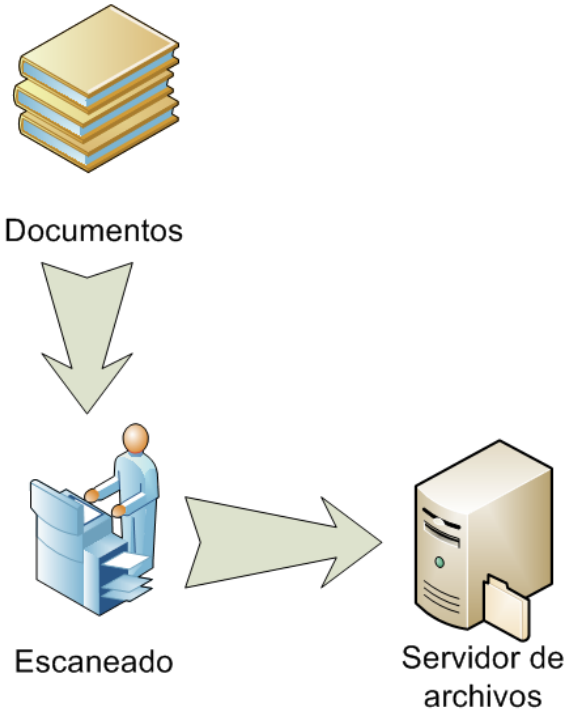

Del eſtylo. Fo

aquella palabra, e den lo por fechor que puede ſer juſtificado, ſegũ dicho es. Mas el querelloſo no le deue matar: e ſi lo mata, deue ſer dado por enemigo de los ſus pa-riẽtes: e pechar el omezillo. Y eſto es por razõ q̃ no gelo dio el alcalde por enemigo, ſegun dize en el titulo de los omezillos: en la ley que comiença. Si aquel enel. &c.
E todo otro hombre que matare ſu enemi go, maguer que lo aya deſafiado con dere

fuero, dar lo an por enemigo de los tes, e que ſalga de la villa, e del term ſobre eſto es a ſaber, que quando manera de defendimiento ſe comi demandar la muerte ſegun el fuero que todo lo que dize en eſte fuero de fazer, e de juzgar deſpues del d miento, q̃ eſſo ſe ha de guarda, e den e juzgar: e no puede mudar la q̃rell demanda en otra manera ſi no ſegu

