

An open diachronic corpus of historical Spanish: annotation criteria and automatic modernisation of spelling

Felipe Sánchez-Martínez, Isabel Martínez-Sempere,
Xavier Ivars-Ribes, Rafael C. Carrasco
Dep. Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03071, Alacant, Spain
fsanchez@dlsi.ua.es

Abstract

The IMPACT-es diachronic corpus of historical Spanish compiles over one hundred books —containing approximately 8 million words— in addition to a complementary lexicon which links more than 10 thousand lemmas with attestations of the different variants found in the documents. This textual corpus and the accompanying lexicon have been released under an open license (Creative Commons BY-NC-SA) in order to permit their intensive exploitation in linguistic research.

Approximately 7% of the words in the corpus (a selection aimed at enhancing the coverage of the most frequent word forms) have been annotated with their lemma, part of speech, and modern equivalent. This paper describes the annotation criteria followed and the standards, based on the Text Encoding Initiative recommendations, used to represent the texts in digital form.

As an illustration of the possible synergies between diachronic textual resources and linguistic research, we describe the application of statistical machine translation techniques to infer probabilistic context-sensitive rules for the automatic modernisation of spelling. The automatic modernisation with this type of statistical methods leads to very low character error rates when the output is compared with the supervised modern version of the text.

Keywords: diachronic corpus; historical Spanish; linguistic annotation; spelling modernisation

1 Introduction

Diachronic corpora are a valuable source of information with which to understand the historical evolution of languages. Unfortunately, diachronic collections are relatively scarce —at least, when compared to the overwhelming availability of resources containing transcriptions of modern text or speech (Kocjančič, 2009; Procházková, 2006; Davies, 2010a; Francis and Kucera, 1979)—, since creating a diachronic corpus is a costly task. The transcription of old texts must be manually reviewed because a number of features —such as spelling variations, old fonts, deprecated characters, and blurred text caused by stains or page transparency— may cause the accuracy of the automatic process of converting

printed text into computer-encoded text (commonly referred to as OCR or Optical Character Recognition) to fall below acceptable rates. In the particular case of Spanish, access to suitable linguistic resources can be challenging since the most renowned on-line diachronic resources, such as the *Corpus Diacrónico del Español* —CORDE (Real Academia Española, s.a.)—, and the *Corpus del Español* (Davies, 2002) can be consulted via Web interfaces¹ which provide limited querying capabilities (Davies, 2010b).

This paper describes the IMPACT-es diachronic corpus of historical Spanish and the accompanying lexicon created by IMPACT,² a research project funded by the European Commission under its seventh Framework Programme and focused on the improvement of the precision of OCR and the access to historical texts. The components developed in IMPACT include:

- Historical language resources for nine European languages which have allowed significant improvements to be made —up to a 30% (de Does and Depuydt, 2012)— in OCR word recall rates for historical documents in addition to increased productivity when used in combination with post-correction tools.
- A toolkit (Depuydt and de Does, 2009) with which to build lexical resources, tools for their deployment —both in OCR and information retrieval applications—, and named entity recognition tools for historical documents.
- Better OCR engines with improved technologies for image enhancement, binarisation, character recognition and algorithms for layout detection.
- A large ground-truth data set (a collection of images mapped onto extremely accurate transcriptions), coupled with a comprehensive evaluation toolkit.
- A framework of service and workflow layers (Neudecker et al, 2011) which enables full flexibility between all the IMPACT components.

In particular, the IMPACT-es corpus contains 107 Spanish texts first printed between 1481 and 1748 and covering a representative variety of creators and genres (prose, theatre, and verse). The digital versions in the corpus are based on early editions or faithful reprints of the early editions. This content is divided into two separate sections:

1. The GT section which compiles the 21 Spanish documents in the ground-truth data set created by IMPACT.
2. The BVC section which compiles 86 texts provided by the *Biblioteca Virtual Miguel de Cervantes*³ digital library and which have been partially annotated as will be described later.

Moreover, the corpus is complemented with a lexicon that links more than 10 thousand lemmas —corresponding to the most frequent word forms—, with a representative sample of attestations in the corpus. The full IMPACT-es corpus and the accompanying lexicon are available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license⁴ and can be downloaded at

¹See <http://corpus.rae.es/cordenet.html> and <http://www.corpusdelespanol.org>, respectively.

²IMproving ACcess to Text, <http://www.impact-project.eu>

³<http://www.cervantesvirtual.com>

⁴<http://creativecommons.org/licenses/by-nc-sa/3.0/>

the IMPACT Centre of Competence Website.⁵ A compact version of the lexicon, with no quotes from the corpus, has also been released under a dual license — Creative Commons Attribution-ShareAlike 3.0⁶ and GNU GPL v3⁷— in order to encourage its integration into free/open-source tools for the analysis and part-of-speech tagging of historical Spanish texts (Sánchez-Marco et al, 2011) such as FreeLing (Carreras et al, 2004).

To the best of our knowledge, this is the first diachronic corpus of historical Spanish distributed under an open license. In addition to the aforementioned CORDE and *Corpus del Español*, some other collections serve specific purposes. For example, the *Corpus de Documentos Españoles Anteriores a 1700* — CODEA (Sánchez-Prieto Borja, 2012)— contains over 1,500 documents, written between the 12th and 17th centuries, mainly collected from archives (for example, letters and administrative, legal and ecclesiastic documents). The project Website⁸ provides an interface through which to access the palaeographic and critical editions on a single-document basis. The *Corpus Histórico del Español en México* —CHEM, (Medina Urrea and Méndez Cruz, 2011)— collects documents written between the 16th and the 20th centuries, but the access to the collection is restricted to text visualisation.⁹

We hope that releasing the corpus under an open license will foster its exploitation and improvement by the linguistic research community in addition to its integration into natural language processing applications. As an illustration of the complementarity between linguistic digital resources and tools, we have explored the application of statistical machine translation techniques to the automatic modernisation of spelling. More precisely, as is described in Section 4, we have applied phrase-based statistical machine translation methods (Koehn, 2010, ch. 5) in order to infer probabilistic context-sensitive rules for the automatic modernisation of spelling. This method could be used as the transcription step of tools such as ToTrTaLe (Erjavec, 2011).

