The SPARTACUS-Database: a Spanish Sentence Database for Offline Handwriting Recognition

Salvador España, María José Castro and José Luis Hidalgo

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
{sespana,mcastro,jhidalgo}@dsic.upv.es

Abstract

In this paper we describe a database that consists of offline handwritten Spanish sentences from four different subtasks. The database includes 1 500 forms produced by the same number of writers. A total of around 100 000 word instances out of a vocabulary of around 3 300 words occur in the collection.

This database is intended to be used for offline handwriting recognition tasks. However, this database is expected to be specially useful for recognition systems that may take advantage of language models of restricted-semantic tasks.

The database also includes a few image-processing procedures for extraction of handwritten text images from the forms and segmentation of the images into lines and words.

Keywords: Handwriting recognition, offline handwriting recognition, databases, Spanish sentences, linguistic knowledge.

1. Introduction

The availability of large amounts of data for training and testing is a fundamental prerequisite for building handwriting recognition systems. The acquisition and distribution of standard databases has therefore become an important issue in the handwriting recognition research community. Examples of widely used databases in the offline domain are CEDAR (Hull, 1994), NIST (Wilkinson et al., 1992), CENPARMI (Suen et al., 1992), IRONOFF (Viard-Gaudin et al., 1999) and IAM (Marti and Bunke, 2002).

This paper describes the first version (as of February 2004) of a new offline handwritten database, the SPARTACUS-database (which stands for *SPAnish Restricted-domain TAsk of CUrsive Script*), that contains full Spanish sentences. There are two main reasons to create this corpus. First of all, none of the above described databases contains Spanish sentences, whereas Spanish is a widespread major language. To our knowledge, the only publicly available Spanish database of offline handwritten text is described in (Juan et al., 2001) and (Toselli et al., 2004) and comprises 485 images of handwritten numbers by 29 writers and contains a total of 2 127 words, which is more than one order of magnitude smaller than the previously cited databases.

Other important motivation has been the availability, in the same corpus, of several levels of difficulty at the language model level. This feature is particularly useful for recognition tasks where linguistic knowledge beyond the lexicon level is used, because this knowledge can be automatically inferred from the underlying corpus. For instance, the HMM-based recognizers may integrate language models in the form of finite-state models (e.g., see, (Nagy, 2000), (Plamondon and Srihari, 2000) and the references therein).

The sentences of the SPARTACUS-database are extracted from four different subtasks. Three of these tasks are considered restricted-semantic tasks. Constrained semantic context tasks cover an important part of handwritten recognition applications, such as filling in forms in a registry office, bank checks, postal address processing, database querying, etc.

The entities at the lowest level in this version are words which have been automatically segmented and manually checked for correctness.

2. Corpus

Our goal was to acquire a restricted-semantic database of handwritten sentences in order to use linguistic knowledge in the recognition process. Also, we wanted to acquire short sentences from one line of written text instead of whole paragraphs to simplify the process of sentence extraction.

It was decided to use four different tasks to collect written text: "NUMERALS", "GDQ", "TRAVELER" and "GENERAL". The NUMERALS task consists of different quantities of numbers and prices printed with digits and expressed as the quantity in letters. Prices are expressed in unities of euro and dollar, some of them with fractional parts. Sentences of this task have been randomly generated and automatically written using the grammatical rules of Spanish numerals. Special care has been taken to limit the maximum length in order to fit every sentence into a single line (see Figure 1, the distribution of the NUMERALS task and some examples in Table 2). Three types of magnitudes were acquired: low prices with fractional part[1] (12,50€), high prices ($3 025 260) and high numbers (39 792 310). Despite the little size of the vocabulary (see Table 1), this corpus may be useful for the prominent task of *bank check reading*.

[1] In Spanish, the fractional part of a number is denoted by a comma instead of a dot used in English style.

