Research

Peer Review:

Copyright:

Open Access:

Digital Preservation:

RESEARCH

What Loanwords Tell Us about Spanish (and Spain)

Adriana Soto-Corominas[1], Javier De la Rosa[2] and Juan Luis Suárez[1]

[1] University of Western Ontario, CA

[2] Stanford University, US

Corresponding author: Adriana Soto-Corominas (asotocor@uwo.ca)

Using one lexicographical tool, the Diccionario Crítico Etimológico Castellano e Hispánico (DECH), and two corpora, the HathiTrust's digital library and the Google Books Ngrams, we tracked the occurrence of thousands of loanwords in Spanish to describe their use, origin, and historical context. In doing so, we used computational methodologies to parse, lemmatize, group, count, and extract the information of all these tools. Results from parsing the etymologies of the DECH dictionary reveal a strong influence of Greek and French on the Spanish lexicon. The results from tracking the occurrence of loanwords in Spanish reveal a clear trend in the use of loanwords over time and support the hypothesis that the lexicon of a language reflects the sociopolitical and sociocultural change that their speakers undergo.

Keywords: Spanish; DECH; HathiTrust; Google Ngrams; loanwords; etymology

En employant l'outil lexicographique Diccionario Crítico Etimológico Castellano e Hispánico (DECH), ainsi que les deux corpus HathiTrust et Google Books Ngrams, nous avons localisé l'occurrence de milliers d'emprunts lexicaux utilisés en espagnol afin de décrire leur usage, leur origine et leur contexte historique. Nous nous sommes ainsi servis des méthodologies informatiques pour faire l'analyse, pour lemmatiser, pour classifier, pour compter et pour récupérer les données venant des outils mentionnés ci-dessus. Les résultats de l'analyse des étymologies du dictionnaire DECH révèlent que le grec et le français ont fortement influencé le lexique espagnol. L'analyse de l'occurrence des emprunts lexicaux en espagnol révèle une tendance claire concernant l'usage d'emprunts lexicaux au cours du temps et appuie l'hypothèse que le lexique de chaque langue reflète les changements sociopolitiques et socioculturels que ses locuteurs vivent.

Mots-clés: L'espagnol; DECH; HathiTrust; Google Ngrams;   emprunts lexicaux; étymologie

---

1. Introduction

Languages change and evolve in many respects but one of the aspects that is most obvious to casual observation is how languages change their word-stock (Trask 2009). We do not need to wait for several generations to go by to realize that new words are being incorporated into a language, whether they are from foreign languages (*loanwords*) or whether they are newly coined in the language. Loanwords have trespassed the boundaries of linguistics into the realm of the public opinion, politics and policy. In some cases, language academies, such as the Académie Française, have issued recommendations to avoid the use of "unnecessary" loanwords (Samuel 2011); in others, journalists have voiced complaints about the "usurpation" of Spanish by English (Villarreal 2014) making the incorporation and use of loanwords a topic of public interest. It is neither exaggerated nor gratuitous, then, to concur with Gómez Capuz's (2005, 7) metaphor of loanwords as *lexical immigrants*, because of their status as "international threats" who arrive to a language and need to live with the threat of expulsion, at least until the point they become conventionalized as part of the host language.

In this study, we aim to describe the use, origin, and historical context of the loanwords that have been incorporated into Spanish over the course of 400 years using two literary corpora made available by HathiTrust Digital Library and Google Books Ngrams, respectively, and a highly comprehensive lexicographical tool, the electronic edition of the Diccionario Crítico Etimológico Castellano e Hispánico dictionary (henceforth, DECH) (Coromines and Pascual 2012).[1] In doing so, we have

---

[1] An earlier version of this study used the Dirae dictionary (Rodríguez Alberich 2014) as a lexicographical tool. Following the suggestion of an anonymous reviewer, this revised version uses the etymological information of the DECH (Coromines and Pascual 2012), which has been reported, repeatedly, as the best, most reliable and most comprehensive source of etymological information for the Spanish language (Haensch and Omeñaca 2004, 143). The DECH, as opposed to the Dirae, is an etymological dictionary (Buchi 2016).

used computational methodologies to parse, lemmatize, group, count, and extract the information of HathiTrust, Google Books Ngram, and the DECH dictionary. This process resulted in a total of almost 160 billion tokens. (A token is an individual unit in the natural language analysis of texts. Words are tokens, as are acronyms, proper names, or temporal marks [*e.g.*, "sXVI" – 16th Century], among others.) This project intends to address four main questions about loanwords in Spanish, namely:

1. What language donated more loanwords to Spanish?
2. Are there any marked peaks of loanwords use?
3. If so, do they tell us anything about the language?
4. Can we establish a connection between the use of loanwords and historical circumstances?

The structure of this article is the following: In section 2, we briefly discuss the history of Spanish as it was formed and became in contact with other languages. In section 3, we provide a background of what loanwords are and of their importance in different fields of study. In section 4, we describe our study in detail. In section 5, we provide a discussion of our results.

2. A Brief Overview of the History of Spanish

Spanish is a member of the Italic branch of the Indo-European family of languages and is currently spoken by over 420 million speakers as a mother tongue and over 90 million as a second language (L2) in several countries in Europe, America and Africa (Lewis, Gary and Fennig 2016).

With the expansion of the Roman Republic and Empire, between 510 BC and 476 AC, Vulgar Latin evolved into a continuum of overlapping varieties, some of which were mutually unintelligible, thus giving rise to many European languages, one of which was Spanish (Penny 2002). Latinization of the Iberian Peninsula began in 218 BC and lasted roughly two centuries. The geographical distribution of the several languages that existed in the Iberian Peninsula in pre-Roman times was complex but many of the inhabitants of the Peninsula are believed to have

been competent bilinguals in Latin and in their own pre-existent languages (Adams 2003, 279–283; Penny 2002, 255). The pre-Roman languages are believed to have left a substrate influence, albeit minimal, on the subsequent Latin used in the Peninsula.

Whereas the later Visigoth conquest of Spain (585AD) had little effect on the Latin spoken in the Peninsula, the Islamic Conquest of the Peninsula in 711AD (whose outcome was a continuous presence of their cultural and linguistic influence that lasted for about 5 centuries) had an enormous impact: numerous words and phrases and even some morphology was taken from Arabic and included into Spanish, proving the significance of the prolonged contact between the two languages (Penny 2002, 16).