The following section provides an overview of the source documents in the IMPACT-es corpus and describes the criteria followed in their annotation. The markup format used to encode the documents that integrate the corpus and the lexicon is presented in Section 3. In Section 4 we illustrate the exploitation of the historical corpus in natural language processing applications. Finally, some concluding remarks are presented in Section 5.

2 Content and annotation criteria

The IMPACT-es corpus was created to assist in the improvement of OCR techniques by allowing the integration of specific vocabularies in the digitisation process in both information retrieval services and comprehensive evaluation frameworks. The selection of content and the annotation criteria have therefore been oriented towards this objective: for instance, the original spelling (even if clearly unintentional) has been preserved by the accurate transcription performed to create the ground-truth documents.

The GT section of the IMPACT-es corpus compiles 21 texts printed between

⁵<http://www.digitisation.eu/tools/language-resources/impact-es/>

⁶<http://creativecommons.org/licenses/by-sa/3.0/>

⁷<http://www.gnu.org/licenses/gpl.html>

⁸<http://demos.bitext.com/codea>

⁹<http://www.iling.unam.mx/chem/>

1543 and 1748, and contains around 6 million (unannotated) words with a transcription accuracy (as compared to the original text) of over 99.95%. The creator, title and dates of the first and source edition of the constituent documents are listed in Appendix A.

In order to ensure 99.95% fidelity to the source texts the transcription followed the standard *acceptance sampling* (Montgomery, 2009, part 6) statistical procedure for quality control: each document was processed in batches (containing between 500 and 1200 pages) in which each page was scanned and the automatic transcription obtained from the scanned image was manually corrected; when the transcription of a whole batch was complete, a sample containing about 4% of the pages was randomly selected and reviewed by an external quality control team; whenever the accuracy of the digitisation was found to be below 99.95%, the whole batch was rejected and its processing restarted.

The BVC section of the corpus compiles 86 texts printed between 1482 and 1990 which are listed in Appendix B together with the dates of the first and source editions. These documents contain approximately 2 million

words and a significant fraction of this content —over 27% of the section—, has been manually annotated with linguistic metadata.

The metadata added to Spanish words are lemma (in modern form), part of speech and modern equivalent. The words originating from other languages (less than 0.1%, and principally Latin) are labelled solely with their language. The part-of-speech categories and their tags, shown in Table 1 together with their relative frequencies, are based on those defined by the Apertium machine translation platform (Forcada et al, 2011) for the dictionaries in the Spanish–Catalan language pair.

The annotation was assisted by the CoBaLT tool (Kenter et al, 2012), a Web-based editor which supports the creation of corpus-based lexicons. The tool allows a common annotation to be assigned to a sequence of consecutive words, and also accepts compound lemmas and compound part-of-speech categories. The annotation of the corpus is based on the 22nd edition of the *Diccionario de la Lengua Española* (Real Academia Española, 2001a) which has served as a primary reference in determining the lemmas. For archaisms, the on-line version of the *Nuevo Tesoro Lexicográfico de la Lengua Española* (Real Academia Española, 2001b), which compiles Spanish historical dictionaries dating from 1726 (up to the 21st edition of the *Diccionario de la Lengua Española*), was employed as a secondary reference.

The corpus is accompanied by a lexicon which links more than 10 thousand entries (simple or compound lemmas) with their attestations in the BVC section of the corpus. The historical variants are classified under lemma, part of speech and modern form (see Figure 4 on page 10 for an example). Each occurrence of a variant is associated with the context in which it appears (10 preceding and 10 trailing words) and a reference to the document that contains it (title, author, dates of the first and source editions). The lexicon has been generated in parallel with the linguistic markup, and an ample coverage of lemmas and word forms was sought. Higher priority for their inclusion in the lexicon has therefore been given to the forms with a greater frequency. Since some words are ubiquitous, at most 500 occurrences per lemma have been annotated and therefore registered in the lexicon.

The following criteria have been applied during the annotation of the BVC section:

Table 1: Part of speech tags and their relative frequencies in the annotated part of the corpus.

Tag	Part of speech	Frequency
abr	abbreviation	0.03%
adj	adjective	10.70%
adv	adverb	2.92%
cnj	conjunction	1.08%
det	determiner	2.64%
ij	interjection	0.19%
n	noun	29.41%
np	proper noun	7.63%
num	numeral	0.39%
pr	preposition	1.59%
prn	pronoun	7.25%
rel	relative pronoun	0.17%
verb	verb	36.00%

- *Modern forms with compound lemmas:* word forms which cannot be associated with a simple lemma —such as verbs with enclitic pronouns— are marked with a compound lemma. For example *arrepentirse* has the compound lemma *arrepentir+se*.
- *Word boundaries:* whenever two or more consecutive words in the transcription correspond to a single modern form (for example *aun que* becomes *aunque*), the word group receives a shared annotation. Conversely, when one form corresponds to a sequence of modern forms (for example *quel* becomes *que el*), they are tagged with a compound lemma and a compound part of speech.
- *Archaisms:* when the reference dictionary (Real Academia Española, 2001a) registers a word as an archaism with a modern equivalent, the modern form has been preferred for the lemma. For example *apercibir* is the lemma assigned to the word form *apercebir*.
- *Contractions:* the modern word forms *al* and *del* have been assigned a compound part-of-speech tag (*pr+det*), but only a single lemma.
- *Numbers:* all cardinal numbers share a single part-of-speech category (called *num*). The original style —alphabetic characters, Arabic or Roman figures— has been preserved in the associated metadata. If they are split, as is the case of the historical variant *diez y seis* rather than the modern *dieciséis*, then they are handled as if they were different words in the sentence.
- *Optional diacritics:* optional diacritics used only for disambiguation (for example, in the words *sólo*, *éste*, *ése*, and *aqué*) have been preserved in the associated metadata (lemma and modern form).
- *Apocopation:* the full form has been preferred as the lemma for words with apocopation. For example *algún* is a form with the lemma *alguno*.
- *Past participles:* past participles have been classified as verbs only if they

```

<fileDesc>
  <sourceDesc>
    <bibl>
      <title>Primera parte de comedias del célebre poeta español,
        Don Pedro Calderón de la Barca</title>
      <author>Pedro Calderón de la Barca</author>
      <date type="first-edition">1685</date>
      <date type="source-edition">1685</date>
    </bibl>
  </sourceDesc>
  <editionStmt>
    <!-- continued (trailing content omitted) -->
  </editionStmt>
</fileDesc>

```

Figure 1: Fragment showing the bibliographical metadata encoded in the header of one of the documents.

follow an auxiliary verb (some cases of *haber* or *ser*), or they are not described as adjectives in the reference dictionary (e.g., *dormido*).

- *Proper nouns*: all lemmas and modern forms, with the exception of proper nouns, are written in lower case letters. The modern form of proper nouns, when available, has been preferred in the annotation to the word form in the source text; for example, the modern form and lemma of *Quixote* is *Quijote*.

3 Markup schema

The IMPACT-es corpus is distributed as a collection of XML files, each of which corresponds to a different work, and organised in one folder per section. The XML standard (World Wide Web Consortium, 2008) specifies how metadata must be inserted in a digital text in the form of tags which serve to identify the nested *elements* that make up the logical structure of the document. The names of the elements and their *attributes* (optional features whose value can make the meaning of the tag more specific) are defined by the annotation schema. In this case, the markup vocabulary follows the Text Encoding Initiative P5 guidelines.¹⁰

The TEI vocabulary is widely and actively used in digitisation projects in the area of humanities, example of which are: Europeana Regia,¹¹ the Perseus Digital Library,¹² or the British National Corpus.¹³ The TEI vocabulary defines a rich variety of elements such as paragraphs, words, and characters, in addition to a number of optional attributes —such as type or language— for each element, and only a reduced subset has therefore been employed in this case. Other corpora released in the scope of the IMPACT project use a similar TEI vocabulary (Erjavec, 2012).

¹⁰<http://www.tei-c.org/release/doc/tei-p5-doc/en/html>

¹¹<http://www.europeanaregia.eu>

¹²<http://www.perseus.tufts.edu>

¹³<http://www.natcorp.ox.ac.uk>

Every document in the corpus is encoded in a single XML file whose root element (with tag *TEI*) contains a header (under the element tag *teiHeader*) — with descriptive metadata of the document (marked as *fileDesc*)—, and a body (*body*) —with the digitised content structured in one or more divisions (elements *div*)—. The descriptive metadata element includes the bibliographical description of the source (under tag *bibl*) together with the information concerning the digital edition (under tag *editionStmt*). The bibliographic descriptions, illustrated in Figure 1, consist of a *title* element, an *author* element, the year of the first edition (as a *date* element with a *first-edition* value of the *type* attribute), and the year of the source edition (as a *date* second element with a *source-edition* value of the *type* attribute).

Figure 2 illustrates the elements and attributes employed in the annotation of the BVC section within a single main division (element with a *div* tag) for each document:

- Anonymous blocks (an element with an *ab* tag) contain one block of text (e.g., a paragraph or header) in the document.
- Every anonymous block contains one or more words (elements with a *w* tag). Foreign words have a single attribute defined, *xml:lang*, which stores the language of the word. In contrast, Spanish words are annotated with the following metadata:
 - The lemma is stored as the value of the *lemma* attribute.
 - The part-of-speech category is the value of the *type* attribute.
 - The original and modern equivalent are provided as the content of *orig* element tags (defined by TEI as an element with which to mark a piece of text as following the original) and *reg* (defined as an element with which to mark a reading which has in some respect been regularised) under the *choice* element tag.
- Punctuation characters (*pc* tag) and other characters (*c* tag) can appear between the words.

The subset of tags described above is similar to that used by Sánchez Marco et al (2009).

The GT section documents contain one division per page (see Figure 3) which accepts some sub-elements:

- Page number, drop capital, footnote and table-of-content entries are tagged as anonymous blocks with the specific values (*page-number*, *drop-capital*, *footnote*, and *TOC-entry*) of their *type* attribute.
- A heading is tagged as a *head* element.
- Paragraphs are marked with a *p* tag.

Other constituents of the source document, such as figures, catchwords —words placed at the foot of a page to anticipate the first word of the following page— or glosses were not digitised owing to the OCR orientation of the corpus.

Because of its specific nature, a different subset of elements defined by the TEI P5 guidelines (module “9 dictionaries”) has been employed for the lexicon (see Figure 4). Indeed, the body of the lexicon document consists of entries (elements with an *entry* tag) which contain:

```

<div type="pb">
  <ab type="p">
    <w lemma="comedia" type="n">
      <choice>
        <orig>Comedia</orig>
        <reg>comedia</reg>
      </choice>
    </w>
    <w lemma="del" type="pr_det">
      <choice>
        <orig>del</orig>
        <reg>del</reg>
      </choice>
    </w>
    <w lemma="príncipe" type="n">
      <choice>
        <orig>Príncipe</orig>
        <reg>príncipe</reg>
      </choice>
    </w>
    <w lemma="Inocente" type="np">
      <choice>
        <orig>Ynocente</orig>
        <reg>Inocente</reg>
      </choice>
    </w>
    <pc>.</pc>
    <w lemma="en" type="pr">
      <choice>
        <orig>En</orig>
        <reg>en</reg>
      </choice>
    </w>
    <w lemma="Madrid" type="np">
      <choice>
        <orig>Madrid</orig>
        <reg>Madrid</reg>
      </choice>
    </w>
    <w>a</w>
    <!-- continued (trailing words omitted) --->
  </ab>
</div>

```

Figure 2: Excerpt showing the content of one block element in the body of a TEI encoded document in the BVC section of the corpus.

```

<div type="page">
  <ab type="page-number">Num. 28.</ab>
  <head>LA GRAN CENOBIA.</head>
  <head>COMEDIA FAMOSA.</head>
  <head>De Don Pedro Calderon de la Barca.</head>
  <head>PERSONAS QUE HABLAN EN ELLA.</head>
  <p>Aureliano.
    Decio.
    Libio, Infante.
    <!-- continued (trailing text omitted) --->
  </p>
  <!-- continued (trailing paragraph and anonymous blocks omitted) --->
  <ab type="drop-capital">E</ab>
  <p>Spera sombra mia, palida imagen de mi
    <!-- continued (trailing text omitted) --->
  </p>
</div>

```

Figure 3: Excerpt showing one page element of a document in the GT section.

- A lemma as an element with a *form* name (a form is defined by TEI as an element which groups all the information concerning the written and spoken forms of one headword) and the *lemma* value of its *type* attribute.
- A number of modern variants of the lemma, labelled as elements with a *form* name and a *modern-form* value of their *type* attribute.

On the one hand, every *form* element of a *lemma* type contains:

- The lemma under the *orth* TEI element (which is defined as the orthographic form of a dictionary headword).
- The part-of-speech category (under the *gram* element that provides grammatical information in a *gramGrp* element).
- The number of annotated occurrences in the collection, given by the content of a *lbl* element with an *occurrences* type.

On the other hand, every *form* element of a *modern-form* type contains:

- The orthographic variant, as the content of an *orth* element.
- One or more *form* elements with *historical-form type* which contains the historical forms (orthographic variant) as the content of an *orth* sub-element and a number of attestations as the content of a *cit* sub-element.

Finally, every attestation contains the following information:

- The bibliographical information (within the *bibl* element), i.e. the reference of the work cited.
- A number of quotes within the *quote* element in which the historical form is labelled with the *oVar* tag.

```

<entry xml:id="id80" n="abrazarle-vblex_lprn">
  <form type="lemma">
    <orth>abrazarle</orth>
    <gramGrp>
      <gram type="pos">vblex</gram>
      <gram type="pos">prn</gram>
    </gramGrp>
    <lbl type="occurrences">30</lbl>
  </form>
  <form type="modern-form">
    <orth>abracele</orth>
    <form type="historical-form">
      <orth>abr cele</orth>
      <cit>
        <bibl>
          <title>Segunda Celestina</title>
          <author>Feliciano de Silva</author>
          <date type="first-edition">1536</date>
          <date type="source-edition">1536</date>
        </bibl>
        <quote><!-- text omitted--> y <oVar>abr cele</oVar> ay, <!-- text omitted--></quote>
      </cit>
    </form>
    <form type="historical-form">
      <orth>abra cele</orth>
      <cit>
        <bibl>
          <title>Viaje del Parnaso</title>
          <author>Miguel de Cervantes Saavedra</author>
          <date type="first-edition">1614</date>
          <date type="source-edition">1614</date>
        </bibl>
        <quote><!-- text omitted--> y <oVar>abra cele</oVar>, en la <!-- text omitted--></quote>
      </cit>
    </form>
  </form>
  <form type="modern-form">
    <orth>abrazandole</orth>
    <form type="historical-form">
      <orth>abra andole</orth>
      <!-- continued (trailing citations omitted) -->
    </form>
    <form type="historical-form">
      <orth>abra andoles</orth>
      <!-- continued (trailing citations omitted) -->
    </form>
  </form>
  <!-- continued (trailing word forms omitted) -->
</entry>

```

Figure 4: Example of an entry in the lexicon.

4 Automatic modernisation of spelling

The lack of normalisation in the spelling of old texts poses a challenge to information retrieval systems, since users cannot include all the possible variants of every word in their queries. It is also difficult for the non-expert to interpret such documents, and modernised and critical editions are therefore often produced to facilitate reading. Automatic modernisation seeks to minimise the cost of creating modern editions by using rules for the updating of spelling which can be either provided by experts in palaeography (following a deductive approach) or can be induced from examples when large corpora are available.

For instance, some deductive methods use phonetic matching techniques in order to generate alternative spellings (Baron and Rayson, 2008). In contrast, inductive methods for language processing assume no prior linguistic knowledge but require a large amount of data to identify a suitable model for the transformation rules (Manning and Schütze, 1999). Updating the spelling of a text can clearly be seen as a particular case of translation or text rewriting (Mihov and Schulz, 2007), although, in contrast to the translation between divergent languages, rules to convert the spelling operate at the character level rather than at the word level. For example, a common modernisation rule replaces every long s (the old character “s”) with a standard s.

We have therefore explored the applicability of inductive machine translation techniques to the task of updating the old forms in a document. The statistical machine translation approach (Koehn, 2010) would appear to be a natural candidate for this task since it can deal with the most important features of modernisation:

- It is an asynchronous process, that is, it cannot be achieved on a letter-by-letter basis (for example, the digraph “ph” usually becomes “f”).
- It is non-deterministic because the replacements can differ even in the same context (for example, the spelling “fijo” must be sometimes preserved, whereas in other cases it must become “hijo”).
- It is essentially a monotonous process, in the sense that the transformation of a character will not show a long range dependence on the context.

Alternative models, such as finite-state transducers (Oncina et al, 1993) are better suited for deterministic transductions. The traditional input-output HMMs are not prepared to handle asynchronous transductions (Bengio and Frasconi, 1994) and, although they can be extended for that purpose, the training phase has considerable computational costs (Bengio and Bengio, 1996). Furthermore, since there are a number of open-source implementations of phrase-based statistical machine translation methods (Koehn, 2010, ch. 5) which can be used to test natural language applications, we have explored the applicability of these methods to the task of updating the spelling.

4.1 Method overview

The phrase-based approach translates a sentence s by maximising the probability of the result t . The probability $p(t|s)$ is defined in terms of a linear combination of a number of *feature functions*:

$$p(t|s) = \exp \sum_k \lambda_k h_k(s, t). \quad (1)$$

Table 2: Size of the training, development and testing subsets (for source documents).

Subset	Words	Characters
Training	599,126	3,739,262
Development	5,000	31,416
Test	5,000	31,239

Typical feature functions $h_k(s, t)$ are the logarithms of source-to-target and target-to-source *phrase*¹⁴ translation probabilities, logarithms of source-to-target and target-to-source token translation probabilities, reordering costs, the output length, the number of phrases used in the translation, and the logarithm of the likelihood of the output as provided by a *target-language model*.

The inference process in the SMT approach consists of the following steps:

1. The feature functions are estimated using a parallel corpus —more precisely, the *training* subset— after token alignment and phrase extraction (Zens et al, 2002).
2. The weights λ_k are tuned in order to optimise the translation quality on a held-out parallel corpus —the *development* subset— using the minimum error rate training (MERT) algorithm (Och, 2003). This optimisation is traditionally oriented towards maximising the popular BLEU score (Papineni et al, 2002), an automatic measure of translation quality.

After training, the translation for a sentence s is selected by looking for the target-language sentence t which maximises $p(t|s)$.

As noted previously, modernisation can be regarded as a type of translation in which sequences of characters that already constitute phrases and words play the traditional role of sentences: A training sample thus consists of pairs of words (the source and the modern forms). For phrase extraction (phrases being groups of characters often transformed together) the characters can be aligned with a very simple procedure: first the longest common sub-sequence (Hirschberg, 1975) is used to discover which characters are identical in both forms and can be aligned without further character swaps; then, the remaining sub-sequences are aligned if their alignment does not cross over the one-to-one alignments obtained in the first step. This simple procedure provides suitable results and avoids the costly training phase required by standard statistical alignment methods (Och and Ney, 2003).

4.2 Experiments and results

The set of samples obtained from the lexicon was split into training, development and testing subsets, with the sizes shown in Table 2. Interestingly, the number of characters in the target subsets (the modern words) is slightly smaller (by only 0.3%) than the number of characters in the source documents, even if they contain an identical number of words.

In order to evaluate its performance, the statistical approach has been compared with four other methods:

¹⁴In the phrase-based approach, any segment of text, even without a syntactic coherence, is called a phrase.

Table 3: Character error rate (as a percentage) for the automatic modernisation of spelling.

Method	CER
No modernisation	5.75%
Naive	0.50%
Spellchecker	5.91%
Spellchecker + dictionary	5.27%
SMT	0.21%

1. The text remains as it is in the source file or, in other words, no modernisation takes place.
2. A naive approach selects the most frequent modernised form for every word which is also in the training subset and preserves the source form in those cases in which no reference is found in the training subset.
3. The source text is updated using the suggested correction provided by a modern spell checker (Ispell version 3.3.02 with the Spanish dictionary version 1.11).
4. The source text is updated with a modern spell checker with the enhanced coverage provided by the list of words in the modern part of the training subset.

The accuracy of these methods is measured using the *character error rate* (CER), defined as the minimal number of characters that need to be modified (inserted, removed or replaced) in order to transform the output into the target, normalised with the total number of characters in the target.

As is shown in Table 3, an average of 5.75% of the characters must be modified in the source text in order to obtain the modernised spelling. It is worth noting that the naive approach achieves a high accuracy by simply selecting the most common spelling, although this method fails with all unseen words that require modernisation (0.34% in the test set), since in these cases the input word is copied verbatim to the output.

In contrast, the replacements suggested by the spell checker cannot be used to modernise the spelling, since the error rate remains comparable to the rate obtained with the unmodified text. This low performance is not originated by an insufficient lexical coverage, since the addition of a specific dictionary does not reduce the number of mistakes. Corrections based on word similarity do not therefore appear to be adequate for the modernisation of spelling.¹⁵

The lowest error rate is clearly obtained with the application of the SMT technique, which reduces the CER to less than one half of that obtained with the naive approach. This accuracy suggests that the statistical method identifies the essential rules needed to transform segments of characters into their modern spelling and it can deal with new, unseen words.

In our settings, the SMT system generated translation rules of any length up to 8 letters, a value which showed the best compromise between the accuracy of results and model complexity.

¹⁵Ispell corrections are based on the Damerau–Levenshtein distance (Lowrance and Wagner, 1975), a measure which enhances the traditional edit distance (Levenshtein, 1965) by permitting the transposition of adjacent characters.

Table 4: Character error rate (CER) for the SMT approach when N characters are added before and after the word in old Spanish whose spelling is to be modernised. The row labelled “unaligned” shows the results when the context characters are left unaligned, the one labelled “aligned” shows the results when these characters become aligned (see running text).

N	0	1	2	3	4	5
Unaligned	0.21%	0.27%	0.36%	0.60%	1.41%	3.41%
Aligned	0.21%	0.28%	0.32%	0.60%	2.19%	7.15%

About one half of the rules learnt only transferred the input to the output verbatim but the others produced transformations like the following: “eio→ejo”, “zys→cís”, “euo→evo”, “ço→zo”, “çe→ce”, “sçe→ce”, “nuio→nvio”, “vbe→ube”, or “xu→ju”. These type of rules are similar to those proposed by experts (Sánchez-Marco et al, 2010).

A set of additional experiments has been performed in order to identify the influence of neighbouring words on the modernisation process, which may be especially important in languages that exhibit external *sandhi* (Matthews, 1997) or similar effects on spelling (such as some Indian and Celtic languages), and may help to disambiguate those cases in which a word has more than one possible modernisation.

Table 4 shows the results obtained when the words in each sample are contextualised by adding, in addition to a special character representing blank spaces, N characters from the previous word as a left context and N characters from the following one as a right context. The left and right context characters receive different codes to the word characters and could remain unaligned or they could be aligned with the initial and last character, respectively. Consider, for example, the training sample with $N = 2$ “e_ll_i fijo d_re_r→hijo”. The characters in this training sample can be aligned in two different ways: one that leaves “e_l”, “l_i”, “d_r”, “e_r” and the blank spaces unaligned, and another that aligns all the characters in the left context to the first letter of the modernised form (“h”), and all the characters in the right context to the last letter of the modernised form (“o”).

The experiments show that, in the case of Spanish, the best results are obtained when no context is added ($N = 0$). The analysis of the output reveals that often, especially in the case of a large N , many words never appear in the training subset with an identical context to the test subset and, in such cases, the quality of the translation deteriorates.

5 Concluding remarks

The IMPACT-es open diachronic corpus of historical Spanish contains approximately 8 million words and has been released under an open license (Creative Commons BY-NC-SA). We have described the criteria applied for the linguistic annotation —of nearly 7% of the words in the corpus— with lemmas, parts of speech and modern equivalents.

The corpus is divided into two sections: the BVC section (from the *Biblioteca Virtual Miguel de Cervantes* digital library), which has been manually annotated, and the GT section (developed by the IMPACT project), which has been digitised with a fidelity of 99.95% to the original. Furthermore, a lexicon has been

extracted from the BVC section. Every entry in the lexicon corresponds to one lemma and part of speech, and contains a sample of variants and their attestations in the corpus.

The application of phrase-based statistical machine translation techniques for the modernisation of spelling has a very high accuracy —the character error rate is below 0.25%— and demonstrates the complementarity of diachronic corpora and linguistic applications. The transformation rules which have been automatically identified are analogous to those identified by experts. However, it is worth noting that the statistical machine translation approach does not apply modernisation rules based only on the probability of the translation rules, but rather on a combination of features such as the target language likelihood. The best accuracy is achieved when the words are modernised without considering the context characters from neighbouring words. This observation is consistent with the fact that 99.11% of the words in the training subset have only one possible modernisation.

This automatic modernisation achieves sufficient accuracy for the development of useful tools to assist in the production of modernised and critical editions, or to alleviate the difficulties that searching and retrieving texts with multiple historical variants creates.

A Content: the GT section

Author: Title	First edition	Source edition
<i>Anonymous:</i>		
Vida de Lazarillo de Tormes	1554	1652
<i>Francisco de Quevedo:</i>		
El Parnasso español	1648	1648
<i>Garcilaso de la Vega:</i>		
Obras de Garcilasso de la Vega con las anotaciones por el Mtro. Francisco Sánchez Brocense	1574	1612
<i>Inca Garcilaso de la Vega:</i>		
Commentarios reales	1609	1609
<i>Jorge Juan:</i>		
Observaciones astronomicas y phisicas hechas de orden de S. M. en los Reynos del Peru	1748	1748
<i>Juan Boscán:</i>		
Las obras de Boscán y algunas de Garcilasso de la Vega repartidas en cuatro libros	1543	1543
<i>Lope de Vega:</i>		
Las comedias del famoso poeta Lope de Vega	1604	1604
<i>Luis de Góngora:</i>		
El Polifemo de Don Luis de Góngora with comments by Don García de Salzedo	1629	1629
<i>Mateo Alemán:</i>		
Vida y hechos del pícaro Guzmán de Alfarache	1599	1681
<i>Miguel de Cervantes Saavedra:</i>		
El ingenioso hidalgo Don Quixote de la Mancha	1605	1605
<i>Pedro Calderón de la Barca:</i>		
Primera parte de comedias del célebre poeta español, Don Pedro Calderón de la Barca	1685	1685
<i>Real Academia Española de la Lengua:</i>		
Diccionario de la lengua castellana [...] Tomo primero. Que contiene las letras A, B	1726	1726
Diccionario de la lengua castellana [...] Tomo segundo. Que contiene la letra C	1729	1729
Diccionario de la lengua castellana [...] Tomo tercero. Que contiene las letras D, E, F	1732	1732
Diccionario de la lengua castellana [...] Tomo cuarto. Que contiene las letras G, H, I, J, K, L, M, N	1734	1734
Diccionario de la lengua castellana [...] Tomo quinto. Que contiene las letras O, P, Q, R	1737	1737
Diccionario de la lengua castellana [...] Tomo sexto. Que contiene las letras S, T, V, X, Y, Z	1739	1739
<i>Ruy López de Sigura:</i>		
Libro de la invención liberal y arte del juego del Axedrez	1561	1561

Continued on next page ...