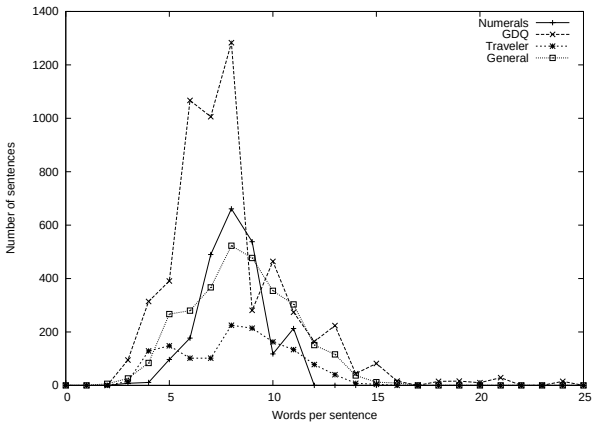

Figure 1: Distribution of the number of words per sentence.

Table 1: Number of sentences and words of each task.

Task	Sentences	Words	Vocabulary
NUMERALS	2 313	18 698	104
GDQ	5 790	46 112	247
TRAVELER	1 362	11 085	645
GENERAL	3 012	25 522	2 607
Total[a]	12 477	101 417	3 288

[a]Note that the size of the vocabulary is lower than the sum of the different vocabularies' sizes due to common words.

The GDQ ("Geographical Data Queries") task is extracted from a task-oriented Spanish speech corpus defined and recorded in the framework of the Albayzin Spanish project (Díaz-Verdejo et al., 1998). The selected application are queries to a geographical database. According to the conceptual scheme of the semantic universe (which can be consulted in (Díaz-Verdejo et al., 1998)), queries can be made about *regions*, *rivers*, *mountains*, *seas*, etc. Despite the fact that the GDQ task had been used in the speech recognition domain, the original acquisition of sentences was carried out by means of a written questionnaire distributed among a large set of people.

The TRAVELER task (Amengual et al., 2000) covers common sentences at a reception desk of a hotel by a traveler. For example, asking for rooms, wake-up calls, keys, the bill, moving the luggage, asking information about rooms, confirming a previous reservation, etc.

The last subtask, GENERAL, is unconstrained but is intended to cover the possible lacks of symbols, sequence of graphemes, and words not included in the other tasks.

The distribution of the number of words per sentence for each task is shown in Figure 1. Some statistics for each corpora and some examples are shown in Table 1 and Table 2, respectively.

3. Forms Design and Acquisition

As the main focus of our research activity is on high-level recognition, we wanted to make the image processing part as easily as possible. For the corpus we wanted to acquire, sentences are bound to a maximum width. This allowed us to acquire them separately in one line to simplify the process of sentence extraction.

An important restriction was the set of people available to fill the forms, mainly university students. Writers could not spent much time in the acquisition process, so nor special material neither long instructions could be imposed. For the same reason, only one single side A4 page forms were designed (with an average of 70 handwritten words per form). We preferred independent sentences in order to avoid errors in copying. Very light rectangles were used to guide the writing and were removed from the scanned image later. We also told the writers to stop writing if there was not enough space. Moreover, two different form types were generated: portrait and landscape forms. Longer sentences are collected into landscape forms in order to avoid getting compressed and deformed handwritten words.

No restrictions were imposed on the writing instrument. Hence, text produced with a number of different writing instruments is included in the database (mostly ink and ballpoint pens).

The forms were composed of a heading with a brief description of the purpose of the acquisition, a reference to the research project and an identifying code. An example of a filled form is illustrated in Figure 2. Every sentence is separated from the others by horizontal rulers. The typographic reference sentence and a guiding area to write into appear between two rulers.

In portrait forms, 10 sentences were acquired (6 GDQ sentences and 4 sentences from the TRAVELER and the GENERAL tasks); in landscape form, only 7 (2 GDQ sentences, 3 NUMERALS sentences and 2 sentences from the TRAVELER and the GENERAL tasks).

Different forms (1 500) were automatically generated by creating a LaTeX document containing the text and the structure of the form. The formatted documents were printed by a HP LaserJet 4100 DTN at a resolution of 600 dpi. The filled forms were scanned in grey level at 300 dpi. with a Scanner Hewlett Packard Scanjet ADF 6300c with automatic sheet feeder.

4. Text Extraction, Segmentation and Labeling

Line extraction from filled forms is easily performed by using horizontal projection thanks to the rulers included in the forms. These rulers may be detected by computing the longest horizontal black run and horizontal projections. The skew and the slant (Slavik and Govindaraju, 2001) have not been corrected.

A dynamic programming scheme has been used to segment every line into words. This algorithm try to align the sequence of words (which is known) to the sequence of ink and blank items obtained with a vertical projection of the line.