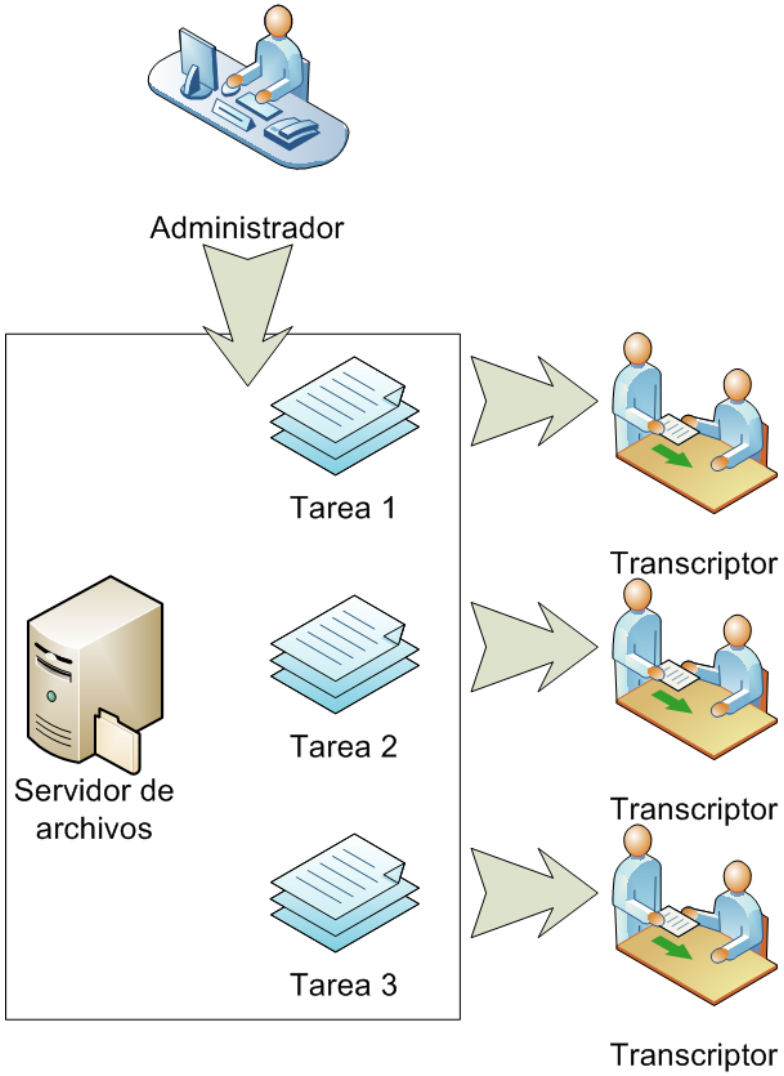

„Mr. Powell finds it easier to take it out of mothers, children and sick people than to take on his vest industries," Mr. Bevan commented icily. „Let us have a full inquiry into the cost of drugs and the pharmaceutical industry." The health of children today owed much to the welfare food scheme. It was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

Documentary resources digitization



- Documents are **scanned**…
- …and the obtained images are uploaded to a **file server.**

Documentary resources digitization



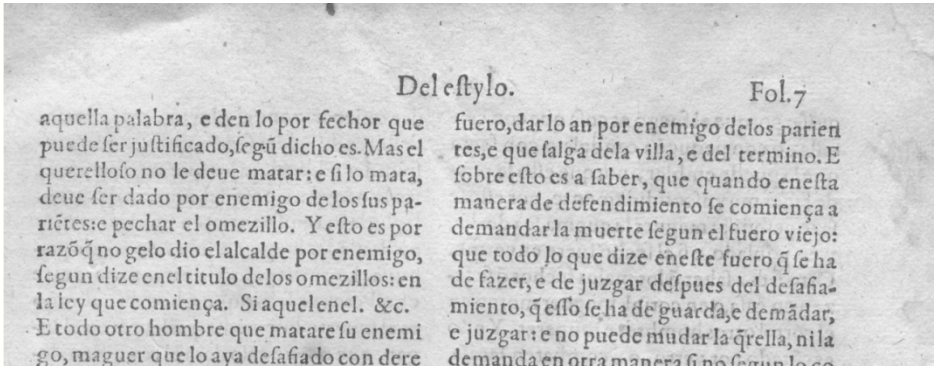
- The **administrator** divides the documents in groups of pages (**tasks**).
- Each task is assigned to a **human transcriber**.

His dutie is to obtain a text file from the original document.

Documentary resources digitization

- Each document is **transcribed with an OCR sytem**.
- 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?



Del eſtylo. Fol.7

aquella palabra, e den lo por fechor que puede ſer juſtificado, ſegũ dicho es. Mas el querelloſo no le deue matar: e ſi lo mata, deue ſer dado por enemigo de los ſus pariẽtes: e pechar el omezillo. Y eſto es por razõq̃ no gelo dio el alcalde por enemigo, ſegun dize enel titulo delos omezillos: en la ley que comiença. Si aquel enel. &c. E todo otro hombre que matare ſu enemigo, maguer que lo aya deſafiado con dere

fuero, dar lo an por enemigo delos parientes, e que ſalga dela villa, e del termino. E ſobre eſto es a ſaber, que quando eneſta manera de defendimiento ſe comiença a demandar la muerte ſegun el fuero viejo: que todo lo que dize eneſte fuero q̃ ſe ha de fazer, e de juzgar deſpues del deſafiamiento, q̃ eſſo ſe ha de guarda, e demãdar, e juzgar: e no puede mudar la q̃rella, ni la demanda en otra manera ſi no ſegun lo co

aquella palabra , e den lo por fechor que
.puede serj 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 quelo 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:

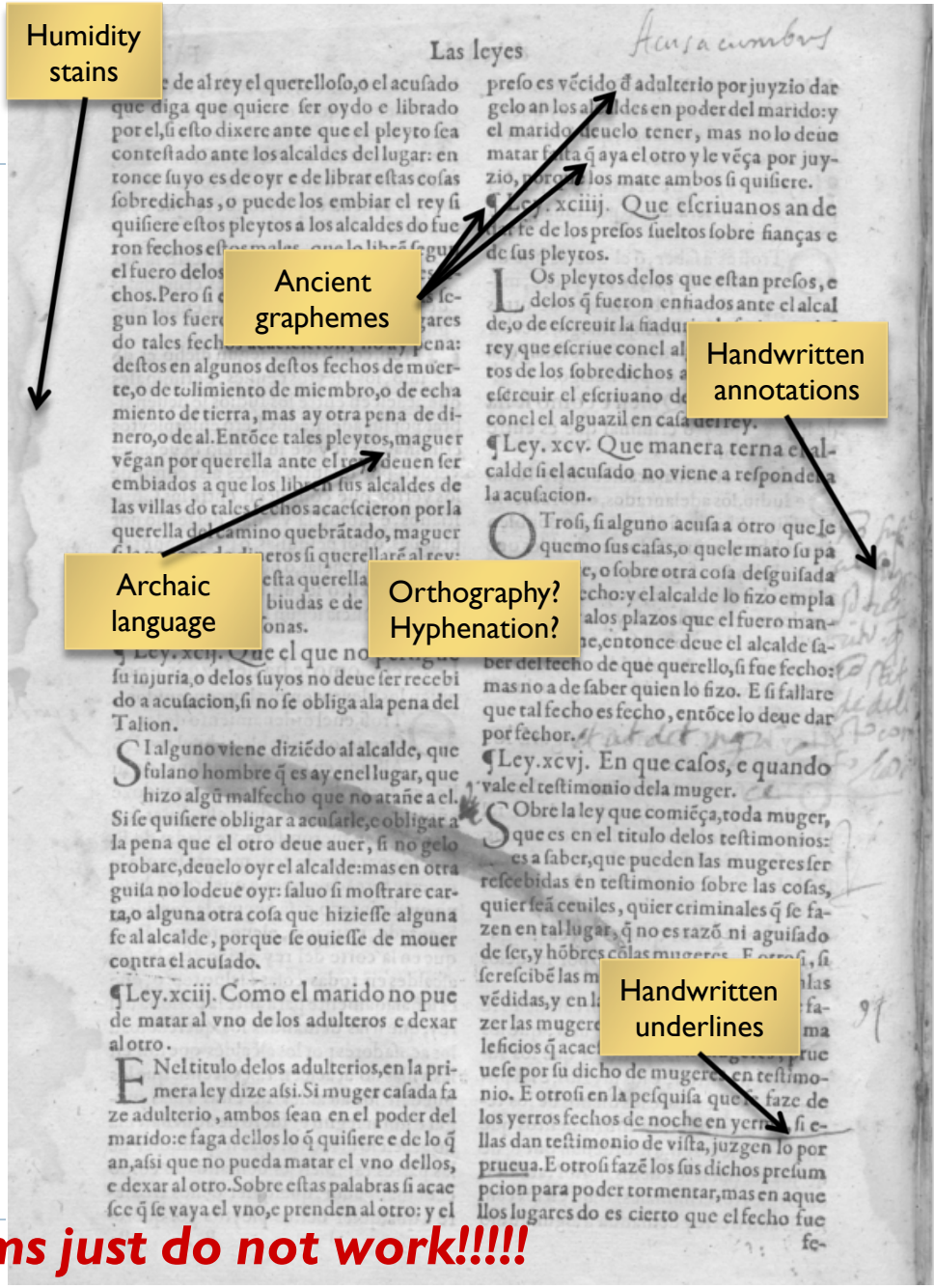
  Text extraction is difficult for OCR systems.

- 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!!!!!

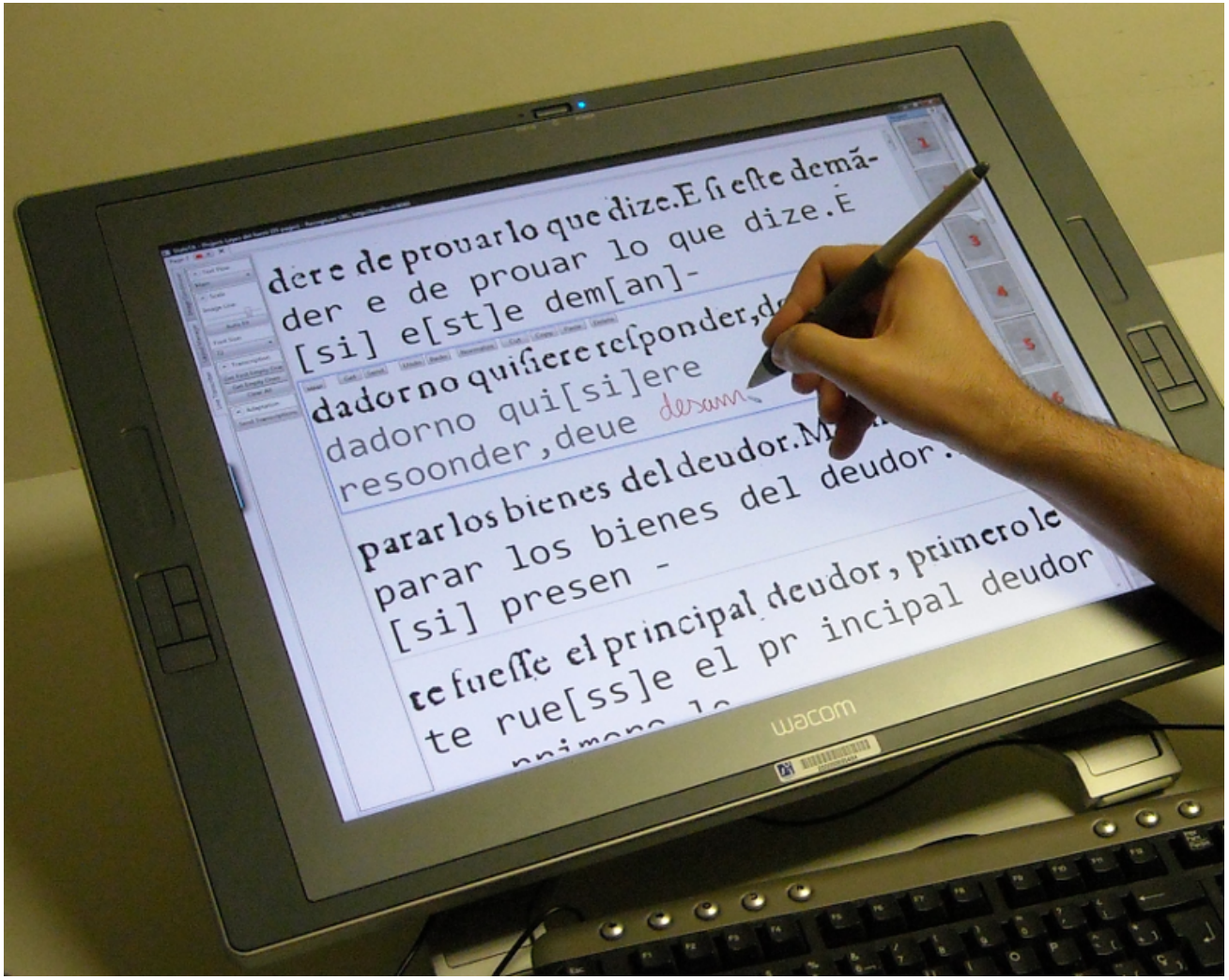
Our aims

- 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.**

Multimodal interaction

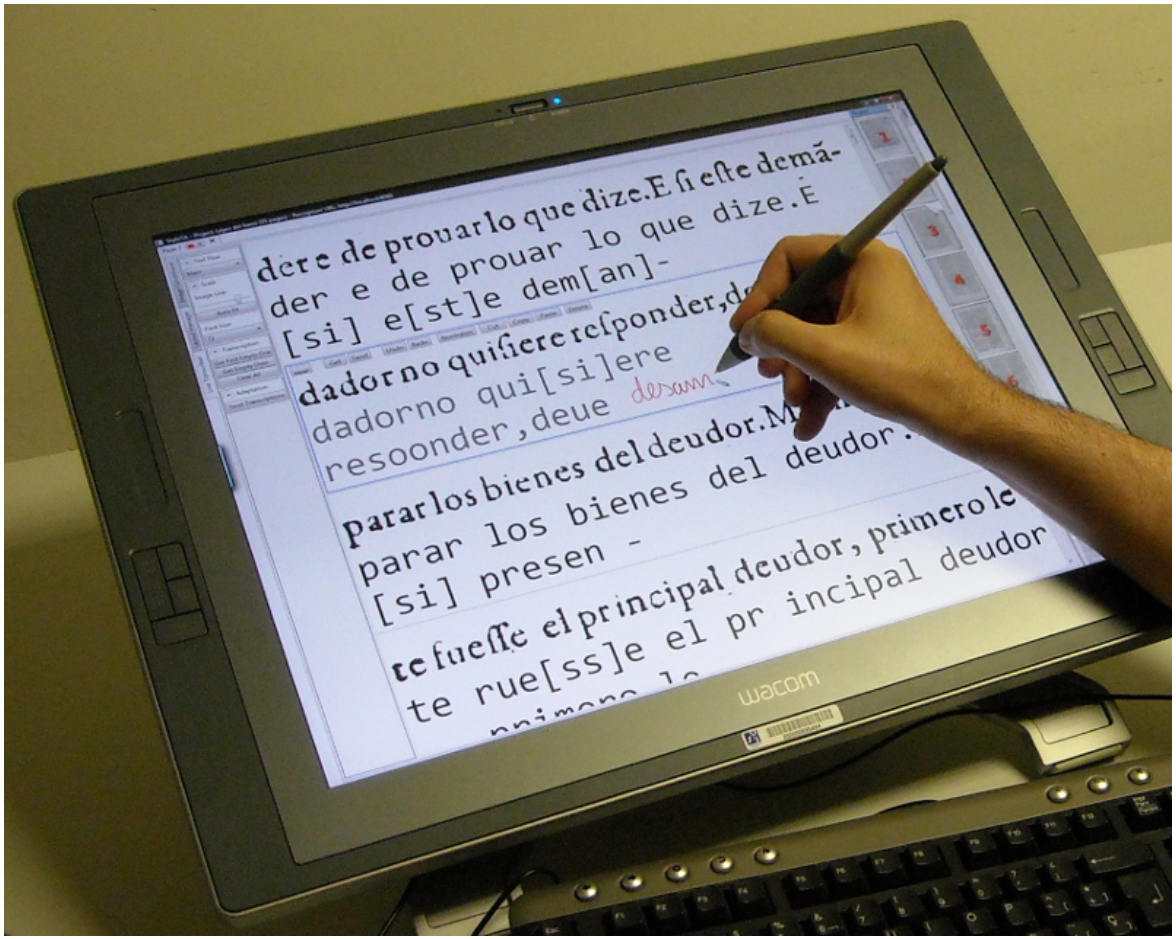
- ***Mouse, keyboard,*** and ***stylus.***
- ***Voice***

STATE: An Assisted Text Transcription System

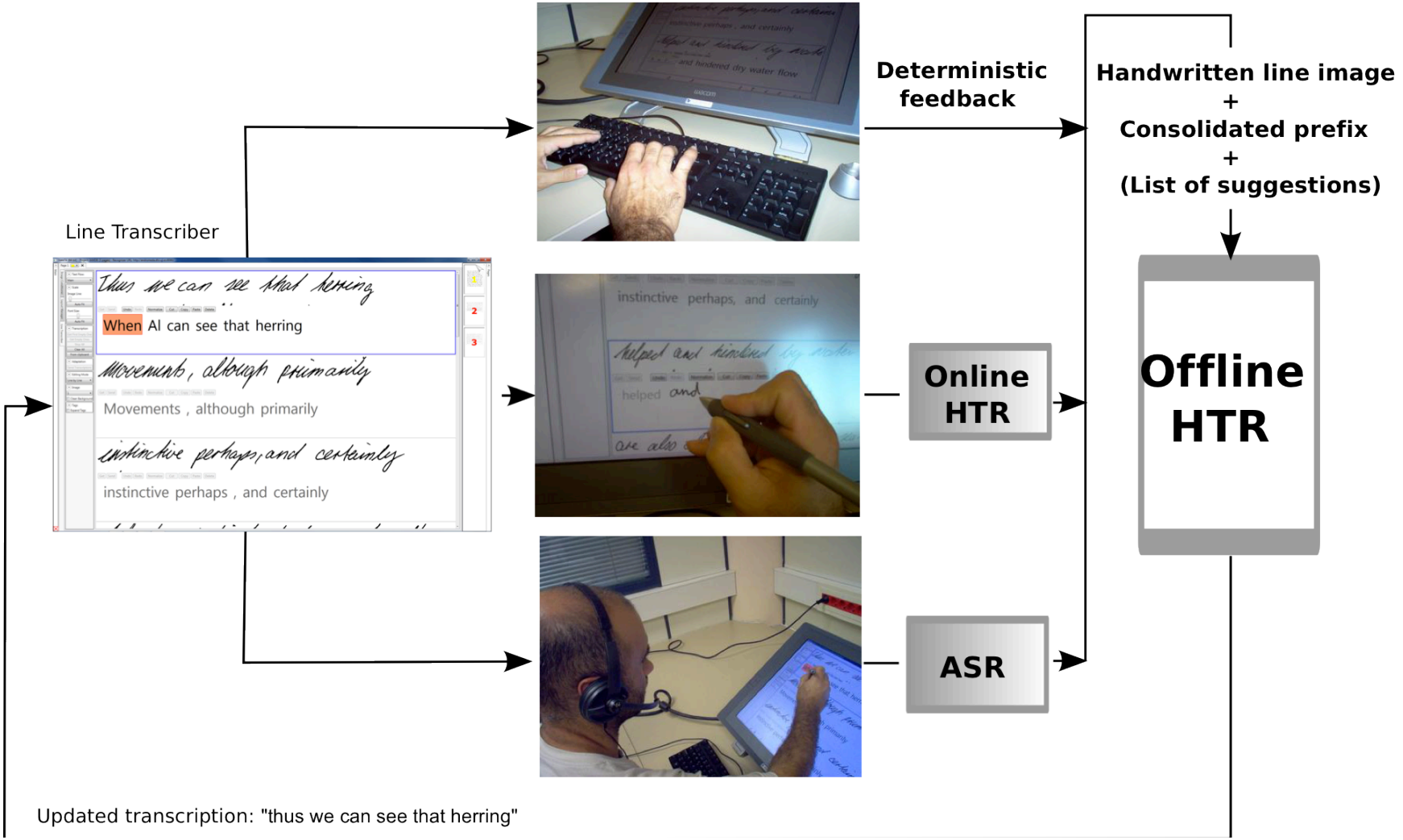


StateTA: Line Transcriber

- ✓ It allows **to edit text** line by line
- ✓ Comfortable **interaction**
- ✓ **Keyboard**, **stylus** or **speech**



User-interaction cycle with STATE via keyboard, stylus, or voice



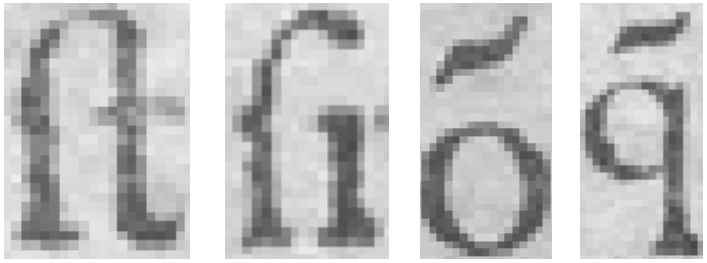
But… Does STATE increase the productivity? System evaluation

der e de prouar lo que dize. E ſi eſte demã-
dador no quiſiere reſponder, deue deſam-
parar los bienes del deudor. Mas ſi preſen-
te fueſſe el principal deudor, primero le
deue demandar a ſu deudor la deuda que
el deue en juyzio, o ſi el deudor otros bie-
nes touieſſe que cumplieſſen al ſu deudo
del demandador, ſaluo ſi los bienes que de
manda fueſſen ſeñaladamente obligados
a eſſa deuda.

¶ Ley. iiij Como no puede hõbre to-
mar los bienes de ſu deudor a otro que los
tenga en ſu poder por ſi miſmo.

MAguer es derecho, q̃ ha poder de to
mar los bienes de ſu deudor aquela
de auer el deudo por el obligamien
to a q̃ ſe obligo: maguer paſſen los bienes a
otro en ſu poder, por qual manera quiere
q̃ paſſen. Pero de coſtũbre ſe guarda aſsi en
caſa del rey: q̃ ſi paſſan los bienes a otro q̃

- **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.



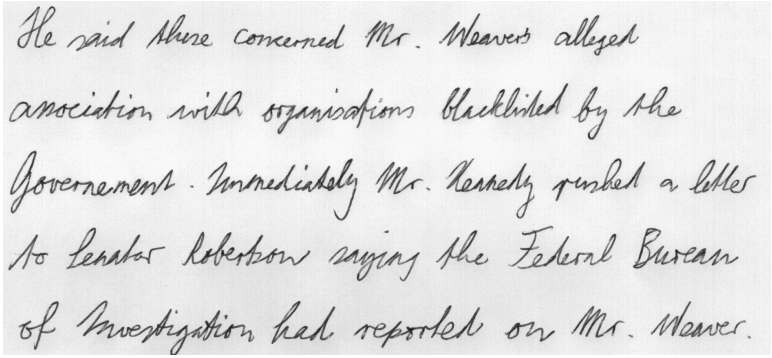
System evaluation: Experiment

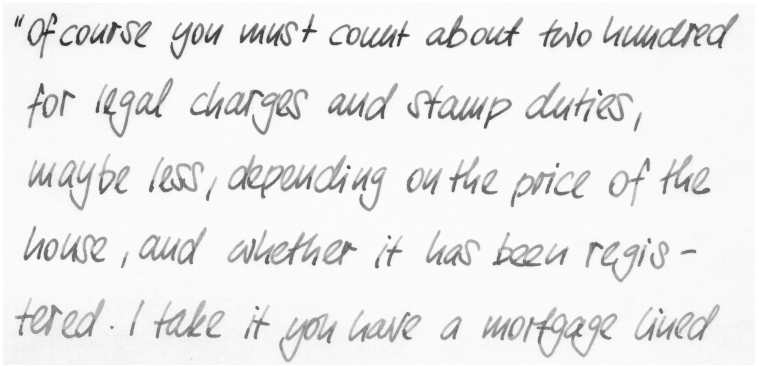
- Time saving when using STATE with respect to using only a text editor.

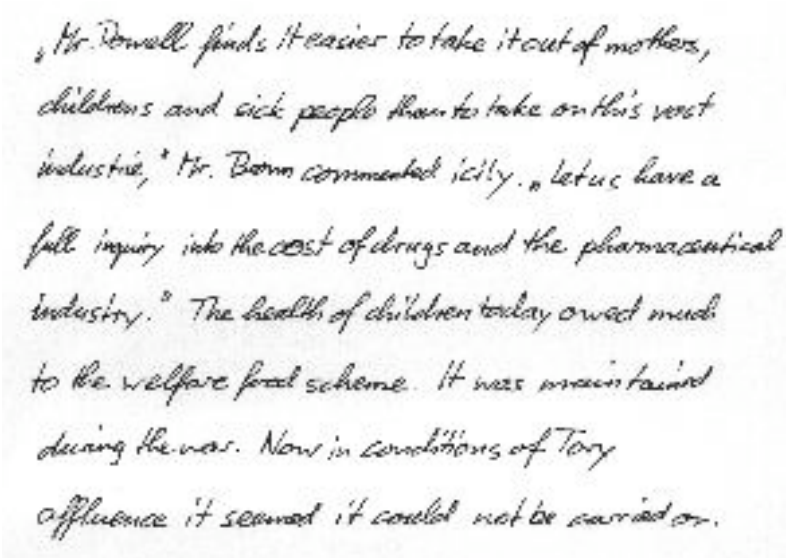
System	Time for each page
Editor only	between **27 and 35** minutes
STATE	around **12** minutes

System evaluation

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







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 ≤ CER < 30	100	69
30 ≤ 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, stylus and voice.
- It can be connected to **different recognition engines** (at the moment: NN, HMM, HMM/ANN).
- It can be easily **adapted to new documents**: it learns from samples obtained from the documents to be transcribed.

Questions?