Author: Title	First edition	Source edition
<i>San Juan de la Cruz:</i>		
Obras del venerable y mistico Dotor F. Joan de la Cruz	1629	1629
<i>Santa Teresa de Jesús:</i>		
Los libros de la Madre Teresa de Jesús	1588	1588
<i>Sor Juana Inés de la Cruz:</i>		
Carta athenagorica	1690	1690

B Content: the BVC section

Author: Title	First edition	Source edition
<i>Baltasar Gracián:</i>		
Oráculo manual y arte de la prudencia	1647	1647
<i>Beato Juan de Ávila:</i>		
Epistolario espiritual	1578	1962
<i>Cristóbal de Castillejo:</i>		
Dialogo de mujeres	1544	1544
Obras morales y de devoción	1542	1958
<i>Diego Sánchez de Badajoz:</i>		
Farsa de Abraham	1554	1554
Farsa de la muerte	1554	1554
Farsa racional del libre alvedrío	1554	1554
<i>Feliciano de Silva:</i>		
Segunda Celestina	1536	1536
<i>Fernando Rojas:</i>		
La Celestina	1499–1502	1499, 1514
<i>Fernán Pérez de Oliva:</i>		
Dialogo de la dignidad del hombre	1585	1586
<i>Francisco de la Torre:</i>		
Poesías	Various	1969
<i>Francisco Delicado:</i>		
La Lozana Andaluza	1528	1528
<i>Gabriel Lobo Lasso de la Vega:</i>		
Tragedia de la honra de Dido restaurada	1587	1587
<i>Guillén de Castro:</i>		
Las Mocedades del Cid	1605–1615	1618
<i>Íñigo de Mendoza:</i>		
Coplas de Vita Christi Frayy	1482	1482
<i>Juan Boscán:</i>		
Obra completa	Various	1917
<i>Juan Cortés de Tolosa:</i>		
El desgraciado	1617	1620
El nacimiento de la verdad	1617	1620
La Comadre	1617	1620
Novela del licenciado periquín	1617	1620
Novela de un miserable llamado Gonzalo	1617	1620
<i>Juan de Encina:</i>		
Égloga representada en la noche postrera de Carnal	1496	1496
Aucto del repelón	1509	1509
Égloga de Cristino y Febea	1509	1509
Égloga de Fileno, Zambardo y Cardonio	1509	1509
Égloga de las grandes lluvias	1507	1507
Égloga de Mingo, Gil y Pascuala	1496	1496
Égloga de Plácida y Vitoriano	1513	1962

Continued on next page ...

Author: Title	First edition	Source edition
Representación sobre el poder del amor	1507	1507
<i>Juan de Mena:</i>		
Laberinto de Fortuna	1481	1505
<i>Juan Ruiz de Alarcón y Mendoza:</i>		
El antichristo	1634	1990
El desdichado en fingir	1628	1990
El dueño de las estrellas	1634	1990
El tejedor de Sevilla	1634	1990
Examen de maridos	1634	1990
Ganar amigos	1634	1990
La amistad castigada	1634	1990
La crueldad por el honor	1634	1990
La cueva de Salamanca	1628	1990
La industria y la suerte	1628	1990
La manganilla de Melilla	1634	1990
La prueba de las promesas	1634	1990
Los empeños de un engaño	1634	1990
Los pechos privilegiados	1634	1990
Mudarse por mejorarse	1628	1990
Todo es ventura	1628	1990
<i>Lope de Vega:</i>		
Comedia del Príncipe Ynocente	1590	1762
<i>Luis Vélez de Guevara:</i>		
La serrana de la Vera	1613	1916
<i>Miguel de Cervantes Saavedra:</i>		
Comedia del cerco de Numancia	1615	1615
Comedia famosa de la casa de los zelos y seluas de Ardenia	1615	1615
Comedia famosa del gallardo español	1615	1615
Comedia famosa del laberinto de amor	1615	1615
Comedia famosa de los baños de Argel	1615	1615
Comedia famosa de Pedro de Vrdemalas	1615	1615
Comedia famosa intitvlada el rvfian Dichoso	1615	1615
Comedia famosa intitvlada la gran svltana doña Catalina de Ouiedo	1615	1615
Comedia llamada Trato de Argel	1615	1615
Don Quijote de la Mancha (1ª parte)	1605	1605
Don Quijote de la Mancha (2ª parte)	1615	1615
Entremes de la cueua de Salamanca	1615	1615
Entremes de la eleccion de los alcaldes de Daganço	1615	1615
Entremes de la guarda cuydadosa	1615	1615
Entremes del juez de los diuorcios	1615	1615
Entremes del retablo de las marauillas	1615	1615
Entremes del rufian viudo, llamado Trampagos	1615	1615
Entremes del viejo zeloso	1615	1615
Entremes del vizcayno fingido	1615	1615

Continued on next page ...

Author: Title	First edition	Source edition
La entretenida	1615	1615
La Española inglesa	1613	1613/1614
La Galatea	1585	1585
Novela de la Fuerça de la sangre	1613	1613/1614
Novela de la Gitanilla	1613	1613/1614
Novela de la Ilustre Fregona	1613	1613/1614
Novela del amante liberal	1613	1613/1614
Novela de las dos Donzellas	1613	1613/1614
Novela de la Señora Cornelia	1613	1613/1614
Novela del Casamiento engañoso	1613	1613/1614
Novela del Licenciado Vidriera	1613	1613/1614
Novela del Zeloso extremeño	1613	1788
Novela de Rinconete y Cortadillo	1613	1788
Novelas exemplares	1613	1613/1614
Novela y coloquio que passó entre Cipion y Bergança, perros del hospital de la Resurreccion	1613	1613/1614
Ocho comedias y ocho entremeses nuevos	1615	1615
Persiles y Sigismunda	1617	1617
Poesías sueltas	1615	1615
Viaje del Parnaso	1614	1614

Acknowledgements

Work funded by the European Commission under the Seventh Framework Programme (FP7) through the IMPACT (IMproving ACcess to Text) project. We thank Mikel L. Forcada for his fruitful suggestions.

