

Building spell-checking facilities for ancient Spanish

Alejandro Bia and Manuel Sánchez-Quero
Miguel de Cervantes Digital Library, University of Alicante

June 13, 2001

Abstract. Being aware of the usefulness of spell-checkers on the correction of modern works, and lacking this facility for ancient texts, we decided to build dictionaries for ancient Spanish. This decision led to new problems and new questions. We have built a time-aware system of dictionaries that takes into account the temporal dynamics of language, to help solve the problem of ancient Spanish spell-checking. In this paper we present the problems we have found, the decisions we have made and the conclusions and results we arrived at.

Keywords: spell-checking, digital libraries, NLP resources, dictionaries, automatic learning

1. Introduction

The huge development of information technology has motivated the appearance of this new type of libraries, called digital libraries (Arms, 2000). The Miguel de Cervantes Digital Library ¹ is one of the most ambitious projects of its kind ever to have been undertaken in the Spanish-speaking world with more than 5000 digital books at present. This enormous collection of digitized works mainly comprises Hispanic classics from the 12th up to the 20th century. The development of these digital books require a lot of care from the point of view of correction and editing, but are processed in a massive uniform way afterwards, so as to produce the different publications formats and services offered to the readers. Concerning human resources involved in the project, the biggest group by far corresponds to correction and markup people (Bia and Pedreño, 2001), who are in charge of the hardest-to-automate part of the production process, which involves reading and correcting digitization errors, structurally marking up the texts, and taking important editing decisions that involve both rendering and functionality of the hypertext documents to be published. These humanists are highly skilled people with at least a bachelor degree in philology, or other humanistic disciplines. We want them to devote their time to higher intellectual tasks like taking editing or markup decisions, or preparing the texts for interesting Internet services (like text analysis or concordance queries), than to spend their energies in the tedious

¹ <http://cervantesvirtual.com>



mechanical task of correction, the main bottleneck in our production workflow (Bia, 2000), and by far the most time-consuming task.

In the case of contemporary works, spell-checkers turned out to be a useful aid to the correction process, but for literary works written in ancient Spanish, commercially available modern spell-checkers may produce more mistakes than they can prevent. The reason for this is that spell-checker-dictionaries include only modern uses of the language, and when they are applied to old texts, the result is that they take correct ancient uses of words for mistakes and try to correct them. Unable to use spell-checking as an aid, correctors have to do a side by side comparison of the original and the digitized texts to detect the errors.

2. Previous work

As Carretero and Rodríguez correctly state (Carretero and Rodríguez, 1996): “Most of the largest operating systems and information retrieval tools do not provide lexical tools for Spanish [let alone for ancient Spanish], which makes it very difficult to check the information fed to the systems. This problem is specially critical in big organizations (libraries, museums, etc.) where information is acquired mechanically by scanning or typing, which adds to the already existing errors, new errors caused by the mechanization process.” This problem gets worse in the case of ancient Spanish.

The need to filter large amount of information written in Spanish led these Spanish researchers to build COES (Rodríguez and Carretero, 1999; López et al., 1999), a complete environment that allows the users to deal with the grammatical problems of modern Spanish, which is based on Ispell, and it is freely distributed since 1994.

3. Spell-checking

Ispell is a good example of open spell-checking software. It is indeed a very old program. The original was written in PDP-10 assembly in 1971, by R. E. Gorin. The C version was written by Pace Willisson of MIT and Walt Buehring of Texas Instruments added the emacs interface and posted it to the net. There are versions of this spell-checking software for the most common operating systems nowadays.

What led us to create dictionaries of ancient Spanish Language was the impossibility to check the spelling of ancient Spanish literary works with commercial software packages. The problem of correcting spelling in older texts has received scarce consideration.

Since we could not find old-Spanish spell-checkers available, ancient texts had to be thoroughly revised and corrected after OCR without any automated aid.