This automatic segmentation is manually supervised by using a graphic tool specially designed for this purpose. This supervised step is necessary because in some cases the handwritten text did not correspond to the printed text.

Table 2: Examples of sentences of each task (the English translation is provided).

NUMERALS	
Doscientos dólares con veintiséis centavos.	*Two hundred dollars and twenty six cents.*
Sesenta y nueve millones veintitrés mil novecientos.	*Sixty nine million and twenty three thousand and nine hundred.*
Cuarenta y siete mil seiscientos treinta euros.	*Forty seven thousand and six hundred and thirty euros.*
GDQ	
Dime el nombre de todas las comunidades que tienen mar.	*Tell me the name of every community which is on the sea.*
Quiero saber los nombres de los ríos más largos de 200 km.	*I want to know the name of rivers longer than 200 km.*
¿En qué comunidad desemboca el río Ebro?	*In which community flows into the Ebro river?*
TRAVELER	
Vamos a marcharnos hoy a las tres en punto de la tarde.	*We are going to leave today at three o'clock in the afternoon.*
Quiero que nos despierte mañana a las cuatro, por favor.	*Please, we want to be woken up tomorrow at four.*
¿Puede llevarnos nuestras bolsas al coche?	*Could you take our luggages to the car?*
GENERAL	
Cada generación tiene su derecho a la nostalgia.	*Every generation owns their rights to nostalgia.*
El ilustre artista estaba encantado.	*The renowned artist was delighted.*
Hoy en día la gente joven se va a trabajar a las ciudades.	*Nowadays young people leave to work in cities.*

ADQUISICIÓN DE ESCRITURA MANUSCRITA. Proyecto TIC-2000-1153 Código: 0490V

Esta muestra de escritura manuscrita servirá para ayudar a realizar y verificar sistemas de reconocimiento de escritura por ordenador. Por favor, escribe utilizando la zona sombreada como referencia, procurando no tocar la frase a copiar ni la línea inferior. Si te falta espacio, no hace falta que termines la frase.

Mares que bañan la comunidad gallega.

Mares que bañan la comunidad gallega.

¿Cuántas comunidades están bañadas por 2 mares?

¿ Cuántas comunidades están bañadas por 2 mares?

¿Cuántos ríos son más largos de 200 km?

¿ Cuántos ríos son más largos de 200 km?

¿En qué comunidad desemboca el río Ebro?

¿ En qué comunidad desemboca el río Ebro?

Comunidad autónoma más grande.

Comunidad autónoma más grande.

Dime las comunidades autónomas.

Dime las comunidades autónomas.

Esta vez no me había preparado en absoluto.

Esta vez no me había preparado en absoluto

¿Nos sube nuestras bolsas al autobús?

¿ Nos sube nuestras bolsas al autobús?

Hubo aclamación general.

Hubo aclamación general.

¿Puede llevarnos nuestras bolsas al coche?

¿ Puede llevarnos nuestras bolsas al coche?

"*This handwritten sample is intended to help the experimentation and testing of computer handwriting recognition. Please, write using the guiding rectangle as reference, trying not to touch the typographic text neither the bottom horizontal rule. If there is not enough space, sentence may be left unfinished.*"

Figure 2: An example of a filled vertical form. The translation of instructions to fill the form appears below.

Figure 3 shows an example of a sentence and the post-processed segmented words by using this tool. The automatic segmentation appears in the display and manual correction can be easily performed by editing the segmentation marks. Substitution, deletion and insertion errors (crossing-out words, for example) can be easily marked. Comments can be included in the sample (typos, for example).

Images of lines and words are stored in PNG format. Both extended ASCII (ISO-8859-1) and XML files containing information of position and labeling are also provided.

A proposal of partitions into several subsets is also provided aimed at defining training, validation and text subsets for each different subtask, or even for cross-validation experiments.

5. Conclusions

A database consisting of handwritten Spanish sentences has been described in this paper. It is mostly built upon semantic-restricted tasks (NUMERALS, GDQ and TRAVELER tasks). In this way, language models can be used to help the recognition process. Also, a few preprocessing and segmentation procedures have been developed together with the database.