References

- Baron A, Rayson P (2008) VARD 2: A tool for dealing with spelling variation in historical corpora. In: Proceedings of the Postgraduate Conference in Corpus Linguistics, Birmingham, UK
- Bengio S, Bengio Y (1996) An EM algorithm for asynchronous input/output hidden Markov models. In: Proceedings of the International Conference on Neural Information Processing, ICONIP, Hong Kong, pp 328–334
- Bengio Y, Frasconi P (1994) An input output HMM architecture. In: Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994], MIT Press, pp 427–434
- Carreras X, Chao I, Padró L, Padró M (2004) FreeLing: an open-source suite of language analyzers. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, pp 239–242
- Davies M (2002) Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *Procesamiento del Lenguaje Natural* 29:21–27

- Davies M (2010a) The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4):447–464
- Davies M (2010b) Creating useful historical corpora: A comparison of CORDE, the Corpus del Español, and the Corpus do Português. In: *Diacronía de las lenguas iberorromances: nuevas perspectivas desde la lingüística de corpus*, Vervuert/Iberoamericana, Frankfurt, Germany/Madrid, Spain, pp 137–166
- Depuydt K, de Does J (2009) *Fons Verborum*. Feestbundel voor prof. dr. A.M.F.J. (Fons) Moerdijk, aangeboden door vrienden en collega’s bij zijn afscheid van het INL, Instituut voor Nederlandse Lexicologie, Leiden/Amsterdam, chap *Computational Tools and Lexica to Improve Access to Text.*, pp 187–199
- de Does J, Depuydt K (2012) Lexicon-supported OCR of eighteenth century Dutch books: a case study. In: *Proceedings of the 20th Document Recognition and Retrieval Conference*, San Francisco, CA USA, (to appear)
- Erjavec T (2011) Automatic linguistic annotation of historical language: ToTr-TaLe and XIX century Slovene. In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, Portland, OR, USA, pp 33–38
- Erjavec T (2012) The goo300k corpus of historical Slovene. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA), Istanbul, Turkey
- Forcada ML, Ginestí-Rosell M, Nordfalk J, O’Regan J, Ortiz-Rojas S, Pérez-Ortiz JA, Sánchez-Martínez F, Ramírez-Sánchez G, Tyers FM (2011) Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* 25(2):127–144
- Francis WN, Kucera H (1979) *Brown corpus manual*. Online at <http://www.hit.uib.no/icame/brown/bcm.html>
- Hirschberg DS (1975) A linear space algorithm for computing maximal common subsequences. *Communications of the ACM* 18(6):341–343
- Kenter T, Erjavec T, Dulmin MZ, Fiser D (2012) Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In: *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, Avignon, France, pp 1–6
- Kocjančič P (2009) Internet y los recursos lingüísticos para la lengua española: diccionarios y corpus. *Verba hispanica: anuario del Departamento de la Lengua y Literatura Españolas de la Facultad de Filosofía y Letras de la Universidad de Ljubljana* 17:145–164
- Koehn P (2010) *Statistical Machine Translation*. Cambridge University Press
- Levenshtein VI (1965) Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR* 163(4):845–848, English translation in *Soviet Physics Doklady*, 10(8), 707–710, 1966.

- Lowrance R, Wagner RA (1975) An extension of the string-to-string correction problem. *Journal of the Association for Computing Machinery* 22(2):177–183
- Manning CD, Schütze H (1999) *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA
- Matthews PH (1997) *The Concise Oxford Dictionary of Linguistics*. Oxford University Press
- Medina Urrea A, Méndez Cruz CF (2011) El corpus histórico del español en México. *Revista Digital Universitaria* 12(7):3–25
- Mihov S, Schulz KU (2007) Efficient dictionary-based text rewriting using sub-sequential transducers. *Natural Language Engineering* 13(4):353–381
- Montgomery DC (2009) *Introduction To Statistical Quality Control*. John Wiley & Sons
- Neudecker C, Schlarb S, Dogan M, Missier P, Sufi S, Williams A, Wolstencroft K (2011) An experimental workflow development platform for historical document digitisation and analysis. In: *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, Beijing, China, pp 161–168
- Och FJ (2003) Minimum error rate training in statistical machine translation. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp 160–167
- Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51
- Oncina J, Garcia P, Vidal E (1993) Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(5):448–458
- Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, Philadelphia, USA, pp 311–318
- Procházková P (2006) *Fundamentos de la lingüística de corpus*. Concepción de los corpus y métodos de investigación con corpus. Available online at http://prochazkova.de/fundamentos_de_la_lingüística_de_corpus.pdf
- Real Academia Española (2001a) *Diccionario De La Lengua Española*, 22nd edn. Espasa Calpe, Madrid, online at <http://lema.rae.es/drae>
- Real Academia Española (2001b) *Nuevo tesoro lexicográfico de la lengua española*, 1st edn. Espasa Calpe, Madrid, online at <http://buscon.rae.es/ntlle/SrvltGUILoginNtlle>
- Real Academia Española (s.a.) Banco de datos CORDE, corpus diacrónico del español. Online at <http://corpus.rae.es/cordenet.html>, last accessed 2012.09.24

- Sánchez Marco C, Boleda G, Fontana JM (2009) Propuesta de codificación de la información paleográfica y lingüística para textos diacrónicos del español. uso del estándar TEI. In: Proceedings of the Congreso Internacional Tradición e innovación: Nuevas perspectivas para la edición y el estudio de documentos antiguos, Madrid, Spain
- Sánchez-Marco C, Boleda G, Fontana JM, Domingo J (2010) Annotation and representation of a diachronic corpus of Spanish. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, La Valleta, Malta, pp 2713–2718
- Sánchez-Marco C, Boleda G, Padró L (2011) Extending the tool, or how to annotate historical language varieties. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland, OR, USA, pp 1–9
- Sánchez-Prieto Borja P (2012) Desarrollo y explotación de un corpus de documentos españoles anteriores a 1700 (CODEA). *Scriptum Digital* 1:5–35
- World Wide Web Consortium (2008) Extensible markup language (XML) 1.0 (fifth edition). Online at <http://www.w3.org/TR/2008/REC-xml-20081126>
- Zens R, Och FJ, Ney H (2002) Phrase-based statistical machine translation. In: Proceedings 25th Annual German Conference on Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol 2479, Springer-Verlag, pp 18–32