Scriptum digital
Vol. 9 (2020), pp. 209-236

CoDiAJe – THE ANNOTATED DIACHRONIC CORPUS OF JUDEO-SPANISH. DESCRIPTION OF A MULTI-ALPHABETIC CORPUS AND ITS TEXTUAL AND LINGUISTIC ANNOTATIONS

ALDINA QUINTANA
The Hebrew University of Jerusalem
Aldina.Quintana@mail.huji.ac.il
ORCID-iD: https://orcid.org/0000-0003-4179-2502

ABSTRACT
Judeo-Spanish differs from late 15th-century Spanish and modern Spanish in several respects, such as its morphology, syntax, and semantics, but the most visible difference is in the alphabet. From the end of the 19th century, Judeo-Spanish has been written in various alphabets –Greek, Cyrillic, and especially Latin–. However, the Hebrew alphabet had been used since ancient times, before it was abandoned finally only in the 1940s. This means that the majority of Judeo-Spanish texts are written in Hebrew characters.
CoDiAJe is an annotated diachronic corpus that includes documents produced from the 16th century up to the present day, developed in TEITOK. The significance of its development is that this tool processes linguistic data in the alphabets mentioned above, allowing users to visualize each text in five orthographic forms (the original version in which it was written, its transcription in Latin characters, an expanded form to complete abbreviations or to correct defective writing, a version in modern Judeo-Spanish, and a version in orthographic modern Spanish). *CoDiAJe* enables the user to conduct searches not only for a specific word, but also for all its linguistic and orthographic variants in the different alphabets. During the annotation process, tags from the EAGLES tagset for Spanish were modified, and others were created: these are simply steps towards the creation of an accurate tagset for Judeo-Spanish. The digitized texts are also enriched with semantic-conceptual information and information on the affiliation of all non-Romance elements.

KEY WORDS: Judeo-Spanish, multi-alphabetic corpus, corpus annotation, linguistic variation, diachrony.

CoDiAJe – CORPUS DIACRÓNICO ANOTADO DEL JUDEOESPAÑOL. DESCRIPCIÓN DE UN CORPUS MULTIALFABÉTICO Y DE SU ANOTACIÓN TEXTUAL Y LINGÜÍSTICA

RESUMEN
El judeoespañol se diferencia del español de finales del siglo XV y del español moderno en varios aspectos que afectan a la fonética y fonología, morfología, sintaxis y semántica. Sin embargo, la diferencia más fácilmente apreciable está en el alfabeto. A finales del siglo XIX se comenzó a escribir con diferentes alfabetos: griego, cirílico y, sobre todo, latino en diferentes versiones. Sin embargo, desde tiempos remotos se utilizó el alfabeto hebreo, y su abandono definitivo solo ocurrió en la década de los cuarenta del siglo pasado, por lo que la mayor parte de los textos escritos en esta lengua están en caracteres hebreos.
CoDiAJe es un corpus diacrónico anotado que incluye documentos creados desde el siglo XVI hasta nuestros días, desarrollado en TEITOK. La importancia de su desarrollo está en que procesa datos lingüísticos en los alfabetos mencionados anteriormente, da al usuario la opción de visualizar cada texto en cinco formas gráficas (la versión original independientemente del alfabeto en el que fue escrita, su transcripción en caracteres latinos, una forma expandida para completar las abreviaturas o corregir la escritura defectuosa, una versión en judeoespañol moderno y una versión en la ortografía del español moderno), y permite realizar búsquedas no solo de una palabra específica sino de todas sus variantes lingüísticas y ortográficas en textos escritos en los diferentes alfabetos. Durante el proceso de anotación se fueron modificando las etiquetas de EAGLES para el español y se crearon algunas nuevas. Significa que, a medida que se van anotando los textos, vamos creando un etiquetador para el judeoespañol. Los textos digitalizados también se enriquecen con información semántico-conceptual e información sobre la filiación de todos los elementos no románicos que se detectan en los textos.

PALABRAS CLAVE: judeoespañol, corpus multialfabéticos, anotación de corpus, variación lingüística, diacronía.

Fecha de recepción: 16/08/2020
Fecha de aceptación: 27/09/2020

ISSN: 2014-640X

0. INTRODUCTION

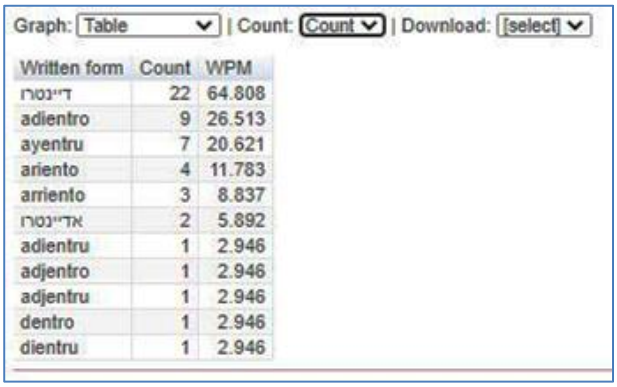
Judeo-Spanish is an autonomous linguistic diasystem made up of a continuum of dialects that have developed without contact with Peninsular and American Spanish. The exception is the North African or *Hakitia* variety, which never ceased to maintain contact with Peninsular Spanish. *CoDiAJe - Corpus diacrónico anotado del judeoespañol* (*The Annotated Diachronic Corpus of Judeo-Spanish*) is a project[1] whose purpose is to build a resource that allows researchers to study in depth the evolution of Judeo-Spanish. This resource, in addition to providing paleographic information about the texts, enriches them with linguistic information (POS tagging and lemmatization), ensuring its easy usage by non-experts in NLP. The resource is easily maintainable and has the possibility of being permanently improved by non-NLP experts, following OLDES (see Janssen *et al*. 2017) — the first model from which its development started. Nevertheless, it should also offer satisfactory solutions to the specific problems that Judeo-Spanish texts pose.

Developed from the popular Spanish spoken in the late 15th century (cf. Arnold in this volume), it is the only historical variety of Spanish that does not conform to its unity. This is not only due to the remarkable differences that Judeo-Spanish exhibits in many respects, such as its phonology, morphology, syntax, and semantics (cf. Penny 2000: 176-192; Cárdenas 2004; Lleal 2004; Bradley and Delforge 2006; Minervini 2006; Varvaro and Minervini 2008; García Moreno 2010; Hualde and Şaul 2011; Hualde 2013), in comparison with Spanish, but primarily because of the different systems used in the graphic representation of the language, and the extent of graphemic variation found in its documents (cf. Quintana 2010; Bunis 2019). Therefore, the problems taken into account before starting the building of *CoDiAJe* can be summarized in the following points:

1. Most of the Sephardic textual heritage, preserved in both printed and manuscript documents, is written in the Hebrew alphabet. It was only in the late 19th century that Sephardim began to make progressive use of the Latin alphabet in various versions (French, Italian, Serbo-Croatian, or Romanian, Turkish and, to a lesser extent, Spanish), together with the Cyrillic and Greek alphabets, depending on their dominant language, and to adapt them to the phonemic characteristics of Judeo-Spanish (cf. Bunis 2019). However, these alphabets did not fully replace the Hebrew alphabet until after World War II, and only in the last few decades has a relatively unified system of writing in Latin characters been imposed. This means that perhaps 90% of Judeo-Spanish texts were written using the Hebrew alphabet. Another consequence of this is that the Judeo-Spanish texts accessible to scholars who have not necessarily specialized in the study of this variety are very few compared to those which make up its textual legacy. They are also not accessible to speakers because they are literate in other languages and, consequently, they cannot read these scripts, and also because Judeo-Spanish works have not been republished for one, two, or more centuries. As for the manuscripts, few are the scholars who have acquired the ability to read and understand them.

[1] *CoDiAJe* is being developed within the framework of two research projects funded by grant 473/11 (completed) and grant 486/19 of the Israel Science Foundation (ISF). This paper was written in the framework of the latter project at the Hebrew University. Other members of the project are Josep M. Fontana (Universitat Pompeu Fabra, Barcelona) and Maarten Janssen (Charles University, Prague).

2. In addition to the graphemic variation that the texts display in each alphabet, one must bear in mind the variation of the language in all its dimensions (diachronic, diatopic, diastratic and diaphasic) as a consequence of the situation of low normative pressure (cf. Quintana 2006), which allowed for a flexible internal development of Judeo-Spanish in keeping with universal tendencies of natural human languages (Trudgill 2011). Particularly in the texts written in the 18th and the early 19th centuries, the language also shows a significant degree of medium-transferability (Lyons 1981: 12), meaning that a high percentage of units of the abstract language system became medium-independent, giving rise to a considerable degree of variation. To illustrate, Figure 1 shows the linguistic and orthographic variants of the lemma *adientro* 'inside' in *CoDiAJe*, which in Judeo-Spanish expresses both situation and movement, unlike in Standard Spanish, and the frequency of each of them[2].



Written form	Count	WPM
דיינטרו	22	64.808
adientro	9	26.513
ayentru	7	20.621
ariento	4	11.783
arriento	3	8.837
אדיינטרו	2	5.892
adientru	1	2.946
adjentro	1	2.946
adjentru	1	2.946
dentro	1	2.946
dientru	1	2.946

Figure 1. Linguistic and orthographic variants of the lemma *adientro* 'inside' with their frequency in *CoDiAJe*.

3. Although the principle of language representation has always been phonemic, texts written or printed in the Hebrew alphabet have the peculiarity of representing the vowels with *matres lectionis*, i.e. consonants that are used to indicate a vowel. In Hebrew and Judeo-Spanish they are א (aleph), ה (he), ו (waw) and י (yod). Without taking into account the representation of diphthongs, which is even more complex, two consonant graphemes <ה-, א> represent the vowel /a/ and two others <י, ו> are used to write the vowels /o, u/ and /e, i/ respectively, except in Hebrew words and expressions, where the etymological form based only on consonants was maintained and generally follows the Sephardic tradition to use *haser* or defective spelling —i.e. only consonants without any indication of vowels— in the Hebrew script. This is because the Hebrew language exhibits a pattern of stems consisting of 3-consonant consonantal roots. Moreover, the affricate consonant phonemes /ʦ/, /ʣ/, /ʧ/, /ʤ/ and the palatal /ʒ/ —which do not exist in Hebrew— are also represented by Hebrew letters bearing diacritical marks. The result is that at some point, the grapheme <'ג> represented up to three phonemes: /ʧ/, /ʤ/ and /ʒ/. Hebrew does not have the voiced palatal nasal /ɲ/, represented by <ני> or <ניי> in Judeo-Spanish. This makes it hard to differentiate geographical variants, such as *nieve* <נייב'י> or *ñeve* <נייב'י> 'snow', since both readings are possible. Another problem lies in reading words that are pronounced with a voiced alveolar tap /ɾ/ or a trill /r/ (e.g. <פירה>, which may

[2] In this paper, all data concerning *CoDiAJe* refer to its standing on September 15, 2020.

have two readings: /'pera/ ‘pear’ or ‘bitch’ and /'pera/ ‘bitch’, since only one grapheme is available for writing the two phonemes still preserved in some varieties). Adjustments to the spelling system imposed by phonological changes also need to be borne in mind.

4. Contact of Judeo-Spanish speakers with Hebrew as the Sephardim’s ethnoreligious language, and the different types of contact with the surrounding Romance and non-Romance languages must also be considered. These languages are Turkish, Greek, Slavic languages, Arabic, Romanian and Italian dialects, to which German, but especially Italian and French —and to a lesser extent Spanish—, as languages of culture since the mid-19th century, must be added[3]. Judeo-Spanish contact with the last three furthered its revival and the re-Romanization of its regional norms. Before that, however, mainly in works belonging to the rabbinical style —which are almost all of them— and private letters, not only are quotations from Hebrew sources embedded in Judeo-Spanish texts, but all kinds of Hebrew nouns in construct state or *smikhut* ‘genitive’ and other words pertaining to all parts of speech may also appear merged with Hebrew inflectional morphemes.

Some examples are shown in Figure 2. While Hebrew single words and nouns in construct state do not pose problems and can be lemmatized as integrated words in Judeo-Spanish, it is impossible to lemmatize words merged with all kinds of Hebrew inflectional morphemes following the rules of Hebrew.

Hebrew word merged with grammatical morphemes	Transcription according to the Sephardic Hebrew pronunciation	Linguistic glosses	Modern Spanish translation
יעקב טודא סו וידה איסטוב'ו בצער	Yakov toda su vida estuvo *be-sar*	Prep.-N	Yacob estuvo *triste* toda su vida
אי סי מאלייורגו מי יצר *הרע*	i se mayorgo mi *yeser a-ra*	N DetArt.-N	y aumentó mi *instinto del mal*
אי אב'יזאן *רבותינו*	I avizan *rabotenu*	N-Poss.	y nos llaman la atención *nuestros señores sabios*…
ואח"כ קאייו מאלו אלא מואירטי	*Ve-ahar kah* kayo malo ala muerte	CC-Adv. Adv.	*Y después* cayó enfermo de muerte
אי אין אקיל אניו *נפטרה* לה סיניורה די לאה	i en akel anyo *niftera* la sinyora de Lea	V(IndPas.3FSg)	Y en aquel año *murió* la señora Leah.
לוש איג'וס די *בלהה וזלפה*	los ijos de *Bila ve-Zilpa*	N CC-N	Los hijos de *Bila y Silpa*
אי *מן הדין* איס *מותר*	*y min a-din* es *mutar*	Prep. DetArt.N V(Passiv.MSg)	y, *según la Ley*, se *puede*
פארה טראטאר *כדרך הסוחרים*	Para tratar *ke-dereh a-soharim*	Adv.N DetArt.N	para tratar *como comerciante*
mizerah amiluha	*mi-zerah a-meluha*	Prep.N DetArt.-N	*de estirpe real*

Figure 2. Hebrew words merged with other morphemes.

[3] For the rules of integration of alien words from different languages into Judeo-Spanish, see Cárdenas (2004).

5. Finally, an essential aspect taken into account in the development of *CoDiAJe* is the ascription of Judeo-Spanish to different cultural traditions, mainly the Hispanic and Jewish ones. All this, naturally, without «concealing the Judeo-Spanish nature of the text, the characteristics, and history of the Sephardic language» (Busse 2005: 105).

In short, a single corpus should

a) process linguistic data in the alphabets mentioned above,
b) allow the visualization of each text in the original version independently of the alphabet in which it was written, its Latinized transcription, and a modern standardized version, and
c) enable the user to conduct searches not only for a specific word but also for all its linguistic and orthographic variants in the different alphabets.