Images corresponding to the whole written form, images corresponding to whole sentences and images corresponding to one word are available. The database described in this paper is freely available to other researchers upon request.

6. Acknowledgments

We would like to thank all the individuals who contributed to the database described in this paper, specially students and teachers from the Universidad Politécnica de Valencia and the Universitat Jaume I de Castelló.

This work is supported by the Spanish CICYT under contract TIC2000-1153.

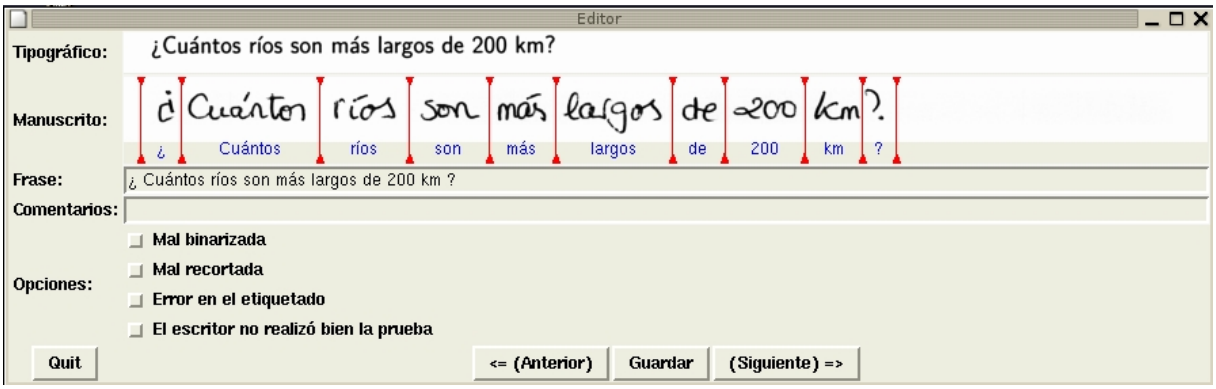

Figure 3: Automatic segmentation tool with manual supervision capability. Example of a sentence and the segmented words.

7. References

Amengual, J. C., J. M. Benedí, F. Casacuberta, A. Castaño, A. Castellanos, V. M. Jiménez, D. Llorens, A. Marzal, M. Pastor, F. Prat, E. Vidal, and J. M. Vilar, 2000. The EUTRANS-I speech translation system. *Machine Translation Journal*, 15:75–103.

Díaz-Verdejo, J. E., A. M. Peinado, A. J. Rubio, E. Segarra, N. Prieto, and F. Casacuberta, 1998. ALBAYZIN: a Task-Oriented Spanish Speech Corpus. In *First International Conference on Language Resources and Evaluation (LREC'98)*. Granada, Spain.

Hull, J. J., 1994. A database for handwritten text recognition research. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(5):550–554.

Juan, A., A.H. Toselli, J. Domnech, J. González, I. Salvador, E. Vidal, and F. Casacuberta, 2001. Integrated Handwriting Recognition and Interpretation via Finite-State Models. Technical Report ITI-ITE-01/1, Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, Valencia, Spain.

Marti, U. V. and H. Bunke, 2002. The IAM-database: an English sentence adatabase for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46.

Nagy, G., 2000. Twenty years of document image analysis in PAMI. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):38–62.

Plamondon, Réjean and Sargur N. Srihari, 2000. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):63–84.

Slavik, Petr and Venu Govindaraju, 2001. Equivalence of Different Methods for Slant and Skew Corrections in Word Recognition Applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(3):323–326.

Suen, C. Y., C. Nadal, R. Legault, T. A. Mai, and L. Lam, 1992. Computer recognition of unconstrained handwritten numerals. *Special Issue of Proc. IEEE*, 7(80):1162–1180.

Toselli, A. H., A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta, 2004. Integrated Handwriting Recognition and Interpretation using Finite-State Models. *International Journal of Pattern Recognition and Artificial Intelligence*.

Viard-Gaudin, Christian, Pierre Michel Lallican, Philippe Binter, and Stefan Knerr, 1999. The IRESTE On/Off (IRONOFF) Dual Handwriting Database. In *Fifth International Conference on Document Analysis and Recognition*.

Wilkinson, R., J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson, 1992. The first census optical character recognition systems conference. In *#NISTIR 4912*. The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD.