



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.

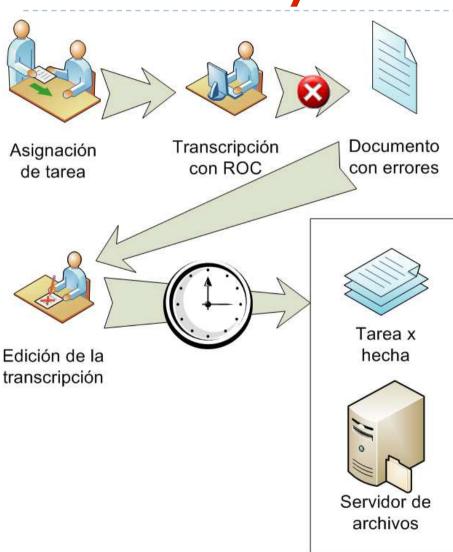
Deleftylo.

aquella palabra, e den lo por fechor que puede ser justificado, segú dicho es Mas el tes, e que salga dela villa, e del term querelloso no le deue matar: e fi lo mata, deue ser dado por enemigo de los sus pariétes:e pechar el omezillo. Y esto es por razoq no gelo dio el alcalde por enemigo, fegun dize enel titulo delos omezillos: en la ley que comiença. Si aquelenel. &c. E todo otro hombre que matare su enemi go, maguer que lo ava defafiado con dere

fuero, dar lo an por enemigo delos fobre esto es a saber, que quando manera de defendimiento se comi demandar la muerte segun el suero que todo lo que dize eneste fuero de fazer, e de juzgar despues del di miento, qesso se ha de guarda, e den e juzgar: e no puede mudar la grell.

. He Dowall finds Heasies to take it out of mothers, childrens and sick people than to take on this veret industrie, " the Born commented jelly , a letus have a full inquiry into the cost of closures and the pharmacentical todaying." The health of dildren today owed much to be welfare first scheme. It was maintained during the was. Now in conditions of Tany Offluence it seemed it could not be carried or.

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?

Deleftylo.

Fol.7

querellofo no le deue matar: e fi lo mata, deue ser dado por enemigo de los sus parietes:e pechar el omezillo. Y esto es por razoq no gelo dio elalcalde por enemigo. fegun dize eneltitulo delos omezillos: en la ley que comiença. Si aquelenel. &c. E todo otro hombre que matare su enemi go, maguer que lo aya defafiado con dere demanda en orra manera fino for

aquella palabra, e den lo por fechor que fuero, dar lo an por enemigo delos parien puede ser justificado, segu dicho es Masel tes, e que salga dela villa, e del termino. E fobre esto es a saber, que quando enesta manera de defendimiento se comiença a demandarla muerte segun el fuero viejo: que todo lo que dize eneste suero q se ha de fazer, e de juzgar despues del desafiamiento, qesso se ha de guarda, e demadar, e juzgar: e no puede mudar la grella, nila

aguella palabra, e den lo por fechor que .puede serj u [st]i[si]cado,seq[un] dicho es.Mas el quer.elloso no le deue matar: e si lo mata, deue ser dado por enemigo de los sus pari[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.

How is text editing performed?

Deleftylo.

Fol.7

querelloso no le deue matar: e si lo mata, deue ser dado por enemigo de los sus parietes:e pechar el omezillo. Y esto es por razoq no gelo dio el alcalde por enemigo, segun dize eneltitulo delos omezillos: en la ley que comiença. Si aquelenel. &c. E todo otro hombre que matare su enemi go, maguer que lo ava desafiado con dere demanda en orra manera fino se

aquella palabra, e den lo por fechor que fuero, dar lo an por enemigo delos parien puede ser justificado, segu dicho es Masel tes, e que salga dela villa, e del termino. E fobre esto es a saber, que quando enesta manera de defendimiento se comiença a demandar la muerte segun el suero viejo: que todo lo que dize eneste suero q se ha de fazer, e de juzgar despues del desafiamiento, q esso se ha de guarda, e demadar, e juzgar: e no puede mudar la grella, nila

aquella palabra, e den lo por fechor que .puede serj u [st]i[si]cado,seq[un] dicho es.Mas el quer.elloso no le deue matar: e si lo mata, deue ser dado por enemigo de los sus pari[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.uquer quelo aya desa[si]ado con dere

- Comparing the page with its transcription is very uncomfortable and difficult.
- The human expert must
 - **jum** p 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.