The development of *CoDiAJe* originated from the need to recover, for research and for the speech community, at least part of the nearly 4,000 Judeo-Spanish printed books, some of them of more than 1,000 pages, and about 250 periodicals, some of which were published for over 30 years. To these one must add thousands of manuscripts that were never published. Most of this textual legacy in Judeo-Spanish is hidden in archives around the world. Their publication would make it possible, in many cases, to piece together fragments of a single document, now scattered over different collections and archives.

The present paper is organized as follows: Section 1 provides the reader with a brief description of the corpus and the current state of its development. Section 2 deals with metadata (2.1), the multi-alphabetic nature of *CoDiAJe* (2.2), the problems raised in the tasks of textual (2.3) and linguistic annotations (2.4) and their solutions. Section 3 addresses the advantages offered by the search options of the corpus and presents examples of the frequency distribution of some of the search results. Finally, Section 4 contains concluding remarks and offers directions for the future development of multi-alphabetic corpora for languages whose texts have similar characteristics to those of Judeo-Spanish.

1. BRIEF DESCRIPTION OF *CODIAJE*

After several attempts using other tools, with unsatisfactory results, *CoDiAJe* was created in TEITOK[4], initially developed at the *Centro de Linguística da Universidade de Lisboa* (Janssen 2016: 4037), and follows the structure of the project *P. S. Post Scriptum*, with the replacement of some features and the addition of others to satisfy the specific requirements of Judeo-Spanish texts.

Taking advantage of TEITOK as a web-based platform for visualizing, searching, and editing TEI/XML-based corpora (Janssen 2018) that combines textual and linguistic annotation within a single TEI-based XML document, *CoDiAJe* is a structured multi-genre diachronic corpus that includes documents produced from the 16th century up to the 21st century —this does not preclude the addition of older texts in the future— enriched with different kinds of textual and linguistic information. Every document is also accompanied by metadata.

[4] For a description of TEITOK, see Janssen (2016, 2018).

CoDiAJe currently contains 74 documents totaling to 352,357 tokens. 16 documents were written in the Hebrew alphabet, 1 in the Cyrillic script, and 57 in Latin characters, representing 98,112, 323, and 253,922 tokens respectively, as shown in Figures 3 and 4[5]. This disproportion derives from the fact that some texts had already been transcribed in Latin characters for the failed attempts to develop a corpus of Judeo-Spanish before starting its development in TEITOK. In the future, we will correct this by returning each of the documents originally written in Hebrew characters to its original alphabet.

24 of the documents (35,624 tokens) have been fully annotated. This still small set of texts is being used for corpus training, which will allow annotations to be made automatically in the future, requiring only their revision by the *CoDiAJe* editors to exclude possible errors.

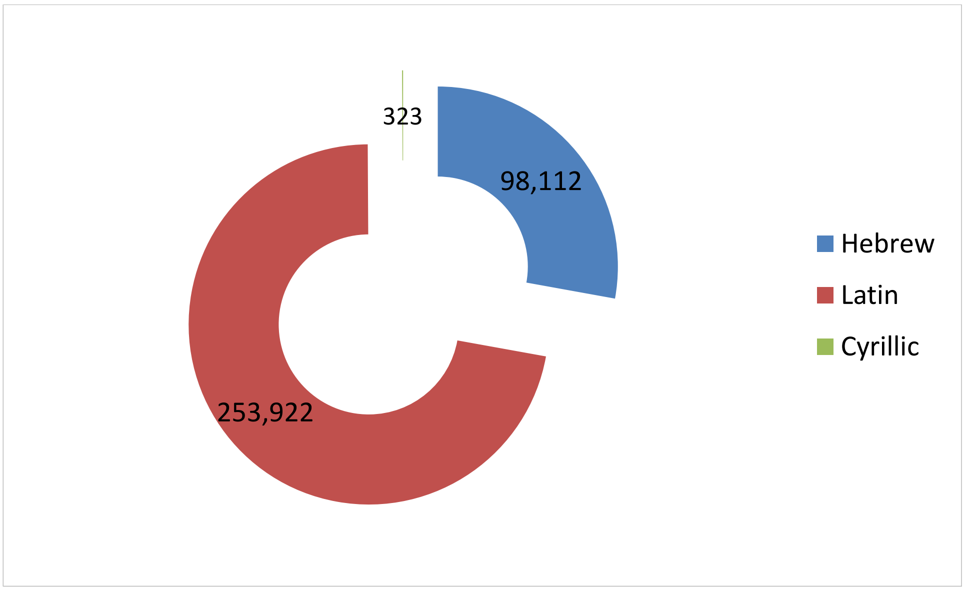
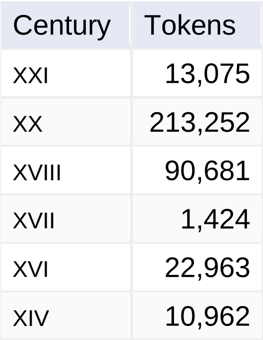


Figure 3. Current number of tokens distributed by alphabet in which the texts were loaded onto *CoDiAJe*.



Century	Tokens
XXI	13,075
XX	213,252
XVIII	90,681
XVII	1,424
XVI	22,963
XIV	10,962

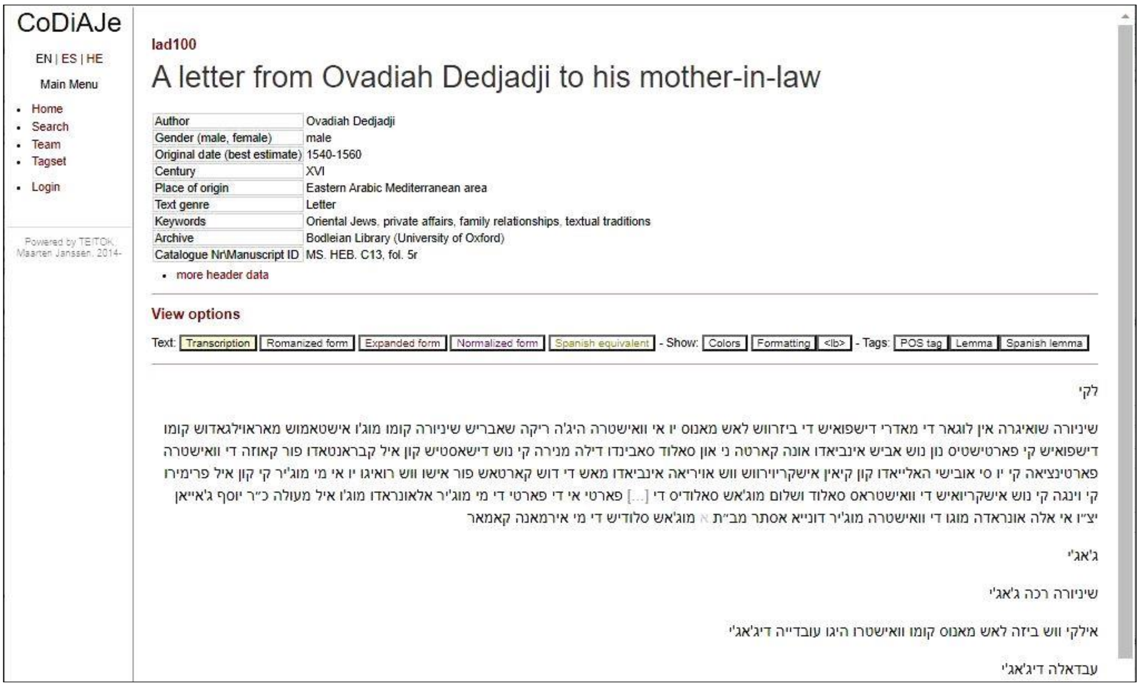
Figure 4. Number of tokens classified by century according to the date of composition of the *CoDiAJe* text collection.

[5] The corpus does not yet have texts originally written in the Greek alphabet.

2. THE STRUCTURE OF *CODIAJE*

2.1. Metadata

The list of metadata has been carefully planned to allow advanced corpus search options, and could be improved in the future. The metadata of each document provide information of scientific interest about the author (name, gender, year and birthplace, place of residence), and about the document, such as date, genre, alphabet, or documentary source, as shown in Figure 5.



lad100

A letter from Ovadiah Dedjadji to his mother-in-law

Author	Ovadiah Dedjadji
Gender (male, female)	male
Original date (best estimate)	1540-1560
Century	XVI
Place of origin	Eastern Arabic Mediterranean area
Text genre	Letter
Keywords	Oriental Jews, private affairs, family relationships, textual traditions
Archive	Bodleian Library (University of Oxford)
Catalogue Nr\Manuscript ID	MS. HEB. C13, fol. 5r

- more header data

לקי

שיניורה שואיגרה אין לוגאר די מאדרי דישפואיש די ביזרווש לאש מאנוס יו אי וואישטרה היג'ה ריקה שאבריש שיניורה קומו מוג'ו אישטאמוש מאראוילגאדוש קומו
דישפואיש קי פארטישטיס נון נוש אביש אינביאדו אונה קארטה ני און סאלוד סאבינדו דילה מנירה קי נוש דישאסטיש קון איל קבראנטאדו פור קאוזה די וואישטרה
פארטינציאה קי יו סי אובישי האלייאדו קון קיאין אישקריוירווש ווש אויריאה אינביאדו מאש די דוש קארטאש פור אישו ווש רואיגו יו אי מי מוג'יר קי קון איל פרימירו
קי וינגה קי נוש אישקריואיש די וואישטראס סאלוד ושלום מוג'אש סאלודיס די [...] פארטי אי די פארטי די מי מוג'יר אלאונראדו מוג'ו איל מעולה כ״ר יוסף ג'אייאן
יצ״ו אי אלה אונראדה מוגו די וואישטרה מוג'יר דונייא אסתר מב״ת א מוג'אש סלודיש די מי אירמאנה קאמאר

ג'אג'י

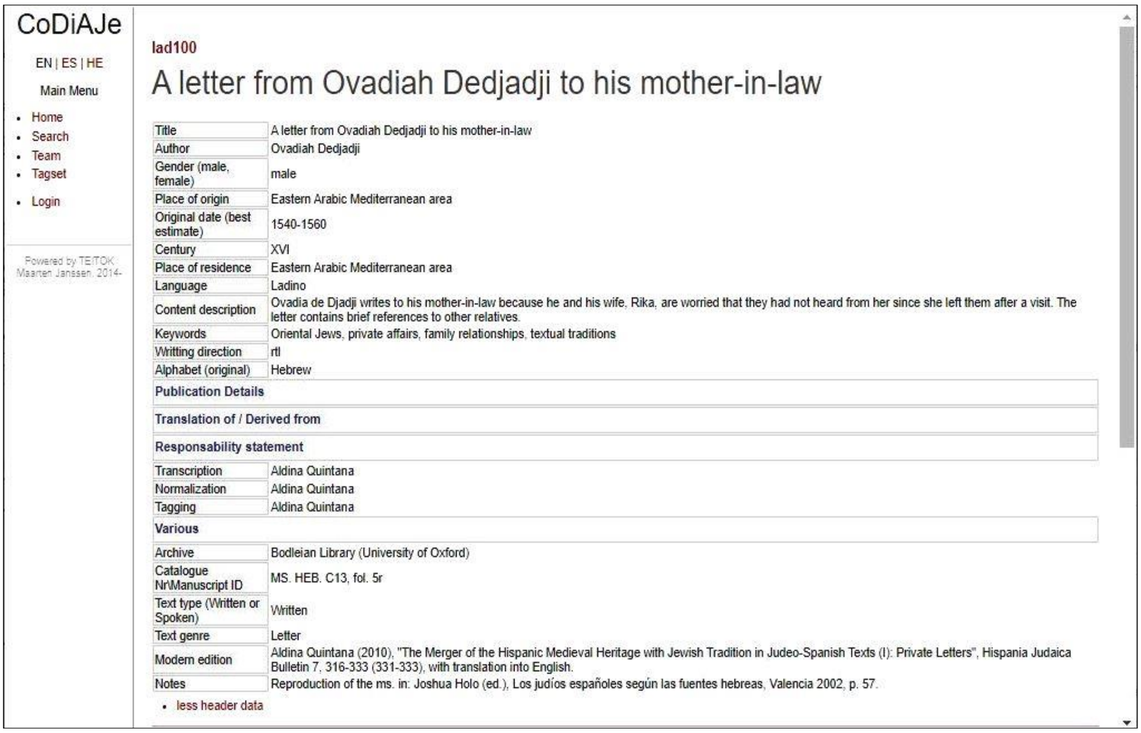
שיניורה רכה ג'אג'י

אילקי ווש ביזה לאש מאנוס קומו וואישטרו היגו עובדייה דיג'אג'י

עבדאלה דיג'אג'י

Figure 5. A letter in *CoDiAJe*, handwritten originally in Hebrew characters, known as *solitreo* script.

There is also the option of viewing more data, which include information about the medium in which the text was created, whether it is an original or a translation, and who was responsible for the tasks of transcription, normalization, and tagging of each text, among other data (Figure 6).



lad100

A letter from Ovadiah Dedjadji to his mother-in-law

Title	A letter from Ovadiah Dedjadji to his mother-in-law
Author	Ovadiah Dedjadji
Gender (male, female)	male
Place of origin	Eastern Arabic Mediterranean area
Original date (best estimate)	1540-1560
Century	XVI
Place of residence	Eastern Arabic Mediterranean area
Language	Ladino
Content description	Ovadia de Djadji writes to his mother-in-law because he and his wife, Rika, are worried that they had not heard from her since she left them after a visit. The letter contains brief references to other relatives.
Keywords	Oriental Jews, private affairs, family relationships, textual traditions
Writting direction	rtl
Alphabet (original)	Hebrew
Publication Details	
Translation of / Derived from	
Responsability statement	
Transcription	Aldina Quintana
Normalization	Aldina Quintana
Tagging	Aldina Quintana
Various	
Archive	Bodleian Library (University of Oxford)
Catalogue Nr\Manuscript ID	MS. HEB. C13, fol. 5r
Text type (Written or Spoken)	Written
Text genre	Letter
Modern edition	Aldina Quintana (2010), "The Merger of the Hispanic Medieval Heritage with Jewish Tradition in Judeo-Spanish Texts (I): Private Letters", Hispania Judaica Bulletin 7, 316-333 (331-333), with translation into English.
Notes	Reproduction of the ms. in: Joshua Holo (ed.), Los judíos españoles según las fuentes hebreas, Valencia 2002, p. 57.

Figure 6. Example of the metadata that accompany each document.
In this case, they correspond to the letter in Figure 5.