In this study, we count the pre-Roman, Latin, Arabic, and Visigoth influences as the ones that shaped the development of Spanish before the establishment of the Crown of Castile in 1230, a political process through the union of the previously independent kingdoms of Castile and Leon. We will count all lexical items that derived directly or indirectly from these languages as baseline Spanish (or core Spanish).[2] The reason why we take 1065 as a cutoff date is because of its importance for the linguistic identity of Castilian Spanish. The battles and unions that took place between 1065 and 1230 culminated (linguistically) with the creation of an early standard Spanish thanks to the work of Alfonso X the Learned, king of Castile and Leon (1252–84) (Fernández-Ordóñez 2004; Penny 2002).

2.1. Spanish in Contact

During its birth and growth, Spanish was not the only language spoken in the Iberian Peninsula. Other Latin-based dialects (Astur-Leonese, Catalan, Galician-Portuguese and Navarro-Aragonese) were also becoming independent, full-fledged languages during the 10th and 11th centuries (Lapesa 1990, 38). These languages together with

---

[2] While it is true that there had been Greek colonies in present-day Spain before the arrival of the Romans, they left very few words: toponyms (Empúries, Roses) and, probably, the word *seta* ('mushroom'). Because of this, Greek is not counted as substrate despite early contacts (Dworkin 2012; Penny 2002). For the same reason, we do not count Greek as baseline Spanish in this study.

Basque, a language isolate spoken in the northern part of the Peninsula, occupied and sometimes shared the geographical space of the Peninsula.

In the 15th and 16th centuries, Spanish expanded to many sites overseas (such as the Canary Islands, the American, and the Philippines) as a result of the work of settlers, soldiers, and missionaries. In linguistic terms, the colonial expeditions of the Spanish to the Americas in 1492 made Spanish more widely spoken and gave rise to more contact scenarios, in this case with various Amerindian languages.

Direct and prolonged contact with all this variety of languages produced several additions to the Spanish vocabulary (Dworkin 2012; Penny 2002). Although these contact situations had been ongoing by the time the first official dictionary of Spanish, the *Diccionario de Autoridades*, was published by RAE in 1726, we will count the lexical items that came from these languages as loanwords as these were not existent in the creation of the aforementioned "standard."

3. Loanwords

We have briefly mentioned that Spanish has been very prolific in terms of borrowing all sorts of linguistic components from the languages it has had contact with, and this language is by no means an exception. Most world languages have borrowed from the languages they have maintained contact with, even when bilingualism in the speakers of the donor and recipient languages has been infrequent (Durkin 2009, 132; Kaufman and Thomason 1988, 47; Sayahi 2011, 86).

This study focuses on the borrowing of lexicon and its associated meaning, which results in *loanwords* (Durkin 2009, 134). We take Tagliavini's (1949, 369) etymological based definition of loanwords: "a word in a language that comes from another language, different from the one that constitutes the basis of the borrowing language or that, if it *does* originate in that language, it is not obtained through regular, continuous, and popular transmission, but borrowed later on."

There is a plethora of possible reasons as to why languages may borrow lexicon from other languages: language economy, fashion, need for (more appropriate) terms (Durkin 2009), clarification, impact, precision, attenuation, snobbism,

prestige (Gerding et al. 2012), identity, adaptation, or dialect/language leveling (Young 1996).

Since borrowing is a highly common process, it has patterns: content words (such as nouns or verbs) are borrowed more frequently than function words (such as prepositions or conjunctions). In turn, nouns are borrowed more than any other part of speech (such as adjectives or adverbs) (Matras 2009). It is thought that both referential transparency and morphosyntactic freedom are factors that ease borrowing of nouns (Matras 2009).

3.1. How Do Languages Borrow?

As we have described before, languages borrow all sorts of linguistic components from other languages. Thomason and Kaufman (1988, 37) define borrowing as "the incorporation of foreign features into a group's native language by speakers of that language: the native language is maintained but is changed by the addition of the incorporated features". A possible way in which languages borrow is proposed by Backus (in Zenner and Christiansen 2014) and Croft (2000, 211): a speaker who speaks language A and, to some extent, language B, has two choices when faced with a communicative task, saying a concept in a familiar way (that is, using language A) or saying it in a creative way (using language B). Making the "creative decision" will imply an innovation, also known as "altered replication". Loanwords (and all types of innovations) very often die soon after they are born and never become part of language A. The more this innovation (or foreign incorporation) is repeated and encountered, the more *entrenched* it becomes. If an innovation becomes entrenched enough, it may end up being conventionalized as part of language A.

If the loanword from language B conveys a meaning that did not exist in language A, loanwords are typically accepted without much resistance. This could be the case of the word *kimono*, which entered English (among many other languages) from Japanese to name a concept that was previously nonexistent and/or unknown to English-speaking populations. It is possible, however, that

the loanword from language B overlaps in a position that was already occupied in language A. When this is the case, both words are in competition and more than just one outcome is possible. In the first place, one of the two may become obsolete, as it is the case of the Old English word *firen*, which was replaced by the French word *crime* (Ringe and Taylor 2014). It is also possible that both words stay on language A with highly similar meanings (such as English *kind* and French loanword *type*). Both words can stay with more obvious differences in meaning (for English, this would be the case for word pairs such as the Germanic word "cow" and the Old French loanword "beef"), or as words used by different groups of people or for different purposes (such as English "ant" and the scientific Latin loanword "formic").

3.2. The Importance of Loanwords

While loanwords played a somewhat peripheral role in the early study of Linguistics (Gómez Capuz 2005, 11), more recently, many studies, especially in the field of language contact and sociolinguistics (Adalar and Tagliamonte 1998; Arroyo and Tricker 2000; Poplack and Dion 2012; Stammers and Deuchar 2011) have investigated the incorporation, use, and popularization of loanwords over time. The study of loanwords provides valuable data for studying language change (Backus in Zenner and Christiansen 2014, 19). Layers of loanwords in a language (such as the bulk of loanwords from Old Norse and French in English) tell us about the past contacts between speakers of the donor and recipient languages, and the kind of words that were borrowed inform us about the nature of the contact (Bynon 1977, 231).

4. Our Study

Our present study focuses specifically on written text. The reasons why we are focusing on written language are several. First of all, written language has been shown to display a higher ratio of lexical items to total running words than spoken language, which is known as *lexical density* (Halliday 1989, 61). A high lexical density is desirable in that it gives us access to a greater number of lexical items than if

we used spoken language. Secondly, written language does not display as much inherent variability of forms as spoken language does (Poplack and Dion 2012) and it leaves out some aspects that are difficult to analyze (i.e. prosodic information or paralinguistic features). In addition, studies that analyze speech samples can only look back to some 80 years of history whereas the study of written samples allows for a much higher retrospective outlook. Equally importantly, writers tend to use the standard language (unless it is to create a particular effect). Therefore, it is reasonable to assume that when words appear in a book, they have been present in the speech for a relatively long amount of time; i.e. they have become entrenched. The last reason, but not less important, is practicality. Written texts are already transcribed (and in this case, digitized), which makes them easier to be processed computationally.

There are some corpora for spoken Spanish: Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC 1992), Proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA 2014), and Corpus de Referencia del Español Actual (CREA 2008). CREA, in fact, has both an oral and a written Spanish corpus, the latter being much larger. But these corpora have something in common, in addition to being very well made and well selected: they are small. (CORLEC has 1,100,000 transliterated words [CORLEC 1992], PRESEEA, in 2006, had 2,000,000 transcribed words [Moreno Fernández 2006], and CREA has almost 9,000,000 transcribed words [CREA 2008].) Because of their size, manual efforts of transcription and annotation have made sense. However, this study seeks to face a new challenge: using billions and billions of words to account for the introduction and diffusion of loanwords.

Differently from other sociolinguistic and language contact studies, our investigation does not focus necessarily on the language production of a bilingual territory. In most cases, authors were monolingual and lived in stable monolingual zones. Therefore, the fact the we see words borrowed from other languages is not necessarily the result of personal or societal bilingualism but of established borrowing processes.

4.1. Research Questions

To remind the reader, in this study we seek to answer four main questions:

1. What language donated more loanwords to Spanish?
2. Are there any marked peaks of loanwords use?
3. If so, do they tell us anything about the language?
4. Can we establish a connection between the use of loanwords and historical circumstances?

4.2. Methodology

In order to collect a dataset as exhaustive and extensive as possible that would allow us to address our aforementioned questions, we resorted to two corpora made publicly available, the HathiTrust Digital Library (henceforth: HT) with its Spanish corpus of 146,635 texts (Capitanu et al. 2015); and the Google Books Ngram dataset V2 (NGram) and its Spanish portion comprising 854,649 texts (Lin et al. 2012). Each text in HT is represented as a compressed JSON file that includes metadata about the volume in which it was originally published, such as the year of publication or the author if known, and a list of pages. Each of these pages contains a frequency map for each of the individual tokens found in an automated part-of-speech (POS) tagging task of the text in the page. Order and structure are lost, and the original text cannot be reconstructed. Although this might represent a problem to other studies, it is not problematic to ours because word counts were the only requirement to address our questions.

To analyze the 33GB of JSON data we used the Apache Spark toolkit (Meng et al. 2015; Zaharia et al. 2010), an upcoming de facto standard for big data analysis, and the Python bindings that it provides, together with some of the scientific stack tooling in the Python ecosystem (mainly Scipy, Numpy, Pandas, Cython, matplotlib, seaborn and Jupyter). Cleaning and aggregating word counts per year for the 22,236,839,455 syntactically annotated tokens (29,105 were unique lemmas) in HT took 16 hours in a 16GB RAM 8-core Intel® Core™ i7 machine. Despite containing several times more tokens (140,363,350,682 1-grams, and

46,067 unique lemmas), the NGram dataset was smaller: around 10GB of clean tabulated data with calculated word counts per year. Processing NGram was easier when compared to the HT format since the NGram dataset only contains isolated information about words (some POS annotated), counts, and years and volume of apparition. An additional problem that the HT database posed was that it only allowed us to use texts up to the copyright start date, that is, volumes that were published in the US before 1923 or outside the US before 1876. Hence, and as we will see in the results sections, our only source for words in the last century comes from the NGram dataset.

Once the first word counts were calculated, we lemmatized them and ignored proper nouns and other constructions (such as initials, numbers, or words with numbers). Lemmatization of Spanish words was carried out using the lexicon built for the VL3 project (Nuñez and Jiménez-Mavillard 2012), which accounted for 1.2 million words and included all inflections of verbs as collected by the 23rd edition of the dictionary of the Royal Academy of Spanish, known as DLE (RAE 2014). After filtering those lemmas out, the HT dataset accounted for 14,079,479,528 tokens (63% of the original number), of which 9,733,355,056 were loanwords, and NGram for 130,170,787,502 (92%) words, of which 89,087,235,626 were loanwords. (It is worth noting that the authors of the NGram dataset reported only 83,967,471,303 words in Spanish).

Identifying which lemmas were loanwords was possible thanks to the etymological information contained in the electronic edition of the DECH. These etymologies, only available for roughly a 21% of the lemmas (19,446) were expressed in natural language, making the extraction of the source language very difficult. We built a grammar to recognise the tree structure of the etymology statements. As an example, consider the etymology for the lemma "crónica":

Del latín chronĭcus, y este del griego χρονικός chronikós; la forma femenina, del latín chronĭca, y este del griego χρονικά [βιβλία] chroniká [biblía] '[libros] que siguen el orden del tiempo'.

"From Latin chronĭcus, and this one from Greek χρονικός chronikós; the feminine form, from Latin chronĭca, and this one from Greek χρονικά [βιβλία] chroniká [biblía] '[books] that follow a sequence of time'."

Our grammar was able to extract the next structure:

[{'from': {'lang': [['latín']], 'lemma': ['chronĭcus']}},
 {'from': {'lang': [['griego']], 'lemma': ['χρονικός', 'chronikós']}},
 {'from': {'lang': [['latín']], 'lemma': ['chronĭca']}},
 {'from': {'lang': [['griego']],
  'lemma': ['χρονικά', '[βιβλία]', 'chroniká', '[biblía]'],
  'meaning': ['[libros] que siguen el orden del tiempo']}}]

However, given the complexity of the language used to express these etymologies, in some cases we had to automatically traverse the resulting tree after parsing the grammar to fix some mistakes.

We classified the 437 language tags that the DECH uses to define etymological origin in 10 categories in order to capture, in a group, a set of languages that would share a geographic and temporal relation with regards to the Spanish language. The groups, with some of their more representative languages in terms of their contribution to the Spanish lexicon, are shown below. Note that we are not including all the source languages that the DECH specifies, but only those that served as donor languages directly to Spanish.[3]

Iberian Peninsula: Andalusian, Aragonese, Asturian, Basque, Caló, Catalan, Galician, Galician-Portuguese, Leonese, Portuguese, Salmantino, Valencian

---

[3] Each group is made up of the etymological sources according to the DECH. Some of these sources are languages (such as Basque or Catalan), some are groupings of languages (such as African languages), some are legally considered dialects (such as Genoese and Gascon), some are very general etymologies in lack of more specific information (as it is the case for Indigenous languages). It should also be taken into account that some etymologies are summarized for the sake of space (*Catalan* in our list represents the denomination *Old Catalan, Vulgar Catalan,* etc).

Amerindian & Philippines: Aimara, Arahuaco, Cumanagoto, Guaraní, K'iche', Mapuche, Mayan, Nahuatl, Quechua, Tagalog, Taíno, Tupi, various indigenous American languages

Western Europe: Dutch, Finnish, French, Gascon, Genoese, German, (Medieval and Modern) Greek, Italian, Napolitan, Norwegian, Occitan, Swedish

Ancient Europe: Frankish, Gaulish, Germanic, Gothic, Greek, (Pre-Roman) Indo European, Nordic, Pre-Celtic

Middle East & India: Aramaic, Bengali, Dravidian, Farsi, Hebrew, Hindi, Malabar, Pahlavi, Sanskrit, Sinhalese, Syriac, Tamil, Turkish, Urdu

Eastern Europe: Hungarian, Dalmatian, Polish, Russian, Armenian, Czech

Other languages: Chinese, Japanese, Malay

Africa: Berber, Quimbundu, African languages

Latin: loanwords: botanical Latin, modern Latin, scientific Latin, scholastic Latin, ecclesiastic Latin

English: English

While, as mentioned earlier, Latin and Arabic are part of what we have considered the baseline in this study, Spanish has borrowed words from both of these languages more recently. In the case of Latin, recent academic terms or learned words have been incorporated into Spanish, such as *referéndum* 'referendum', which was first attested in Spanish in 1829 according to Dirae (Rodríguez Alberich 2014). Some words from Arabic have also been borrowed recently, especially those related with Islam, such as *ayatolá* 'ayatollah', first attested in Spanish in 1977 (Rodríguez Alberich 2014). Examples such as these two, by all means, should be counted as loanwords, and not as core Spanish. The DECH has a way of marking some recent loanwords from Latin (under categories such as *latín botánico* 'botanical Latin' or *latín científico* 'scientific Latin') but the category *latín* 'Latin' and *arabic* includes both baseline items and loanwords. To be conservative, we counted these items as belonging to the baseline. Therefore, we acknowledge that the number of loanwords from Latin is higher in reality but, unfortunately, this is a limitation of this otherwise highly specific and thorough dictionary.

English was left as its own group because of its unique relationship to Spanish. This language has influenced Spanish at different stages from different parts of the

world. If English had been grouped together with other languages, the difference in loanword borrowing from British English and American English to Spanish might have been masked.

4.3. Results
4.3.1. Donor Languages

Out of the 65,357 lemmas (and definitions) included in the DECH dictionary, 33,529 counted with etymology information we were able to extract. And although 19,446 were found to be loanwords, only 6,827 appeared in our corpora and came from languages other than Latin and Arabic, which again, were counted as baseline Spanish. When looking at the donor languages for these lemmas, shown in **Figure 1**, we found that Greek and French were the most prolific languages in donating lexical items to Spanish: 2,191 words come from Greek and 1,447 from French.[4] The counts that appear in **Figure 1** disregard the notion of frequency in the language. That is,
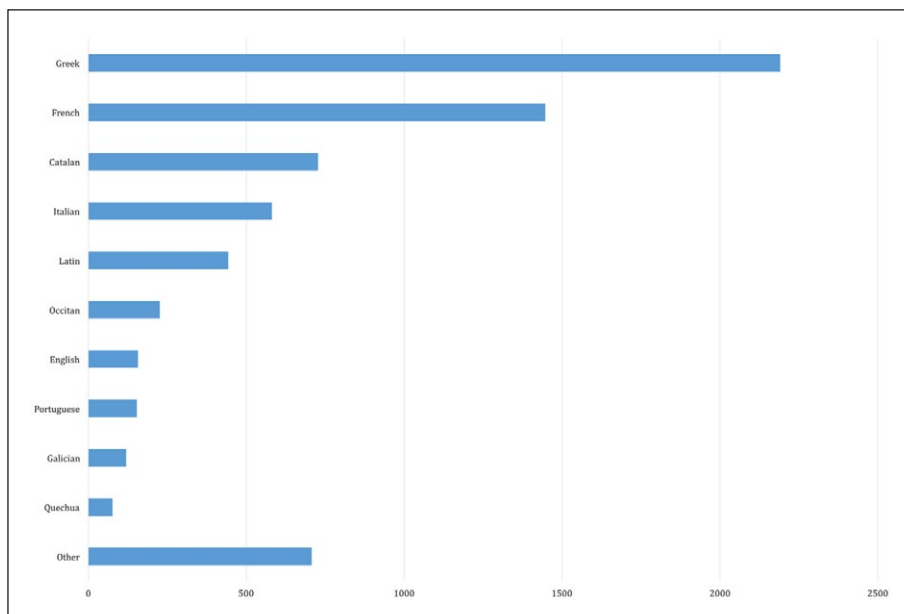

Figure 1: Number of loanwords per donor language.

---

[4] Note that for this specific subsection we are not using grouped languages but individual ones to gain a more specific understanding of which languages were most often resorted to by Spanish in need for new lexicon.

low frequency words such as *ñacundá* ('Nancunda nighthawk') are equivalent to one word, just as high-frequency lemmas such as *maíz* ('corn').

The majority of the most prolific donor languages in **Figure 1** are not surprising. After all, Spanish (in all Spanish-speaking nations) shares a geographic relation or a historical or political link with most of these languages. One of the notable exceptions is Greek, which appears as the top donor to Spanish. The great number of Greek loanwords may be unexpected at first. Spain and Greece have not established direct relations through colonial expansion, wars, or trading in the past centuries (other than the one described in Footnote 2). Upon careful observation of the Greek loanwords, we realize that the borrowing process for this language is unlike the one for the other languages. It could be deemed as being more "artificial". Ancient Greek has been resorted to in need for scientific and technical terms from ancient times until nowadays (Dworkin 2012; Gutiérrez Rodilla 1998). Besides the great amount of words that originated in Greek and made their way through Spanish via the Vulgar Latin and the Arabic spoken in the Iberian Peninsula before the 20th century (which, again, have not been counted as Greek loanwords in this study), we find a great amount of Ancient Greek loanwords in Spanish that were borrowed, unmediated, after the 18th century (Fernandez Galiano 1980). Examples of this type are *didáctico* ('didactic'), *asfixia* ('asphyxia'), or *fase* ('phase'). In the 19th and 20th century, Ancient Greek, as the first internationally prestigious language in history (Bergua Cavero 2002), was often resorted to due to the need to keep up with the rapidly evolving fields of science and technology. New inventions and discoveries, growing by leaps and bounds, required names. While the common vocabulary has its limits in terms of speed and growth, the scientific and specialized vocabulary required immediate additions, and Ancient Greek was often used as a tool (Fernández-Sevilla 1982; Gutiérrez Rodilla 1998). It is precisely because loanwords from Ancient Greek are specialized vocabulary that investigations that focus on casual speech or press corpora might overlook the fact that a great amount of Spanish words were loaned by this language. These type of words are simply not common in more casual types of texts but are nevertheless frequent in literature

and specialized treaties such as the ones contained in the HathiTrust and Google Books Ngrams databases.