Being aware of the usefulness of spell-checkers on the correction of modern works, and lacking this facility for ancient texts, we decided to build dictionaries for ancient Spanish. This decision led to new problems and new questions. As there is no such thing as ancient Spanish, but instead a dynamically evolving language that changes throughout the centuries, how many old-Spanish dictionaries should we build? Should we set arbitrary chronological limits?

In this paper we present the problems we have found, the decisions we have made and the conclusions and results we arrived at. We have also been able to extract statistical information on the evolution of the Spanish language through time.

4. Time periods in Spanish language

Constructing just one dictionary to cover all ancient Spanish language made no sense due to the scarce normalization of the language in ancient times, and above all, the many different ways in which words could be written in different epoches. This is briefly but well explained on *The Spanish Language Home-Page* (Zamora, 2001).

We can distinguish four big time periods in the evolution of Spanish language which also involve important changes in spelling.

- Medieval age (up to 1500)
- Golden age (16th-17th centuries)
- 18th and early 19th centuries
- late 19th and 20th centuries

According to the *Ortografía de la Lengua Española* (Real Academia Española, 1999), the orthographic normalization of the Spanish language is the result of a long process of adaptation and simplification of the varied and variable ancient uses. It can be said that such normalization dates back to 1713, when the *Real Academia Española de la Lengua* was created, but the definitive settlement of the language took effect in 1844, with the publication of the *Prontuario de ortografía de la lengua castellana*. Therefore, works from 18th and 19th centuries, although not being modern Spanish, are closer to it, tending to use normalized Spanish which is more precise concerning spelling.

According to Rafael Cano (Cano-Aguilar, 1988), “with the 18th century, it can be said that the main constitutive historic processes of the Spanish language conclude. From then on, we have not only ‘modern Spanish’, but, above all, a language that has reached its stability: on one hand concerning geographical spread, because Spanish has not acquired new territories but has obtained a firm settlement in those it had already possessed (specially the American continent); on the other hand, the main lines of language structure had not varied: neither in the phonetic nor in the morphosyntactic field can any fundamental alteration be outlined in this period, and the basic vocabulary keeps being, in general, the same”.

5. The corpora used

Taking advantage of the 5,000 books already digitized and corrected at the Miguel de Cervantes Digital Library (University of Alicante, 2001), as a corpus covering several centuries of Spanish writings, and using also the TESO (Chadwyck-Healey, 1997) corpus with more than 800 works on Golden-Age Spanish drama, we have built a set of dictionaries taking into account the temporal dynamics of language, to help solve the problem of ancient Spanish spell-checking.

6. A *Golden-Age* dictionary

In our digital library, we encountered the necessity to correct and publish a significant amount of texts from the 16th and 17th centuries (Golden Age). In a parallel way, there are good corpora of Spanish literary texts from this age available on the web to be taken as raw material to create the spell-checker dictionary. This is why we first focused just on the language of these two centuries. The corpus we used to extract the entries for the Golden Age dictionary is the TESO (*Teatro Español del Siglo de Oro*).

We started by building a dictionary from *Golden-Age* drama texts downloaded from the TESO ².

We have processed 12,316,640 words, 113,572 numbers, 1,344,962 sentences, 7,038,024 tags, from 842 HTML files, 120 Mb in size from 16 different authors:

- Calderón de la Barca, Pedro (1600–1681)

² <http://teso.chadwyck.com/>

- Castro, Guillén de (1569–1630)
- Cervantes Saavedra, Miguel de (1547–1616)
- Cueva, Juan de la (1550–1609?)
- Diamante, Juan Bautista (1625–1687)
- Matos Fragoso, Juan de (1608?–1689)
- Moreto, Agustín (1618–1669)
- Pérez de Montalbán, Juan (1602–1638)
- Quiñones de Benavente, Luis (1593?–1652)
- Rojas Zorrilla, Francisco de (1607–1648)
- Rueda, Lope de (1510–1565)
- Ruiz de Alarcón y Mendoza, Juan (1581–1639)
- Solís, Antonio de (1610–1686)
- Téllez, Fray Gabriel (Tirso de Molina) (1579–1648)
- Vega, Lope de (1562–1635)
- Zamora, Antonio (1662?–1728)

6.1. PREPROCESSING OF THE SOURCE TEXTS

Before starting to build the Golden-Age Spanish dictionary there are some preprocesses we cannot skip. The first aspect we had to consider was deciding which kind of text format we were going to use: ASCII, ANSI, ISO entities, etc. Then we analyzed whether the corpus was in HTML, XML or just plain text. In the case of the drama texts extracted from TESO all the texts were retrieved in HTML format.

Most of the times, the texts used as a source to build the word lists need to be preprocessed. This is particularly true in Golden-Age texts, since they have remarkable peculiarities compared to modern Spanish.

One of the preparation tasks necessary to build a dictionary is the analysis of character encoding to see which characters are used, and how they should be translated for the target use.

Golden-Age Spanish contains ancient symbols that had to be coded as computer characters following an arbitrary convention as for instance:

- The intervowel stressed *s*, coded as β in TESO, which we prefer to code as *ss* in our texts.
- The ε , which we prefer to code as *ae*.
- The use of contractions like q' or t' , forbidden in modern Spanish, were common in the 16th and 17th centuries. So this words should count as *que* and *te* for word frequency purposes.
- The use of $\hat{}$ as in *nôbres* to represent an *n* or *m* sound following the consonant, in the example *nombres*.³ We preferred to leave this ones unmodified.
- Another ancient character no longer used is ζ , equivalent to modern *z*, as in *açumbres*⁴.
- The use of accents is completely anarchic in these centuries, finding acute and grave accents used indistinctively, or simply omitted.
- The same chaotic usage is found in the use of umlauts $\ddot{}$, which appear in unexpected places (e.g. *crüel* instead of *cruel* may be used to mark an unusual accent to accomplish a certain metric).
- Finally, the indistinct use of *u* and *v*, or of *b* and *v* in this epoch, produces many variations of the same words, which enlarge the dictionary.

7. Rendering, Structure and Semantics

A cleaning process was also necessary to free the TESO texts from some fragments which were written in modern Spanish, like page markers, which could be identified by the following regular expression:

”[Página”[0-9]*”]”

This process consisted of two well-defined parts. The first one was cleaning the TESO texts from some fragments which were written in modern Spanish. To perform this, we used some HTML tags as a pointer to establish which parts of the whole TESO text should be removed and which not. As the rendering was very consistent throughout the whole collection, all modern Spanish fragments could be detected and eliminated.

³ This appeared for instance in “*Hazer del contrario amigo (1861)*”

⁴ In “*El licenciado vidriera (1676)*”

In the second part of the process we had to take out all the HTML marks ($\langle p \rangle$, $\langle head \rangle$, etc.) from the text, since they should not appear as entries in our dictionary.

One of the problems we had was having to separate drama speaker names from the rest of the text. These names were usually rare abbreviations of real names which we decided should not appear in the dictionary. As speaker names repeatedly appear in a drama text, the frequency of these words is not low, getting mixed with common words instead of appearing as rare unusual words. In a 160.000 words list, this makes it difficult to detect them by hand. A different case would have been if speaker names were marked as such, as it is done with highly structural markup schemes like TEI (Sperberg-McQueen and Burnard, 1994). In our case we used a combination of HTML rendering marks and capitalization to detect speaker names.

There is nothing more unrelated, we may think, as rendering and semantics. But in a few cases, rendering may give us a reliable hint on structure, which also may imply some semantic information. This was the case of speaker names in the TESO files (Chadwyck-Healey, 1997). In these, we applied the same method successfully used in (Bia and Muñoz, 2000) to extract information from poorly marked up Galician language texts based in part on rendering markup. In our case, Speaker names could be reliably detected with a regular expression like:

```
"<BR><STRONG><EM>"[ ]?({U}|{N})?{L}{0,7}[. ,]?
([ ]?({U}|{N})?{L}{0,7}[. ]?)?{0,3}"</EM></STRONG>"
```

where L (lowercase letters), U (uppercase letters) and N (numbers) are:

```
L [a-zñçáéíóúäëïöüâêîôû ...]
U [A-ZÑÇÁÉÍÓÚÄËÏÖÜÂÊÎÔÛ ...]
N [0-9]
```

This allowed us to detect speaker names as: “*Inigo.*”, “*Rom.*”, “*Ellos*”, “*1. Cor. Mus.*”, “*2. Cor.*”, “*Maria Ies.*”, “*Guevara.*”, “*D. Isab.*”, “*Los 4.*”, “*Niñ.*”, “*Cos.*” y “*Bez.*”, “*loc. Mus. y Po.*”, “*Ley de Grac.*”, etc.

The number of speaker names not accurately detected, or detected words that were not speakers was very low, and the program also highlighted these cases for human correction. We solved this problem by fixing HTML markup in the source files for these few cases. This approach was much simpler than having to detect and eliminate all speaker names by hand.

For the texts from the Miguel de Cervantes DL corpus was not necessary so much preprocessing work, since on one hand most of the problems mentioned above are due to Golden-Age spelling and are not found in the following centuries, and on the other hand working directly with the XML sources made it easier to separate modern Spanish comments and notes from the ancient text. Figure 1 shows the process of

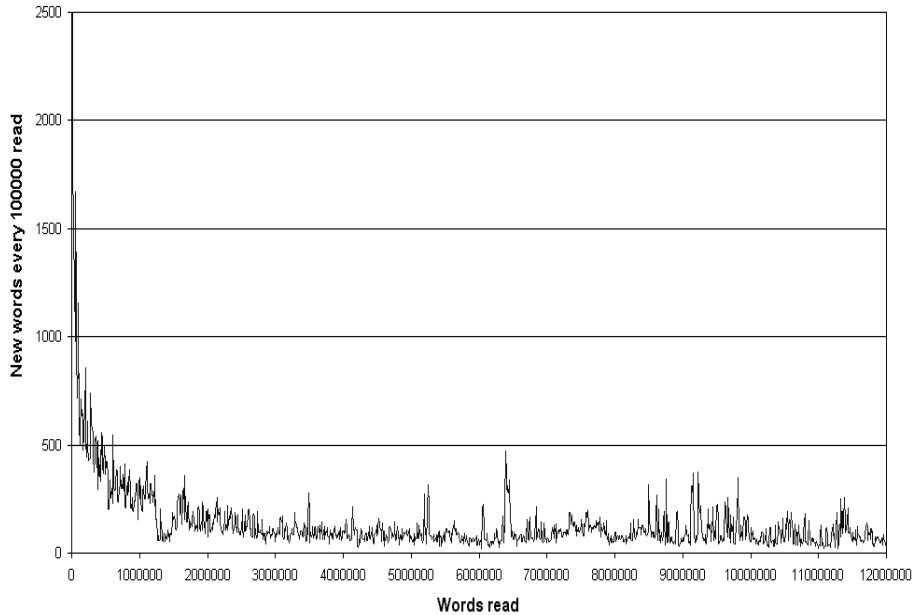


Figure 1. As more words are read, new words are more difficult to find

new-word acquisition from the corpus. Counting the new words that appeared every 100,000 words read, we realized that at the beginning most of the words were new, but as the program kept on reading words it became harder to find new ones. From time to time there is a peak in the graphic line, due to the appearance of a literary work with a richer vocabulary that adds new words all of a sudden.

8. Results

We have produced many dictionary formats, with or without frequency counts, ordered alphabetically, by word-size-first, or by frequency, and ignoring or considering letter case.

Table I shows the sizes of each corpus used and the corresponding dictionaries created.

Table I. Dictionaries built in this research work

Texts source	Centuries	Corpus words	Dict. words (case-sensitive)	Dict. words (all lowercase)
TESO	16 th -17 th	12,316,640	200,920	166,648
Cervantes DL	18 th	4,478,694	154,917	133,372
Cervantes DL	19 th	29,062,010	366,372	307,083
Cervantes DL	20 th	2,590,567	117,153	103,544

The word-size ordered dictionary is successfully used for spell-checking using the WinEdt text editor, that supports open dictionaries of this format.

According to Barnbrook (Barnbrook, 1996): “A frequency list for a corpus shows you the words that occur in it and the relative proportion that each contributes towards it. If your corpus is properly representative, this information can give you a reasonably accurate picture of the language as a whole. For the lexicographer this means that the words to be included in the dictionary can be selected objectively with their frequency of occurrence giving a guide to their relative importance.”

9. Conclusions: A Historical Surprise

M.J. Woods in a recent article (Woods, 2001) has compared the “word frequencies of the few most common words in Spanish as revealed by a modern corpus of over five [million] words with a corpus of Golden-Age Spanish texts of over a million words, and finds that although *de* is by far the most common word in contemporary Spanish, in the 16th and 17th Centuries it was considerably less frequent, and in many texts was less frequent than *y* or *que* for which shared very similar frequency figures. It is argued that this significant change in the Spanish language comes about in the 20th Century.”

These values as well as ours can be seen in table II. We confirmed Woods’ observation that the word “*que*” is more frequently used in the Golden-Age than in the 20th century. However, *de* is the second runner in our Golden-Age dictionary based on the TESO corpus of 12,316,640 words. Woods suggests that “as a wider range of electronic texts in Spanish becomes available it will be interesting to try to pinpoint when the change in characteristic frequencies comes about.” According to our results (see figure 2), the switch in position between these two words occurs somewhere between the 17th and 18th centuries, long before

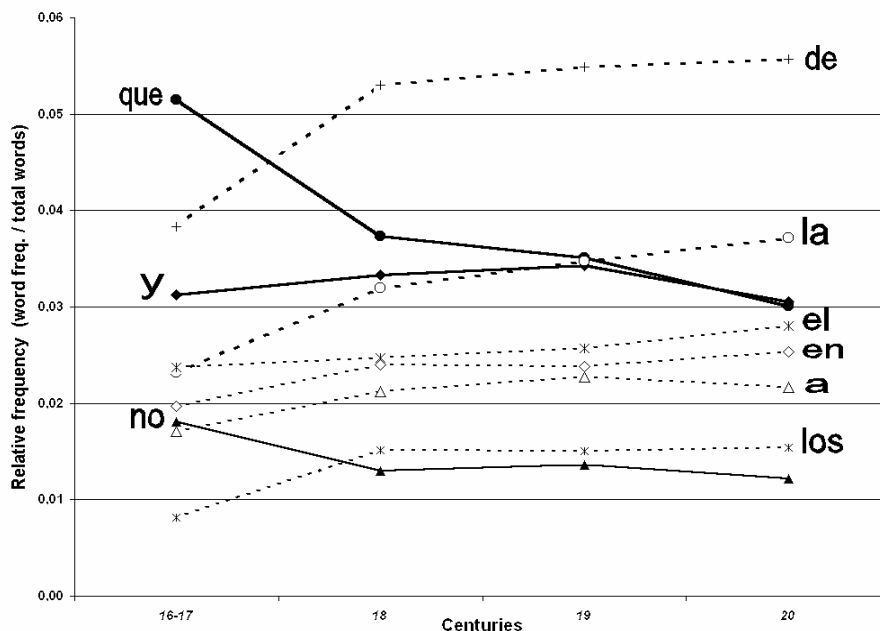


Figure 2. Evolution of the nine most frequent words in Spanish

Woods' hypothesis: "the frequency of *de* ... may be a reliable guide as to whether or not it was composed before the twentieth century." He sustains that "to judge by the statistics for Galdos' novel *Nazarín* (1895), the real change may not have come about until this century", but we believe that as Galdos was a very unusual writer with a vocabulary by far richer than that of his contemporaries and a very personal style, he should not be taken as a norm for statistical comparison.

From the graph, we can see that some words as "*que*", "*y*" and "*no*" descend in the frequency ranking, while others like "*de*" and "*la*", and in a minor degree "*el*", "*en*", "*a*", and "*los*" gain ground. This information can be used for instance as an additional criterion to take into account in automatic document classification-by-time approaches.

Another curious observation resulting from figure 2, is the stable behaviour of those nine most frequent words from the 18th century on. This coincides with the normalization of the Spanish language by the *Real Academia Española de la Lengua*.

Table II shows the 20 most frequent words in each of our dictionaries along with the figures used by Woods in his article.

10. Future work

10.1. CONSIDERING GEOGRAPHICAL ZONES

After having carried out these experiments we have come to the conclusion that it is necessary for the purposes of a Digital Library like ours to consider geographical zones or areas apart from time periods. The reason for this is that old Spanish is very irregular both in spelling and in its use, which varies depending on the geographical zones of linguistic influence, and this makes very wide the range of possibilities when spell-checking. A mistake in a text can be overlooked because in other texts it is perfectly correct, as it is the right spelling in other zone. This can be the case of the verb “*hacer*” which can be written in old Spanish almost in every way: “*fazer*”, “*faer*”, “*faser*”, “*acer*”, “*azer*”, “*hazer*”, “*haer*”, etc.

Moreover, spell-checking dictionaries aware of geographical zones are useful because the language although with some differences is quite normalized within an area.

Besides, for a project like cervantesvirtual.com, which is very focused in Latin American Literature, it can be more profitable the creation of a dictionary with native Latin American words and expressions which are not included in conventional Spanish spell-checkers.

10.2. AUTOMATIC MODERNIZATION OF TEXTS

Automatic modernization of texts is another possible objective, since we usually produce modernized transcriptions of facsimiles along with the ancient normal transcription based literarily on the original .

10.3. BETTER SPELLING USING NLP

According to Barnbrook (Barnbrook, 1996) “the spell checking routines found in most major word processing packages would benefit greatly from an ability to disambiguate the senses of the words being verified. If a misspelling in a document is the correct form of another word the spell checker will normally accept it, since they usually operate purely on the basis of a comparison of the forms found in the text with a list of acceptable forms”. Usual spell-checking software does not detect, for instance, the incorrect use of Spanish words like *número*, *numero* and *numeró*, or *fórmula* and *formula*, where the only difference in spelling is an accent, but lexically some are nouns and other verb forms. Barnbrook continues saying that “the use of a semantic disambiguator within the spell checking software would make it possible for

the computer to carry out something much closer to automatic proof-reading”. Automatic sense disambiguation is an interesting research subject and a problem not yet fully solved.

Fortunately, in the case of our digital library, most texts are first generated by OCR output, which make the type of errors very distinct and different in most cases from human mistypings. In this environment a standard spell-checking approach, but based on an ancient dictionary of the time of the text, has given excellent practical results, reducing the time and effort necessary to correct ancient texts.

Acknowledgements

Special thanks to Elena Pellus, Sergio Galindo and Javier Fresnillo for their linguistic advice and encouragement.

Notes

¹ This is an example of an endnote.

² And another one.

References

- Arms, W.: 2000, *Digital Libraries*. Cambridge, Massachusetts: MIT Press.
- Barnbrook, G.: 1996, *Language and Computers*, Edinburgh Textbooks in Empirical Linguistics. 22 George Square, Edinburgh EH8 9LF: Edinburgh University Press.
- Bia, A.: 2000, ‘Automating the Workflow of the Miguel de Cervantes Digital Library’. Poster at the ACM 2000 Digital Libraries conference (Fifth ACM Conference on Digital Libraries).
- Bia, A. and R. Muñoz: 2000, ‘Aplicación de Técnicas de Extracción de Información a Bibliotecas Digitales (Applying Information Extraction Techniques to DLs)’. In: M. V. Ferro (ed.): *Proceedings of the XVI Conference of the SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural)*, Vol. 26. University of Vigo, Spain, pp. 207–214. (published in: *Procesamiento del Lenguaje Natural*, journal of the SEPLN).
- Bia, A. and A. Pedreño: 2001, ‘The Miguel de Cervantes Digital Library: The Hispanic Voice on the WEB’. *LLC (Literary and Linguistic Computing) journal*, *Oxford University Press* **16**(2), 161–177. Presented at ALLC/ACH 2000, The Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, 21/25 July 2000, University of Glasgow.
- Cano-Aguilar, R.: 1988, *El Español a través de los tiempos.*, Vol. 1. Juan Bautista de Toledo 28, Madrid: Arco Libros S.A.
- Carretero, J. and S. Rodríguez: 1996, ‘Building Lexical Tools to Manage Information Written in Spanish’. *Journal of Information Science* **22**(5), 391–399.

- Chadwyck-Healey: 1997, 'TESO (Teatro Español del Siglo de Oro)'. <http://teso.chadwyck.com/>. A corpus of Spanish drama from the XVIIth and XVIIth centuries, Copyright 1997-98 Chadwyck-Healey Spain.
- López, M., S. Rodríguez, and J. Carretero: 1999, 'Herramientas ortográficas de libre distribución para la lengua castellana.'. In: *II Congreso Hispalinux*.
- Real Academia Española: 1999, *Ortografía de la Lengua Española*. R.A.E. Edition revised by the different Academies of the Spanish Language.
- Rodríguez, S. and J. Carretero: 1999, 'Corrector ortográfico de libre distribución basado en reglas de derivación.'. In: *Primer encuentro del grupo de usuarios de TeX hispanohablantes. EGUTH'99*. pp. 44-52.
- Sperberg-McQueen, C. M. and L. Burnard (eds.): 1994, *Guidelines for Electronic Text Encoding and Interchange (Text Encoding Initiative P3), Revised Reprint, Oxford, May 1999*. Chicago - Oxford: TEI P3 Text Encoding Initiative.
- University of Alicante: 1999-2001, 'Biblioteca Virtual Miguel de Cervantes Saavedra (Miguel de Cervantes Digital Library)'. <http://cervantesvirtual.com>.
- Woods, M. J.: 2001, 'Spanish Word Frequency: A Historical Surprise'. *Computers and the Humanities* **35**, 231-236.
- Zamora, S.: 2001, 'Historia del idioma español. Los orígenes y el desarrollo de nuestra lengua.'. <http://www.el-castellano.com/historia.html>. Site a.k.a. *The Spanish Language Home-Page*.

Table II. The 20 most frequent words in different epoches

Corpus:	TESO 16 th /17 th		MCDL 18 th		MCDL 19 th		MCDL 20 th		Woods 16 th /17 th		CCS 20 th	
Num. of words:	12,316,640		4,478,694		29,062,010		2,590,567		1,268,445		5,250,351	
Position	word	freq.	word	freq.	word	freq.	word	freq.	word	freq.	word	freq.
1 st	que	634157	de	237430	de	1595974	de	144264	que	65361	de	363662
2 nd	de	472486	que	167250	que	1019034	la	96273	y	61934	la	223127
3 rd	y	385487	y	149251	la	1008611	y	79118	de	58754	que	153859
4 th	el	292388	la	143027	y	996536	que	78058	en		en	153844
5 th	la	286148	el	110637	el	747822	el	72727			y	149580
6 th	en	242671	en	107668	en	693400	en	65738				
7 th	no	222879	a	94895	a	659758	a	56186				
8 th	a	210547	los	67626	los	438081	los	39836				
9 th	es	150291	no	58242	no	394910	se	34442				
10 th	mi	143810	se	54190	se	344748	no	31558				
11 th	me	124711	las	44760	las	296571	un	29493				
12 th	con	113930	con	44249	por	292357	con	27530				
13 th	por	110648	por	42106	con	289002	las	26947				
14 th	si	100945	su	40872	su	277218	por	25593				
15 th	los	100414	del	37259	del	263395	del	25322				
16 th	se	92276	es	35314	un	252288	su	24119				
17 th	su	86327	un	31449	al	203291	una	21674				
18 th	lo	84978	lo	28836	es	198226	al	16833				
19 th	yo	82076	al	26916	lo	191499	lo	16718				
20 th	pues	81261	para	23161	una	179087	para	15601				