2.2. The multi-alphabetic corpus of Judeo-Spanish

The digitization of the texts is carried out with the ABBYY OCR software, except for the texts written in Hebrew characters, where *Transkribus* is used after being trained for the recognition of texts in Judeo-Spanish[6]. The first version usually contains a large number of errors that must be corrected manually. In spite of this, the task is exceptionally profitable, since, in the time required to copy a page manually, it is possible to check between 35 and 40 pages once the optical character recognition has been completed.

The documents are incorporated into *CoDiAJe* using the XML-TEI format in order to enter all the necessary metadata for further processing.

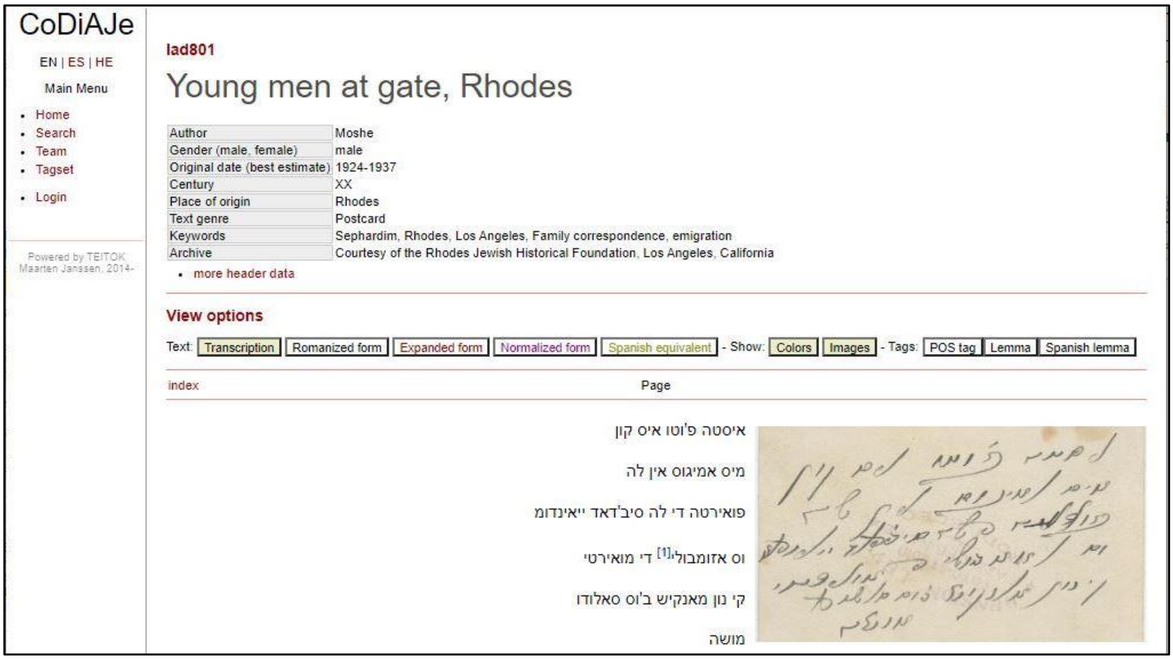
A particularly innovative aspect of *CoDiAJe* is the possibility of incorporating texts in the alphabets in which they were originally written or published and visualizing them correctly. Therefore, the first step in the development of *CoDiAJe* consisted in adapting TEITOK to the peculiar multi-alphabetic character of the Judeo-Spanish documents. The possibility of including orthographic forms in the Hebrew alphabet raised the difficulty of writing and reading in a right-left direction without disturbing the opposite directionality of the Latin, Greek and Cyrillic alphabets. After expanding the potential of *CoDiAJe* to allow the inclusion of texts written in different alphabets and the coexistence of multiple orthographies, it was necessary to determine the layers required to encode the multiple orthographies and alphabets in which a word can be written. At present *CoDiAJe* has a total of five different orthographic realizations:

1) An original spelling (*Transcription*), which is an exact copy of the text (Figure 7);

[6] My thanks to Matan Stein, researcher of *CoDiAJe*, for the digitization of the texts in Hebrew script using *Transkribus*, and to Sinai Rusinek, who adapted this tool for the digitization of Judeo-Spanish in the framework of her project *DiJeST: Digitizing Jewish Studies*.

2) A transcription in Latin characters (*Romanized form*), available only when the *Transcription* is written in non-Latin characters (Figure 8);
3) an *Expanded form*, very useful for expanding the numerous abbreviations in the Judeo-Spanish texts, and correcting defective writings, very frequent in manuscripts and documents in Hebrew script (cf*.* the tokens *hashem* and *vegomer* in the *Expanded form* after the completion of the abbreviations *h´* and *vego´* shown in the *Romanized form* in Figure 11);
4) a *Normalized form* in which each word appears standardized according to the spelling rules (cf. Álvarez López 2017) authorized by the National Authority of Ladino on August 13, 2018, and to the standard characteristics of the Judeo-Spanish spoken in Istanbul, which is the variety currently spoken by the largest number of Judeo-Spanish speakers (Figure 9);
5) A hispanized form (*Spanish equivalent*) that in the future will allow for a visualization of the texts in modern Spanish spelling.

These results can be shown in different layers —including the original spelling of the TEI-based XML file as shown in Figure 7, and a Romanized version when the file is not in the Latin alphabet (see Figure 8)—, which allows for a visualization of the diversity of variants. All the orthographic options can be visualized by clicking on the corresponding buttons on top to switch between the various layers. Figures 7, 8, and 9 show different orthographic realizations of the text next to its facsimile image, in this case the reverse of a postcard sent from Rhodes to Los Angeles at the end of the 1920s, listed as lad801 in *CoDiAJe*.


CoDiAJe
EN | ES | HE
Main Menu
Home
Search
Team
Tagset
Login
Powered by TEITOK
Maarten Janssen, 2014-

lad801

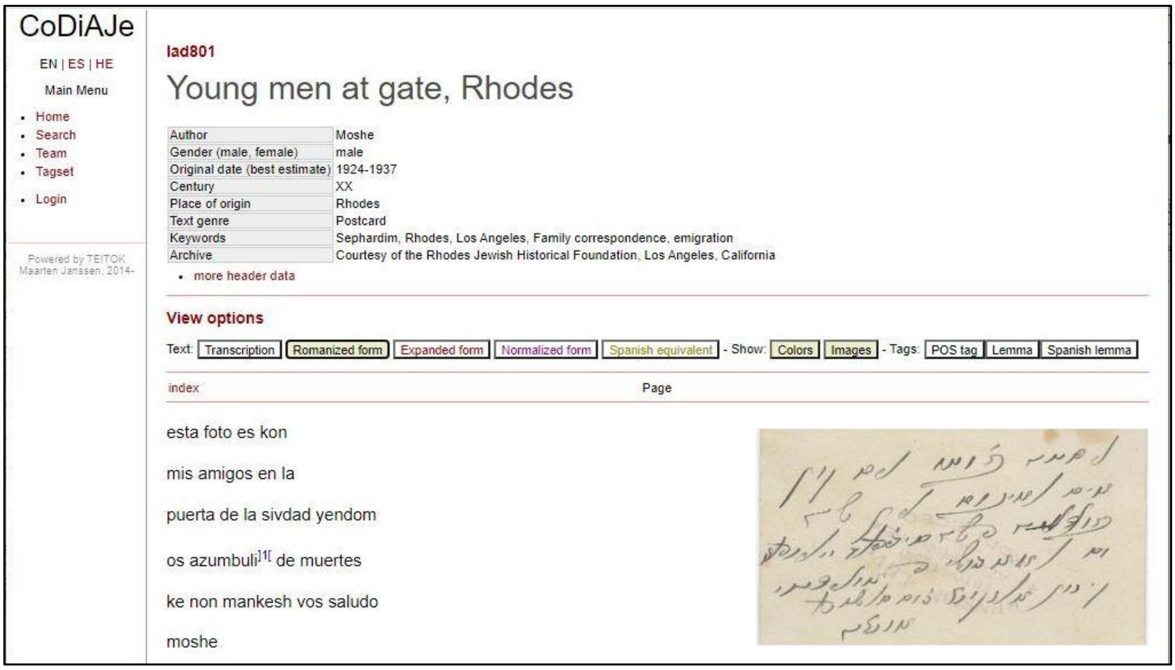
Young men at gate, Rhodes

Author	Moshe
Gender (male, female)	male
Original date (best estimate)	1924-1937
Century	XX
Place of origin	Rhodes
Text genre	Postcard
Keywords	Sephardim, Rhodes, Los Angeles, Family correspondence, emigration
Archive	Courtesy of the Rhodes Jewish Historical Foundation, Los Angeles, California

more header data
View options
Text: Transcription Romanized form Expanded form Normalized form Spanish equivalent - Show: Colors Images - Tags: POS tag Lemma Spanish lemma
index
Page

איסטה פ'וטו איס קון
מיס אמיגוס אין לה
פואירטה די לה סיב'דאד ייאינדומ
וס אזומבולי,[1] די מואירטי
קי נון מאנקיש ב'וס סאלודו
מושה

Figure 7. Original orthographic visualization (*Transcription*).



lad801

Young men at gate, Rhodes

Author	Moshe
Gender (male, female)	male
Original date (best estimate)	1924-1937
Century	XX
Place of origin	Rhodes
Text genre	Postcard
Keywords	Sephardim, Rhodes, Los Angeles, Family correspondence, emigration
Archive	Courtesy of the Rhodes Jewish Historical Foundation, Los Angeles, California

- more header data

View options

Text: Transcription | Romanized form | Expanded form | Normalized form | Spanish equivalent - Show: Colors | Images - Tags: POS tag | Lemma | Spanish lemma

index Page

esta foto es kon

mis amigos en la

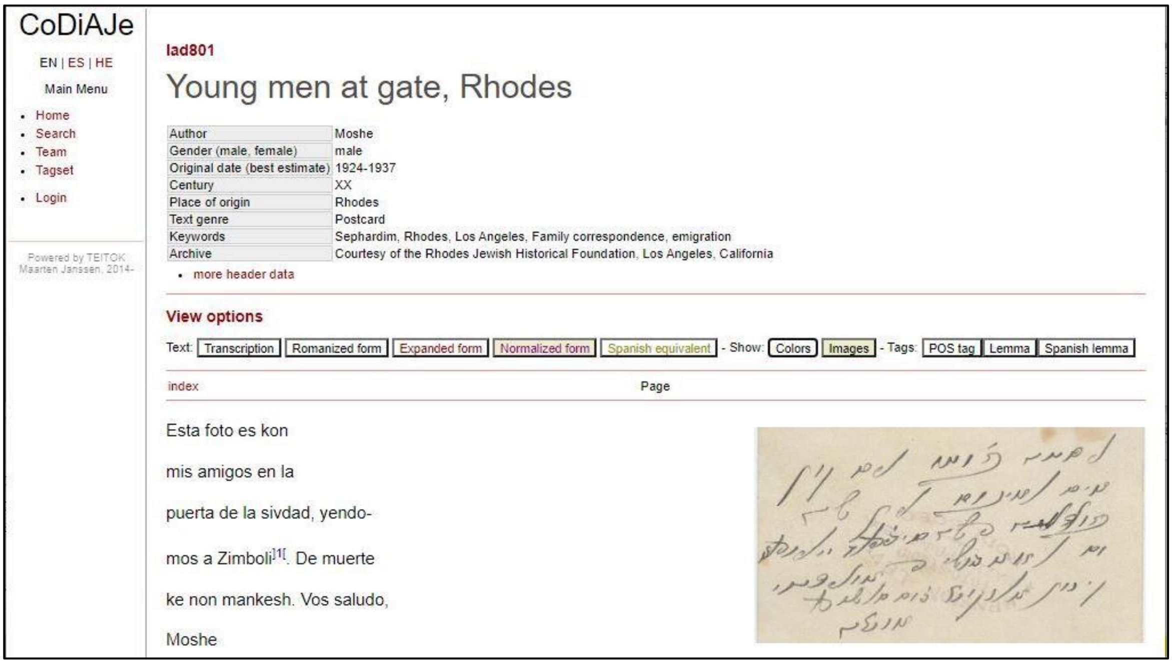
puerta de la sivdad yendom

os azumbuli]1[de muertes

ke non mankesh vos saludo

moshe

Figure 8. Transcription in Latin characters (*Romanized form*).



lad801

Young men at gate, Rhodes

Author	Moshe
Gender (male, female)	male
Original date (best estimate)	1924-1937
Century	XX
Place of origin	Rhodes
Text genre	Postcard
Keywords	Sephardim, Rhodes, Los Angeles, Family correspondence, emigration
Archive	Courtesy of the Rhodes Jewish Historical Foundation, Los Angeles, California

- more header data

View options

Text: Transcription | Romanized form | Expanded form | Normalized form | Spanish equivalent - Show: Colors | Images - Tags: POS tag | Lemma | Spanish lemma

index Page

Esta foto es kon

mis amigos en la

puerta de la sivdad, yendo-

mos a Zimboli]1[. De muerte

ke non mankesh. Vos saludo,

Moshe

Figure 9. The same text once normalized (*Normalized form*).

It should be stressed that due to the specific characteristics of Judeo-Spanish, which derive from a low standardization level, the aim here is to achieve a structured approach to the visualization of diversity, focusing on variant detection. The multi-alphabetic character of *CoDiAJe* does therefore not affect the search function. One of *CoDiAJe*'s achievements is that a single query can yield results for all variants without losing track of the original spelling form, as shown in Figure 10.

Transcription	*Romanized form*	*Normalized form*	*Spanish equivalent*	Modern Judeo-Spanish	Modern Spanish
vežežirije		viejejeria	vejedumbre	vejedumbre	
חאב אל עזיז	ḥabb el ʿaziz	hab el aziz	jab el asis	havachichi	chufa
לשון הרע	lashon hara	lashon a-ra	lašón ha-rá		difamación

Figure 10. Possible visualizations of the Judeo-Spanish diatopic variants *vežežirije* ‘old age’, חאב אל עזיז ‘Cyperus esculentus’, and the Hebrew constructus לשון הרע (H.) ‘defamation’.

Other Judeo-Spanish texts involve greater difficulties than the one shown in Figures 7, 8, and 9. For example, quotations from other texts, often in Hebrew, are usually not printed with a distinctive typeface. Their normalization in capital letters not only facilitates their quick location in the *Normalized form*, but also in the *Transcription* with the original spelling, as can be verified in the brief extract of a text first printed in 1730 (Figure 11), included in *CoDiAJe*.