As many philological studies point out (Curell Aguilà 2006; Dworkin 2012; Holmlid 2014; Varela Merino 2009), French has influenced Spanish from the Middle Ages to the present moment due to Spain and France's aristocratic and political relationships, population movements, and economic, cultural, and commercial exchanges. Having loaned high-frequency words such as *cine* ('cinema, movie theater'), *ciclista* ('cyclist'), or *caloría* ('calorie'), French is typically found as the most common donor language by those studies that look at established (i.e. included in the dictionary) loanwords in common speech or in the press (Holmlid 2014). Even before the use of technology, Spanish philologists were well aware of the influence of French on the Spanish lexicon. In the 18th Century there was a heated debate among writers regarding the widespread use of foreign words in the language, with some advocating in favor of purity of Spanish (and against loanwords), such as Gaspar Melchor de Jovellanos (1744–1811) or José Caldoso (1741–81) and some in favor of the expansion of vocabulary through loanwords, such as Benito Jerónimo Feijoo (1676–1764) (Dworkin 2012, 135–136).

English, occupying a seventh place in the donor list deserves a special mention. The "high" amount of English loanwords in Spanish has recently generated backlash from linguistic authorities, concerned journalists, and self-aware speakers who are worried about the purity (or lack thereof) of the language (see Pérez Tamayo [2002] for a discussion of English calques; Villareal 2014). While it is true that English has loaned high-frequency words such as *gol* ('goal') or *gay* ('gay') that are found in the literature, the reality is that our results need to be taken in the context of the corpora we have used. Two main reasons make the results for English only partially representative of the linguistic reality of Spanish nowadays. First of all, as we mentioned, literature tends to be more conservative in terms of lexicon used. Therefore, English loanwords that are being used in Spanish casual speech may have still not made it into the literature. Loanwords such as *selfie* or *online* are increasingly common, but are not equally present in Spanish literature. Secondly, and parallel

to that, linguistic authorities such as the Real Academia Española (RAE) are more cautioned in terms of which words are included in the dictionary. Going back to *selfie* and *online*, neither of them is included in a dictionary (probably because they did not make the cut in terms of use and validity in the Real Academia Española's last edition of the dictionary) while a quick Google Search for Spanish websites including these words yields thousands of results. For these reasons, studies that investigate the use of loanwords which have not been incorporated to dictionaries in casual speech or in texts that describe current affairs find a greater presence of English loanwords (Esteban Asencio 2008; Gerding et al. 2012). In other words, our methods have analyzed words that have been present in the language for a long time and have been conventionalized into Spanish. While it is true that nowadays there are many loanwords that are fashionable and highly common, we cannot foresee for how long they might last.

4.3.2. Frequencies

As mentioned in Section 3, the 437 reported donor languages were divided into 10 groups with only English being left by itself due to the difficulty of associating it with other languages. Our two databases, HathiTrust and Google Books NGrams, have different data and, because of this, their results will be presented separately.

4.3.2.1. HathiTrust

In the stacked percentage area chart shown in **Figure 2** (and **3** below), the 100% on the Y axis refers to the total of loanwords used in the corpus for that year (X axis) in the Spanish texts annotated in HathiTrust. That is, this figure does not inform the reader of how many loanwords were used (which is discussed below in Section 4.3.3), only of how loanwords were divided according to language origin.

In the 146,635 texts in Spanish available in HathiTrust between years 1700–1923, there were 22,236,839,455 syntactically annotated tokens, out of which we lost the information of 65%. This percentage contained a variety of cases: words that were not included in the DECH dictionary because they were loanwords (or other neologisms) that had not been normalized in the language yet, because they were initials, acronyms or proper nouns, or because the OCR failed to transcribe
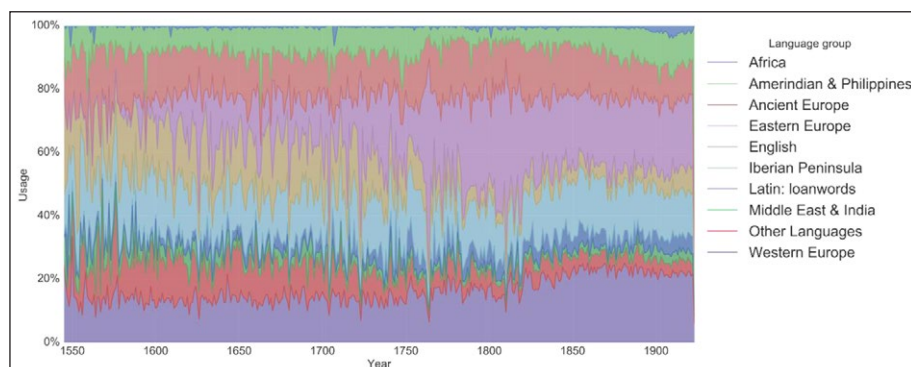
Figure 2: Frequencies of loanwords per language group in HathiTrust corpus.

them correctly. Besides that, a certain degree of noise is observed before 1840, showing that the amount of data is insufficient to draw firm/definite conclusions.

What can be observed in **Figure 2** is, first of all, the preponderance of certain language groups, namely of Eastern and Western Europe and, to a lesser extent, of Ancient Europe and Iberian Peninsula. Leaving aside the question of Eastern European languages momentarily, the predominance of the other three groups reiterates the importance of geographical proximity and the fact that Greek and French were not only prolific in donating lexicon to Spanish, but that their loanwords were frequently used in the literature and in the scientific texts collected by HathiTrust.

Now, the frequency of loanwords donated by Eastern European languages might be slightly misleading at first. There has not been any direct contact, historically, between Spanish and these languages, and Eastern European languages were not resorted to in order to coin new terms for the sciences. Upon closer inspection, the two most common terms donated by this group of languages are *coche* 'car', from Hungarian, and *soviet* 'soviet', from Russian. However, the first word appears 476,206 times in total, while the latter, the closest in frequency, only appears 15,673 times. Therefore, it is easy to see that the graph reflects the high frequency of *coche.*

One anonymous reviewer posed the question of what is more informative, the number of loanwords that a language borrows from a donor language or the

frequency of such loanwords. The case of Eastern European languages is perfect to speak to this question. According to the DECH, Hungarian has only donated two terms to Spanish, one of them being *coche*. French, on the other hand, has donated more than 1,400 lexical items, many of which are not as frequent in Spanish as *coche*. Therefore, we can claim that both the number of borrowed words and the frequency of such borrowed words inform us about different aspects of the contact and relationship between languages.

4.3.2.2. Google NGrams

The Google Books Ngram dataset was comparatively larger at 854,649 texts in Spanish with 140,363,350,682 1-grams between years 1500–2000. As apparent in **Figure 3**, we also find a significant degree of noise in the data, specifically before 1780.

The results of Google Ngrams largely correlate with those of HathiTrust, especially in terms of the importance of Ancient and Western European languages. What this database allows us to see, due to the increased years of its span, is a steady increase of English loanwords during the second half of the 20th century. Therefore, these results correlate with those that claim that, while English has influenced Spanish vocabulary for centuries, it is only over the past 50 years that there has been a massive influx and use of these loanwords (de la Cruz Cabanillas et al. 2007).
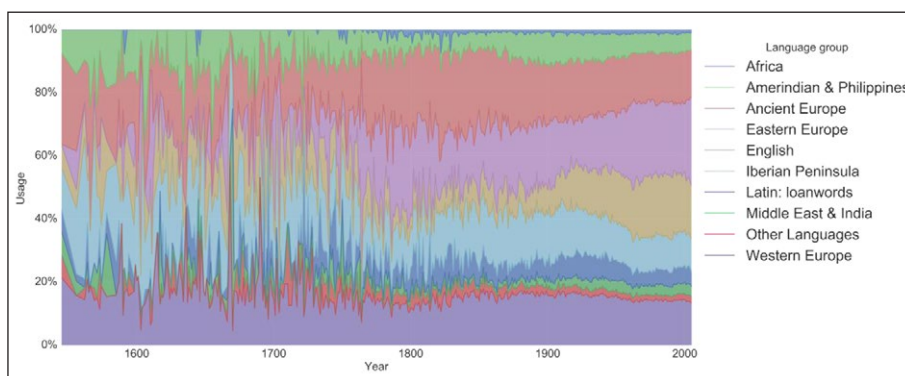

Figure 3: Percentage of loanwords per language group in Google Books Ngrams corpus.

Table 1: Average number of loanwords and core language words per century as well as the percentage of loanwords over the total of that century.

	Loanwords	Core Language	Proportion
16th Century	4.581638	16.960933	27.01288%
17th Century	4.078132	15.236676	26.7652%
18th Century	25.019335	148.065570	16.8974%
19th Century	71.059728	401.765578	17.68686%
20th Century	397.999064	1943.034384	20.48337%
21th Century	931.629326	4483.081108	20.781005%

In addition, it should be noted that this study [or rather the etymological sources it makes use of] does not make a distinction between British and American English and although we can assume that they have influenced Spanish at different times, we cannot prove it with data 3.3 Loanwords vs. "Core" Spanish.

In Section 2, we discussed how the words entering Spanish through certain languages (namely, Latin, Arabic, and Iberian Peninsula substrate) would be considered "core" Spanish for this study. **Table 1** shows the number of loanwords and core language words per century as well as the percentage of loanwords over the total of that century. Apart from the 16th-century data for which the corpora only contain 25 years, the average peak of loanwords happens during the 17th Century, with a 26.76% of words originating in donor languages. This means that more than 1 out of 4 words contained in Spanish cultured language, that is, the type of language used in literature, law, academic and scientific books, was taken from other languages. In order to draw final conclusions about the significance of the data, a comparison with the level of borrowing in other similar languages would be required. However, it can be claimed with confidence that an average of 26.76% is a very high degree of influence of foreign languages and cultures on a given language, and that a sustained relation of this kind over a long period of time would be indicative of either a great permeability of the recipient language or of a pronounced anxiety in the local elite classes (to which the authors writing these books belong), to conform their knowledge and expression to that of the most prestigious (international) producers of knowledge of the time.

After the 17th Century, the data shows a steep decline that takes the percentage of loanwords to only 16.9% in the 18**th** Century. We believe that this decline is the direct result of the control over the use of prescriptive Spanish exercised by the Real Academia Española starting this century, a control that has permeated society through the educational system, television and newspapers and is still very much present in public debates nowadays. In general terms, this control means that *good* written Spanish is to use as much vocabulary from Spanish as possible and, in consequence, the fewest amount of loanwords. Theoretically, the best Spanish would be that which draws the most lexical units from classical Spanish literary authors, specially from the Golden Age (16 and 17**th** centuries), the Silver Age (from 1914 up to the so called literary generation of 1927) or the best authors of the last part of the 20**th** Century.

After the 18**th** Century, and as we will expand in the following section, we see a small yet continuous upwards trend in the use of loanwords. This progressive use of more loanwords after the 18**th** Century shows the pressures of the current times.

4.3.3. Highest Use of Loanwords: The Peninsular War

Regardless of the position of the Real Academia Española about what it means to write *good* Spanish, it is clear that there is a relationship between international affairs and the level of loanwords present in Spanish at certain historical moments. This is more evident when we descend to the scale of years in our data. In **Figure 4**, we can see the percentage of loanword usage per year as compared to that "core" Spanish on a year by year basis. The peak of loanword use, at 20.53%, occurred in the year 1809, although the following years maintained an overall high rate in the use of loanwords. When we look closely at what the most frequently used loanwords where that year, we are faced with a plethora of war-related words, such as *guerra* ('war', from Germanic), *buque* ('ship', from French), *artillería* ('artillery', from French), *tropa* ('troop', from French), or *batalla* ('battle', from French).

Upon closer inspection of the history of the time, 1809 was one of the early years of *la Guerra de Independencia* or "the Peninsular War" (1808–1814). In this military conflict, Napoleon's Empire and Spain (together with Britain and Portugal) fought for the control of the Iberian Peninsula. It is not accidental, then, that in this
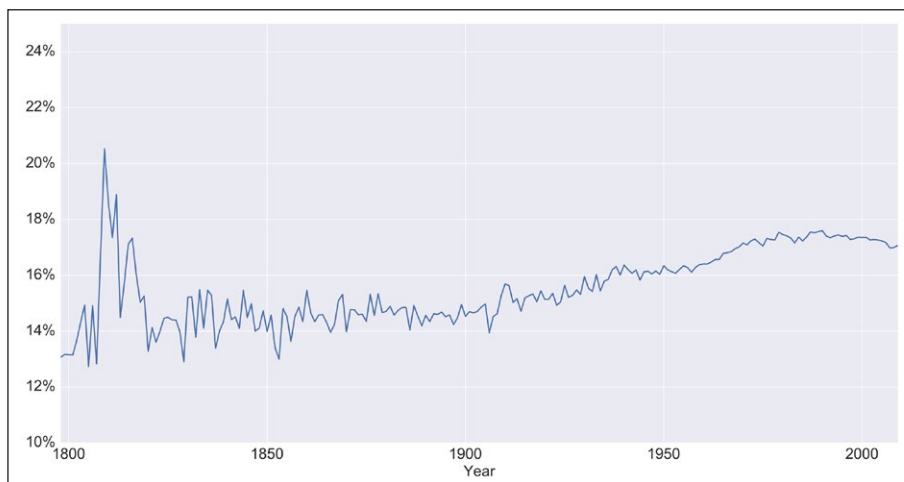
Figure 4: Frequencies of total loanwords in Spanish per year.

time of international accords and interactions, Spanish used an equally international vocabulary. The fact that, during this time of conflict against France, French was the most popular donor language (supplying Spanish with a 39% out of this 20.53% percent of loanwords) can also be considered evidence of how related languages and historical and political interactions are.

5. Discussion and Conclusions

As Fernández-Sevilla (1982, 21) argued, some words are testimonies of social, sociopolitical and sociocultural change. Loanwords in Spanish have proven this hypothesis right: they told us a story, not only the story of when they were borrowed or of how they were used, also of what the parties involved in these sociocultural changes were and what the nature of the sociocultural exchange was. This study supports the claim that language does not stand on its own, but that it rather is the product of a combination of political, social, and economic factors (Lapesa 1990, 35). None of the loanword donor languages and none of the peaks, as we have seen, were fortuitous.

Some of our results might seem counterintuitive at first, the first one being that Greek is the most common source language for Spanish loanwords. However, as we have showed, there is a reason for that: the native vocabulary has its limits in

terms of growth, whereas borrowing does not (Fernández-Sevilla 1982; Gutiérrez Rodilla 1998. See comment above about the proportion of Latin loanwords.) The other counterintuitive result is, perhaps, the importance of Eastern European languages, which through the case of *coche* show the weight that technology might have in penetrating the lexical fabric of a borrowing language.

The dominance (in terms of their frequency of use) of words originated in Western and Ancient European languages (French and Greek, respectively) in HathiTrust and Google NGrams were easily justifiable — both languages gave Spanish a plethora of words that abound in these types of texts.

The progressive upward trend in the use of loanwords observed in **Figure 4**, despite the work of the Real Academia Española to "control" the amount of loanword use, reflects the permeability of the language and its need to resort to the use of this new lexical items. This need, especially apparent in cultured texts, such as the ones contained in HathiTrust and Google Ngrams, far from stopping, keeps increasing as political and cultural exchanges, trade and globalization affect the shape of human communication and make languages more porous to external influences.

As we saw, the increase of English loanwords towards the end of the 20th century correlates with the public opinion: Spanish is using words borrowed from English more frequently (which probably correlates with the fact that Spanish is borrowing more and more words from English). However, as we showed in section 4.3.2.2, those numbers are still very far from reaching the widespread use of established French and Greek loanwords.

For the emerging area of cultural analytics, our study shows a promising path of enquiry into inter-cultural relations. The data of the year with highest presence of loanwords in Spanish shows a strong relation between historical periods in which Spain was very open to foreign. In order to draw stronger conclusions, two elements are needed. First, we would need a point of comparison with which we would be able to establish levels of loanwords presence in target languages when the subject nations are entangled in military and/or diplomatic conflicts of large proportions. For instance, is there a quantifiable increase in the use of Japanese words in Korean

between 1910–1945? Second, the data needs to be of enough quality at the year level so that we can pinpoint specific years in which the influence is felt and recorded, which due to rhythms of cultural integration and production, it rarely coincides with years of publication of books.

In the case of the highest use of loanwords in Spanish, it has to be noted that the years surrounding 1809 are especially significative in the case of Spain. In addition to the already mentioned War of Independence between Spain and the Napoleonic armies, two other historical facts must be taken into account. First, we have the strong and continued cultural influence of France through the Enlightenment. Second, the years following the re-installation of King Ferdinand VII in the Spanish throne are immediately followed by the independence of the Spanish territories in the Americas. The intellectual revolutions that inspired those wars of independence and led to the birth of Latin American countries have a great debt with French and North American ideas and authors.

6. Limitations of this Study

The main limitation of this study is the same one as our main point of support: our total reliance on previous etymological information. Etymology is not an exact science and it needs to make use of whatever resources loanwords have to offer. Deciding whether a word entered Spanish via Latin, or French, or Italian is extremely difficult if the word does not come from pre-Roman languages (Varela Merino 2009). In addition, while the DECH is continuously credited as an extremely reliable source of etymological information, its specificity has its limits. Words that were loaned by Latin in modern times (that is, well after the creation of "standard" Spanish) were overlooked (and thus not counted as loanwords) due to the fact they were not always marked as modern loanwords in our lexicographical tool.

The second limitation is our data. Whatever was in HathiTrust and Google Books NGrams, that is what we worked with. Their data made up an unprecedented amount of tokens that we could analyze, but it also came with its limitations: some written works were included, and some were not. Had we included everything, our results could have changed. However, we do believe these results to be representative.

The last limitation is, perhaps, the decisions we had to make, specifically, what counts as "core" Spanish and what does not. We decided on Latin, Iberian substrate and Arabic as forming it, although some linguists and philologists might disagree on the latter.

Whereas the amount of data to be processed did not impose a limitation by itself, although it was challenging at times, decisions were made for practical reasons that might have impacted the results. We tried to minimize the extent of code and algorithm effects while preserving the accuracy and fidelity of the data. Unfortunately, some of these decisions were arguably not ideal: when extracting the etymology of a word, after converting to lowercase, we first checked the word, and then the lemma, giving more importance to nouns than conjugated verbs. Moreover, if a number was found in between the word, such as "se3n" in HT, we removed the word. Lemmatization was also imperfect, and disambiguation was not taken into account as part of the process, therefore proper nouns were considered regular nouns and language of the first definition in the dictionary was used.

Competing Interests

The authors have no competing interests to declare.

References

Adalar, Nevin, and **Sali Tagliamonte.** 1998. "Borrowed Nouns; Bilingual People: The Case of the "Londrali" in Northern Cyprus." *International Journal of Bilingualism* 2(2): 139–159. DOI: https://doi.org/10.1177/136700699800200202

Adams, James Noel. 2003. *Bilingualism and the Latin language.* Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511482960

Arroyo, José Luis Blas, and **Deborah Tricker.** 2000. "Principles of Variationism for Disambiguating Language Contact phenomena: The Case of Lone Spanish Nouns in Catalan Discourse." *Language Variation and Change* 12(2): 103–140. DOI: https://doi.org/10.1017/S095439450012201X

Asencio, Laura Esteban. 2008. "Uso, origen y procesos de creación de neologismos en prensa española." *Círculo de lingüística aplicada a la comunicación* 33(1).

