Preparació de corpus i desenvolupament de prototips en reconeixement de text manuscrit

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Tesi de Master
realitzada per Dani Pérez i Cardona
dirigida per Dr. Alfons Juan i Císcar i Dr. Moisès Pastor i Gadea

30 de novembre de 2009

Agraïments

Work supported by the EC (FEDER/FSE) and the Spanish MCE/MICINN under the MIPRCV "Consolider Ingenio 2010" programme (CSD2007-00018), the iTransDoc project (TIN2006-15694-CO2-01) and the FPU scholarship AP2007-02867.

Índex

Capítol 1

Introducció

Hi ha enormes col·leccions de documents històrics en les biblioteques, museus i arxius que a sovint estan siguent digitalitzats per propòsits de preservació i per a fer-los disponibles al llarg de tot el mon mitjançant llibreries digitals on-line. L'objectiu principal, en canvi, no és simplement el proporcionar accés a les imatges crues dels documents digitalitzats, sinó anotar-lo amb contingut informatiu real i, en particular, amb transcripcions de text i, si és convinent, amb les traduccions també. Aquest treball pretén contribuir en el desenvolupament de tècniques avançades i interfícies per a l'anàlisi, transcripció i traducció d'imatges de documents d'arxius antics, seguint una aproximació interactiva-predictiva. La nostra hipòtesi es que aquest objectiu no es pot acomplir de manera completament fidel utilitzant tècniques automàtiques; en canvi, es pot seguir un model de col·laboració persona-màquina per a produir interpretacions de documents precisos amb d'una manera efectiva. Per a mostrar aquesta hipòtesi, s'ha desenvolupat i avaluat una eina de sofre. Noti's que el treball aportat ací s'ha dut a terme dins del marc de treball del projecte de recerca Espanyol "Interactiva Transcriptor andà Translatiu F Ols Text Documents (iTransDoc)" [13].

Més específicament, les contribucions descrites en aquest treball són les següents:

- **GERMANA**: es presenta la base de dades de GERMANA i alguns resultats preliminars. La transcripció i digitalització de documents antics és molt important per a la recerca en detecció de blocs i de línies i en reconeixement de text manuscrit. Així doncs, en aquest treball hem col·laborat en la preparació de la base de dades GERMANA que es un manuscrit Espanyol de 764 pàgines i data de l'any 1891 i la majoria de pàgines només contenen text cal·ligrafiat escrit en pàgines pautades amb línies ben separades. Està escrit completament en Castellà fins la pàgina 180. En canvi, altres parts contenen escriptira en altres llengües com Català, Llatí, Francès, etc. GERMANA ha sigut publicada i està disponible on-line. Està decrit en un artícle en una conferència internacional:
  - **ICDAR-2009**: D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos and A.

Juan. The GERMANA database. *Proceedings of the 10th ICDAR*. Barcelona (Spain). July 2009.

- **GIDOC: Gimp-based Interactive transcription of old text DOCuments**. Aquest prototip s'ha desenvolupat per a proporcionar a l'usuari suport integrat per a l'anàlisi d'estructura de pàgina (layout analysis) de manera interactiva-predictiva, detecció de línies i transcripció de text. Aquest treball s'ha publicat en els següents articles:

  - **WEBIST-2010**: N. Serrano, L. Tarazón, D. Pérez, O. Ramos-Terrades and A. Juan. The GiDOC Prototype (submitted). *Proceeding of WEBIST 2010*. Valencia (Spain). April 2010.

- **Mesures de confiança en RTM i Adaptació**: respecte les mesures de confiança, s'explica com usar aquestes per a detectar i corregir errors durant la transcripció interactiva de text manuscrit. D'altra banda, respecte a l'adaptació de models, es proposa una aproximació interactiva-predictiva on les transcripcions produïes successivament, es van reutilitzant per a re-entrenar els models. Aquests treballs s'han publicat respectivament en els següents dos artícles en conferències internacionals:

  - **ICIAP-2009**: L. Tarazón, D. Pérez, N. Serrano, V. Alabau, O. Ramos Terrades, A. Sanchis and A. Juan. Conficence Measures for Error Correction in Interactive Transcription of Handwritten Text. *Proceedings of the 15th ICIAP*. Vietri sul Mare (Italy). September 2009.
  - **ICMI-MLMI-2009**: N. Serrano, D. Pérez, A. Sanchís and A. Juan. Adaptation from Partially Supervised Handwritten Text Transcriptions. *In Proceedings of the Intelignet User Interface 2010*. Hong-Kong, China. February 2010.

Noti's que les contribucions descrites dalt són el resultat d'un treball en col·laboració amb altres autors i, en particular, amb autors que també presenten la seva Tesis de Màster per a l'especialitat "d'Intel·ligència Artificial, Reconeixement de Formes i Imatge Digital". No obstant, comparant el treball aportat ací i aquell aportat en les altres Tesis de Màster, l'autor d'aquesta Tesi es deuria considerar l'autor líder del treball aportat en els articles indicats dalt com ICDAR-2009 i WEBIST-2010. D¡acord amb açò, el treball realitzat en aquests articles es descriu detalladament en les Capítols 2 (ICDAR-2009) i 3 (WEBIST-2010). D'altra banda, el treball realitzat en la resta dels articles es descriu de manera breu en els Capítols 4.1 (ICIAP-2009) i 4.2 (ICML-MLMI 2009). Per a una informació més detallada d'aquestes parts, es referència a les Tesis de Màster de L. Tarazón [3] i N. Serrano [23].

Noti's que el prototip de GIDOC bàsic ha estat desenvolupat per l'autor d'aquesta Tesi junt amb N. Serrano i L. Tarazón, amb el mateix nivell d'esforç i dedicació i, així, tots aquests tres autors podrien ser considerats com els autors líders de l'article WEBIST-2010.

Per a més claredat, en la Taula 1.1 vegem de manera resumida la correspondencia entre els capitols i els artícles junt amb el seu corresponent indicador de qualitat de conferència (CORE rank).

Article	Estat	Indicador de qualitat	Contribució	Capítol
ICDAR-2009	Publicat	A	Autor líder	2
ICIAP-2009	Publicat	A	Autor líder	4.1
ICML-MLMI 2009	Publicat	B	Co-autor	4.2
WEBIST-2010	Presentat	C	Co-autor	3

Taula 1.1: Articles obtinguts a partir del treball realitzat i descrit a aquest document marcant la puntuació del congrés “Indicador de qualitat” i el capítol de la tesis on s’explica.

Capítol 2

La base de dades GERMANA

2.1 Introducció

Hi ha col·leccions de documents històrics enormes en biblioteques, museus i arxius que a sovint estan sent digitalitzats per propòsits de conservació i per a fer que estiguen disponibles en tot el mon mitjançant llibreries digitals online. L'objectiu principal, en canvi, no és simplement proporcionar l'accés a les imatges crues dels documents digitalitzats, sinó anotar-les amb el seu vertader contingut informatiu i, en particular, amb transcripcions de text.

Lamentablement, l'extracció de línies o segmentació i el reconeixement de text manuscrit encara és un problema d'investigació obert [5, 29].

En aquest paper, presentem una base de dades de text manuscrit, GERMANA, per a facilitar comparacions empíriques entre diferents aproximacions d'extracció de línies de text i reconeixement de text manuscrit off-line.

GERMANA és el resultat de digitalitzar i anotar un manuscrit Espanyol de 764 pàgines titulat "Noticias y documentos relativos a Doña Germana de Foix, última Reina de Aragó" i escrit en 1891 per Vicent Salvador, el marqués de Cruïlles. Aquest té aproximadament 21K línies de text manualment marcades i transcrites per experts paleogràfics.

Germana no és una tasca particularment dificultosa per diferents raons. Primerament, el llibre ha sigut escrit per un sol autor en un domini limitat: la vida de Germana de Foix (1488-1538), neboda del Rei Lluís XII de França i segona esposa de Ferran el Catòlic d'Aragó. A més, el manuscrit original està ben conservat i la majoria de pàgines només contenen text cal·ligrafiat en pàgines pautades amb línies ben separades. També cal dir que, el manuscrit comprèn sobre 217K paraules (running words) amb un vocabulari de 30K paraules les quals, aparentment, són una quantitat raonable de paraules manuscrites per a un sol autor i per al modelat de llenguatge.

D'altra banda, cal dir que l'extracció de línies de text i reconeixement de text manuscrit off-line en GERMANA no és una tasca particularment fàcil. GERMANA té caracte-

rístiques típiques de documents històrics que dificulten l'objectiu com: taques al paper, escriptura de la cara anterior de la pàgina que es transparenta en l'altra cara, caràcters i paraules inusuals, etc. A més, el manuscrit inclou varies notes i documents afegits que estan escrits en llenguatges diferents al Castellà com el Català, Francès, Llatí, etc.

Per tot açò, pensem que GERMANA suposa una compensació apropiada entre complexitat i quantitat de dades. Que nosaltres sapiguem, aquesta és la primera base de dades pública disponible per a la recerca de text manuscrit, majoritàriament escrit en Castellà i comparable en grandària a altres bases de dades estàndards com IAM [34, 27]. Degut a l'estructura seqüencial del llibre, aquest es també apropiat per a avaluacions realistes de sistemes de reconeixement de text manuscrit interactius [32]. A més a més, també es pot utilitzar per a avaluar aproximacions per a identificar llenguatges i adaptació de text manuscrit d'un únic autor.

A continuació, primerament descriurem el manuscrit i la base de dades al Capítol 2.2 i 2.3, respectivament. Després, en el Capítol 2.4, es presentaran alguns resultats preliminars utilitzant un reconeixedor basat en Models Ocults de Markov estàndard. Finalment, es discutiran les conclusions i el treball futur referents a aquest apartat al Capítol 2.5.

2.2 El manuscrit

Tal i com s'ha dit en la introducció, GERMANA és el resultat de digitalitzar i anotar un manuscrit Espanyol de l'any 1891 sobre la vida de Germana de Foix. El manuscrit original es conserva en la "Nicolau Primitiu Collectio" a la llibreria Valenciana [1]. Aquest està composat per 764 pàgines de volum les quals, d'acord amb el seu índex que es troba a la pàgina 728, es divideix en 17 seccions.

Per a simplificar, nosaltres distingirem sols 7 parts del manuscrit:

1. Part inicial (pp 1-6): un subtítol, un títol i el retrat de Doña Germana de Foix.

2. Els capítols (pp 7-180): 174 pàgines dividides en 6 capítols, cadascun dels quals està basat en un període distint de la vida de Germana.

3. Notes (pp 181-282): 290 notes enumerades referenciades en els diferents capítols.

4. Notes biogràfiques (pp 283-302): de 8 persones rellevants mencionades en la segona part.

5. Documents (pp 303-540): copies de text manuscrit de 71 documents històrics relacionats en la vida de Germana.

6. Il·lustracions (pp 541-716): 4 documents amb el seu propi peu d'imatge al final.

7. Part final (pp 717-764): varis índex i imatges.

La majoria de pàgines només contenen text manuscrit alineat amb un pautat horitzontal en una plantilla simple de 24 línies (pp 1-180 and 729-764) o 32 línies (pp 181-728). Com a exemple dels dos tipus de plantilles més usuals que hi ha al GERMANA mostrem la pàgina 66 de 24 línies en la Figura 2.1 i la pàgina 335 de 32 línies en la Figura 2.2. Vegem com l'escriptura manuscrita és fàcilment llegible i alineat al pautat horitzontal de manera precisa.

Mostrem també altre tipus de pàgines del Germana en la Figura 2.3 on es poden veure títols, notes, pàgines amb il·lustracions, etc.

Fins la pàgina 180, el manuscrit està escrit únicament en llengua Castellana. A partir d'aquesta, en canvi, el lector pot trobar també text en Català, Francès, Llatí i, amb menor grau, en Alemany i Italià. En la tercera part, hi ha 33 notes la majoria d'elles escrites en Català (4, 47, 50, 73, 78, 79, 81, 82, 84, 85, 87-91, 94-96, 134,177,194, 205, 209, 214, 227, 229, 236, 238, 261, 266-268 i 270); 18 en Francès (1, 2, 15, 22, 23, 25, 29, 44-46, 71, 109, 110, 119, 155, 170, 257 i 280); i 1 en Alemany (180). També, hi ha 24 documents en la quinta part que estan escrits en Català (7, 8, 27, 29, 31-33, 36-40, 44, 48-54, 59, 64, 68 i 69); 10 en Llatí (2, 4-6, 12, 24, 34, 42, 43, 70); 1 en Francès (7); 1 en Alemany (25); i 1 en Italià (65). La biografia, les notes i les il·lustracions estan escrites en Castellà primerament, encara que hi ha també algun contingut en Català (un breu passatge de 13 línies començant a la última línia de la pàgina 300; les notes 39, 47 i 61 de la il·lustració C; i la nota 17 de la il·lustració D).

Al lector interessat es remet a [4] per a un estudi més profund del manuscrit des del punt de vista d'un historiador.

2.3 La base de dades

El manuscrit va ser cuidadosament escanejat per experts de la Biblioteca Valenciana a 300dpi en color real. Com es fa amb documents històrics en general, les pàgines escanejades tenen efecte soroll com punts, taques d'aigua, decoloració de la tinta i transparència del costat de darrere. El document també ens mostra un lleu efecte "warping" degut al enquadernat del llibre. En canvi, el manuscrit es pot llegir fàcilment i per això decidírem no aplicar cap preprocessat per a corregir-ho a l'hora de marcar les línies de text.

L'anotació de les línies de text de GERMANA consisteix en dos parts. Primerament, tots els blocs de text foren marcats amb rectangles de mínima inclusió i, dins de cadascun d'aquests blocs de text, cada línia de text fou marcada amb línies base (rectes). Açò es va fer de manera semiautomàtica mitjançant l'ajuda del programa GNU Image Manipulation Program (GIMP) [2] i alguns plugins de GIMP que desenvoluparem específicament per a la detecció de blocs i línies del GERMANA. Tots els blocs i línies detectats automàticament foren també supervisats manualment i corregits en cas de ser necessari.

Per a detectar les línies es van implementar dos plugins diferents. El primer d'ells (Figura 2.4), útil només per al GERMANA o corpus on estiga marcat el pautat de les línies on s'escriu el text, tracta de trobar aquest pautat per a posteriorment tractar d'ajustar cada línia buscant l'inici i final del text. L'altre plugin de detecció de línies (Figura 2.5), tracta de buscar les línies mitjançant l'ús d'histogrames en els quals es basa en les projeccions horitzontals i verticals per a detectar-les [15]. Aquestes projeccions es poden estimar, per a cada línia de píxels, amb el conteig del nombre de píxels en "negre" o amb el conteig del nombre de transicions "negre/blanc" de la línia. Açò es pot indicar a les opcions de preferències. Evidentment, per a considerar un píxel blanc o negre, anteriorment s'ha calculat un llindar mitjançant la tècnica *d'Otsu* el que ens permet binaritzar una imatge (passar la imatge a blanc i negre).

D'altra banda, el manuscrit complet fou transcrit línia per línia per experts paleogràfics. El procés de transcripció no va començar des de zero, sinó des d'una transcripció parcial produïda per experts de la Biblioteca de València durant l'any 2002. Aquesta transcripció parcial cobria la majoria del manuscrit (76%), però aquesta no fou directament aplicada per a recerca de text manuscrit, principalment perquè aquest no incloïa ni els salts de pàgina ni els salts de línia originals. Així doncs, per a produir la transcripció final, aquesta versió parcial fou revisada primerament i després completada. Açò es va fer més recentment, concretament durant l'any 2007. Es va fer una altra vegada per experts paleògrafs d'acord amb les següents regles de transcripció:

- Els salts de pàgina i de línia es copien exactament igual.
- L'espai en blanc només s'utilitza per a separar paraules.
- No es corregeixen els errors d'ortografia.
- No se fa cap canvi d'accentuació.
- Els signes de puntuació es copien exactament com apareixen al text.
- Les abreviatures es copien, en un principi, textuals, excepte els subíndex i superíndex, els quals s'escriuen amb anotació LATEX com _{sub} i ∧{super}, respectivament. A continuació, s'escriu la paraula corresponent entre claus. Així, per exemple, D^a. es transcriu per D∧{a}. [Doña].

També, per a facilitar un procés dependent del llenguatge del manuscrit, cada línia transcrita va ser manualment etiquetada d'acord amb la llengua predominant d'aquesta. El temps total que un expert requereix per a transcriure a ma tot el manuscrit sencer fou estimat en 232 hores; açò és aproximadament una mitja de 30 minuts per pàgina.

La Taula 2.1 conté algunes estadístiques bàsiques extretes de les nostres transcripcions del GERMANA. Aquestes estadístiques foren executades després d'aplicar els següents passos de preprocessat:

1. Substitució de les abreviatures per les seves corresponents paraules.
2. Concatenació de les paraules de final de línia que acaben amb guió amb la resta de paraula.
3. Aïllar els signes de puntuació.

Llenguatge	Pàgines	Línies	Paraules (K)	Lèxic Talla (K)	Lèxic 1-ocurr (%)	Conjunt Caràcters
Castellà	595	16599	176.8	19.9	55.6	111
Català	87	2417	26.9	4.6	63.2	86
Llatí	29	951	8.3	3.4	69.2	87
Francès	8	266	3.0	1.1	71.1	82
Alemany	8	228	1.5	0.6	52.7	71
Italià	2	68	0.8	0.3	67.3	59
Cap	35	0	0.0	0.0	0.0	0
Tots	764	20529	217.2	27.1	57.4	115

Taula 2.1: Estadístiques bàsiques del GERMANA (1-ocurr=1 ocurrència, paraules que apareixen 1 vegada al text)

S'observa que la part escrita en Castellà del GERMANA conté al voltant de 17K línies de text i 177K paraules (running words) amb un lèxic de 20K paraules, el qual es comparable en talla a una base de dades estàndard com és l'IAM [34, 27]. Una dada significant que també s'observa és que el 56% de les paraules només apareixen una vegada al text (1-ocurrencia). En quan a les altres parts no escrites en Castellà, està clar que no són el suficientment grans per a estimar, de manera fiable, models per a elles (HMMs i models de llenguatge d'n-grames). En canvi, seria molt interessant veure com models entrenats amb diferents dades es poden adaptar a aquestes. En particular, Models Ocults de Markov de caràcter entrenats amb la part Castellana, podria ser ben bé reutilitzada sense fer canvis significatius.

La base de dades està disponible al portal web del PRHLT (prhlt.iti.es) per a recerca de propòsit no comercial. A més, es pot trobar una transcripció impresa del manuscrit en [4] encara que, com aquest no estava previst per a la recerca de text manuscrit, se li va aplicar un format per a que fora més llegible.

2.4 Experiments

Com ja s'ha dit en la introducció, GERMANA és podria utilitzar tant per a avaluar mètodes d'extracció de línies de text com per a avaluar tècniques de reconeixement de text

manuscrit off-line. Especificant més, el nostre objectiu és simplement proporcionar resultats de referència (baseline) per a estudis futurs, utilitzant tècniques i eines estàndards; per exemple, modelat d'imatges basats en HMM i models de llenguatge d'n-grames [32].

Deguda a l'estructura seqüencial del llibre, la tasca bàsica del GERMANA és transcriure'l línia per línia, des del principi fins al final. Assumim que s'utilitza un sistema de transcripció automàtica i que, cada transcripció automàticament obtinguda de cadascuna de les línies, és supervisada i, en cas de ser necessari, corregida per un expert. Clarament, després de processar un bloc de línies o pàgina, totes les transcripcions supervisades podrien ben be ser usades per re-entrenar el sistema de transcripció automàtic. Açò podria ajudar a millorar la precisió del sistema, almenys en la transcripció de les primeres pàgines del GERMANA. Afortunadament, les primeres dues parts del GERMANA estan escrites solament en Castellà i així, almenys, la carència de dades d'entrenament no es combina amb la dificultat de tindre text en múltiples llengües. En la Figura 2.6 s'observa com va incrementant-se el nombre de línies de cada llengua al llarg de les pàgines del GERMANA el que pot ser indicatiu, en part, de la dificultat del procés de RTM.

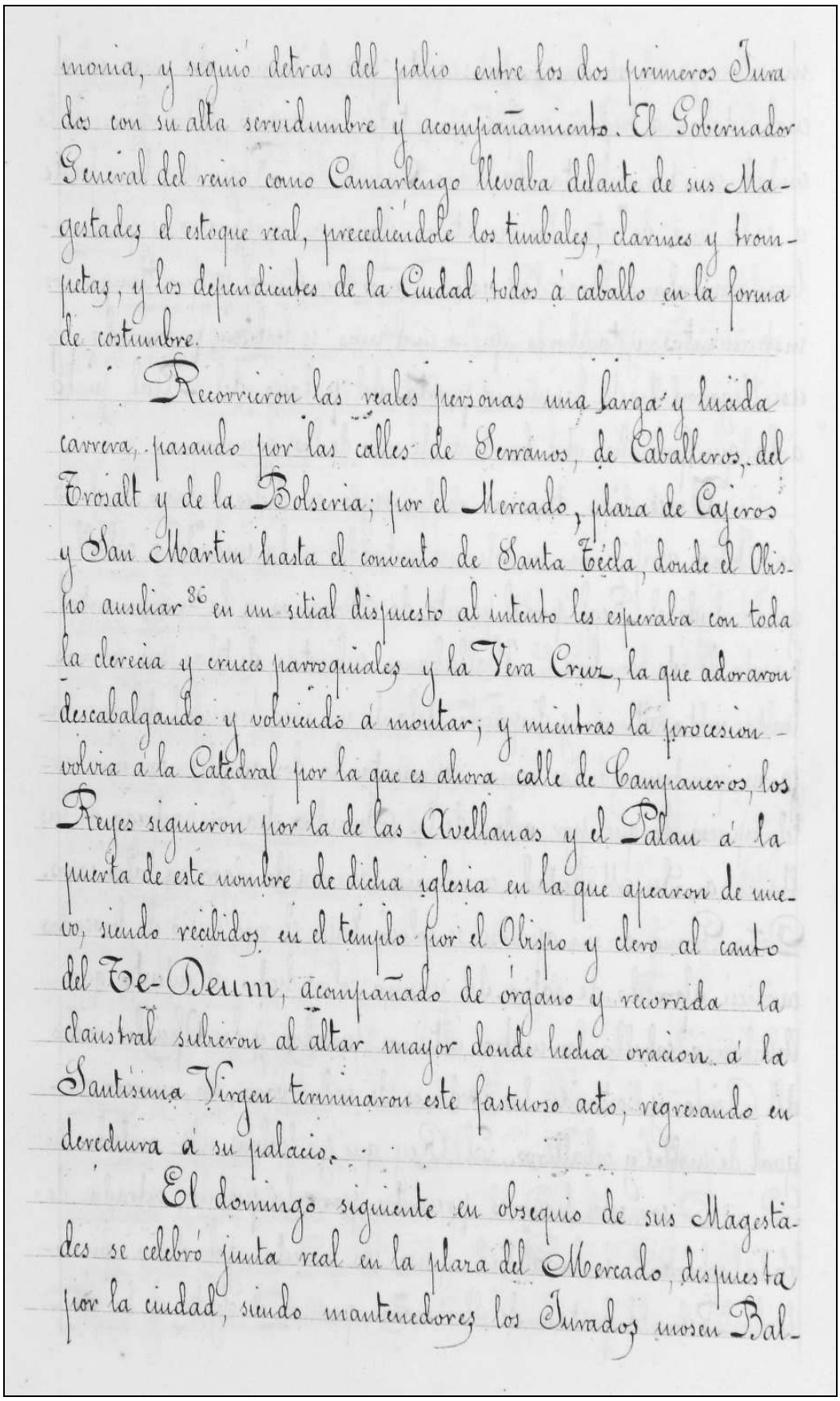
Tenint en compte el que hem dit anteriorment, decidírem provar només amb les transcripcions del GERMANA de les primeres dues parts, és a dir, fins la pàgina 180. Començant des de la pàgina 3, dividírem GERMANA amb 9 blocs consecutius de 20 pàgines cadascuna (3-22, 23-42, ..., 163-180). Després, a partir del bloc 2 i fins al 9, cada bloc va ser automàticament transcrit pel sistema entrenat amb els blocs precedents. Com ja hem apuntat, vàrem utilitzar tècniques i eines per a preprocessat , extracció de característiques, modelat d'imatges basat en HMM, i modelat de llenguatge [32]. Els resultats es mostren en la Figura 2.7, en termes de percentatge d'error de paraules (WER) per bloc. Tal i com s'esperava, el WER va disminuint a mesura que la quantitat de dades per a entrenar els HMMs augmenta. En concret, el sistema aconsegueix al voltant del 37% de WER per als 2 últims blocs, el qual no està malament per a transcripcions assistides per ordinador efectives. Encara que pensem que podria ser millor, s'ha d'observar que la majoria d'errors succeeixen com a conseqüència de no estar la paraula en el vocabulari (out-of-vocabulary, OOV). Açò es pot observar en la Figura 2.7, on una corba traçada mostra la part de WER degut a paraules fora de vocabulari. Observem que el 54% d'error per al primer bloc transcrit correspon a paraules fora de vocabulari, mentre que per a l'últim bloc, la taxa d'error obtingut degut a paraules fora de vocabulari correspon al 64% del WER total. A més, açò podria augmentar inclús més en les parts restants del GERMANA degut a la seva naturalesa multilingüe.

2.5 Conclusió i treball futur

S'ha presentat una nova base de dates de text manuscrit, GERMANA, per a facilitar la comparació empírica entre diferents aproximacions d'extracció de línies de text i reconeixement de text manuscrit off-line. Dintre del nostre coneixement, aquesta és la primera

base de dades pública per a la recerca de text manuscrit, escrita majoritàriament en Castellà i comprable en grandària a altres bases de dades estàndard. S'han presentat alguns resultats empírics preliminars, utilitzant tècniques estàndards i eines per al preprocessat, extracció de característiques, modelat d'imatge basat en HMM i modelat de llenguatge. Encara que pensem que el resultat és millorable, el WER obtingut és acceptable per a usar en sistemes de traducció assistides per ordinador.

Actualment estem completant l'experimentació preliminar aportada ací, que correspon amb les transcripcions completes del GERMANA, el qual implica identificació de llenguatge i adaptació degut a la naturalesa multilingüe del GERMANA.

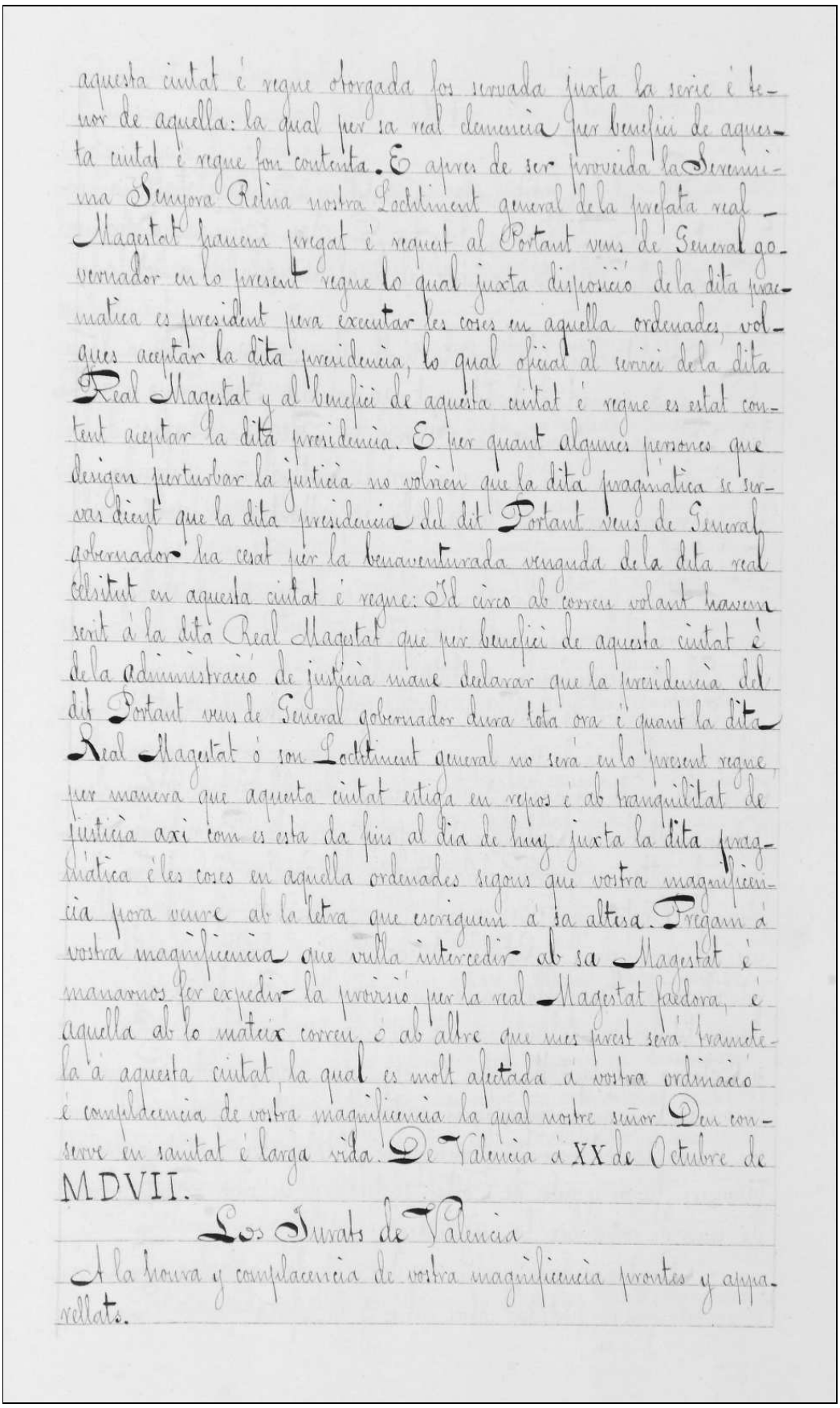


monia, y siguio detras del palio entre los dos primeros Jura dos con su alta servidumbre y acompañamiento. El Gobernador Geniral del reino como Camarlengo llevaba delante de sus Magestades el estoque real, precediendole los timbales, clarines y trompetas, y los dependientes de la Ciudad todos á caballo en la forma de costumbre.

Recorrieron las reales personas una larga y lucida carrera, pasando por las calles de Serranos, de Caballeros, del Trosalt y de la Bolseria; por el Mercado, plaza de Cajeros y San Martin hasta el convento de Santa Tecla, donde el Obispo ausiliar[86] en un sitial dispuesto al intento les esperaba con toda la clerecia y cruces parroquiales y la Vera Cruz, la que adoraron descabalgando y volviendo á montar; y mientras la procesion volvia á la Catedral por la que es ahora calle de Campaneros, los Reyes siguieron por la de las Avellanas y el Palau á la puerta de este nombre de dicha iglesia en la que apearon de nuevo, siendo recibidos en el templo por el Obispo y clero al canto del Te-Deum, acompañado de organo y recorrida la claustral subieron al altar mayor donde hecha oracion á la Santisima Virgen terminaron este fastuoso acto, regresando en derechura á su palacio.

El domingo siguiente en obsequio de sus Magestades se celebró junta real en la plaza del Mercado, dispuesta por la ciudad, siendo mantenedores los Jurados mosen Bal-

Figura 2.1: Pàgina 66 del GERMANA amb 24 línies de text.



aquesta ciutat è regne otorgada los servada juxta la serie è te-
nor de aquella: la qual per sa real demencia per benefici de aques-
ta ciutat è regne fon contenta. E apres de ser provehida la Serenissi-
ma Senyora Reina nostra Lochtinent general de la prefata real
Magestat fonsem pregat è requert al Portant veus de General go-
vernador en lo present regne lo qual juxta disposicio de la dita prag-
matica es president pera executar les coses en aquella ordenades, vol-
gues acceptar la dita presidencia, lo qual oficial al servici de la dita
Real Magestat y al benefici de aquesta ciutat è regne es estat con-
tent acceptar la dita presidencia. E per quant algunes persones que
desigen perturbar la justicia no volrien que la dita pragmatica se ser-
vas dient que la dita presidencia del dit Portant veus de General
gobernador ha cesat per la benaventurada vinguda de la dita real
Esbitut en aquesta ciutat è regne: Hi circo al correu volant havem
scrit à la dita Real Magestat que per benefici de aquesta ciutat è
de la administracio de justicia mane declarar que la presidencia del
dit Portant veus de General gobernador dura tota ora è quant la dita
Real Magestat ò son Lochtinent general no sera en lo present regne
per manera que aquesta ciutat estiga en repos è ab tranquilitat de
justicia axi com es esta da fins al dia de huy juxta la dita prag-
matica è les coses en aquella ordenades segons que vostra magnifecen-
cia pora veure ab la letra que escriguem à sa altesa. Pregam à
vostra magnificencia que vulla intercedir ab sa Magestat è
manarnos per expedir la provisio per la real Magestat faedora è
aquella ab lo matex correu, è ab altre que mes prest sera trameto-
la à aquesta ciutat la qual es molt afectada à vostra ordinacio
è complidencia de vostra magnificencia la qual nostre senyor Deu con-
serve en sanitat è larga vida. De Valencia à XX de Octubre de
MDVII.

Los Jurats de Valencia

A la honra y complacencia de vostra magnificencia prontes y appa-
rellats.

Figura 2.2: Pàgina 335 del GERMANA amb 32 línies de text.

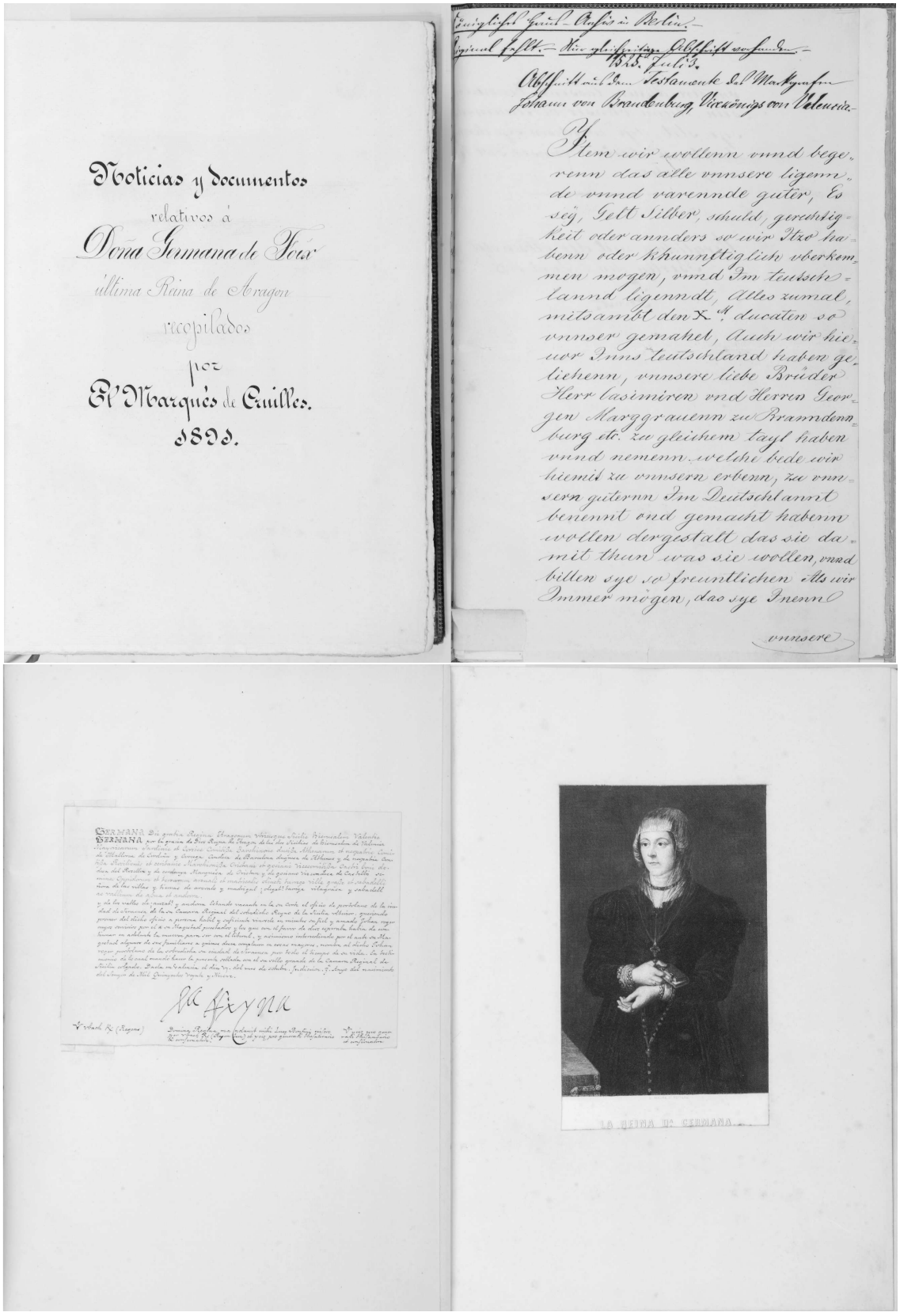
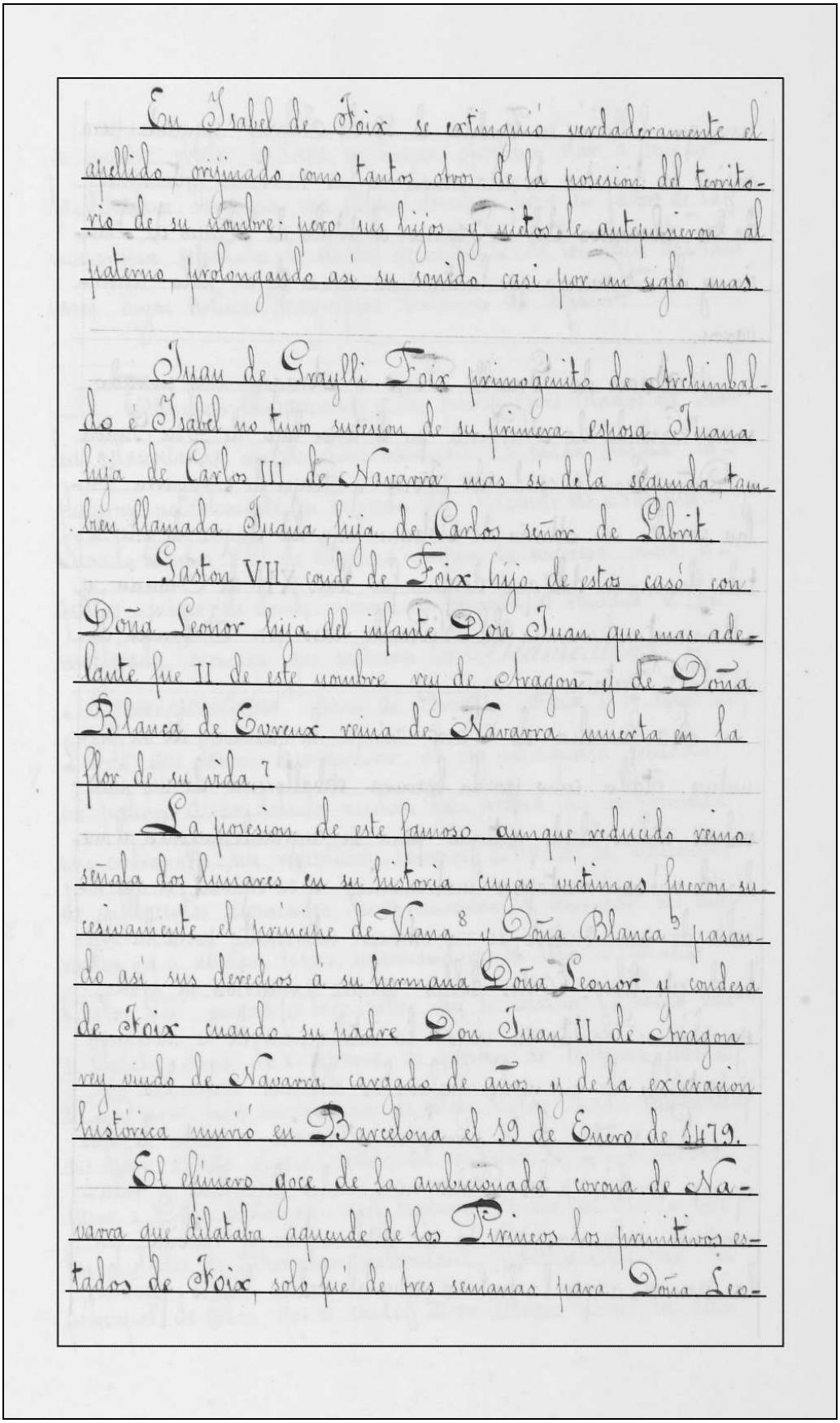


Figura 2.3: Altres tipus de pàgines.



En Isabel de Foix se extinguió verdaderamente el apellido originado como tantos otros de la posesion del territorio de su nombre: pero sus hijos y sucesores le antepusieron al paterno prolongando así su sonido casi por un siglo mas.

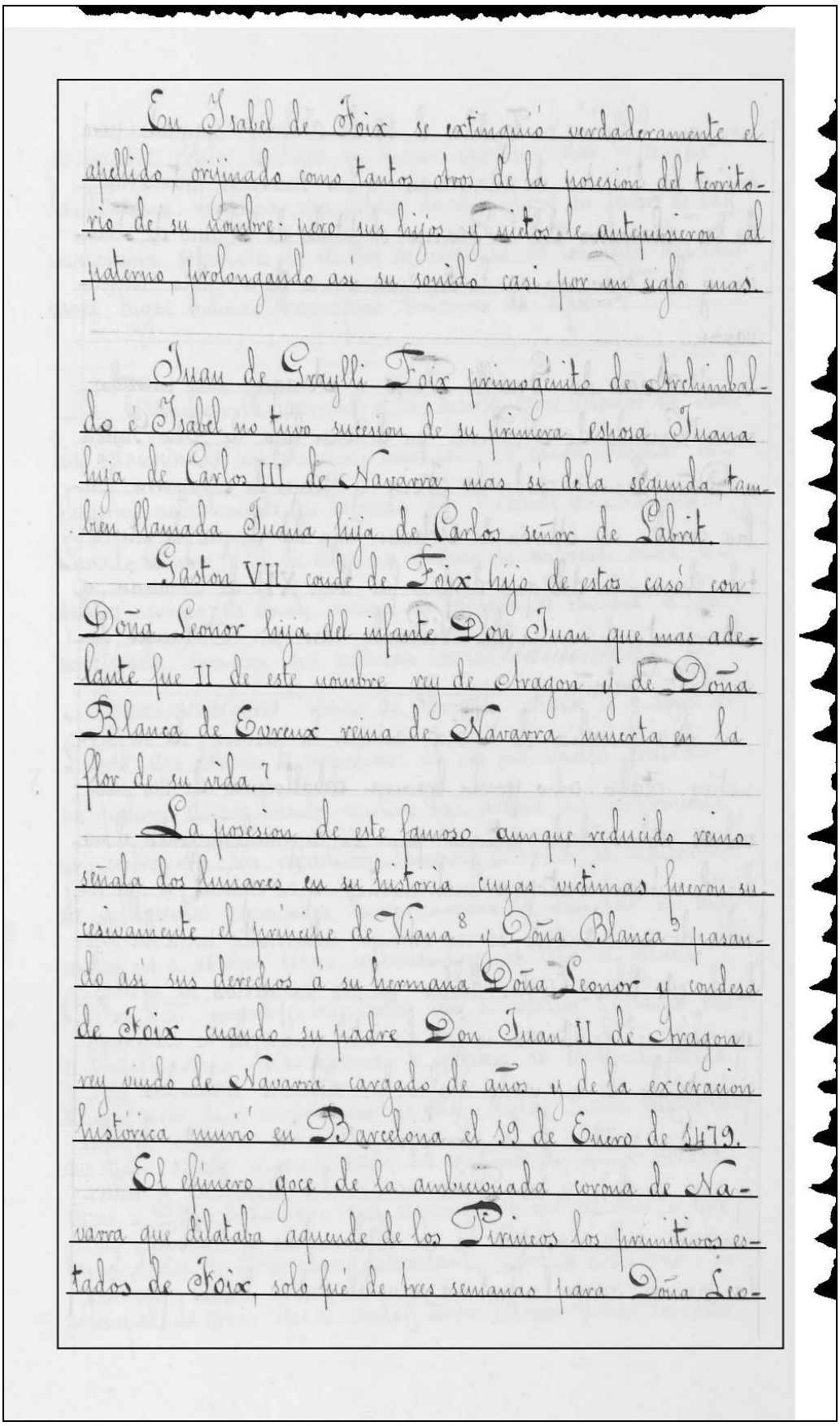
Juan de Grailli Foix primogénito de Archimbaldo é Isabel no tuvo sucesion de su primera esposa Juana hija de Carlos III de Navarra, mas sí de la segunda tambien llamada Juana hija de Carlos señor de Labrit.

Gaston VII conde de Foix hijo de estos casó con Doña Leonor hija del infante Don Juan que mas adelante fue II de este nombre rey de Aragon y de Doña Blanca de Evreux reina de Navarra muerta en la flor de su vida.[7]

La posesion de este famoso aunque reducido reino señala dos lunares en su historia cuyas víctimas fueron sucesivamente el príncipe de Viana[8] y Doña Blanca[9] pasando así sus derechos á su hermana Doña Leonor y condesa de Foix cuando su padre Don Juan II de Aragon rey viudo de Navarra cargado de años y de la execracion histórica murió en Barcelona el 19 de Enero de 1479.

El efímero goce de la ambicionada corona de Navarra que dilataba aquende de los Pirineos los primitivos estados de Foix, solo fué de tres semanas para Doña Leo-

Figura 2.4: Mètode de detecció de línies utilitzant el pautat de les pàgines del GERMANA.



En Isabel de Foix se extinguió verdaderamente el apellido, originado como tantos otros de la posesion del territorio de su nombre; pero sus hijos y sucesores le antepusieron al paterno, prolongando así su sonido casi por un siglo mas.

Juan de Grailli Foix primogénito de Archimbaldo é Isabel no tuvo sucesion de su primera esposa Juana hija de Carlos III de Navarra, mas sí de la segunda, tambien llamada Juana hija de Carlos señor de Labrit.

Gaston VII conde de Foix hijo de estos casó con Doña Leonor hija del infante Don Juan que mas adelante fue II de este nombre rey de Aragon y de Doña Blanca de Evreux reina de Navarra muerta en la flor de su vida.[7]

La posesion de este famoso aunque reducido reino señala dos lunares en su historia cuyas víctimas fueron sucesivamente el príncipe de Viana[8] y Doña Blanca[9] pasando así sus derechos á su hermana Doña Leonor y condesa de Foix cuando su padre Don Juan II de Aragon rey viudo de Navarra cargado de años y de la excecracion histórica murió en Barcelona el 19 de Enero de 1479.

El efímero goce de la ambicionada corona de Navarra que dilataba aquende de los Pirineos los primitivos estados de Foix, solo fué de tres semanas para Doña Leo-

Figura 2.5: Mètode de detecció de línies basant-se en les projeccions horitzontals i verticals del text.

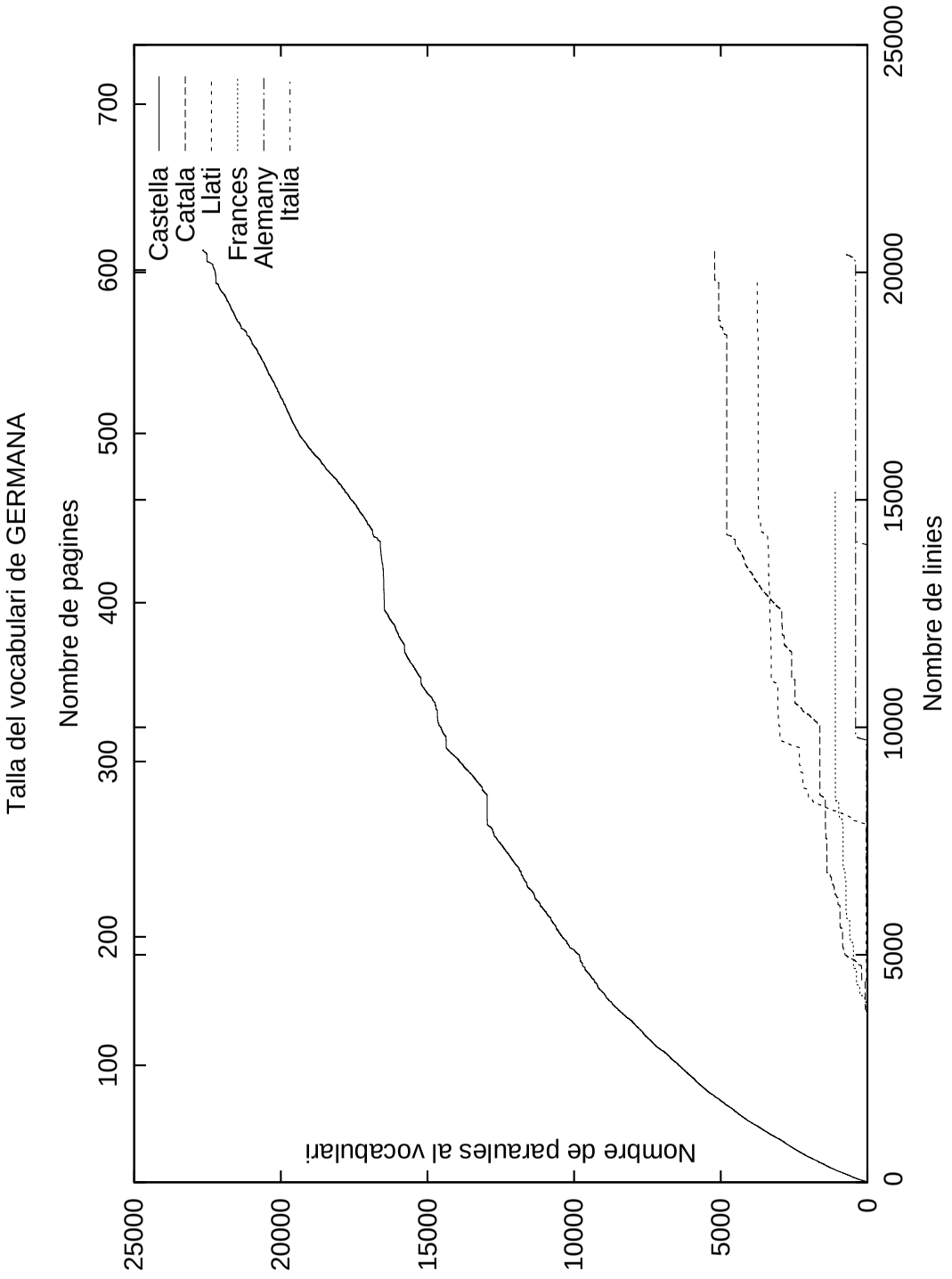


Figura 2.6: Nombre de línies en cada llengua a mesura que avancem en les pàgines del GERMANA.

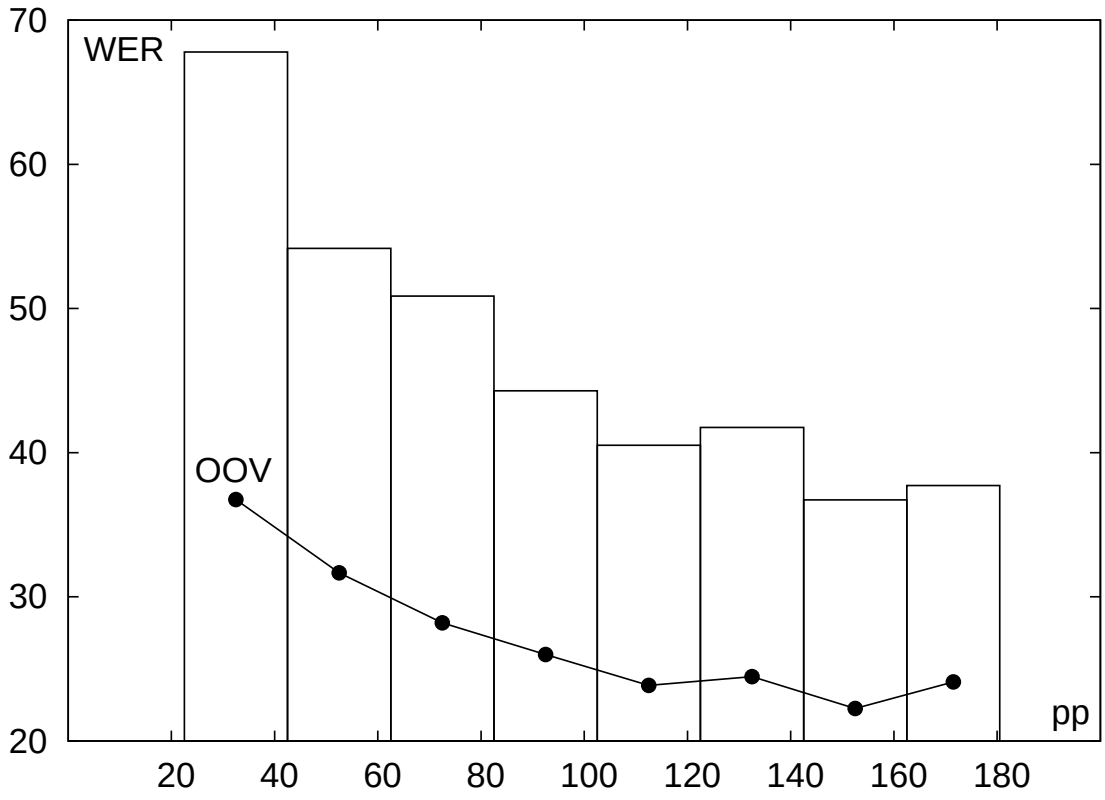


Figura 2.7: Taxa d'error de paraula de la transcripció en el GERMANA com una funció del bloc de les pàgines transcrites (pp). Per a cada bloc, la transcripció del sistema s'entrena amb totes les pàgines dels blocs precedents. També es mostra la part del WER degut a les ocurrències fora de vocabulari (OOV).

Capítol 3

El prototip GIDOC

3.1 Introducció

La transcripció de text manuscrit de documents (antics) és una tasca important per a les biblioteques digitals. Açò es podria dur a terme, primerament processant totes les imatges del document, i després supervisant de manera manual totes les transcripcions a editar en les parts incorrectes. En canvi, les tecnologies de l'estat de l'art per a l'anàlisi automàtic de l'estructura de les pàgines (layout analysis), detecció de línies de text i reconeixement de text manuscrit estan encara lluny de la perfecció [30, 15, 5], i no està clar si és millor post editar l'eixida generada automàticament pel sistema o simplement ignorar-la.

Una aproximació més efectiva per a transcriure documents antics consta en seguir un paradigma interactiu-predictiu en el qual el sistema és guiat pel supervisor humà, i el supervisor humà és assistit pel sistema per a completar la tasca de transcripció tan eficientment com sigui possible. Seguint aquesta aproximació, s'ha desenvolupat un prototip anomenat GIDOC (transcripció Interactiva basada en Gimp de DOCuments antics) per a proporcionar a l'usuari suport integrat per a l'anàlisi d'estructura de pàgina (layout analysis) de manera interactiva-predictiva, detecció de línies i transcripció de text manuscrit [21, 24].

GIDOC s'ha dissenyat per a treballar amb (grans) col·leccions de documents homogenis, és a dir, de similar estructura i estils d'escriptura. Aquestes s'han anotat de manera seqüencial, per hipòtesis (parcialment) supervisades obtingudes a partir de models estadístics que són constantment actualitzats amb un nombre incremental de documents anotats disponibles. L'anotació es fa a diferents nivells. Per exemple, a nivell d'anàlisi d'estructura de pàgina (layout analysis), GIDOC utilitza un mètode nou de detecció de bloc de text en el qual es milloren les tècniques convencionals sense memòria prèvia amb un model "d'història" de les posicions del bloc de text [21]. Similarment, a nivell de transcripció de línies de text, GIDOC inclou un reconeixedor de text manuscrit el qual es millora regularment amb un nombre creixent de transcripcions (parcialment) supervisades

[24].

Ací presentem una descripció comprensible del prototip de GIDOC, emfasitzant aquelles parts que no s'han descrit anteriorment [21, 24]. Després d'una visió general de GIDOC en la Secció 3.2, es descriuen les seves funcions principals en la Secció 3.3 (detecció de blocs i línies), 3.4 (entrenament amb HTK) i 3.5 (transcripció). A continuació es mostren uns experiments fets en la base de dades GERMANA i uns altres amb IAM en les Seccions 3.6 i 3.7 respectivament. Finalment, es discuteixen les conclusions en la Secció 3.8.

3.2 Visió general del sistema

Tal i com indica el seu nom, GIDOC s'ha implementat en el conegut Programa de Manipulació d'Imatge de GNU (GIMP). Al igual que GIMP, GIDOC té la llicencia baix GNU General Public License, i es port descarregar des de [12]. Per a executar GIDOC, primerament hem d'executar GIMP i obrir una imatge d'un document. GIMP vindrà amb la seva excel·lent interfície d'usuari, la qual a sovint és configurada per a mostrar només la caixa d'eines principal.(amb els diàlegs incrustats) i amb una finestra d'imatge. Es pot accedir a GIDOC a partir de la barra de menú de la finestra de la imatge (veure Figura 3.1).

Com es veu en la Figura 3.1, el GIDOC inclou sis entrades:

- Advanced options
- 0: Preferències (Figura 3.2 i 3.3). Aquestes ens van a servir de paràmetres a l'hora d'executar les diferents funcions de GIDOC. Les podem dividir en 4 subgrups que aporten diferent informació; Informació sobre el projecte, informació per al preprocessat, informació per a l'entrenament i informació per al reconeixement. A més, hi ha dos botons que ens permeten crear un projecte nou o obrir-ne un existent.
- 1: Block Detection
- 2: Line Detection
- 3: HTK Training
- 4: Transcription

Les opcions avançades (Advanced options) és un menú de segon nivell on s'agrupen operacions experimentals o d'altres que es realitzen implícitament en algunes de les funcions principals. Per exemple les funcions de preprocessat que es realitzen per a la detecció de blocs o de línies, per a l'entrenament i per al reconeixement també estarien en aquest menú, al igual que els mètodes d'extracció de característiques implementats. Algunes d'aquestes eines de preprocessat s'han extret de [11].

Les preferències (Preferences) obre un diàleg per a configurar opcions globals, o d'altres més específiques per al preprocessat, entrenament i reconeixement. Algunes d'aquestes s'explicaran més avall junt amb altres entrades del menú que van a continuació de les preferències.

3.3 Detecció de blocs i de línies

Durant el desenvolupament, GIDOC s'ha avaluat amb un llibre antic el qual, la majoria de les pàgines només contenen text cal·ligrafiat escrit sobre fulles pautades amb línies ben separades, com es mostra en l'exemple de la Figura 3.1. Com s'ha dit en la introducció, GIDOC està dirigit per a treballar amb documents homogenis i, és més, trau partit de la seva homogeneïtat. En particular, en l'entrada del menú de GIDOC "Detecció de blocs" utilitza un mètode de detecció de bloc de text nou en el qual es milloren les tècniques convencionals sense memòria prèvia amb un model "d'història" de les posicions del bloc. Veure [21] per a més informació.

Donat un bloc de text, l'entrada del menú GIDOC de "Detecció de Línies", detecta totes les línies base (baselines) les quals es marquen amb línies rectes. El resultat es pot observar clarament en l'exemple de la Figura 3.1. Encara que cada línia base té manejadors per a corregir la seva posició gràficament, cal observar que el mètode de detecció de línies implementat funciona prou be, almenys en pàgines com les de l'exemple. És un mètode estàndard basat en projeccions [15]. Primerament, es calcula horitzontalment el promig dels valors dels píxels o les transicions negre/blanc i es projecten verticalment. Després, el histograma vertical resultant es suavitza i analitza per a localitzar les línies base amb precisió. En Preferències s'inclouen dos opcions de preprocessat (Figura 3.2), una per a decidir el tipus d'histograma (valors de píxel o transicions negre/blanc), i l'altra per a definir el màxim nombre de línies base que es pot trobar al document.

3.4 Entrenament amb HTK

GIDOC està basat en tècniques estàndards i eines per al preprocessat de text manuscrit i extracció de característiques, modelat d'imatge basat en HMM i modelat de llenguatge [30]. El preprocessat de text manuscrit aplica un filtrat de soroll a la imatge, correcció de la inclinació del text respecte a l'eix vertical (slant) i normalització vertical de la talla de la imatge de línia de text. Es mostra un exemple il·lustratiu en la Figura 3.4 i 3.5. Tot açò es pot configurar mitjançant les opcions de preprocessat de Preferences. Hi ha una opció per a utilitzar en compte d'un procés adaptat, i dos opcions per a definir els (marges) per a localitzar la part de dalt i de baix de la imatge de la línia de text. Aquests marges van amb respecte la línia base.

L'extracció de característiques per a modelar HMM consisteix en transformar la imatge preprocessada en una seqüència de (dimensió fixa de) vectors de característiques. Hi ha disponible dos mètodes d'extracció de característiques coneguts en GIDOC. El mètode que hi ha per defecte, primerament divideix la imatge preprocessada en una graella de cel·les quadrades les quals tenen una grandària que és una fracció més petita de l'altura de la imatge (e.g 1/20). Després, cada cel·la és caracteritzada pel seu nivell de gris normalitzat i, opcionalment, per la seva derivada horitzontal i vertical del nivell de gris. Veure Figura 3.4 per a un exemple i [30] per a més detalls. El mètode alternatiu mou una finestra simple d'una columna d'esquerra a dreta sobre la imatge, i extrau nou característiques geomètriques de cada posició [5] (Figura 3.5).

El modelat d'imatge mitjançant HMM es du a terme amb el conegut Hiden Markov Model Toolkit (HTK), que està disponible de manera lliure [39]. De manera similar, el modelat de llenguatge està implementat a través del programa lliure SRI Language Modeling Toolkit (SRILM) [26].

L'entrenament amb HTK llegeix el directori de treball on estan les imatges del document i, per cadascuna d'aquestes, extrau totes les línies de text transcrites, en cas d'haverne, junt amb les seves corresponents imatges.

Primerament, les transcripcions es preprocessen per a aïllar caràcters especials (principalment signes de puntuació) i expandir abreviacions (e.g. S.M. s'expandeix per Su Magestad). Després, es construeix un model de llenguatge d'n-grames a partir de les transcripcions preprocessades utilitzant un comandament d'SRILM el qual, per defecte, genera un model de llenguatge de 2-grames amb descompte Kneser-Ney. D'altra banda, es preprocessen les imatges de les línies extretes i es transformen en seqüències de vectors de característiques per a entrenar, utilitzant les seves corresponents transcripcions i HTK, HMMs a nivell de caràcter d'esquerra a dreta amb (Gaussianes) de densitat contínua.

3.5 Transcripcions

L'entrada de *Transcription* en el menú GIDOC obre el diàleg de transcripció interactiu de GIDOC (veure Figura 3.1). Aquest consisteix en dos seccions principals: la secció d'imatge, a la part central, i la secció de transcripció, a la part de baix. En la secció d'imatge, es mostra un nombre d'imatges de línies de text junt amb les seves transcripcions, en cas d'estar disponibles, en caixes de text editables separades dintre de la secció de transcripció. La línia actual a transcriure o simplement supervisada es selecciona situant el cursor en la caixa editable apropiada. La línia base corresponent es remarca (en color blau) i, sempre que sigui possible, GIDOC mourà les imatges de línies i les seves transcripcions de manera que es mostri la línia actual en la part central tant en la secció de la imatge com en la de transcripció. S'assumeix que l'usuari transcriu o supervisa les línies de text de dalt cap a baix (o en qualsevol ordre desitjat), introduint text i movent el cursor d'edició amb les fletxes de direcció o el ratolí.

Cada caixa de text editable té un botó associat a la seva esquerra, el qual està etiquetat amb el seu nombre de línia corresponent. Polsant sobre aquest, s'extrau la imatge de línia associada, es preprocessa, es transforma en una seqüència de vectors de característiques, i es descodifica per Viterbi utilitzant HTK i els models entrenats anteriorment durant l'entrenament amb HTK. D'aquesta manera, no es necessita introduir la transcripció completa de la línia actual, sinó sols aquelles correccions, en cas d'haver-ne, sobre l'eixida del descodificador. Clarament, açò sols és possible si, primer, les línies de text es detecten correctament i, segon, els HMM i models de llenguatge s'entrenen adequadament, a partir d'una suficient quantitat de dades d'entrenament. Així doncs, s'assumeix que part de la transcripció es du a terme en etapes més primerenques dins d'una tasca de transcripció, per tal de poder entrenar els models, i després el sistema assisteix l'humà com s'ha descrit anteriorment.

3.6 Experiments amb GERMANA

Durant el seu desenvolupament, un paleògraf expert ha utilitzat GIDOC per a anotar blocs, línies de text i transcripcions en un conjunt de dades anomenat GERMANA [18]. GERMANA és el resultat de digitalitzar i anotar un manuscrit Espanyol de 764 pàgines de l'any 1891, en el qual la majoria de pàgines només conté text cal·ligrafiat escrit en fulls pautats amb línies ben separades. L'exemple mostrat en la Figura 3.1 correspon a la pàgina 144. El GERMANA està escrit només en Castellà fins la pàgina 180. després, el manuscrit inclou algunes parts que estan escrites en llengües diferents al Castellà com Català, Francès i Llatí.

Degut a l'estructura seqüencial del llibre, la tasca més bàsica del GERMANA és transcriure'l d'inici a fi, encara que ací només considerem les transcripcions fins a la pàgina 180. Començant a partir de la pàgina 3, dividírem el GERMANA en 9 blocs consecutius de 20 pàgines cadascuna (18 en el bloc 9). I, de mitja, hi ha 417 línies i 4687 paraules (running words). Després, a partir del bloc 2 (pàgines 23-42) i fins al 9 (pàgines 163-180), cada bloc va ser transcrit automàticament per GIDOC anteriorment entrenat per tots els blocs precedents. Els resultats es mostren en la Figura 3.6, en termes d'índex d'error de paraula (WER). Per a permetre fluctuacions degut a la variació de la complexitat del conjunt, el WER va ser calculat per a un bloc fixat (bloc 9) després de cada re-entrenament amb GIDOC, i la corba de WER resultant ha estat afegida a la Figura 3.6. També es mostra la part de WER deguda al nombre d'ocurrències de paraules fora de vocabulari (OOV).

Com s'esperava, el WER decreix a mesura que la quantitat de dades d'entrenament augmenta. En particular, GIDOC aconsegueix al voltant d'un 34% de WER per als últims dos blocs, el qual no està mal per a transcripció assistida per ordinador efectiva. La corba de WER per al bloc 9 no presenta diferències significatives respecte la corba de WER per al bloc següent, encara que pareix que el bloc 9 és un poc més complex que totes els altres

blocs precedents (excepte el bloc 7). Observant les corbes d'OOV, es veu clarament que una fracció considerable d'errors de transcripció es degut a les ocurrències de paraules no vistes. Amb més detall, la quantitat de paraules no vistes és aproximadament d'un 50% dels errors de transcripció.

S'ha d'observar que els resultats de WER preliminars (només per al bloc següent) ja ha estat aportat en [18] per a acompanyar la descripció del GERMANA. A diferencia de l'altre, les corbes de WER i OOV aportades ací són lleugerament millors amb un promig de (5.4% i 6.4%, respectivament). Açò es degut principalment a un millor modelat de les abreviacions i signes de puntuació. També, hem utilitzat una versió actualitzada de les línies base de GERMANA les quals estan ajustades amb més precisió.

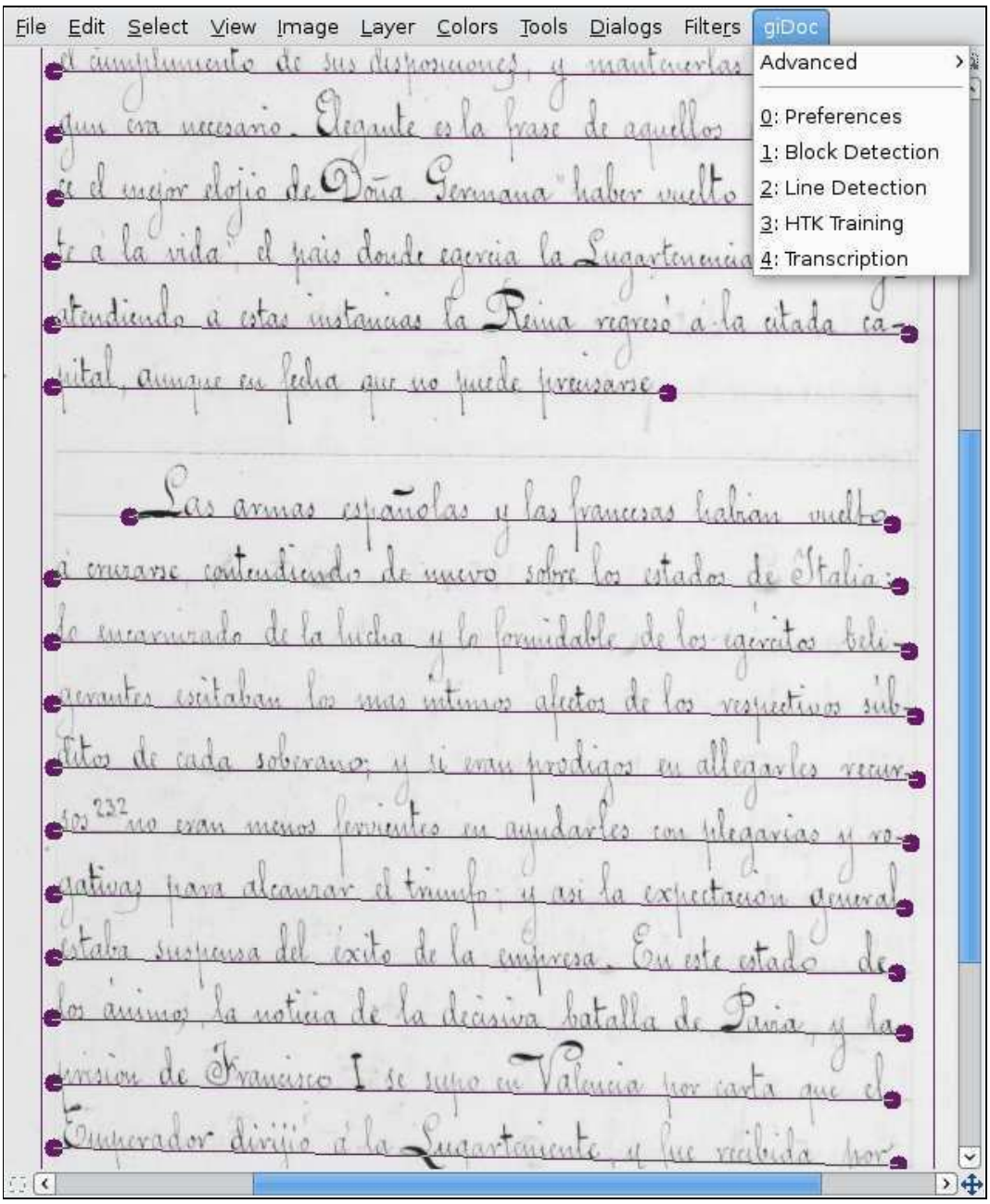


Figura 3.1: Diàleg de transcripció interactiu sobre una finestra d'imatge mostrant el menú de GIDOC.

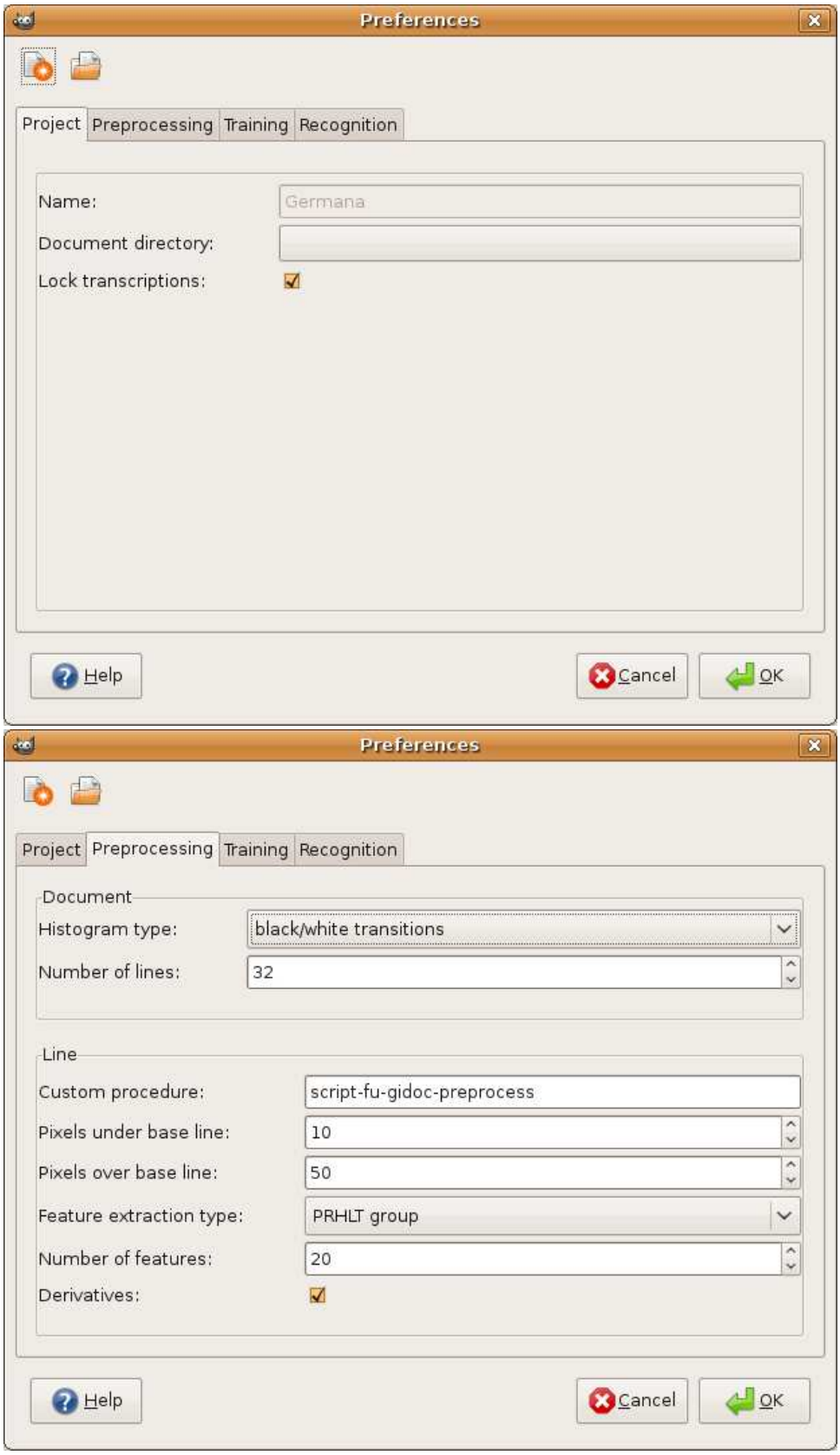


Figura 3.2: Diàleg de preferències de GIDOC. Pestanyes de "Project i Preprocessing".

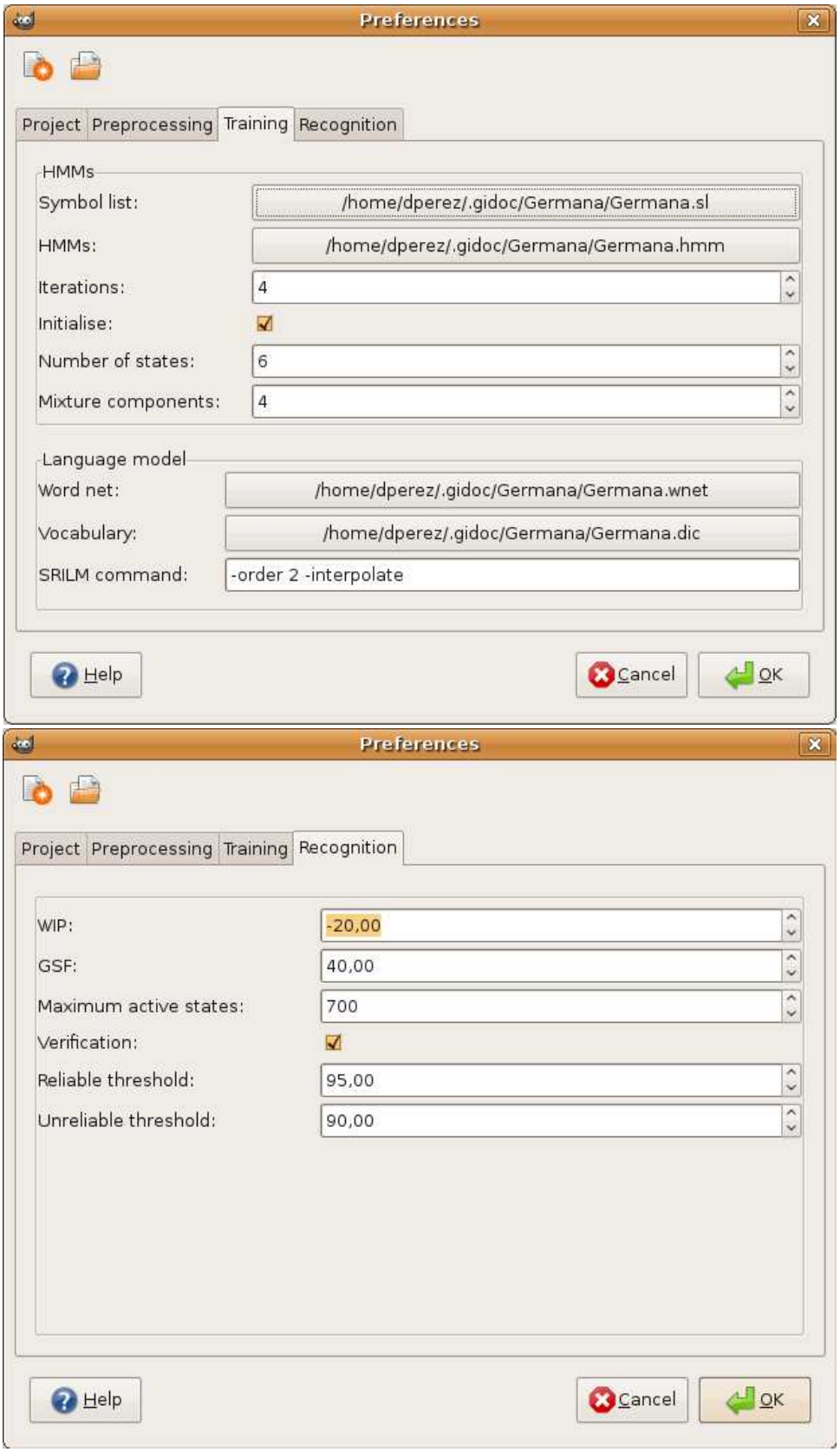


Figura 3.3: Diàleg de preferències de GIDOC. Pestanyes de "Training i Recognition".

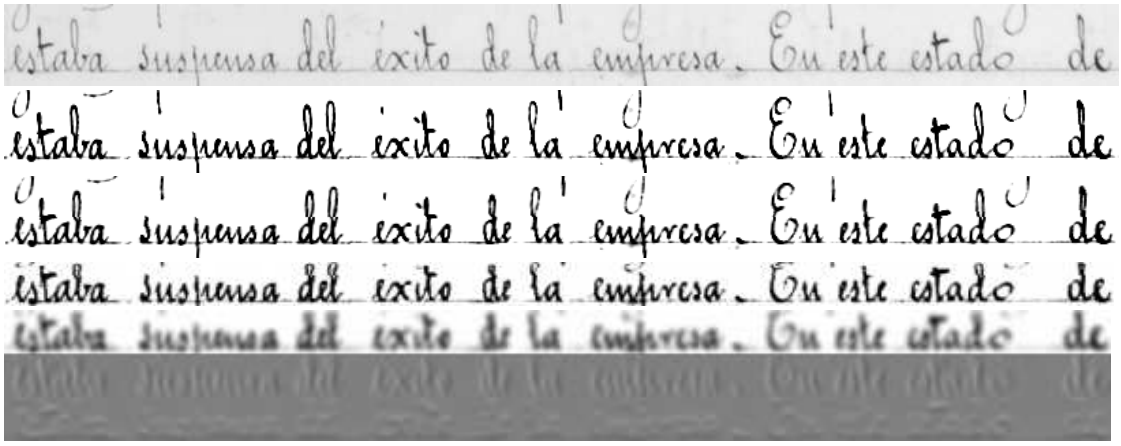


Figura 3.4: Preprocessat i extracció de característiques d'una línia de text de GERMANA. De dalt cap a baix: imatge original, filtrat de soroll, correcció d'slant, normalització vertical de la grandària del text i extracció de característiques corresponent al grup d'investigació *Pattern Recognition and Human Lenguage Technology,* PRHLT del *Departament de Sistemes Informàtics i Computació de València*.

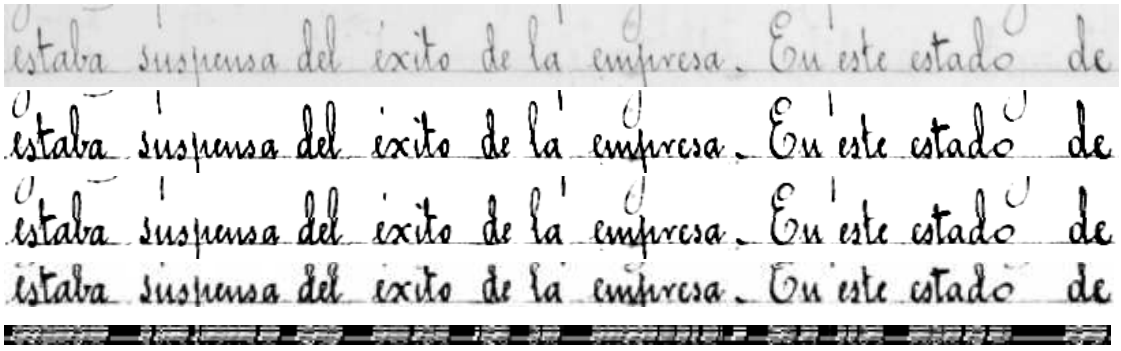


Figura 3.5: Preprocessat i extracció de característiques d'una línia de text de GERMANA. De dalt cap a baix: imatge original, filtrat de soroll, correcció d'slant, normalització vertical de la grandària del text i extracció de característiques corresponent al grup d'investigació *Computer Vision and Artificial Intelligence (FKI)* del institut *Computer Science and Applied Mathematics (IAM)* de Berna.

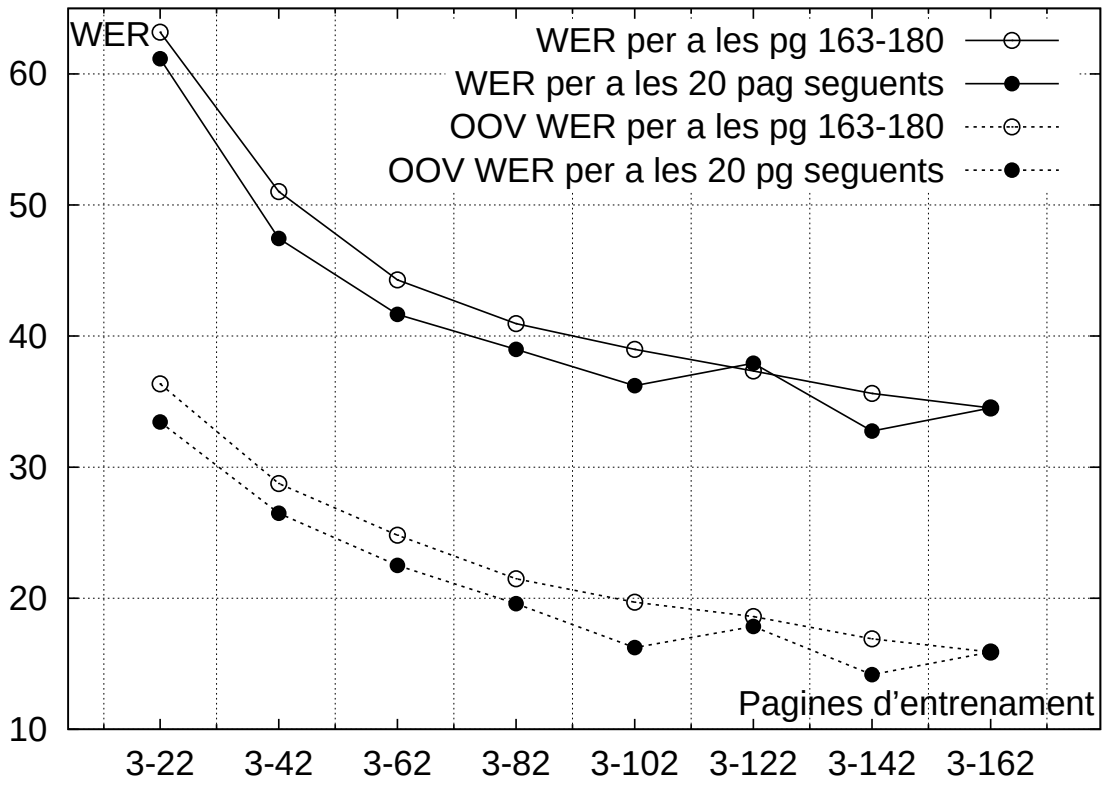


Figura 3.6: Taxa d'error de paraula de transcripció (WER) en GERMANA com una funció de les pàgines ja supervisades i per tant disponibles per a entrenar (pagines d'entrenament). El WER es calcula tant per a les 20 pàgines següents a les supervisades (línia sòlida amb cercles negres), com per al conjunt fixat que compren les pàgines 163-180 (línia sòlida amb cercles blancs). També es mostra la part de WER degut a les ocurrències fora de vocabulari (OOV) (línies discontínues).

3.7 Experimentació amb IAM

3.7.1 Introducció

Per tal de provar l'eficàcia del prototip, també s'ha realitzat experimentació amb la base de dades d'IAM, ja que la gran majoria d'articles que aporten resultats de reconeixement de text manuscrit, ho fan utilitzant aquesta base de dades i d'aquesta manera ens permet comparar l'efectivitat del prototip.

En la Secció 3.7.2, s'inclou una breu descripció d'aquest corpus junt amb algunes estadístiques.

A més, s'ha tractat de reproduïr l'extracció de característiques usada pel grup FKI (veure Secció 3.7.3) per a intentar arribar als resultats de l'estat de l'art per a posteriorment aportat algunes millores que ens ajudin a superar-lo.

3.7.2 La base de dades d'IAM

IAM database és un corpus de text anglès manuscrit no restringit [34, 35, 17, **?**]. Aquest corpus ha sigut desenvolupat pel grup FKI del *Institut für Informatik und angewandte Mathematik*(IAM). La versió del corpus que hem utilitzat és la 3.0 que es pot baixar en (`http://www.iam.unibe.ch/fki/databases/iam-handwriting-database`).

El corpus està composat per un total de 1539 formularis amb text manuscrit. Cada formulari té una capçalera amb un identificador i el text imprès a escriure, i després un gran espai en blanc entre dos línies on es troba el text manuscrit. Un total de 657 escriptors van participar en l'adquisició del corpus. Les dos úniques restriccions que van tindre van ser: que empraren plantilles per a escriure les línies rectes, amb una separació aproximada entre línies de $1.5cm$, i que si una paraula no cabia en l'actual línia que canviaren de línia, abans que tindre que escriure-la atapida contra el marge. Per el demés es va donar total llibertat a l'hora d'escriure tant d'estil com d'instrument emprat, de fet, se'ls va demanar que escrigueren amb el seu estil d'escriptura habitual. Els formularis van ser escanejats amb una resolució de $300dpi$ en imatges en escala de grisos. Respecte als textos que apareixen en els formularis, tots tenen al voltant de 6 frases i almenys 50 paraules. Aquests fragments de text van ser extrets del corpus *Lancaster - Oslo/Bergen* (LOB) [**?**]. El corpus LOB és una recopilació de textos reals del 1961, escrits en anglès britànic. LOB té al voltant del milió de paraules, i es poden trobar fins a 15 temàtiques diferents: reportatge, religió, humor, etc. Per a cada formulari el corpus IAM proporciona un identificador d'escriptor i una referència al text original en el corpus LOB. En les Figures 3.7 i 3.8 es poden veure alguns exemples de formularis.

A més dels formularis, el corpus IAM conté tres subcorpus obtinguts a partir dels formularis mitjançant tècniques de segmentació automàtiques. Aquestos subcorpus són: un subcorpus de línies, un de frases i un altre de paraules. En aquest treball hem utilitzat el subcorpus de línies ja que és el que actualment estan utilitzant els grups més punters en la

matèria i ens permet comparar-nos amb ells. En la Figura 3.1 vegem algunes estadístiques d'aquesta partició.

	Entr.	Valid.	Test
Pàgines	747	116	336
Línies	6161	920	2781
Paraules (K)	53.8	8.7	25.4

Taula 3.1: Estadístiques d'IAM.

3.7.3 L'extracció de característiques

El procés de reconeixement de text manuscrit es pot dividir en tres fases: preprocessat i extracció de característiques, entrenament del models (models morfològics i model de llenguatge) i reconeixement. En aquests apartats, anem a aportar o reproduir alguns mètodes que ens permeten arribar a l'estat de l'art i en alguns casos millorar-lo. En concret, s'ha implementat l'extracció de característiquese que usa el grup FKI en els quals s'aconsegueixen els millors resultats en reconeixement de text manuscrit online en l'actualitat. Així doncs, s'ha pres com a referència l'article [10] on es detalla aquesta part en concret i aporta algunes pinsellades a les altres etapes de la resta de procés de RTM.

Com ja s'ha comentat, un dels passos que es realitza en el RTM consisteix en l'extracció de característiques en la qual es transforma una imatge en una sèrie de vectors numèrics d'una determinada dimensió i que ens permeten processar-la i així entrenar els models.

Així doncs, després d'aplicar el preprocessat corresponent a les imatges (eliminació del soroll, correcció d'slant, normalització de l'altura del text i normalització de l'altura de la imatge) s'ha aplicat una extracció de característiques en la qual una finestra igual a una columna de la imatge, recorre aquesta d'esquerra a dreta per a extraure, per cadascuna de les posicions, 9 característiques d'origen morfològic (veure Figura 3.5). Aquestes característiques es poden obtenir fent ús de les formules descrites en la Figura 3.9.

En la imatge es mostra com obtenir cadascuna de les característiques donada una columna de la imatge **j** d'm píxels d'altura. Així doncs les característiques descriuen els següents aspectes de la imatge:

1. Moment central d'ordre 0 (massa).
2. Moment central d'ordre 1 (centre de massa).
3. Moment central d'ordre 2.
4. Posició del primer píxel en negre de la columna.

5. Posició de l'últim píxel en negre de la columna.

6. Mitjana entre les posicions del primer píxel en negre de la columna anterior i posterior.

7. Mitjana entre les posicions de l'últim píxel en negre de la columna anterior i posterior.

8. Nombre de transicions de negre/blanc en la columna.

9. Massa de la columna llevant el primer i últim píxel en negre de la columna.

Considerem píxel negre aquell que superi un llindar estimat. Aquest llindar es pot calcular per exemple mitjançant la tècnica d'Otsu en la qual es troba un valor que minimitza la variància intragrup i maximitza les de grups diferents.

3.7.4 Resultats de referència

Com ja s'ha dit en la *Introducció*, per a realitzar aquest apartat, hem pres com a referència [10] en els quals es poden aconseguir els millors resultats en RTM. En aquests articles cal destacar l'extracció de característiques ja que tenen un origen prou diferent a altres utilitzats (e.g grup **PRHLT** [30]) en el que també s'aconsegueixen resultats competitius.

En aquest tipus d'extracció de característiques realitzat pel grup FKI, es recorre la imatge amb una finestra d'una columna i per a cadascuna de les posicions es calculen 9 característiques d'origen morfològic. En la Figura 3.9 vegem 9 fórmules per a calcular cadascuna de les característiques. Podem veure la representació gràfica d'aquesta extracció de característiques en lúltim pas de l'exemple mostrat en la Figura 3.5.

Pel contrari, en les característiques utilitzades pel grup PRHLT una finestra quadrada de 20 píxels recorre la imatge calculant en cada posició el valor de grisos i opcionalment les derivades verticals i horitzontals. En l'últim pas de la Figura 3.4 també podem veure de manera gràfica una representació d'aquesta extracció de característiques.

Per tant, l'extracció de característiques FKI transforma les imatges en una seqüència de vectors de característiques de 9 dimensions i el del PRHLT en vectors de 60 característiques (valor per defecte). Per a aquest apartat hem decidit experimentar amb característiques FKI ja que actualment s'aconsegueixen millors resultats en RTM.

Respecte el preprocessat aplicat abans d'extraure les característiques, cal dir que s'apliquen una sèrie de mètodes de manera seqüencial per tal de millorar l'aspecte del text abans d'aplicar l'extracció: eliminació del soroll, correcció d'slant o inclinació del text respecte l'eix vertical, normalització de la grandària vertical del text (Figura 3.5).

Durant l'entrenament dels Models de Markov el procés es centra en ajustar el nombre d'estats que s'utilitza per a modelar els caràcters i el nombre de gaussianes en cadascun

dels estats el qual ens permet modelar diferents formes de representar un mateix caràcter (o fragment d'aquest). Açò es fa utilitzant el toolkit d'HTK.

Respecte a l'estimació del model de llenguatge, el que es fa per a aconseguir aquests resultats de 3.2 és: agafant els corpus WELLINGTON, BROWN i LOB (llevant els conjunts de validació i test de l'IAM ja que el LOB conté aquest corpus), s'utilitzen les 20000 paraules més freqüents per a estimar el model de bi-grames. Per a més detalls veure [10].

Finalment durant el reconeixement, simplement el que es fa és una descodificació per Viterbi i en aquesta fase s'intenten ajustar dos paràmetres, amb el conjunt de validació, que són: el Factor d'Escalat de la Gramàtica o model de llenguatge (GSF) i la Penalització d'Inserció de Paraula (WIP).

Així doncs, segons [10] després de realitzar el preprocessat i extracció de característiques, entrenar els HMMs i mòdel de llenguatge, s'obtenen els següents resultats al reconèixer el conjunt de validació i test:

Conjunt	**WER**
Validació	30.1
Test	35.5

Taula 3.2: Resultats de referència en termes de WER sobre el conjunt de Validació i Test.

Aquests resultats obtinguts són els que ens van a servir de referència per als apartats següents.

3.7.5 Resultats preliminars

Primerament, s'ha tractat d'obtenir resultats similars reproduint tot el procés que realitza el grup FKI. Així doncs, partint d'unes característiques obtingudes segons s'ha explicat anteriorment, s'han estimat els models, s'han ajustat els paràmetres amb el conjunt de validació i finalment s'ha reconegut el conjunt de test i s'ha obtingut una taxa d'error.

El model de llenguatge emprat és un model de bigrames amb les 20000 paraules més freqüents del corpus BROWN, WELLINGTON i LOB (excloent el conjunt de validació i test de l'IAM) considerant com a paraules els signes de puntuació i diferenciant entre paraules amb majúscula i amb minúscula.

Després d'entrenar els models, s'han obtingut els següents resultats de reconèixement sobre els conjunts de validació i test.

Conjunt	GSF	WIP	WER
Validació	20	-1	30.7
Test	20	-1	37.0

Taula 3.3: Resultats preliminars en termes de WER sobre el conjunt de Validació i Test. Es mostra el *Grammar Scale Factor* i *Word Insertion Penalty* utilitzat durant el reconeixement per a obtenir el WER indicat.

Com es pot observar, els resultats obtinguts són molt similars als obtinguts en 3.2, especialment en el conjunt de validació on no s'observen diferències significatives amb respecte les de referència.

No obstant, en el conjunt te test s'observa un empitjorament d'1.5% de taxa d'error. Aquesta diferència podria ser deguda a que, durant l'entrenament del HMMs, nosaltres hem duplicat en cada iteració el nombre de gaussianes i el grup FKI les incrementa en 1 en cada pas i,per tant, realitza un major nombre d'iteracions. Encara que açò no tindria perquè millorar els resultats, es possible que per a aquesta base de dades si que influeixi i d'aquesta manera les gaussianes s'ajusten millor a les dades.

3.7.6 Millores al model de llenguatge

Fins ara, en els apartats anteriors, s'ha estat utilitzant un model de llenguatge estimat amb un diccionari amb les 20000 paraules més freqüents del corpus LOB (llevant la partició de validació i test de l'IAM), WELLINGTON i BROWN.

No obstant, hem considerat que seria apropiat provar amb altres valors de grandària de diccionari *N* per a veure el comportament del model de llenguatge sobre el reconeixement i així veure si es pot millorar en termes de WER.

Per aquest motiu, s'ha decidit variar el valor d'*N* per a crear el diccionari i el model de llenguatge i, posteriorment, reconèixer la partició de validació i test. El criteri que s'ha establert per a augmentar *N* no ha sigut el d'incrementar aquest en un nombre fixat en cada pas sinó basant-nos en la freqüència d'aparició de les paraules. Açò és, s'agafaran les paraules que apareguen, com a mínim, un nombre X de vegades en la base de dades (composta pels tres corpus).

Així, en aquest cas hem utilitzat el mateix que en el punt anterior però canviant el diccionari i model de llenguatge de les 20000 paraules més freqüents pel nou estimat en el que, per a estimar el diccionari i model de llenguatge, s'agafen aquelles paraules que apareixen com a mínim 4 vegades als textos ja que és el que millor resultat ens ha donat. Recordem que s'han considerat com a paraules, els signes de puntuació i les paraules amb capitalització diferent.

El resultat obtingut després d'ajustar els paràmetres apropiats els vegem a la Taula 3.4.

Conjunt	GSF	WIP	WER
Validació	21	9	28.0
Test	21	9	34.5

Taula 3.4: Resultats obtinguts en termes de WER sobre el conjunt de Validació i Test al cambiar el model de llenguatge. Es mostra el *Grammar Scale Factor* i *Word Insertion Penalty* utilitzat durant el reconeixement per a obtenir el WER indicat.

En aquest cas, observem que els resultats milloren (respecte als de la Taula 3.3) amb 2.7 punts de WER en el conjunt de validació i 2.5 punts de WER en el conjunt de test, simplement pel fet de canviar el model de llenguatge al explicat anteriorment. Respecte als resultats de referència (Taula 3.2), es supera amb 2.1 punts de WER en el conjunt de validació i 1 punt en el conjunt de test.

3.7.7 Millores als HMMs

En aquest cas, també s'ha pres com a punt de partida allò realitzat en el punt anterior i s'han realitzat alguns canvis, en aquest cas els HMMs.

Així doncs, s'ha utilitzat el model de llenguatge explicat en l'anterior punt però hem canviat la metodologia emprada per a entrenar els Models de Markov i en compte d'ajustar un nombre d'estats fix per a cadascun dels HMMs que representin els caracters, s'ha tractat d'estimar un nombre d'estats específic per a cadascun dels models. Açò s'ha fet mitjançant la tècnica de la segmentació forçada [10] en la que donat unes imatges de línies de text i les seves corresponents transcripcions, es tracta d'estimar la mitja del nombre d'estats necessaris per a representar cada caràcter. D'aquesta manera, es podran detectar més fàcilment els caràcters durant la descodificació per Viterbi. A més d'entrenar cada HMM amb un nombre d'estats variable, em canviat l'estratègia emprada per a incrementar el nombre de gaussianes en cada iteració. En aquest cas ho hem fet de la mateixa manera que ho fa el grup FKI i, en cada iteració, incrementarem amb 1 el nombre de gaussianes (corresponent a la gaussiana més probable), en compte de duplicar-les com ho fèiem fins ara.

Vegem els resultats obtinguts al realitzar aquests canvis a la Taula 3.5.

Conjunt	GSF	WIP	WER
Validació	19	30	25.7
Test	19	30	31.7

Taula 3.5: Resultats obtinguts en termes de WER sobre el conjunt de Validació i Test al cambiar el model de llenguatge i els Models de Markov de cadascun dels caracters. Es mostra el *Grammar Scale Factor* i *Word Insertion Penalty* utilitzat durant el reconeixement per a obtenir el WER indicat.

Observem que aquests canvis també aporten certa millora. Així per exemple en el conjunt de validació tenim un 25.7 de WER contra el 28.0 (Taula 3.4) que teníem en el punt anterior i els 30.1 de referència (Taula 3.2). D'altra banda, en el conjunt de test aconseguim un 31.7% de WER contra els 34.5% del punt anterior (Taula 3.4) i 35.5% de referència (Taula 3.2).

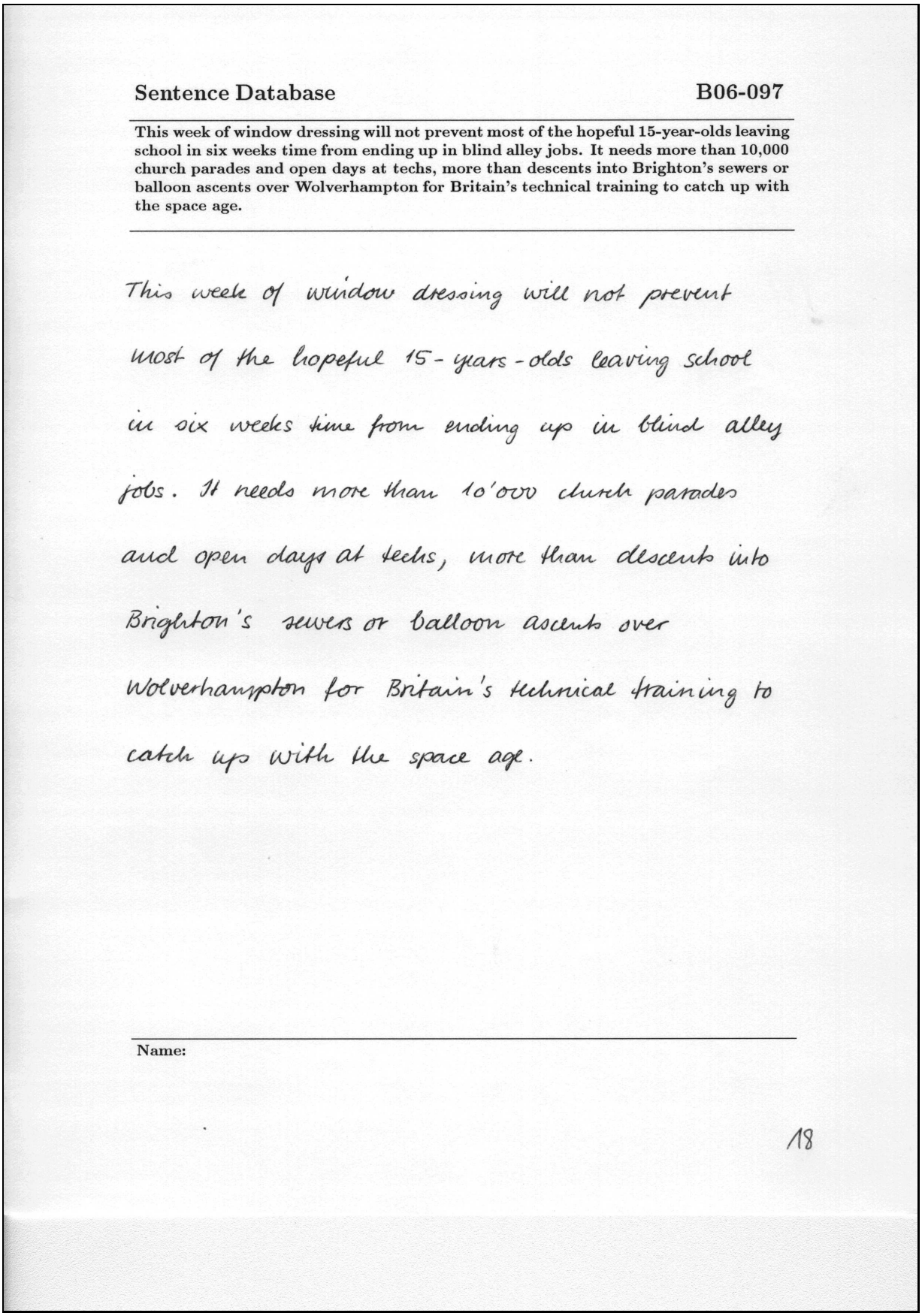
Aquest és el millor resultat obtingut en l'experimentació en el que s'aconsegueix pràcticament un 4% de millora respecte als resultats de referència sobre el conjunt de test.

3.8 Conclusions

S'ha presentat un prototip de transcripció assistida per ordinador per a documents antics de text manuscrit anomenat GIDOC. GIDOC és un primer intent de proporcionar suport integrat per a anàlisis d'estructura de pàgina (layout analysis) interactiva-predictiva, detecció de línies de text i transcripció de text manuscrit. Les eines estan integrades al menú superior de GIMP, i utilitza tècniques estàndards i eines per al preprocessat de text manuscrit i extracció de característiques, modelat d'imatge basat en HMM, i modelat de llenguatge. Al igual que GIMP, GIDOC té la llicencia baix *GNU General Public License*, i es pot descarregar de manera gratuïta per Internet. L'efectivitat de GIDOC s'ha demostrat empíricament en la base de dades de GERMANA, la qual està també públicament disponible en Internet. D'altra banda, també s'han presentat uns experiments realitzats amb aquest prototip amb la base de dades d'IAM en els quals s'ha tractat de reproduir el procés realitzat pel grup FKI i aportar algunes millores. Aquestes millores suposen un 4% de WER de millora respecte als resultats de referència obtinguts amb el conjunt de test.

Cal tindre en compte que recentment s'han aconseguit millors resultats de RTM utilitzant xarxes recurrents **??** on s'aconsegueix un 25.9% de WER.

El treball futur inclou la integració de noves eines per al modelat de HMM i descodificació per Viterbi, identificació del llenguatge, entrada de text multi-modal, incorporar xarxes recurrents al prototip, etc.



Sentence Database **B06-097**

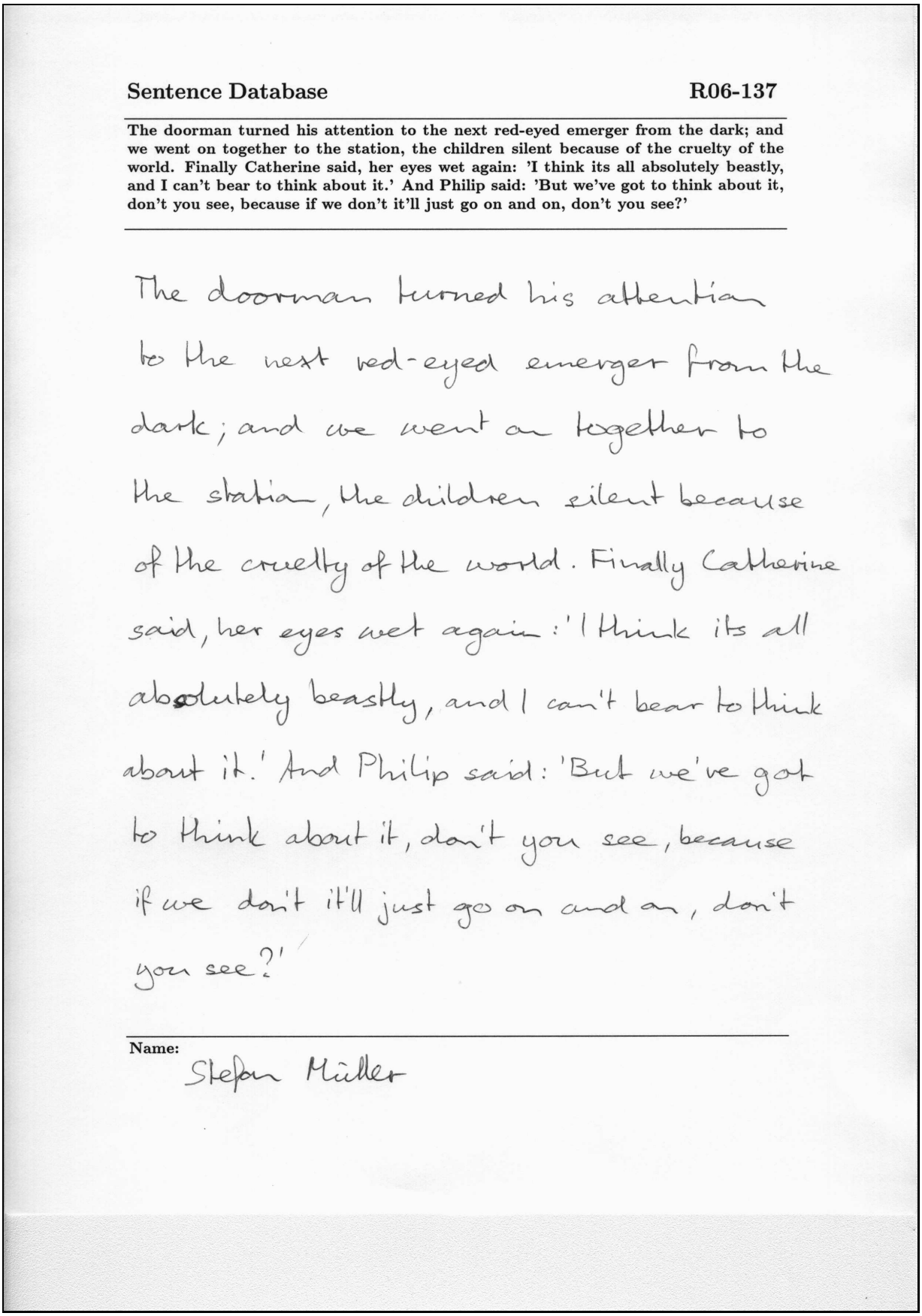
This week of window dressing will not prevent most of the hopeful 15-year-olds leaving school in six weeks time from ending up in blind alley jobs. It needs more than 10,000 church parades and open days at techs, more than descents into Brighton's sewers or balloon ascents over Wolverhampton for Britain's technical training to catch up with the space age.

This week of window dressing will not prevent most of the hopeful 15-years-olds leaving school in six weeks time from ending up in blind alley jobs. It needs more than 10'000 church parades and open days at techs, more than descents into Brighton's sewers or balloon ascents over Wolverhampton for Britain's technical training to catch up with the space age.

Name:

18

Figura 3.7: Exemple de formulari del corpus IAM. Formulari *B06-097*.



Sentence Database **R06-137**

The doorman turned his attention to the next red-eyed emerger from the dark; and we went on together to the station, the children silent because of the cruelty of the world. Finally Catherine said, her eyes wet again: 'I think its all absolutely beastly, and I can't bear to think about it.' And Philip said: 'But we've got to think about it, don't you see, because if we don't it'll just go on and on, don't you see?'

The doorman turned his attention to the next red-eyed emerger from the dark; and we went on together to the station, the children silent because of the cruelty of the world. Finally Catherine said, her eyes wet again: 'I think its all absolutely beastly, and I can't bear to think about it.' And Philip said: 'But we've got to think about it, don't you see, because if we don't it'll just go on and on, don't you see?'

Name:
Stefan Müller

Figura 3.8: Exemple de formulari del corpus IAM. Formulari *R06-137*.

ngr: píxel negre; **blc**: píxel blanc

$$c_1(j) = \sum_{i=1}^{m} I(i,j)$$

$$c_2(j) = \frac{1}{c_1(j)} \sum_{i=1}^{m} i \cdot I(i,j)$$

$$c_3(j) = \frac{1}{c_1(j)} \sum_{i=1}^{m} (i - c_2(j))^2 \cdot I(i,j)$$

$$c_4(j) = min(i | I(i,j) = \mathbf{ngr})$$

$$c_5(j) = max(i | I(i,j) = \mathbf{ngr})$$

$$c_6(j) = \frac{c_4(j+1) - c_4(j-1)}{2}$$

$$c_7(j) = \frac{c_5(j+1) - c_5(j-1)}{2}$$

$$c_8(j) = NT_{ngr \xrightarrow[c_4(j) \leq i \leq c_5(j)]{} blc}(I(i,j))$$

$$c_9(j) = \sum_{c_4(j) < i < c_5(j)} I(i,j)$$

Figura 3.9: Descripció de les 9 característiques que componen els vectors de característiques emprats.

Capítol 4

Altres aportacions

4.1 Mesures de confiança per a correcció d'errors en transcripció interactiva de text manuscrit

4.1.1 Introducció

Una manera efectiva de transcriure documents antics és seguint un paradigma interactiu-predictiu en el qual el sistema es guiat pel supervisor humà i, a la vegada, el supervisor es ajudat pel sistema per a completar la tasca de transcriure de la manera més eficient possible. En el cas d'iDoc, s'ha desenvolupat un prototip de sistema CAT anomenat GIDOC (Gimp-based Interactive transcription of old text DOCuments) per a proporcionar a l'usuari, suport integrat per a l'anàlisi de l'estructura de pàgines (layout analysis) de manera interactiva-predictiva, detecció de línies de text i transcripció de text manuscrit (Secció 3). En aquest punt ens anem a centrar en la part del reconeixement de text manuscrit de GIDOC.

Els HMM i el model de llenguatge s'entrenen a partir de línies de text transcrites manualment durant etapes anteriors de la tasca de transcripció. Després, cada imatge corresponent a una línia de text és processa, primerament es prediu la transcripció més probable i després es localitzen i editen els errors del sistema. En els experiments realitzats en la secció **??**, per exemple, es considera una tasca de transcripció en la qual GIDOC aconsegueix al voltant d'un 37% de WER en (test) després de transcriure 140 pàgines del document d'un total de 764 (18%). Encara que un 37% de WER no està malament per a sistemes de CAT efectius, també s'ha de tenir en compte l'esforç humà que s'empra a l'hora de localitzar i editar els errors del sistema, i açò succeeix en tasques de transcripció de text manuscrit en general.

En aquest treball, tornem a fer ús de la tecnologia emprada en reconeixement de text manuscrit, en particular en les mesures de confiança (a nivell de paraula) [14, 22], les quals es proposen per a la localització i correcció d'errors en transcripció de text manus-

crit interactiu. Encara que l'ús de les mesures de confiança en reconeixement de línies de text manuscrit offline no es nou [6], ací anem un poc més enllà i proposem les mesures de confiança per a guiar el supervisor humà en la localització de possibles errors del sistema i en la decisió de com procedir. D'altra banda i en contrast amb altres treballs [6], ací les mesures de confiança es basen en la probabilitat de paraula a posteriori estimada a partir de grafs de paraules ja que, almenys en el cas de reconeixement de la veu, evidències experimentals ens mostren clarament que aquestes superen les mesures de confiança alternatives, i inclús les probabilitats de paraula estimades a partir de les llistes d'N-millors [14, 22].

D'altra banda, com ja s'ha explicat en la secció 3, el menú de GIDOC inclou sis entrades. En aquest cas només es va a descriure breument l'última que és l'entrada de Transcription. Aquesta obre un diàleg de transcripció interactiu, el qual consisteix en dos seccions principals: la secció d'imatge i la secció de transcripció.

Cadascuna de les caixes de text editable té un botó associat a la seva esquerra el qual està etiquetat amb el seu corresponent nombre de línia. Polsant sobre aquest, s'extrau la imatge de la línia associada, es fa el preprocessat, és transforma en una seqüència de vectors de característiques i es fa la descodificació per Viterbi usant HMM i un model de llenguatge prèviament entrenats. Les paraules de la línia actual les quals el sistema no està altament segur, són remarcades en color roig tant en la secció d'imatge com en la de transcripcions. Després el usuari supervisarà l'eixida del sistema completament o, simplement les paraules marcades en roig. Aquest tindrà que acceptar, editar o descartar la transcripció de la línia actual donada pel sistema.

En la següent secció 4.1.2 es descriu de manera breu l'experimentació realitzada i els resultats obtinguts. Per a veure de manera més detallada tot aquest apartat corresponent a la secció 4.1 veure la Tesi de Master del company L.Tarazón [3].

4.1.2 Experiments

Base de dades

Per a realitzar l'experimentació s'ha utilitzat la base de dades IAM-DB 3.0 [17] vista en la secció 3.7.2. L'extracció de característiques ha sigut realitzada utilitzant el mètode explicat a la secció 3.7.3. Els HMMs tenen una tipologia lineal composta de 7 estats amb mixtures de 16 gaussianes per cada estat. Amb açò havem aconseguit un WER de 35.5% per al conjunt de test del corpus IAM.

Pel que fa a la base de dades del GERMANA (explicat en la Secció 2) cal recordar que el 68% de les paraules del model de llenguatge apareixen només una vegada, i les abreviacions apareixen de moltes maneres diferents. A més, el 33% de les paraules estan incompletes ja que apareixen al principi o al final de la línia. En aquest cas, els HMMs tenen també una tipologia lineal composta de 6 estats amb mixtures de 64 gaussianes per cada estat. Havem aconseguit un 42% de WER en el conjunt de test. Veure [18] per a una

descripció detallada.

Resultats

Per a mostrar els resultats es va a fer ús de les corbes ROC (veure [3]). Calcular la Precisió i la Supervisió com a una funció de la corba ROC permet avaluar el impacte de les mesures de confiança sobre la compensació entre esforç-precisió.

L'aproximació proposada ha sigut provada utilitzant el paquet d'eines de GIDOC amb els corpus d'IAM i GERMANA, descrits en seccions anteriors.

Per als dos corpus, s'han obtingut un model de llenguatge de 2-grames i HMMs a nivell de caràcter utilitzant el conjunt d'entrenament de cadascun dels corpus. En els dos casos, els signes de puntuació van ser modelats com paraules separades. El conjunt de validació s'ha utilitzat per a ajustar els factors de reconeixement de GSF (Grammar Scale Factor) i WIP (Word Insertion Penalty). Per a l'estimació de la confiança, també s'ha optimitzat usant el conjunt de validació un paràmetre per a escalar les probabilitats del model de llenguatge. Els paràmetres optimitzats s'han usat en la fase de test.

La referència en la precisió de les transcripcions (sense supervisió) per a la base de dades d'IAM és sobre 69%. L'estimació de la confiança permet millorar la precisió fins un 80% supervisant només el 15% de les paraules reconegudes. Per a obtenir un 100% de precisió, necessitem supervisar el 69% de les paraules reconegudes. Per a aquesta base de dades, açò suposa una reducció de l'esforç humà de 7K paraules. Una altra manera de veureu es que quan es pot tolerar un xicotet nombre d'errors en les transcripcions, l'ús de les mesures de confiança pot ajudar a reduir dràsticament l'esforç de supervisió. Per exemple, en el cas d'aquesta base de dades, supervisant la meitat de les paraules reconegudes (12K) (la meitat de l'esforç humà) produeix una precisió en les transcripcions de 97%.

En el cas de la base de dades de GERMANA, s'han obtingut resultats similars. La precisió de referència es de 67% i supervisant el 16% de les paraules reconegudes la millorem fins al 80%. Supervisant la meitat de les paraules reconegudes (4K), obtenim una precisió del 96%.

Conclusions i treball futur

Hem presentat una estimació de confiança per a reduir l'esforç en la supervisió de transcripció interactiva de text manuscrit. S'han utilitzat probabilitats a posteriori a partir de grafs de paraula com a mesures de confiança. L'aproximació proposada ha estat provada utilitzant GIDOC amb les bases de dades de IAM i GERMANA. Hem mostrat com el ús de les mesures de confiança ens poden ajudar a reduir dràsticament el esforç de supervisió millorant la precisió de les transcripcions. Els resultats obtinguts en l'experimentació ens mostra que la precisió en la transcripció pot ser major del 95% reduint l'esforç humà a la meitat. El treball futur podria ser explorar nous camins d'ús de les

mesures de confiança en el paradigma interactiu. Es pot utilitzar diferents criteris per a validar les paraules que es reconeixen com a possibles errors.

4.2 Adaptació a partir de transcripcions parcialment supervisades

4.2.1 Introducció

La transcripció de text manuscrit de documents (antics) és una tasca important per a les biblioteques digitals. Açò es podria dur a terme, primerament processant totes les imatges del document, i després supervisant de manera manual totes les transcripcions a editar en les parts incorrectes. En canvi, les tecnologies de l'estat de l'art per a l'anàlisi automàtic de l'estructura de les pàgines (layout analysis), detecció de línies de text i reconeixement de text manuscrit estan encara lluny de la perfecció [30, 18, 5], i no està clar si és millor post editar l'eixida generada automàticament pel sistema o simplement ignorar-la.

Una aproximació més efectiva per a transcriure documents antics consta en seguir un paradigma interactiu-predictiu en el qual el sistema és guiat pel supervisor humà, i el supervisor humà és assistit pel sistema per a completar la tasca de transcripció tan eficientment com sigui possible. Aquesta aproximació de transcripció assistida per ordinador (CAT) ha estat probada.

Així doncs, hem usat de nou les mesures de confiança explicades en punts anteriors i els grafs de paraules per a reduir l'esforç humà a l'hora de localitzar els errors en les frases d'eixida que ens proporciona el sistema. Tot açò s'ha fet també mitjançant l'ús de GIDOC (Secció 3).

A continuació, una vegada el supervisor ha corregit algunes transcripcions, aquestes es podrien utilitzar per a millorar els models morfològics (HMMs) i model de llenguatge de la tasca, per exemple re-entrenant-los a partir de les dades anteriors i les noves ja corregides i validades pel supervisor. Per contra, si les transcripcions només s'han supervisat parcialment, aleshores l'error de reconeixement podria no millorar i tindre un efecte negatiu en l'adaptació del model. Ací anem a estudiar aquest efecte com una funció de grau de supervisió en dos tasques reals de text manuscrit de complexitat considerable. També anem a considerar tres estratègies d'adaptació (re-entrenament): a partir de totes les dades, sols a partir de les dades amb una confiança alta, i sols amb les dades de les parts supervisades. El re-entrenar a partir de les parts amb una confiança alta està inspirat en el treball de Wessel i Ney [36], en el qual les mesures de confiança s'usaven de manera satisfactòria per a restringir el aprenentatge no supervisat de models acústics per al reconeixement de veu continu amb vocabularis grans. En aquest treball, en canvi, les parts amb confiança alta inclouen tant les paraules no supervisades que estan per damunt d'un cert llindar de confiança com les paraules ja supervisades. Aquestes també s'utilitzen per a re-entrenar els HMMs i el model de llenguatge d'n-grames. D'altra banda, amb l'objectiu de simular les accions de l'usuari a diferents graus de supervisió, proposem un model d'interacció d'usuari simple però realista.

En les següents seccions 4.2.2 i 4.2.3 es descriu de manera breu l'experimentació

realitzada i els resultats obtinguts. Per a veure de manera més detallada tot aquest apartat corresponent a la secció 4.1 veure la Tesi de Master del company N.Serrano [23].

4.2.2 Model d'interacció d'usuari

Com ja s'ha dit en la introducció, proposem un model d'interacció d'usuari simple però realista per a simular les accions d'usuari a diferents graus de supervisió. El grau de supervisió es modela com el màxim nombre de paraules reconegudes (per línia) que són supervisades: 0 (no supervisat), 1, ..., ∞ (completament supervisat). S'assumeix que les paraules reconegudes són supervisades en un ordre de confiança no decreixent.

Amb l'objectiu de predir les accions d'usuari associades amb cada supervisió de paraula, primer havem calculat una distancia mínima d'edició (Levenshtein) entre la transcripció reconeguda i la real d'una línia de text donada. Com és usual, es consideren tres operacions d'edició elementals: substitució (d'una paraula reconeguda per una altra diferent), esborrat (d'una paraula reconeguda) i inserció (d'una paraula perduda en el reconeixement de la transcripció).

4.2.3 Experiments

Durant el seu desenvolupament, GIDOC ha estat utilitzat per un expert paleògraf per a anotar els blocs, línies de text i transcripcions en el nou conjunt de dades anomenat GERMANA (vist en la Secció 2).

Degut a l'estructura seqüencial del llibre, la tasca més bàsica en GERMANA consisteix en transcriure des del principi fins al final, encara que en aquest cas només considerem les transcripcions fins a la pàgina 180. Començant a partir de la pàgina 3, dividim GERMANA en 9 blocs consecutius de 20 pàgines cadascun (18 pàgines en el bloc 9). Els dos primers blocs (pp. 3-42) foren utilitzats per a entrenar un model d'imatge i un model de llenguatge inicials a partir de transcripcions totalment supervisades. Després, del bloc 3 al 8, cada bloc nou va ser reconegut, supervisat parcialment i afegit al conjunt d'entrenament del blocs precedents. Considerem tres graus de supervisió: 0 (no supervisat), 1 i 3 (nombre de paraules supervisades per línia). També, com s'ha indicat a la introducció, hem considerat tres estratègies d'adaptació (re-entrenament): a partir de totes les dades, sols amb les parts amb una confiança alta, i sols amb les parts supervisades.

Amb els resultats obtinguts, pareix clar que els models de referència (baseline) es poden millorar per adaptació a partir de transcripcions parcialment supervisades, encara que es requereix un cert grau de supervisió per a obtenir millores significatives. En particular, supervisant 3 paraules per línia ens duu a una reducció de més d'un 10% de WER respecte a aprenentatge no supervisat (models de referència), encara que el resultat es millorable ja que amb supervisió completa s'aconsegueix reduir un 5% més de WER (34%). L'estratègia d'adaptació, d'altra banda, té un menor efecte relatiu en el resultat. En canvi, pareix

millor no re-entrenar a partir de totes les dades, sinó sols amb les parts amb confiança alta, o simplement amb parts supervisades.

A part de l'experiment de dalt amb el GERMANA, férem un experiment similar en la coneguda base de dades d'IAM, utilitzant una partició estàndard per amb conjunt d'entrenament, validació i test [5]. El conjunt d'entrenament va ser dividit amb tres sub-conjunts; el primer s'usà per a entrenar els models inicials, mentre que les dos restants foren reconegudes, parcialment supervisades (4 paraules per línia) i afegides al conjunt d'entrenament. Els resultats obtinguts en termes de WER per al conjunt de test són: 42.6%, utilitzant sols el primer sub-conjunt; 42.8% després d'afegir el segon subconjunt; i 42.0% utilitzant també el tercer sub-conjunt. A diferencia del GERMANA, no hi ha reducció significativa en termes de WER després d'afegir dades parcialment supervisades al conjunt d'entrenament. Pensem que aquest resultat és degut a la naturalesa més complexa de la tasca d'IAM comparada amb GERMANA, el qual fa molt més complicat quan només una fracció del conjunt d'entrenament està disponible amb supervisió completa.

4.2.4 Conclusions

S'ha estudiat l'adaptació de models morfològics i models de llenguatge a partir de dades supervisades en el context de sistemes d'ajuda a la transcripció de text manuscrit. S'ha descrit un prototip anomenat GIDOC en el qual s'utilitzen mesures de confiança, estimades a partir de grafs de paraula, per a ajudar a l'usuari a localitzar errors de sistema. També s'ha proposat un model d'interacció d'usuari realista per a simular accions d'usuari a diferent nivell de supervisió. Finalment, s'han presentat els resultats empírics de dos tasques de dificultat considerables.

Capítol 5

Conclusions

En aquest treball s'han aportat varies contribucions en el camp de reconeixement de text manuscrit.

En el Capítol 2 s'ha presentat una base de dades de text manuscrit, GERMANA, per a facilitar comparacions empíriques entre diferents aproximacions d'extracció de línies de text i reconeixement de text manuscrit off-line:

ICDAR-2009: D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos and A. Juan. The GERMANA database. *Proceedings of the 10th ICDAR*. Barcelona (Spain). July 2009.

A continuació, al Capítol 3 s'ha presentat un prototip de transcripció assistida per ordinador per a documents antics de text manuscrit anomenad GIDOC. GIDOC és un primer intent de proporcionar suport integrat per a l'anàlisis d'estructura de pàgina (layout analysis) interactiva-predictiva, detecció de línies de text i transcripció de text manuscrit. L'efectivitat de GIDOC s'ha demostrat empíricament en la base de dades de GERMANA, la qual està també públicament disponible en Internet al igual que GIDOC. D'altra banda, també s'han presentat uns experiments realitzats amb aquest prototip amb la base de dades d'IAM en els quals s'ha tractat de reproduir el procés realitzat pel grup FKI i aportar algunes millores:

WEBIST-2010: N. Serrano, L. Tarazón, D. Pérez, O. Ramos-Terrades and A. Juan. The GiDOC Prototype (submitted). *Proceeding of WEBIST 2010*. Valencia (Spain). April 2010.

Per a finalitzar, les contribucions amb menor mesura es presenten al Capítol 4. En la primera d'elles Capítol 4.1, s'ha presentat una estimació de confiança per a reduir l'esforç en la supervisió de transcripció interactiva de text manuscrit. S'han utilitzat probabilitats a posteriori a partir de grafs de paraula com a mesures de confiança. L'aproximació proposada ha estat provada utilitzant GIDOC amb les bases de dades de IAM i GERMANA

provant empíricament que es por reduir l'esfoç humà. En l'altra contribució "menor" Capítol 4.2 s'ha estudiat l'adaptació de models morfològics i models de llenguatge a partir de dades supervisades en el context de sistemes d'ajuda a la transcripció de text manuscrit. També s'ha proposat un model d'interacció d'usuari realista per a simular accions d'usuari a diferent nivell de supervisió. Finalment, s'han presentat els resultats empírics per a les tasques IAM-DB i GERMANA, de dificultat considerables. Aquests són, respectivament, caduscún dels artícles corresponents a aquestes dues contribucions:

ICIAP-2009: L. Tarazón, D. Pérez, N. Serrano, V. Alabau, O. Ramos Terrades, A. Sanchis and A. Juan. Conficence Measures for Error Correction in Interactive Transcription of Handwritten Text. *Proceedings of the 15th ICIAP*. Vietri sul Mare (Italy). September 2009.

ICMI-MLMI-2009: N. Serrano, D. Pérez, A. Sanchís and A. Juan. Adaptation from Partially Supervised Handwritten Text Transcriptions. *In Proceedings of the Intelignet User Interface 2010*. Hong-Kong, China. February 2010.

Bibliografia

[1] Biblioteca Valenciana. http://bv.gva.es/.

[2] GNU Image Manipulation Program (GIMP). http://www.gimp.org/.

[3] Lionel Tarazón Alcocer. *Confidence Measures in Interactive Handwritten Text Transcription*. PhD thesis, Dep. de Sistemas Informàticos i Computación, Valencia, Spain, Dec 2009. Advisor(s): A. Sanchís and A. Juan.

[4] E. Belenguer, editor. *Germana de Foix, última reina de Aragón*. Univ. de València, 2007.

[5] R. Bertolami and H. Bunke. Hidden Markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition*, 41:3452–3460, 2008.

[6] R. Bertolami, M. Zimmermann, and H. Bunke. Rejection strategies for offline handwritten text recognition. *Pattern Recognition Letter*, 27:2005–2012, 2006.

[7] R.M. Bozinovic and S.N. Srihari. Off-line cursive script word recognition. 11(1):68–83, January 1989.

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 369–376, New York, NY, USA, 2006. ACM.

[9] Simon Günter. *Multiple Classifier Systems in Offline Cursive Handwriting Recognition*. PhD thesis, Institut für Informatik und angewandte Mathematik, Universität Bern, Bern, Switzerland, Jan 2004. Advisor: H. Bunke.

[10] Simon Günter and Horst Bunke. HMM-based handwritten word recognition: on the optimization of the number of states, training iterations and Gaussian components. *Pattern Recognition*, 37:2069–2079, 2004.

[11] Moisés Pastor i Gadea. *Aportaciones al reconocimiento automático de texto manuscrito*. PhD thesis, Dep. de Sistemes Informàtics i Computació, València, Spain, Oct 2007. Advisors: E. Vidal and A.H. Tosselli.

[12] `prhlt.iti.es/projects/handwritten/idoc/content.php?page=gidoc.php`., 2009.

[13] iTransDoc: Interactive Transcription and Translation of Old Text Documents. `prhlt.iti.es/itransdoc.php`., 2010.

[14] A. L. Koerich, R. Sabourin, and C. Y. Suen. Large vocabulary off-line handwriting recognition: A survey. *Pattern Analysis Applications*, 6(2):97–121, 2003.

[15] L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *IJDAR*, 9:123–138, 2007.

[16] L.M. Lorigo and V. Govindaraju. Offline arabic handwriting recognition: A survey. 28(5):712–724, May 2006.

[17] U.V. Marti and H. Bunke. The iam-database: an english sentence database for offline handwriting recognition. 5(1):39–46, 2002.

[18] D. Pérez et al. The GERMANA database. In *Proc. of ICDAR*, pages 301–305, Barcelona (Spain), 2009.

[19] Réjean Plamondon and Sargur N. Srihari. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Trans. on PAMI*, 22(1):63–84, 2000.

[20] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.

[21] O. Ramos, N. Serrano, and A. Juan. Interactive-predictive detection of handwritten text blocks. In *Proc. of DRR XVII*, San Jose, CA (USA), 2010.

[22] A. Sanchis. *Estimación y aplicación de medidas de confianza en reconocimiento automático del habla*. PhD thesis, Univ. Politécnica de Valencia, Spain, 2004.

[23] Nicolás Serrano Martínez Santos. *Adaptation and Interaction on Handwritten Recognition*. PhD thesis, Dep. de Sistemes Informàtics i Computació, València, Spain, December 2009. Advisor: Dr. Alfons Juan Císcar.

[24] N. Serrano, D. Pérez, A. Sanchis, and A. Juan. Adaptation from partially supervised handwritten text transcriptions. In *Proc. of ICMI-MLMI*, Cambridge, MA (USA), 2009.

[25] J.C. Simon. Off-line cursive word recognition. 80(7):1150–yy, July 1992.

[26] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Lang. Proc.*, pages 901–904, 2002.

[27] T. Su, T. Zhang, and D. Guan. Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text. *Int. J. of Document Analysis and Recognition*, 10:27–38, 2007.

[28] C.Y. Suen, C.P. Nadal, R. Legault, T.A. Mai, and L. Lam. Computer recognition of unconstrained handwritten numerals. 80(7):1162–1180, July 1992.

[29] L. L. Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *CoRR*, 9(2):123–138, April 2007.

[30] A. H. Toselli, A. Juan, et al. Integrated handwriting recognition and interpretation using finite-state models. *Int. J. of Pattern Rec. and Artif. Intell.*, 18(4):519–539, 2004.

[31] A. H. Toselli, A. Juan, J. González, I. Salvador, E. Vidal, F. Casacuberta, D. Keysers, and H. Ney. Integrated handwriting recognition and interpretation using finite-state models. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4):519–539, 2004.

[32] A. H. Toselli, V.Romero, L. Rodríguez, and E. Vidal. Computer Assisted Transcription of Handwritten Text. In *Proc. of ICDAR 2007*, pages 944–948, 2007.

[33] Alejandro Héctor Toselli. *Reconocimiento de Texto Manuscrito Continuo*. PhD thesis, Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia, Valencia (Spain), March 2004. Advisor(s): Dr. E. Vidal and Dr. A. Juan (in Spanish).

[34] U. v. Marti and H. Bunke. A full english sentence database for off-line handwriting recognition. In *In Proc. Int. Conf. on Document Analysis and Recognition*, pages 705–708, 1999.

[35] U. v. Marti and H. Bunke. Handwritten sentence recognition. In *In Proc. Int. Conf. on Pattern Recognition*, pages 467–470, 2000.

[36] F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Trans. on Speech and Audio Processing*, 13(1):23–31, 2005.

[37] K. Y. Wong, R. G. Casey, and F. M. Wahl. Document analysis system. *IBM Journal of Research and Development*, 26(6):647–656, 1982.

[38] Hanhong Xue and Venu Govindaraju. Hidden Markov Models Combining Discrete Symbols and Continuous Attributes in Handwriting Recognition. *IEEE Trans. on PAMI*, 28:458–462, 2006.

[39] S. Young et al. *The HTK Book*. Cambridge University Engineering Department, 1995.

[40] M. Zimmermann, J.C. Chappelier, and H. Bunke. Offline grammar-based recognition of handwritten sentences. 28(5):818–821, May 2006.

Índex de figures

Índex de taules