Transcription	**קאפיטוּלוּ 1** וירא אליו ה' באלוני ממרא וגו'. ייא אב'יזי אריב'ה אין לה פרשה די לך לך, קאפיטולו 6, קי סי אקונסיג'ו אברהם קון סוס טריס אמיגוס ענר, אשכול אי ממרא פור קואינטו די איל סירקוסיר...
Romanized form	**kapitulu 1** vayera elav h´ beelone mamre vego´. ya avizi ariva en la perasha de leh leha, kapitulu 6, ke se akonsejo avraham kon sus tres amigos aner, eshkol i mamre por kuento de el sirkusir…
Expanded form	**kapitulu 1** vayera elav hashem beelone mamre vegomer. ya avizi ariva en la perasha de leh leha, kapitulu 6, ke se akonsejo avraham kon sus tres amigos aner, eshkol i mamre por kuento de el sirkusir…
Normalized form	**Kapitolo 1** VAYERA ELAV H´ BE-ELONE MAMRE VE-GOMER. Ya avizi ariva en la perasha de LEH LEHA, kapitolo 6, ke se akonsejo Avraam kon sus tres amigos Aner, Eshkol i Mamre por kuento de el sirkusir…
Spanish equivalent	**Capítulo 1** VAYERÁ ELAV HA-ŠEM BE-ELONÉ MAMBRÉ VE-GOMER. Ya avisé arriba en la perašá de LEJ LEJÁ, capítulo 6, que se aconsejó Abrahán con sus tres amigos Aner, Escol y Mambré por cuento de el circucir…

Figure 11. Text view options in *CoDiAJe*.

The material provided in these five layers constitutes the textual annotation. This means that each token contains the orthographic and linguistic variants that can be displayed on the layers, as can be seen in Figure 12.

```
<tok roman="avizi" nform="avizi" spa="avisé">אב'יזי</tok>
<tok roman="perasha" nform="perasha" spa="perašá">פרשה</tok>
```

Figure 12. <tok> of the verb form *אב'יזי* ‘(I) reported’, and of the noun *פרשה* ‘pericope’.

2.3. Textual annotation: Problems and solutions

Certain forms in the process of grammaticalization or lexicalization may present problems when making textual annotations in historical texts. Good examples in Judeo-Spanish texts are some adverbs ending in *-mente*. Forms written as orthographic variants such as סולא מינטי or סולה מינטי (= *sola mente*) ‘only’ or קואל מינטי (= *kual mente*) ‘also’ emerge in texts from the

16th to the 18th century, while in modern Judeo-Spanish they are written as one word. TEITOK offers the possibility of merging two or more adjacent words in a single token, while preserving their original orthography in two or more words, as shown in Figure 13.

```
<tok roman="kual mente" nform="kualmente" spa="cualmente">קואל מינטי</tok>
```

Figure 13. <tok> of the adverb *קואל מינטי* ‘also’.

Both the same and the opposite happen in sequences formed by some prepositions, such as *a* or *de*, followed by the definite article *el*, which may appear written separately (*a el*, *de el*) or contracted (*al*, *del*). In this case, we follow the mixed approach of TEITOK (cf. Janssen 2016: 4038), according to which contractions are annotated as one orthographic <tok> with two grammatical <dtoks>, but preserve their original spelling, whether they are written together or separately. The same procedure is followed in *ala*, *alas*, *alos*, *dela*, *delas*, *delos*, *enlos* and other similar tokens, always annotated in one orthographic <tok> with two grammatical <dtoks> (Figure 14). Complex forms, such as verbs with enclitic pronouns are also annotated as one orthographic <tok> with the corresponding grammatical <dtoks>.

Original form	
אל	<tok roman="al" nform="al" spa="al">אל<dtok form="a"/><dtok form="el"/></tok>
del	<tok nform="del" spa="del">del<dtok form="de"/><dtok form="el"/></tok>
ala	<tok nform="a la" spa="a la">ala<dtok form="a"/><dtok form="la"/></tok>
דילה	<tok roman="dela" nform="de la" spa="de la">דילה<dtok form="די"/><dtok form="לה"/></tok>
דילוס	<tok roman="delos" nform="de los" spa="de los">דילוס<dtok form="די"/><dtok form="לוס"/></tok>

Figure 14. <tok> and <dtoks> of the original forms *אל*, *del*, *ala*, *דילה* and *דילוס*.

These features of TEITOK are used in *CoDiAJe* to solve similar problems that may affect other Judeo-Spanish elements of Romance origin, but also to address more difficult matters related in particular to the textual annotations of Hebrew nouns in construct state ‘genitive’, Hebrew words merged with inflected and other grammatical morphemes, or quotations from Hebrew sources.

The vast majority of nouns in construct state borrowed from Hebrew are elements of the Judeo-Spanish lexicon, in which they are collocations, although morphologically they preserve their Hebrew structure. In this case, they are annotated as one orthographic <tok>. For example, the two parts of *divre tora* ‘words of the Torah’, *אומות העולם* ‘nations of the world’, *ביקור חולים* ‘visiting the sick’, *בית דין* ‘rabbinical court’ or *בית החיים* ‘cemetery’, which fulfil the conditions of non-compositionality, non-substitutability and non-modifiability[7] (Manning and Schütze 1999: 184) are considered collocations. The same is true for אנשי כנסת הגדולה ‘The Men of the Great Assembly’, which, in addition to the two nominal parts, also contains an adjective (Figure 15).

[7] These characteristics are highlighted in the creation of new simple lexies, such as *amares* (sg.), *amareses* (pl.) ‘ignorant, especially in matters of Jewish law and custom; boorish, unlettered person’ (Bunis 1993: 368, #3169), arising from the contraction of the two parts of the noun in the construct state עם הארץ (= *am a-areṣ*).

Original form	
divre tora	<tok nform="divre tora" spa="dibré torá">divre tora</tok>
אומות העולם	<tok roman="umot haolam" nform="umot a-olam" spa="umot ha-olam">אומות העולם</tok>
ביקור חולים	<tok roman="BIKUR ḤOLIM" nform="BIKUR HOLIM" spa="BICUR JOLIM">ביקור חולים</tok>
בית דין	<tok roman="bet din" nform="Bet Din" spa="Bet Din">בית דין</tok>
בית החיים	<tok roman="bet haḥayim" nform="bet a-hayim" spa="bet ha-jayim">בית החיים</tok>
אנשי כנסת הגדולה	<tok roman="anshe keneset hagedola" nform="Anshe Keneset a-Gedola" spa="Anšé Keneset ha-Gedolá">אנשי כנסת הגדולה</tok>

Figure 15. Textual annotations of characteristic Hebrew nouns
in construct state that in Judeo-Spanish became collocations.

Lexical units formed by two or more Hebrew words, such as עבודה זרה 'idolatry', צדיק גמור 'just among the righteous' אבר מן החי 'organ of a living animal' are annotated in one <tok>, as shown in Figure 16, because, although they are not nouns in the construct state, they also make up one lexical unit.

Original form	
עבודה זרה	<tok roman="avoda zara" nform="avoda zara" spa="avodá zará"> עבודה זרה</tok>
צדיק גמור	<tok roman="ṣadik gamur" nform="sadik gamur" spa="ṣadiq gamur"> צדיק גמור</tok>
אבר מן החי	<tok roman="ever min haḥay" nform="ever min a-hay" spa="éver min ha-jay">אבר מן החי</tok>

Figure 16. <tok> of lexical units of two or more Hebrew words.

Characteristic of Judeo-Spanish is the combination of *ser* 'to be' —less frequently also *aver* 'to have' and *tener* 'to have'— and a Hebrew participle, in which the two verb forms can be inflected (cf. Muñoz Jiménez 1997; Bunis 2009). Although these combinations constitute fixed sequences, only the participle provides the lexical meaning, while the sole contribution of the finite patrimonial verb is grammatical information. Therefore, these two-word sequences cannot be considered locutions, and each form is annotated as one <tok>. It should be noted that these Hebrew participles are treated like Romance participles because they have acquired the grammatical function of Romance participles, but retain the Hebrew form and lexical meaning (see Figure 17).

Original form		
fueron *gozerim*	<tok nform="hueron" esp="fueron">fueron</tok>	<tok nform="gozerim" esp="goserim">gozerim</tok>
איש *מותר*	<tok roman="es" nform="es" esp="es">איש</tok>	<tok roman="mutar" nform="mutar" esp="mutar">מותר</tok>

Figure 17. Textual annotation of two Hebrew participles in Judeo-Spanish.

Different approaches are followed in relation to Hebrew words merged with Hebrew inflected morphemes and tokens involving more than one word that are not included in the patterns already discussed. If they are frequently used forms in Judeo-Spanish texts, albeit sometimes only as diaphasic variants, they are considered elements of its lexicon: for example, *rabotenu* (N+Poss.) lit. 'our wise lords', when preceded by *verba dicendi* (*dezir* 'to say', *avizar* 'to point out', and expressions of similar meaning), belongs to the rabbinical style, while other genres display more standardized phrases such as *muestros hahamim* or *muestros sinyores savios*. However, in this context, *rabotenu* limits the reference to the sages of Israel from the time of the Mishnah and the Talmud, and it is understood as such. Since Judeo-Spanish also has the word *ribi* (lit. 'my lord'), which is a term of respect and courtesy that comes before the names of men, and also before the names of the wise, the two forms could reasonably be considered to belong to the same nominal Judeo-Spanish paradigm of which *ribi* is the unmarked singular form and *rabotenu* the plural.

One of the several lexical units involving more than one word borrowed from Hebrew is בצער (= beṣar), lit. 'with regret', composed of the preposition *be-* —with which the noun to which it is linked can also be adverbialized— and the common noun *ṣar* 'pain, grief, sadness, suffering'. In Judeo-Spanish, *besar* belongs to the class of adverbs. Therefore, it is annotated in a single tok.

Hebrew forms that are only found once or have a low frequency in *CoDiAJe*, with no other evidence of their use in Judeo-Spanish, are annotated as alien elements and, as we saw in §2.2, their linguistic affiliation is the only information that is tagged. In multiple-word sequences in other languages, such as Hebrew *ve-ahar kah* 'and after this', *ke-dereh a-soharim* 'like the merchants', *mizerah amiluha* 'of royal lineage', the textual annotation appears in one <tok>, and they lack standard linguistic annotations.

Frozen idioms, sayings and proverbs, and similar multiple-word units, as well as quotations from Hebrew sources or other languages, are annotated in one <tok> without further linguistic information (see *Ish bitiren para* 'what puts an end to the work is the money' in Figure 18). As we will see later, these sequences, whether in Hebrew, Turkish or another language, are assigned a special mark to indicate their linguistic affiliation, in view of the fact that they are not part of the Judeo-Spanish lexicon.

The task of annotation of alien words that are not integrated into the Judeo-Spanish system is an arduous undertaking that involves a careful analysis of their function in the system and their frequency in the documents, which often necessitates modifying their textual and linguistic annotations more than once.

Original form		
בצער	<tok roman="beṣar" nform="besar" spa="beṣar">בצער</tok>	
רבותינו	<tok roman="rabotenu" nform="rabotenu">רבותינו</tok>	
ואח''כ	<tok roman="vaḥ''k" fform="ve-aḥar kaḵ" nform="Ve-ahar kah" spa="Ve-ajar caj">ואח''כ</tok>	
כדרך הסוחרים	<tok roman="kedereḵ hasoḥarim" nform="ke-dereh a-soharim" spa="que-dérej ha-sojarim">כדרך הסוחרים</tok>	
mizerah amiluha	<tok nform="mi-zerah a-meluha" spa="mi-zéraj a-melujá">mizerah amiluha</tok>	
(היא) מצאה חן בעיני הנז'	<tok roman=" maṣea ḥen beene" nform="matsea hen beene" spa="maṣeá jen be-ené">מצאה חן בעיני</tok>	<tok roman="hnz´" fform="hanizkar" nform="a-nizkar" spa="ha-nizcar">הנז'</tok>
Ish bitiren para	<tok nform="Ish bitiren para" spa="Iš bitirén pará">Ish bitiren para</tok>	

Figure 18. Textual annotations of Hebrew words merged with Hebrew inflected morphemes.

2.4. Linguistic annotation

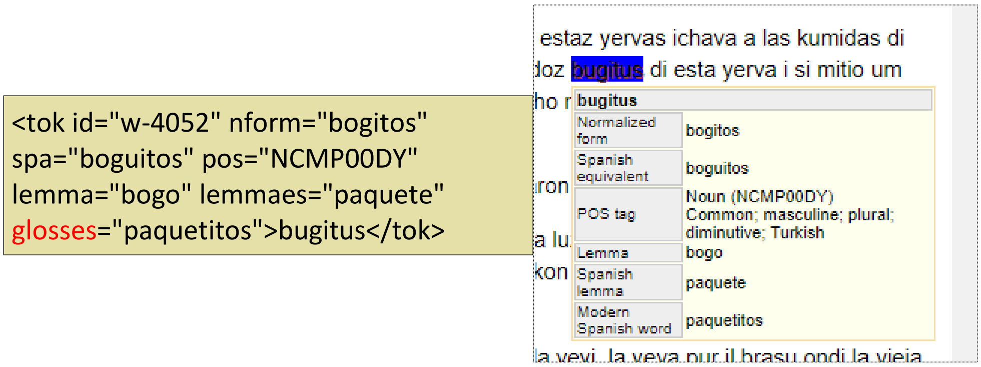
The exploitation potential offered by *CoDiAJe* would not have been achievable without accurate and careful linguistic annotations (POS and lemma), and without the patience needed for the lemmatization of poorly standardized languages, in the absence of ancillary resources, such as a good dictionary, especially when their speakers are unaware of vast portions of the lexicon used by previous generations.

Initially this corpus was tagged using a custom-built version of Freeling (Padró 2011; Padró *et al*. 2010) for Old Spanish (Sánchez-Marco *et al*. 2010; Sánchez-Marco *et al.* 2011; Sánchez-Marco *et al.* 2012) and the EAGLES tagset for Spanish adapted for Old Spanish within the framework of the *OntoSEM* project. Although the level of reliability of the automatic tagging was very high (approximately 85%) when using this tool for 15th–17th-century Sephardic texts manually transcribed in Latin alphabet, the incorporation of documents from later centuries transcribed in an adaptation of the modern orthography of Judeo-Spanish or copied in Hebrew characters revealed that the efficiency of automatic tagging was not sufficient.

Since Judeo-Spanish differs from late 16th-century Spanish and modern Spanish in a number of respects, such as in its morphology, syntax, and semantics, in addition to the orthographic representations, POS tagging is done manually using the EAGLES tagset for Spanish. When the corpus contains a considerably greater number of incorporated texts, the intention is that these tasks will be carried out directly with NeoTag, trained on the already tagged files. This training is necessary since the POS tagger NeoTag uses lexical smoothing to detect grammatical neologisms in a corpus, and therefore tags and lemmatizes known and unknown words alike (cf. Janssen 2012).