Humidity stains

Las leves e de al rev el querellofo,o el acufado preso es vécido d'adulterio por juyzio das

geloan los al aldes en poder del marido: y que diga que quiere fer oydo e librado por el,fi esto dixereante que el pleyto sea el marido deuclo tener, mas no lo deuc ita g aya el otro y le veça por juy-los mate ambos fi quifiere. contellado ante los alcaldes del lugar: en tonce luvo es de oyr e de librar estas cosas sobredichas, o puede los embiar el rey si xciiij. Que elcriuanos ande quifiere eftos pleytos a los alcaldes do fue te de los prefos fueltos fobre fianças e ron fechosef el fuero delos Ancient chos Pero fi e graphemes gun los fuera de,o de escreuir la fiadus do tales fech rey que elcriue conel al destos en algunos destos fechos de muertos de los sobredichos a te,o de tolimiento de miembro,o de echa escreuir el escrivano de miento de tierra, mas ay otra pena de diconclel alguazil en cafa uerrey. nero, o de al. Entoce tales pleytos, maguer Ley. xcv. Que manera terna vegan por querella ante el regidenen fer calde fiel acufado no viene a respond embiados a que los libren fus alcaldes de las villas do tales tenos acacicieron por la

eros fi querellara al reve esta querella Archaic Orthography? biudas e de language Hyphenation?

Ley xen wae el que na paraga fu injuria o delos fuyos no deue fer recebi do a acufacion, fi no fe obliga ala pena del

querella del camino quebratado, maguer

Lalguno viene diziédo al alcalde, que folano hombre q esay enellugar, que hizo algumalfectio que no arañe a el vale el teltimonio dela muger.

Si se quistere obligar a scularie, cobligar a

C Obre la ley que comieça, toda muger, la pena que el otro deue auer, fi no gelo probare, deuelo oyr el alcalde: mas en orra guita no lodeue oyr: falno fi mostrare carta,o alguna otra cofa que hiziesse alguna fe al alcalde, porque se ouiesse de mouer contra el aculado.

Ley.xciij. Como el marido no pue de matar al vno de los adulteros e dexar

Neltitulo delos adulterios, en la primera ley dize alsi. Si muger cafada fa ze adulterio, ambos fean en el poder del marido:e faga dellos lo q quifiere e de lo q an,aísi que no pueda matar el vno dellos, e dexar al otro. Sobre estas palabras si acae sce que vaya el vno, e prenden al otro; y el llos lugares do es cierto que el fecho fue

Os pleytos delos que estan presos, e delos q fueron entiados ante el alcal Handwritten annotations

Hensacumbers

Trofi, fi alguno acufa a otro que le quemo fus cafas,o quele mato fu pa e, o sobre otra cota desguisada echo:yelalcalde lo fizo empla alos plazos que el fuero manse entonce deue el alcalde (a-

per del techo de que querello, fi fue fecho: mas no a de faber quien lo fizo. E fi fallare que tal fecho es fecho, entoce lo deue dar porfechor.

Ley.xcvj. En que casos, e quando

que es en el titulo delos testimonios: es a faber, que pueden las mugeres fer refeebidas en testimonio sobre las cosas, quier lea ceuiles, quier criminales q fe fazen en tal lugar, q no es razo ni aguifado de fer,y hobrescolas mugaeras E serescibé las m Handwritten védidas, v en l zer las mugere underlines leficios q acae uese por su dicho de mugeres en restimonio. E otrofi en la pesquisa que la faze de los yerros fechos de noche en yern fi ellas dan testimenio de vista, juzgen lo por prueua. E otroli faze los fus dichos prefum pcion para poder tormentar, mas en aque

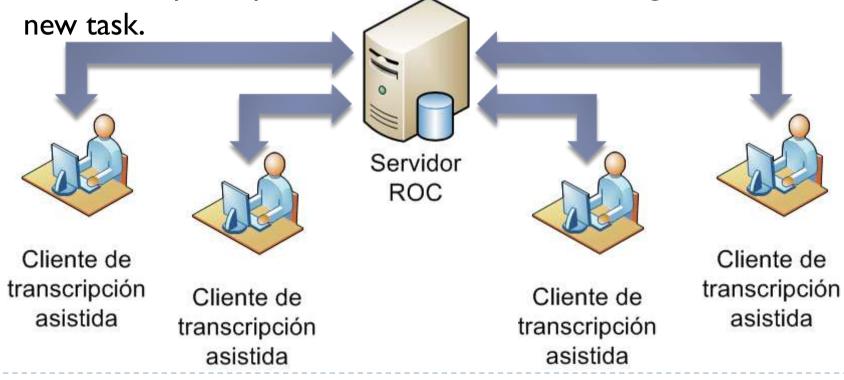
Our aims

- 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.

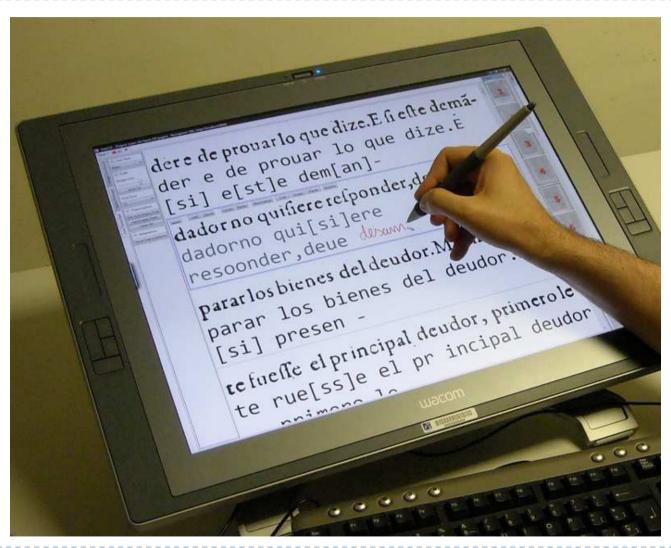
Adaptive learning recognition system

We have designed a **client/server** architecture: what the system learns from a user, all users benefit inmediately.

New glyphs, lexicon entries or LMs can be added during the transcripcion process to adapt baseline recognizers to a



STATE: An Assisted Text Transcription System



No. 10 Avignon 2010/12/06

STATE: An Assisted Text Transcription System

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

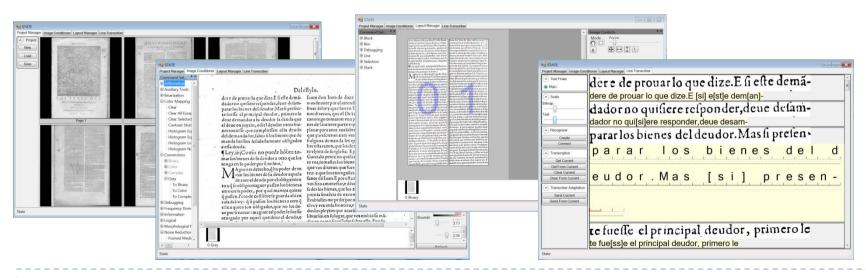
- A text Recognition Engine
 - ▶ For ancient printed documents
 - ▶ For handwritten documents

StateRE_NN
StateRE_HMM/ANN

State TA: State Transcription Assistant

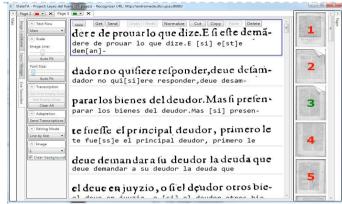
▶ 4 components:

- 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 ...



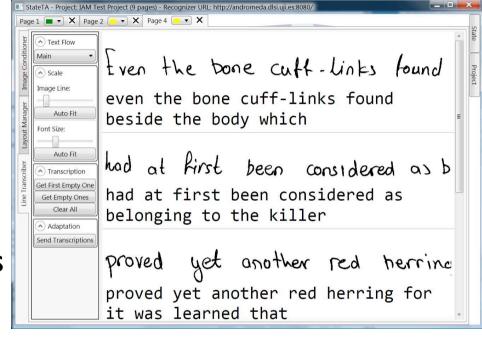
- ...and an application manager to
 - **Enlarge the corpus** to learn.
 - **Re-Train** the recognition engines. Pending lines Processed lines Corpus Recognizer About (A) File בת בכו בכו בכו בכו בנו ביו ביו בו בכון בכו בכו בכו בכו בכו בכו בכו בכו בכו dadorno quisiere responder, deus pararlos bienes del deudor. Mas l essed lines Corpus Recognizer About te fuesse el principal deudor, pr aaaaaaaaaaaaaaa pararlos bienes del deudor. Mas fi PPPPPPPPPPPP

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 prouar lo que dize. E si este demadador no quisiere responder, deue desamparar los bienes del deudor. Mas si presente suesse 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 cumpliessen al su deudo del demandador, salvo si los bienes que de manda suessen seña deuda.

Ley.iiijComo no puede hobre tomar los bienes de su deudora otro que los tenga en su poder por si mismo.!

Aguer es derecho, q ha poder de to mar los bienes de su deudor aquela de auer el deudo por el obligamien to a q se obligo: maguer passen los bienes a otro en su poder, por qual manera quiere q passen. Pero de costúbre se guarda assien casa delrev: a sir passan los bienes a otro q

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

No. 16 Avignon 2010/12/06

System evaluation

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

He said shere concerned Mr. Weavers alleged a mocistion with organisations blacklished by the Jovernement . Instructionally Mr. Kennedy pushed a letter to Lenator lobertron saying the Federal Bureau of Investigation had reported on Mr. Meaver.

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

it's Dowell finds It easies to take it out of mothers, childrens and cick people than to take on this sent inclustrie, " tir. Born commented icity. " let us have a full improve into the cost of drugs and the pharmacentical industry." The health of children today or not much to the welfare find scheme. It was a main tained during the was. Now in conclitions of Tory offluence it seemed it could not be carried or.

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

No. 18 Avignon 2010/12/06

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.

No. 19 Avignon 2010/12/06

STATE in use



Welcome

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

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.

13 R. La Baldi-1189- 52 de

19/11/2010 🖂 🌁 - Archives and Library / The Montserrat Tarradellas i Macià Archive / Documentary resources Català Documentary resources Castellano 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

STATE in use



ARXIU VIRTUAL JAUME I

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

1237, juliol 9. El Puig

Jaume I dóna 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

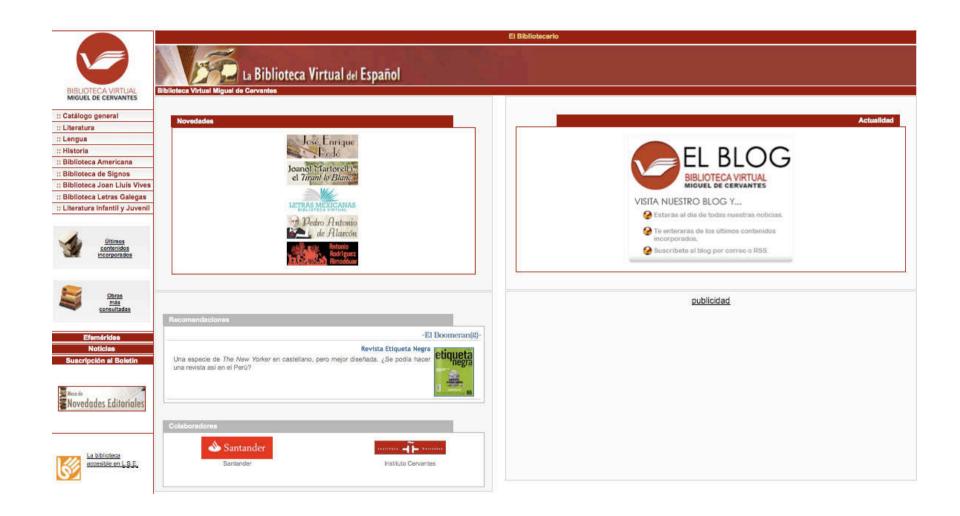
[castellano]

Pàgina principal
Presentació
Notícies
Comentaris a documents
Obres de referència
Documents extensos



Contactar

STATE in use



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
```

> 23 Avignon 2010/12/06

Perspectives and future work

Inmediate 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!!!