The EAGLES tagset had to be modified in many cases, and, in addition, new tags have been created to describe all Judeo-Spanish forms accurately. The digitized texts are also enriched with semantic-conceptual information: it is possible to identify expressions made up of one or more words, automatically classified as names of people, places, institutions, titles, and names assigned to God. Quotations in other languages, expressions and all kinds of forms of one or more words that do not belong to Judeo-Spanish are also tagged. Finally, all non-Romance forms have been enriched with information about their linguistic affiliation. This information is combined with the linguistic tagging of the texts. This means that the semantic-conceptual information and the language affiliation are part of the POS, together with the morpho-syntactic information. Figure 19 shows the full tagging of the geographic variant *bugitus* ‘little packages’, a common noun (NC) in masculine (M), plural (P), and diminutive form (D) of the lemma *bogo*. The final Y of the POS also provides information about the language from which this word was borrowed in Judeo-Spanish, in this case Turkish.

As already explained in §2.1, complex forms of two or more words, such as verbs with postponed clitics, are annotated in a single orthographic form but separated according to their morphological characteristics. Figure 20 contains the textual and linguistic information of מאלסינארלו ‘to denounce him, to slander him’[8].



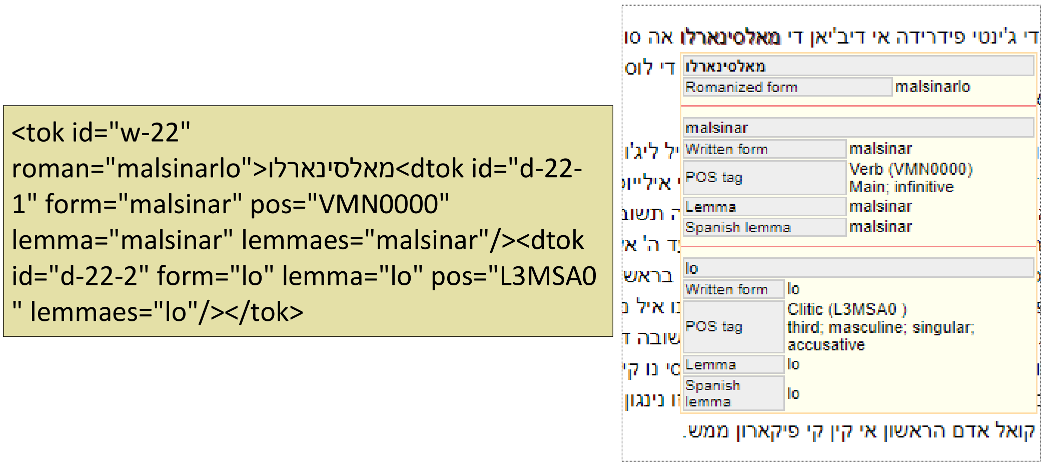
```
<tok id="w-4052" nform="bogitos"
spa="boguitos" pos="NCMP00DY"
lemma="bogo" lemmaes="paquete"
glosses="paquetitos">bugitus</tok>
```

estaz yervas ichava a las kumidas di
loz bugitus di esta yerva i si mitio um

bugitus	
Normalized form	bogitos
Spanish equivalent	boguitos
POS tag	Noun (NCMP00DY) Common; masculine; plural; diminutive; Turkish
Lemma	bogo
Spanish lemma	paquete
Modern Spanish word	paquetitos

la yevi, la yeva pur il brasu ondi la vieja

Figure 19. Annotations for the word *bugitus* ‘little packages’.



```
<tok id="w-22"
roman="malsinarlo">מאלסינארלו<dtok id="d-22-
1" form="malsinar" pos="VMN0000"
lemma="malsinar" lemmaes="malsinar"/><dtok
id="d-22-2" form="lo" lemma="lo" pos="L3MSA0
" lemmaes="lo"/></tok>
```

די ג'ינטי פידרידה אי דיב'יאן די מאלסינארלו אה סו

מאלסינארלו	
Romanized form	malsinarlo

malsinar	
Written form	malsinar
POS tag	Verb (VMN0000) Main; infinitive
Lemma	malsinar
Spanish lemma	malsinar

lo	
Written form	lo
POS tag	Clitic (L3MSA0) third; masculine; singular; accusative
Lemma	lo
Spanish lemma	lo

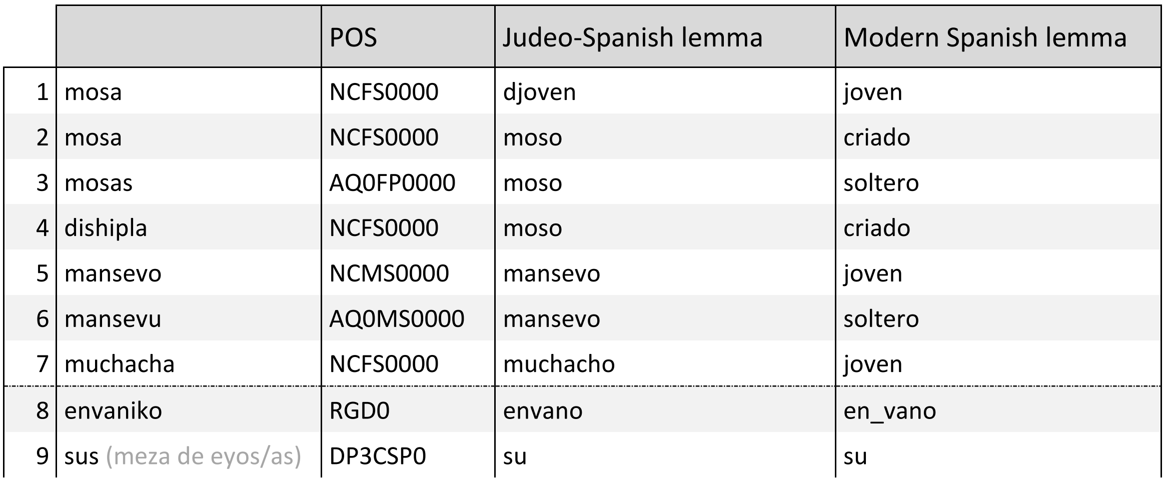
קואל אדם הראשון אי קין קי פיקארון ממש.

Figure 20. Full tagging of *malsinarlo*.

[8] Although this verb is a derivate of the Hebrew noun *malšīn*, it is not clear to which medieval social group its formation should be assigned. Therefore, the verb *malsinar* is here considered a patrimonial word transferred from the Iberian Peninsula and not a Hebrew word.

The lemmatization is done according to the orthographic rules of modern Judeo-Spanish and the norm of Istanbul as much as possible, since that is the variety with the most speakers in the world, and the one most frequently used in writing. The corpus is also lemmatized in modern Spanish. As surprising as that may seem, it can be of great help in the lemmatization of *CoDiAJe*, a task that sometimes becomes difficult due to the dialectal variation. With a quick search by the Spanish lemma, possible errors in the lemmatization of a Judeo-Spanish token can be easily detected. Consequently, its lemmatization can be unified for all the geographical variants belonging to the lemma of the standardized variety —in this case, that of Istanbul. A good example may be the geographical variant *dishipla* (Bitola, Salonika, Sarajevo) 'maid', while in Istanbul, a maid is called *mosa*. In Bitola however, *mosa* means 'young woman', *djovena* in Istanbul, and in Salonika, it can be used with both meanings. The problem is solved by assigning to the lemma *djoven* 'young' all the textual occurrences with this meaning, like *djovena*, *mosa, moso, moçuelo* (Spanish lemma: *joven*), and to the lemma *moso* 'servant' (Spanish lemma: *criado*) all those that have the same meaning, like *mosa* and *dishipla.* If we search for the Spanish lemma *criado*, the results show in the KWIC line only occurrences of *mosa* or *moso* meaning 'servant' in the texts, with the Judeo-Spanish lemma *moso*, while when *mosa* or *moso* have the meaning of 'young', they would always appear under the Spanish lemma *joven*, while *djoven* must also be the Judeo-Spanish lemma. Otherwise, the search results would contain errors. Therefore, it is essential to ensure uniformity in the task of the Judeo-Spanish lemmatization, following as much as possible the variety spoken in Istanbul, selected as the standard variety for the lemmatization in *CoDiAJe*, and the lemmatization in Spanish can help in this task.

The problem arises when there is not a total equivalence of meaning between cognate forms such as the Spanish noun *joven* and the Judeo-Spanish *djoven*. Judeo-Spanish words such as *muchacha* and *manseva* may also appear in the KWIC line on a search for the Spanish lemma *joven* as a noun, because the former has the meaning of 'young' in the Sarajevo variety and the latter is used with the meaning 'young' in all Judeo-Spanish varieties. It has been proved that lemmatization in Spanish can also facilitate the query when the user lacks knowledge of Judeo-Spanish, because it is possible to extract from the texts all the words and variants that in Judeo-Spanish have the same meaning as the Spanish lemma. Figure 21 contains the linguistic annotations of 38 forms included in *CoDiAJe*, tagged according to the exposed criteria.



		POS	Judeo-Spanish lemma	Modern Spanish lemma
1	mosa	NCFS0000	djoven	joven
2	mosa	NCFS0000	moso	criado
3	mosas	AQ0FP0000	moso	soltero
4	dishipla	NCFS0000	moso	criado
5	mansevo	NCMS0000	mansevo	joven
6	mansevu	AQ0MS0000	mansevo	soltero
7	muchacha	NCFS0000	muchacho	joven
8	envaniko	RGD0	envano	en_vano
9	sus (meza de eyos/as)	DP3CSP0	su	su

10	sus (livros de eyos/eyas)	DP3CPP0	su	su
11	vinyendo	VMM02C0	venir	venir
12	unrada	TMSFS0	onrar	honrar
13	estan	TMPCS0	estar	estar
14	zoher	T0NMSH	zoher	digno
15	(tratando)se	L3CS00	se	se
16	(alevantando)sen	L3CP00	se	se
17	ken	PT0CN0000	ken	quién
18	ken	PR0CN0000	ken	quien
19	Erets Israel	NP000G00	israel	israel
20	Albert Einstein	NP000P00	albert_einstein	albert_einstein
21	Neve Şalom	NP000O00	Neve_shalom	nevé_šalom
22	*Regimiento dela Vida*	NP000T00	rejimiento_de_la_vida	regimiento_de_la_vida
23	Alto i Potente	NP000V00	alto_i_potente	alto_y_potente
24	go’el	NP000V0H	goel	redentor
25	henozo	AQ0MS000H	henozo	agraciado
26	kafuy tova	AQ0CN000H	kafuy_tova	ingrato
27	afilu	RG0H	afilu	incluso
28	beṣar	RG0H	besar	con_dolor
29	yine	RG0Y	yene	nuevamente
30	haver	NCMS000H	haver	amigo
31	chanta	NCFS000Y	chanta	bolso
32	ḥabb el ʿaziz	NCMS000A	havachichi	chufa
33	ama	CCY	ama	pero
34	ע"פ (= al pi)	SPS00H	al_pi	según
35	bre	IK	bre	vamos
36	mizerah amiluha	H	mi_zerah_a_meluha	de_estirpe_real
37	Ish bitiren para	Y	ish_bitiren_para	iš_bitirén_pará
38	וירא אליו ה' באלוני ממרא וגו'	H	vayera_elav_a_shem_be_ elone_mamre_ve_gomer	vayerá_elav_ha_šem_be_ eloné_mambré_ve_gomer

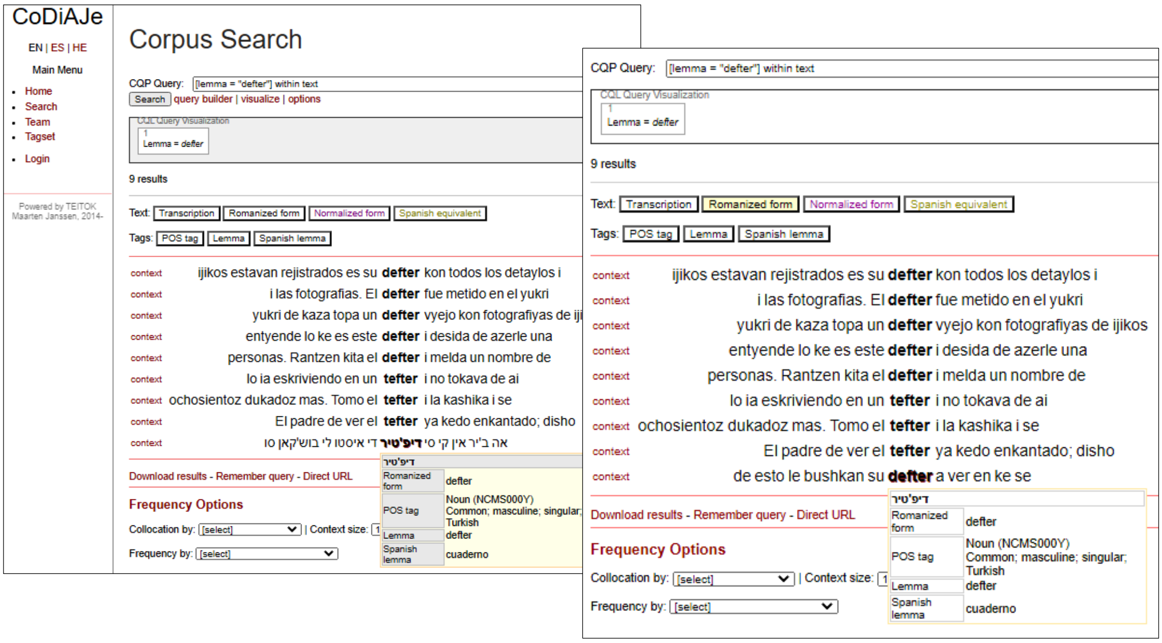
Figure 21. Different kinds of orthographic forms tagged according to the criteria for *CoDiAJe*.

The first 7 lines show the linguistic annotation of the occurrences discussed in the previous paragraph. Lines 8-18 contain tags that needed to be modified or created. From line 19 to 24, we can see POS with semantic-conceptual information (G = geographical names; P = private or fictional names of people; O = institutional names; T = titles of works; V = epithets referring to God) in position 5 of the tag. Capital letters in the end position of the POS offer information concerning the linguistic affiliation of words or groups of words borrowed from non-Romance languages that are part of the Judeo-Spanish lexicon (for instance, H indicates a Hebrew origin of the term; Y refers to Turkish, K to Greek, and A to non-Hispanic Arabic, as seen in lines 25-35). The same tags without any other information indicate only terms, expressions, sayings, proverbs, blessings, and curses, and quotations in non-

Romance languages interpolated in Judeo-Spanish texts (lines 36-38). With this information, it is possible to retrieve the complete list of terms from Hebrew or Turkish or other non-Romance contact languages documented in *CoDiAJe* in a single query. A summarized description of the corpus tagset is available in the main menu of *CoDiAJe*.

3. QUERY

CoDiAJe is searchable via TEITOK, which interfaces with a local CQP. With the annotated corpus, adequately indexed for exploitation via the CQP search engine, it has become possible to conduct searches not only for a specific word and any of its variants through XML files directly using the CQP query, but also to query using CQL directly on the website. This allows for the conduct of searches for all kinds of variants, and for specific sequences of expressions, grammatical categories and combinations thereof, and makes it easy to carry out various types of quantitative analyses (e.g. relative frequencies, distribution). This is very important in order to draw statistical inferences about the degree of incidence of several variants exposed to linguistic changes, to map out the relative frequency of each form, and to draw conclusions from the use of variants in Judeo-Spanish.



CQP Query: [lemma = "defter"] within text

1
Lemma = defter

9 results

context ijikos estavan rejistrados es su **defter** kon todos los detaylos i
context i las fotografias. El **defter** fue metido en el yukri
context yukri de kaza topa un **defter** vyejo kon fotografiyas de ij
context entyende lo ke es este **defter** i desida de azerle una
context personas. Rantzen kita el **defter** i melda un nombre de
context lo ia eskriviendo en un **tefter** i no tokava de ai
context ochosientoz dukadoz mas. Tomo el **tefter** i la kashika i se
context El padre de ver el **tefter** ya kedo enkantado; disho
context די איסטו לי בושקאן סו **דיפ'טיר** אה ב'יר אין קי סי

דיפ'טיר	
Romanized form	defter
POS tag	Noun (NCMS000Y) Common; masculine; singular; Turkish
Lemma	defter
Spanish lemma	cuaderno

CQP Query: [lemma = "defter"] within text

1
Lemma = defter

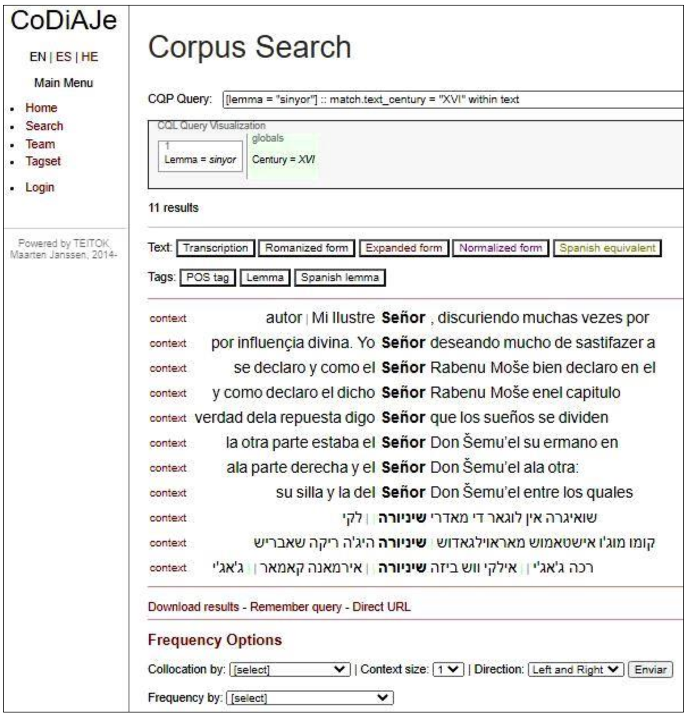
9 results

context ijikos estavan rejistrados es su **defter** kon todos los detaylos i
context i las fotografias. El **defter** fue metido en el yukri
context yukri de kaza topa un **defter** vyejo kon fotografiyas de ijikos
context entyende lo ke es este **defter** i desida de azerle una
context personas. Rantzen kita el **defter** i melda un nombre de
context lo ia eskriviendo en un **tefter** i no tokava de ai
context ochosientoz dukadoz mas. Tomo el **tefter** i la kashika i se
context El padre de ver el **tefter** ya kedo enkantado; disho
context de esto le bushkan su **defter** a ver en ke se

דיפ'טיר	
Romanized form	defter
POS tag	Noun (NCMS000Y) Common; masculine; singular; Turkish
Lemma	defter
Spanish lemma	cuaderno

Figure 22. Variants of the lemma *defter* ‘notebook, account book’, queried by lemma and visualized by the orthographic transcription and the Romanized transcription.

One of *CoDiAJe*’s achievements is that a single query can yield results for all variants without losing track of the original spelling form. For example, a search for the lemma *defter* yields results for all the orthographic and grammatical forms of this lemma in the corpus. The results appear in the browser, showing the KWIC line for each of them. Displaying them in the Romanized transcription is also possible, which allows for visualization of all variants in the transcription in Latin characters (Figure 22). When the purpose of the query is not the variation, the results can be displayed in modern standardized Judeo-Spanish or modern Spanish spelling. It is also possible to search in a single text or in a group of texts, predetermined in the query. Figure 23 shows the result of the search for the lemma *sinyor* in 16th-century texts.


CoDiAJe
EN | ES | HE
Main Menu
Home
Search
Team
Tagset
Login
Powered by TEITOK
Maarten Janssen, 2014-
Text: Transcription | Romanized form | Expanded form | Normalized form | Spanish equivalent
Tags: POS tag | Lemma | Spanish lemma
Download results - Remember query - Direct URL
Collocation by: [select] | Context size: 1 | Direction: Left and Right | Enviar
Frequency by: [select]

Corpus Search

CQP Query: [lemma = "sinyor"] :: match.text_century = "XVI" within text

CQL Query Visualization
1
Lemma = *sinyor*
globals
Century = *XVI*

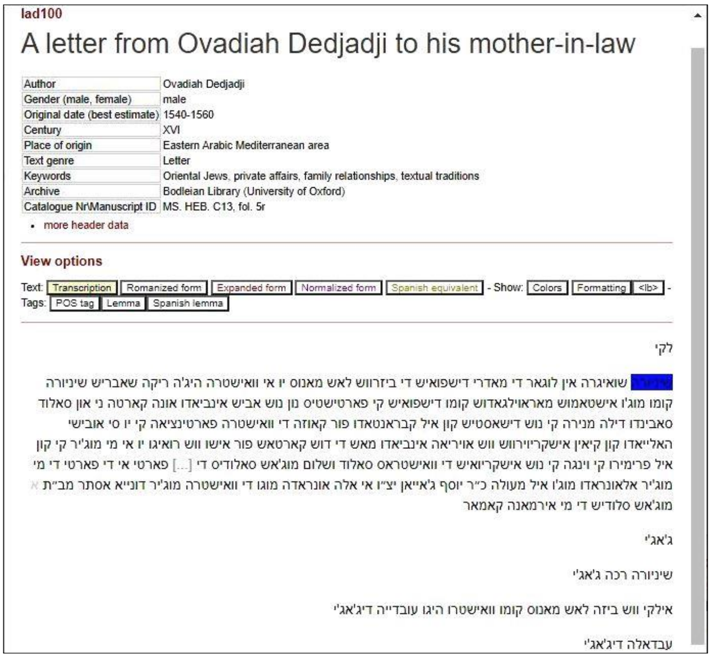
11 results

context autor | Mi Ilustre **Señor** , discuriendo muchas vezes por
context por influençia divina. Yo **Señor** deseando mucho de sastifazer a
context se declaro y como el **Señor** Rabenu Moše bien declaro en el
context y como declaro el dicho **Señor** Rabenu Moše enel capitulo
context verdad dela repuesta digo **Señor** que los sueños se dividen
context la otra parte estaba el **Señor** Don Šemu'el su ermano en
context ala parte derecha y el **Señor** Don Šemu'el ala otra:
context su silla y la del **Señor** Don Šemu'el entre los quales
context שואיגרה אין לוגאר די מאדרי **שיניורה** | לקי
context קומו מוג'ו אישטאמוש מאראוילגאדוש | **שיניורה** היג'ה ריקה שאבריש
context רכה ג'אג'י | אילקי ווש ביזה **שיניורה** | אירמאנה קאמאר | ג'אג'י

Frequency Options

Figure 23. The lemma *sinyor* in 16th-century texts.

By clicking on the word context in each line of the KWIC, it is possible to display the word form of the lemma in the text in the selected orthographic layer (Figure 24) or switch between the various options on the top menu. Further, the visualization of the text without or with the linguistic annotations is possible by clicking on the *Tags* buttons at the top (Figure 25).


View options
Text: Transcription | Romanized form | Expanded form | Normalized form | Spanish equivalent - Show: Colors | Formatting | <lb> -
Tags: POS tag | Lemma | Spanish lemma
more header data

lad100

A letter from Ovadiah Dedjadji to his mother-in-law

Author	Ovadiah Dedjadji
Gender (male, female)	male
Original date (best estimate)	1540-1560
Century	XVI
Place of origin	Eastern Arabic Mediterranean area
Text genre	Letter
Keywords	Oriental Jews, private affairs, family relationships, textual traditions
Archive	Bodleian Library (University of Oxford)
Catalogue Nr\Manuscript ID	MS. HEB. C13, fol. 5r

לקי

שיניורה שואיגרה אין לוגאר די מאדרי דישפואיש די ביזרווש לאש מאנוס יו אי וואישטרה היג'ה ריקה שאבריש שיניורה
קומו מוג'ו אישטאמוש מאראוילגאדוש קומו דישפואיש קי פארטישטיס נון נוש אביש אינביאדו אונה קארטה ני און סאלוד
סאבינדו דילה מנירה קי נוש דישאסטיש קון איל קבראנטאדו פור קאוזה די וואישטרה פארטינציאה קי יו סי אובישי
האלייאדו קון קיאין אישקריוירווש ווש אויריאה אינביאדו מאש די דוש קארטאש פור אישו ווש רואיגו יו אי מי מוג'יר קי קון
איל פרימירו קי וינגה קי נוש אישקריואיש די וואישטראס סאלוד ושלום מוג'אש סאלודיס די [...] פארטי אי די פארטי די מי
מוג'יר אלאונראדו מוג'ו איל מעולה כ״ר יוסף ג'אייאן יצ״ו אי אלה אונראדה מוגו די וואישטרה מוג'יר דונייא אסתר מב״ת א
מוג'אש סלודיש די מי אירמאנה קאמאר

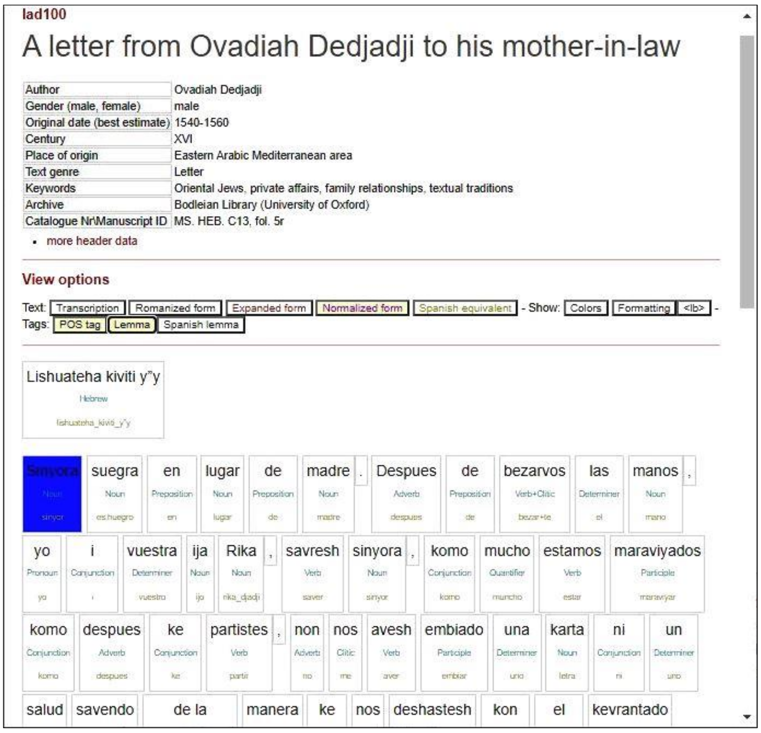
ג'אג'י

שיניורה רכה ג'אג'י

אילקי ווש ביזה לאש מאנוס קומו וואישטרו היגו עובדייה דיג'אג'י

עבדאלה דיג'אג'י

Figure 24. Visualization of the form שיניורה (occurrence in KWIC line 9 in Figure 23) in the text, displayed in the original orthographic form.



lad100

A letter from Ovadiah Dedjadji to his mother-in-law

Author	Ovadiah Dedjadji
Gender (male, female)	male
Original date (best estimate)	1540-1560
Century	XVI
Place of origin	Eastern Arabic Mediterranean area
Text genre	Letter
Keywords	Oriental Jews, private affairs, family relationships, textual traditions
Archive	Bodleian Library (University of Oxford)
Catalogue Nr\Manuscript ID	MS. HEB. C13, fol. 5r

- more header data

View options

Text: Transcription | Romanized form | Expanded form | Normalized form | Spanish equivalent - Show: Colors | Formatting | <lb> -
Tags: POS tag | Lemma | Spanish lemma

Lishuateha kiviti y"y
Hebrew
lishuateha_kiviti_y'y

Sinyora	suegra	en	lugar	de	madre	.	Despues	de	bezarvos	las	manos	,
Noun	Noun	Preposition	Noun	Preposition	Noun		Adverb	Preposition	Verb+Clitic	Determiner	Noun	
sinyor	eshuegro	en	lugar	de	madre		despues	de	bezar+te	el	mano	

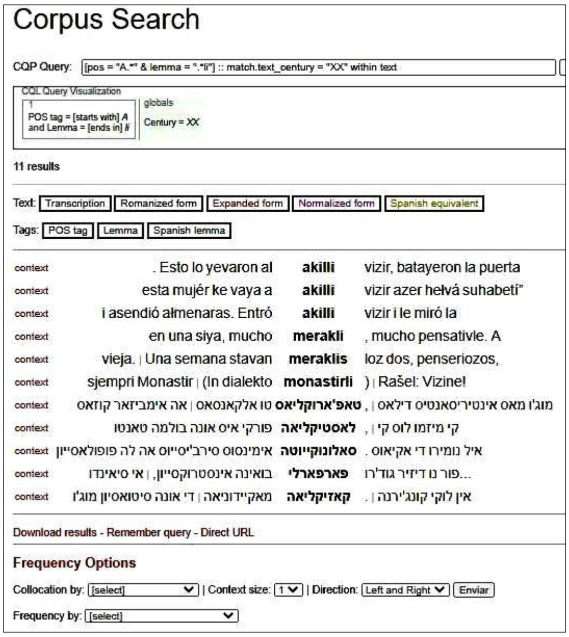
yo	i	vuestra	ija	Rika	,	savresh	sinyora	,	komo	mucho	estamos	maraviyados
Pronoun	Conjunction	Determiner	Noun	Noun		Verb	Noun		Conjunction	Quantifier	Verb	Participle
yo	i	vuestro	ijo	rika_djadji		saver	sinyor		komo	muncho	estar	maraviyar

komo	despues	ke	partistes	,	non	nos	avesh	embiado	una	karta	ni	un
Conjunction	Adverb	Conjunction	Verb		Adverb	Clitic	Verb	Participle	Determiner	Noun	Conjunction	Determiner
komo	despues	ke	partir		no	me	aver	embiar	uno	letra	ni	uno

salud savendo de la manera ke nos deshastesh kon el kevrantado

Figure 25. Visualization of the original orthographic form שיניורה in modern standardized Judeo-Spanish, displayed with lemma and POS tags.

In TEITOK, the CWB files are created directly from the XML files (Janssen 2018). It is therefore possible to define more complex queries, combining filters of all levels. In Figure 26 the search is limited to adjectives whose lemma ends in the morpheme *-li* in texts written in the 20th century. Their distribution and frequency by author are shown in Figure 27.



Corpus Search

CQP Query: [pos = "A.*" & lemma = ".*li"] :: match.text_century = "XX" within text

CQL Query Visualization
1
POS tag = [starts with] *A* and Lemma = [ends in] *li*
globals
Century = XX

11 results

Text: Transcription | Romanized form | Expanded form | Normalized form | Spanish equivalent
Tags: POS tag | Lemma | Spanish lemma

context	. Esto lo yevaron al	**akilli**	vizir, batayeron la puerta
context	esta mujér ke vaya a	**akilli**	vizir azer helvá suhabetí"
context	i asendió almenaras. Entró	**akilli**	vizir i le miró la
context	en una siya, mucho	**merakli**	, mucho pensativle. A
context	vieja. | Una semana stavan	**meraklis**	loz dos, penseriozos,
context	sjempri Monastir | (In dialekto	**monastirli**	) | Rašel: Vizine!
context	קוזאס אה אימביזאר | טו אלקאנסאס	**טאפ'ארוקליאס**	, | מוג'ו מאס אינטיריסאנטיס דילאס
context	פורקי איס אונה בולמה טאנטו	**לאסטיקליאה**	, | קי מיזמו לוס קי
context	פופולאסיון אה לה סירב'יסייוס אימינסוס	**סאלונוקייוטה**	. איל נומירו די אקיאוס
context	בואינה אינסטרוקסיון, | אי סיאינדו	**פארפארלי**	...פור נו דיזיר גוד'רו
context	מאקיידוניאה | די אונה סיטואסיון מוג'ו	**קאזיקליאה**	. | אין לוקי קונג'ירנה

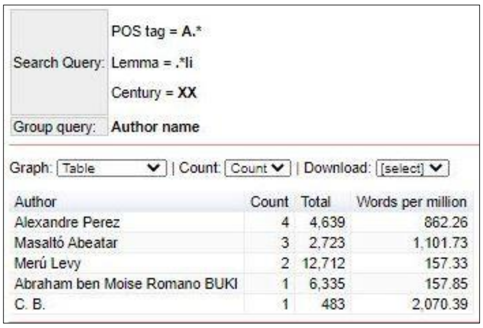
Download results - Remember query - Direct URL

Frequency Options

Collocation by: [select] | Context size: 1 | Direction: Left and Right | Enviar

Frequency by: [select]

Figure 26. Adjectives with lemma ending in *-li* in texts written in the 20th century.



Search Query:	POS tag = **A.***
	Lemma = .*li
	Century = **XX**

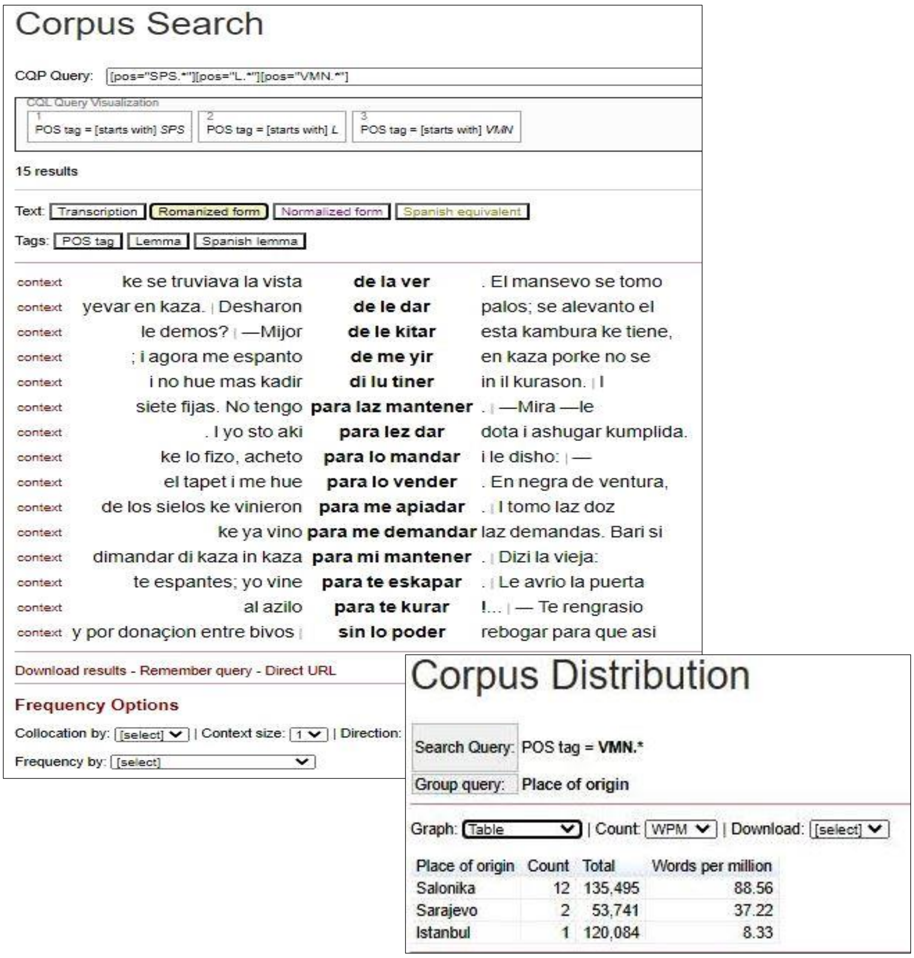
Group query:	Author name

Graph: Table | Count: Count | Download: [select]

Author	Count	Total	Words per million
Alexandre Perez	4	4,639	862.26
Masaltó Abeatar	3	2,723	1,101.73
Merú Levy	2	12,712	157.33
Abraham ben Moise Romano BUKI	1	6,335	157.85
C. B.	1	483	2,070.39

Figure 27. Distribution and frequency by author of the results obtained in the previous search.

Figure 28 shows the Romanized visualization of the results of the search for any infinitive preceded by a clitic personal pronoun after a preposition in the text collection of *CoDiAJe*, and their distribution by the place in which the texts with this word order were written. As Figure 29 shows, it is also possible to retrieve morphological information, for example on the gender of feminine Hebrew nouns ending in *-ut* borrowed into Judeo-Spanish, and confirm that they only emerge with masculine determiners.



Corpus Search

CQP Query: [pos="SPS.*"][pos="L.*"][pos="VMN.*"]

CQL Query Visualization
1 POS tag = [starts with] *SPS*
2 POS tag = [starts with] *L*
3 POS tag = [starts with] *VMN*

15 results

Text: Transcription | Romanized form | Normalized form | Spanish equivalent
Tags: POS tag | Lemma | Spanish lemma

context	ke se truviava la vista	**de la ver**	. El mansevo se tomo
context	yevar en kaza. | Desharon	**de le dar**	palos; se alevanto el
context	le demos? | —Mijor	**de le kitar**	esta kambura ke tiene,
context	; i agora me espanto	**de me yir**	en kaza porke no se
context	i no hue mas kadir	**di lu tiner**	in il kurason. | I
context	siete fijas. No tengo	**para laz mantener**	. | —Mira —le
context	. I yo sto aki	**para lez dar**	dota i ashugar kumplida.
context	ke lo fizo, acheto	**para lo mandar**	i le disho: | —
context	el tapet i me hue	**para lo vender**	. En negra de ventura,
context	de los sielos ke vinieron	**para me apiadar**	. | I tomo laz doz
context	ke ya vino	**para me demandar**	laz demandas. Bari si
context	dimandar di kaza in kaza	**para mi mantener**	. | Dizi la vieja:
context	te espantes; yo vine	**para te eskapar**	. | Le avrio la puerta
context	al azilo	**para te kurar**	!... | — Te rengrasio
context	y por donaçion entre bivos |	**sin lo poder**	rebogar para que asi

Download results - Remember query - Direct URL

Frequency Options

Collocation by: [select] | Context size: 1 | Direction:
Frequency by: [select]

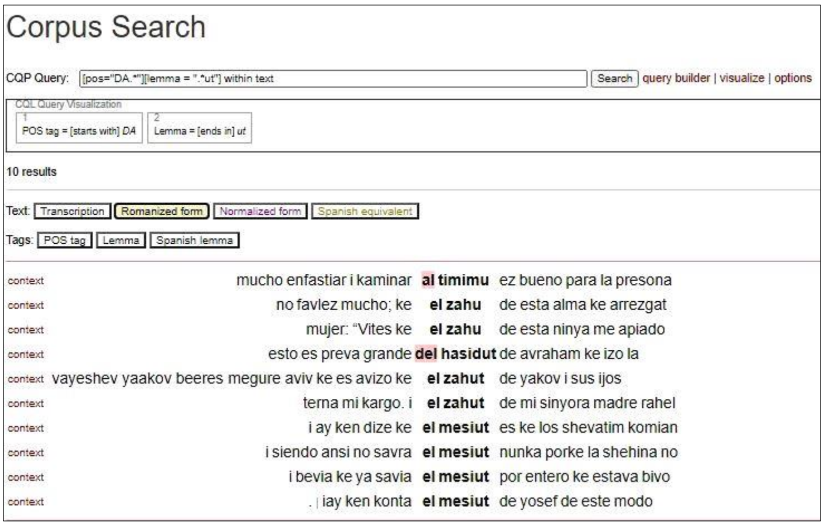
Corpus Distribution

Search Query:	POS tag = **VMN.***
Group query:	Place of origin

Graph: Table | Count: WPM | Download: [select]

Place of origin	Count	Total	Words per million
Salonika	12	135,495	88.56
Sarajevo	2	53,741	37.22
Istanbul	1	120,084	8.33

Figure 28. Output of the preverbal position of clitic personal pronouns before an infinitive, and their distribution by place.



Corpus Search

CQP Query: [pos="DA.*"][lemma = ".*ut"] within text

Search
query builder | visualize | options

CQL Query Visualization
1 POS tag = [starts with] *DA*
2 Lemma = [ends in] *ut*

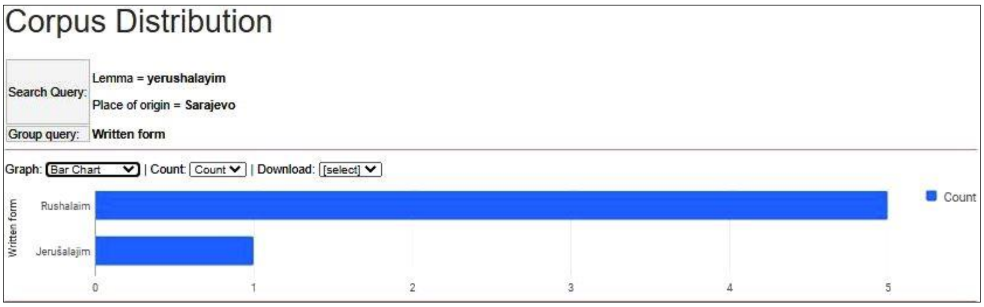
10 results

Text: Transcription | Romanized form | Normalized form | Spanish equivalent
Tags: POS tag | Lemma | Spanish lemma

```
context                                         mucho enfastiar i kaminar  al timimu  ez bueno para la presona
context                                              no favlez mucho; ke   el zahu    de esta alma ke arrezgat
context                                                   mujer: "Vites ke  el zahu    de esta ninya me apiado
context                                           esto es preva grande  del hasidut de avraham ke izo la
context  vayeshev yaakov beeres megure aviv ke es avizo ke  el zahut   de yakov i sus ijos
context                                                terna mi kargo. i  el zahut   de mi sinyora madre rahel
context                                                   i ay ken dize ke  el mesiut  es ke los shevatim komian
context                                            i siendo ansi no savra  el mesiut  nunka porke la shehina no
context                                              i bevia ke ya savia  el mesiut  por entero ke estava bivo
context                                                  . | iay ken konta  el mesiut  de yosef de este modo
```

Figure 29. Search for information about the grammatical gender
of borrowed nouns, such as those ending in *-ut*.

As already noted in §2.2, *CoDiAJe* allows searching for specific semantic-conceptual information, such as place names mentioned in the collection of texts or some of them in particular. For example, the bar chart in Figure 30 presents the distribution of the two variants of the lemma *yerushalayim* found in texts from Sarajevo.



Corpus Distribution

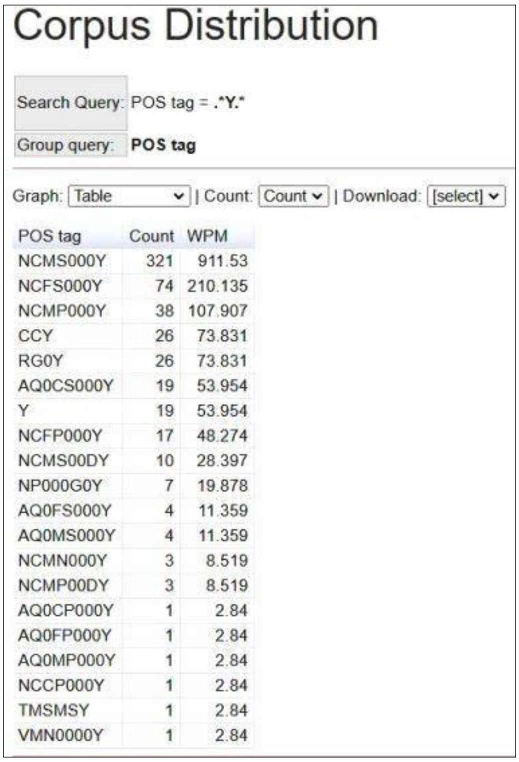
Search Query: Lemma = **yerushalayim**
Place of origin = **Sarajevo**

Group query: **Written form**

Graph: Bar Chart | Count: Count | Download: [select]

Figure 30. Distribution of local variants
of the lemma *yerushalayim* 'Jerusalem' in Sarajevo texts.

A search containing, for example, H or Y in the POS will retrieve all forms borrowed from Hebrew, Turkish, or another non-Romance language with which the Judeo-Spanish speakers were in contact, or quotations, sayings, and proverbs in these languages, as well as other lexical elements not integrated into the Judeo-Spanish system. Figure 31 shows the two kinds of lexical units: all the conventional POS tags ending in Y correspond to integrated words borrowed from Turkish, while the isolated Y tag refers to Turkish lexicon that sometimes appears inserted in Judeo-Spanish texts.



Corpus Distribution

Search Query: POS tag = .*Y.*

Group query: **POS tag**

Graph: Table | Count: Count | Download: [select]

POS tag	Count	WPM
NCMS000Y	321	911.53
NCFS000Y	74	210.135
NCMP000Y	38	107.907
CCY	26	73.831
RG0Y	26	73.831
AQ0CS000Y	19	53.954
Y	19	53.954
NCFP000Y	17	48.274
NCMS00DY	10	28.397
NP000G0Y	7	19.878
AQ0FS000Y	4	11.359
AQ0MS000Y	4	11.359
NCMN000Y	3	8.519
NCMP00DY	3	8.519
AQ0CP000Y	1	2.84
AQ0FP000Y	1	2.84
AQ0MP000Y	1	2.84
NCCP000Y	1	2.84
TMSMSY	1	2.84
VMN0000Y	1	2.84

Figure 31. Distribution by POS tags of all Turkish forms
occurring in Judeo-Spanish texts already incorporated in *CoDiAJe* and tagged.

In the near future, *CoDiAJe* will have other resources that are included in TEITOK, such as (a) a dictionary with the vocabulary of the corpus, and (b) a module for the development of linguistic maps. Later, the module for syntactic annotation (c) will be added.

4. CONCLUDING REMARKS

CoDiAJe now has the potential of functionality allowed by TEITOK's basic design, to which several new features have been added to adapt it to the specific needs of a Judeo-Spanish corpus. First, to the three standard attributes of each tok (transcription, an expanded, and a normalized form), two new attributes were added (a Romanized form, and a Spanish equivalent), yielding up to five orthographic forms of the same word. This addition allows us to visualize each text in five different orthographic forms and, not least, to retrieve all the orthographic and linguistic variants of a lemma through the query.

In more recent times –since the 1940s– Judeo-Spanish documents have been written in Latin characters, albeit using various versions of orthographic systems. Some documents have also been written in the Cyrillic and Greek alphabets. However, until the late 19th century all documents were written in Hebrew characters. Consequently, a requisite in the development of *CoDiAJe* was to obtain a tool with the option of incorporating texts in all the alphabets in which they were originally written or published, and correctly visualizing them without interfering with the search task in the corpus. This paper shows that all these requirements were achieved thanks to Maarten Janssen, TEITOK developer and a collaborator of this project.

CoDiAJe —like any other corpus created in TEITOK— is easy to use by editors who are not experts in NLP, and by users. Other important advantages lie in the ease with which errors detected in the corpus can be corrected or necessary changes can be made promptly or at any time. It is well known that textual and linguistic annotation tasks in a corpus of a limitedly standardized language —such as Judeo-Spanish— are not always easy. The annotation in *CoDiAJe* requires continuous revision every time that forms emerge in the texts for which the EAGLES tagset for Spanish —and for other European languages as well— has no tags. Tags modified or especially created for Judeo-Spanish are simply steps in the creation of an accurate tagset for Judeo-Spanish. The tagging work itself and the test of the adequacy of each annotation through searching offer the possibility of analyzing the language and, frequently, reveal the need to modify the annotation of a particular token, when morphological features not mentioned in the secondary literature are detected.

An unlimited number of new texts can be incorporated into *CoDiAJe*, and the annotation can be improved at any time. The most immediate goal is to incorporate new texts in Hebrew characters that have already been extracted from the image and converted into Word text, as a result of which the corpus will soon have more than three million words.

CoDiAJe also meets other conditions required in corpus linguistics: the possibility of attaching the facsimile alongside the text, descriptive statistics, and the complete download of documents in XML format and plain text. Moreover, it offers the possibility of adding other resources, including (a) a dictionary with the vocabulary of the corpus, (b) a module for the development of linguistic maps, and (c) a module for syntactic annotation.

In view of *CoDiAJe*'s potential for text-processing and its capacity to store an endless number of documents in a single virtual library, a significant part of the Sephardic documentary heritage will be made available to a wide number of scholars from different fields and readers who do not have access to it at present.

Despite being still in its development phase, *CoDiAJe* should serve as a guide in the development of diachronic corpora of majority or minority languages that have been written in different alphabets throughout their history, and for historical varieties of languages that have been written in a different alphabet from the standardized variety, for example, part of the *morisco* texts, or documents written in other Judeo-Romance varieties. A corpus like this makes it easier to detect the linguistic variation that characterizes such languages and linguistic varieties. The multi-alphabetic nature of *CoDiAJe* additionally allows visualizing the texts in their original orthography, regardless of the alphabet in which they were written, and in their Romanized and modernized versions. This would contribute to disseminating these cultural heritages and would provide linguists and philologists with the possibility of analyzing each variety of the language, taking into account the set of all its other varieties, and scholars in other fields could benefit from the information frozen in the endless list of documents still hidden in archives and, in many cases, unintelligible for one reason or another.

REFERENCES

ÁLVAREZ LÓPEZ, Cristóbal José (2017): *Estudio lingüístico del judeoespañol en la revista "Aki Yerushalayim"*. Sevilla: Universidad de Sevilla. PhD. dissertation directed by José Javier Rodríguez Toro and Aitor García Moreno.

ARNOLD, Rafael D. (in this volume): «La digitalización del fichero del *Diccionario del Español Medieval* (*DEM*): una nueva fuente para la historia del español y del judeoespañol», in Miriam Bouzouita and Antoine Primerano (eds.), *Lingüística de corpus e historias de las lenguas iberorrománicas: Nuevas propuestas y últimos desarrollos*, *Scriptum digital*, 9, pp. 191-207.

BRADLEY, Travis G. and Ann Marie DELFORGE (2006): «Phonological Retention and Innovation in the Judeo-Spanish of Istanbul», in Timothy L. Face and Carol A. Klee (eds.), *Selected Proceedings of the 8th Hispanic Linguistics Symposium*. Somerville, MA: Cascadilla Proceedings Project, pp. 73-88.

BUNIS, DAVID M. (1993): *A Lexicon of Hebrew and Aramaic Elements in Modern Judezmo*. Jerusalem: The Magnes Press/Misgav Yerushalayim.

BUNIS, David M. (2009): «Judezmo Analytic Verbs with a Hebrew-Origin Pariciple: Evidence of Ottoman Incluence», in David M. Bunis (ed.), *Languages and Literatures of Sephardic and Oriental Jews*. Jerusalem: Misgav Yerushalayim/The Bialik Institute, pp. 94-166.

BUNIS, David M. (2019): «La ortografia de Aki Yerushaluyim: Un pinukolo en la estoria de la romanizasión del djuezmo (djudeo-espanyol)», *Aki Yerushaluyim*, 101. http://www.akiyerushalayim.com/ay/101/101_03_ortografia.htm [Accessed: 08/07/2020].

BUSSE, Winfried (2005): «Rashí. Transliteración, transcripción y adaptación de textos aljamiados», *Neue Romania*, 34 (= *Judenspanisch*, IX), pp. 97-107.

CÁRDENAS, John (2004): «Judeo-Spanish and the Lexicalist Morphology Hypothesis: A Vindication of Inflectional and Derivational Morphology», *California Linguistic Notes*, 29, 1, pp. 1-23. http://english.fullerton.edu/publications/clnArchives/pdf/cardenas_jslmh.pdf [Accessed: 30/06/2020].

DiJeSt = RUSINEK, Sinai: *DiJeSt: Digitizing Jewish Studies*. http://dijest.net/ [Accessed: 15/08/2020].

GARCÍA MORENO, Aitor (2010): «El judeoespañol II: Características». Madrid: Liceus. https://aprende.liceus.com/producto/judeoespanol-ii-caracteristicas/ [Accessed: 10/09/2019].

HUALDE, José Ignacio (2013): «Language Contact and Change in the Sound System of Judeo-Spanish», in Mahir Şaul (ed.), *Judeo-Spanish in the Time of Clamoring Nationalisms*. Istanbul: Libra kitap, pp. 151-178.

HUALDE, José Ignacio and Mahir ŞAUL (2011): «Istanbul Judeo-Spanish», *Journal of the International Phonetic Association*, 41, 1, pp. 89-110.

JANSSEN, Maarten (2012): «NeoTag: A POS Tagger for Grammatical Neologism Detection», in Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*. Istanbul: ELRA (European Language Resources Association), s. p. http://maarten.janssenweb.net/Papers/neotag-lrec.pdf [Accessed: 26/09/2020].

JANSSEN, Maarten (2016): «TEITOK: Text-Faithful Annotated Corpora», in Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, Paris: ELRA (European Language Resources Association), pp. 4037-4043. http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf [Accessed: 09/07/2020].

JANSSEN, Maarten (2018): «TEITOK as a Tool for Dependency Grammar», *Procesamiento del Lenguaje Natural*, 61, pp. 185-188. doi:http://dx.doi.org/10.26342/2018-61-28 [Accessed: 09/07/2020]

JANSSEN, Maarten, Josep AUSENSI and Josep M. FONTANA (2017): «Improving POS Tagging in Old Spanish Using TEITOK», in Gerlof Bouma and Yvonne Adesam (eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language.* Gothenburg: Linköping University Electronic Press (*NEALT Proceedings Series,* 32), pp. 2-6. https://www.aclweb.org/anthology/W17-0502.pdf [Accessed: 09/07/2020].

LLEAL GALCERÁN, Coloma (2004): «El judeoespañol», in Rafael Cano (coord.), *Historia de la lengua española*. Barcelona: Ariel, pp. 1139-1167.

LYONS, John (1981): *Language and Linguistics*. Cambridge: Cambridge University Press.

MANNING, Christopher D. and Hinrich SCHÜTZE (1999): *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

MINERVINI, Laura (2006): «El desarrollo histórico del judeoespañol», *Revista Internacional de Lingüística Iberoamericana*, 8, pp. 13-34.

MUÑOZ JIMÉNEZ, Isabel (1997): «Perífrasis verbales híbridas en judeoespañol literario», *Revista de Filología Románica*, 14, pp. 363-390.

OntoSem = GLIF (Grupo de Lingüística Formal): *OntoSem Corpora*. http://corptedig-glif.upf.edu/ontosem-corpora/ [Accessed: 15/08/2020].

PADRÓ, Lluís (2011): «Analizadores Multilingües en FreeLing», *Linguamatica*, 3, 2, pp. 13-20. http://www.lsi.upc.edu/~nlp/papers/padro11.pdf [Accessed: 06/07/2020].

PADRÓ, Lluís, Miquel COLLADO, Samuel REESE, Marina LLOBERES and Irene CASTELLÓN (2010): «FreeLing 2.1: Five Years of Open-Source Language Processing Tools», in *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC 2010)*. La Valletta: ELRA. http://www.lrec-conf.org/proceedings/lrec2010/pdf/14_Paper.pdf [Accessed: 06/07/2020].

PENNY, Ralph (2000): *Variation and Change in Spanish*. Cambridge: Cambridge University Press.

Post Scriptum = CLUL (ed.) (2014): *P. S. Post Scriptum: Arquivo Digital de Escritura Quotidiana em Portugal e Espanha na Época Moderna*. http://ps.clul.ul.pt/pt/index.php? [Accessed: 09/07/2020].

QUINTANA, Aldina (2006): *Geografía lingüística del judeoespañol: Estudio sincrónico y diacrónico*. Berna: Peter Lang.

QUINTANA, Aldina (2010): «El judeoespañol, una lengua pluricéntrica al margen del español», in Paloma Díaz-Mas and María Sánchez Pérez (eds.), *Los sefardíes ante los retos del mundo contemporáneo. Indentidad y mentalidades*. Madrid: Consejo Superior de Investigaciones Científicas, pp. 33-54.

SÁNCHEZ-MARCO, Cristina, Gemma BOLEDA, Josep Maria FONTANA and Judith DOMINGO (2010): «Annotation and Representation of a Diachronic Corpus of Spanish», in *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC 2010)*. La Valletta: ELRA. https://upcommons.upc.edu/bitstream/handle/2117/10373/535_Paper.pdf?sequence=1&isAllowed=y [Accessed: 08/07/2020].

SÁNCHEZ-MARCO, Cristina, Gemma BOLEDA and Lluís PADRÓ (2011): «Extending the Tool, or How to Annotate Historical Language Varieties», in Kalliopi Zervanou and Piroska Lendvai (eds.), *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Stroudsburg (PA): Association for Computational Linguistics, pp. 1-9. http://dl.acm.org/citation.cfm?id=2107637&CFID=979433322&CFTOKEN=43121501 [Accessed: 08/07/2020].

SÁNCHEZ-MARCO, Cristina, Josep Maria FONTANA and Judith DOMINGO (2012): «Anotación automática de textos diacrónicos del español», in Emilio Montero and Carmen Manzano (coords.), *Actas* *del VIII Congreso Internacional de Historia de la Lengua Española, Santiago de Compostela, 14-18 de septiembre de 2009*, vol. 2. Santiago de Compostela: Meubook: Asociación de Historia de la Lengua Española (AHLE), pp. 1709-1720.

TRUDGILL, Peter (2011): *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

VARVARO, Alberto and Laura MINERVINI (2008): «Orígenes del Judeoespañol (II): comentario lingüístico», *Revista de Historia de la Lengua Española (RHLE),* 3, pp. 149-195.