Bergua Cavero, Jorge. 2002. "Introducción al estudio de los helenismos del español." *Zaragoza: Ediciones del Departamento de Ciencias de la Antigüedad. Área de Filología Griega.* Universidad de Zaragoza.

Buchi, Éva. 2016. "Etymological Dictionaries." In *The Oxford Handbook of Lexicography*, edited by Philip Durkin, 338–349. Oxford: Oxford University Press.

Bynon, Theodora. 1977. *Historical linguistics.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139165709

Capitanu, Boris Ted Underwood, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, and **J. Stephen Downie.** 2015. "Extracted Feature Dataset from 4.8 Million HathiTrust Digital Library Public Domain Volumes (0.2) [Dataset]." HathiTrust Research Center.

Capuz, Juan Gómez. 2005. *La inmigración léxica* 84. Madrid: Arco libros.

CORLEC. 1992. "Corpus Oral de Referencia de la Lengua Española Contemporánea". http://www.lllf.uam.es/ESP/Info%20Corlec.html.

Coromines, Joan, and **José Antonio Pascual.** 2012. *Diccionario crítico etimológico castellano e hispánico.* Madrid: Editorial Gredos.

Croft, William. 2000. *Explaining language change: An evolutionary approach.* New York: Longman.

Curell Aguilà, Clara. 2006. "La influencia del francés en el español contemporáneo." In *La cultura del otro: español en Francia, francés en España*, edited by Manuel Bruña Cuevas, et al., 785–792. Sevilla: Universidad de Sevilla.

de la Cruz Cabanillas, Isabel, Cristina Tejedor Martínez, Mercedes Díez Prados, and **Esperanza Cerdá Redondo.** 2007. "English Loanwords in Spanish Computer Language." *English for Specific Purposes* 26(1): 52–78. DOI: https://doi.org/10.1016/j.esp.2005.06.002

Dworkin, Steven N. 2012. *A History of the Spanish Lexicon: a Linguistic Perspective.* Oxford: Oxford University Press on Demand. DOI: https://doi.org/10.1093/acprof:oso/9780199541140.001.0001

Fernández-Ordóñez, Inés. 2004. "Alfonso X el Sabio en la historia del español." In *Historia de la lengua española,* edited by Rafel Cano Aguilar, 381–422. Barcelona: Ariel.

Fernández-Sevilla, Julio. 1982. *Neología y neologismo en español contemporáneo.* Universidad de Granada: Editorial Don Quijote.

Gerding, Constanza, Mary Fuentes, Lilian Gómez, and **Kotz Gabriela.** 2012. "El préstamo en seis variedades geolectales del español: Un estudio en prensa escrita." *Revista signos* 45(80): 280–299. DOI: https://doi.org/10.4067/S0718-09342012000300003

Gutiérrez Rodilla, Bertha María. 1998. *La ciencia empieza en la palabra: análisis e historia del lenguaje científico.* Barcelona: Ediciones Península.

Haensch, Günther, and **Carlos Omeñaca.** 2004. Los diccionarios del español en el siglo XXI. Salamanca: Universidad de Salamanca.

Halliday, Michael A. 1989. *"Spoken and Written Language."* Oxford: Oxford University Press.

Holmlid, Julia. 2014. *"Los préstamos léxicos en el español peninsular. Un estudio de los préstamos léxicos en la prensa periódica en el año 2012".* Thesis, University of Gothenburg.

Kaufman, Terrence, and **Sarah Grey Thomason.** 1988. *"Language Contact, Creolization, and Genetic linguistics."* Berkeley CA: University of California.

Lapesa, Rafael. 1990. "Historia de la lengua e historia de la literatura". In *Historia de la literatura española,* 35–76. Madrid: Cátedra.

Lewis, M. Paul, Gary F. Simons, and **Charles D. Fennig,** (eds.). 2016. *Ethnologue: Languages of the World, Nineteenth edition.* Dallas, Texas: SIL International. http://www.ethnologue.com.

Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and **Slav Petrov.** 2012. "Syntactic Annotations for the Google Books Ngram Corpus." In *Proceedings of the ACL 2012 System Demonstrations,* 169–174. Association for Computational Linguistics.

Matras, Yaron. 2009. *Language Contact.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511809873

Meng, Xiangrui, Joseph Bradley, B. Yuvaz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, et al. 2015. "Mllib: Machine Learning in Apache Spark." *JMLR* 17(34): 1–7.

Moreno Fernández, Francisco. 2006. "Información básica sobre el "Proyecto para el Estudio Sociolingüístico del Español de España y de América"-PRESEEA (1996–2010)." *Revista española de lingüística:* órgano *de la Sociedad Española de Lingüística* 36: 385.

Nuñez, Camelia G, and **Antonio Jiménez Mavillard.** 2012. "The VL3: A Project at the Crossroads between Linguistics and Computer Science." *Paper presented at Digital Humanities 2012.* Hamburg, Germany.

Penny, Ralph John. 2002. *A History of the Spanish Language.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511992827

Pérez Tamayo, R. 2002. "Neologismos: ¿contaminación o enriquecimiento de la lengua española?" *Panace* 3(9–10): 3–4.

Poplack, Shana, and **Nathalie Dion.** 2012. "Myths and Facts about Loanword Development." *Language Variation and Change* 24(3): 279–315. DOI: https://doi.org/10.1017/S095439451200018X

PRESEEA. 2014. "Corpus del Proyecto para el estudio sociolingüístico del español de España y de América". Alcalá de Henares: Universidad de Alcalá. http://preseea.linguas.net.

RAE. 2008. "CREA. Corpus de referencia del español actual." http://www.rae.es/recursos/banco-de-datos/crea.

RAE. 2014. *Diccionario de la lengua española. Vigésimotercera edición.* Planeta Publishing.

Ringe, Don, and **Ann Taylor.** 2014. *The Development of Old English* 2. Oxford: OUP.

Rodríguez Alberich, Gabriel. 2014. DIRAE. http://dirae.es/.

Samuel, Henry. 2011. "France's Académie Française Battles to Protect Language from English." *The Telegraph,* October 11, 2011. Accessed December 20, 2016. http://www.telegraph.co.uk/news/worldnews/europe/france/8820304/Frances-Academie-francaise-battles-to-protect-language-from-English.html.

Sayahi, Lotfi. 2011. Contacto y préstamo léxico: El elemento español en el árabe actual. *Revista Internacional De Lingüística Iberoamericana* 9 2(18): 85–99.

Stammers, Jonathan R, and **Margaret Deuchar.** 2011. "Testing the Nonce Borrowing Hypothesis: Counter-Evidence from English-Origin Verbs in Welsh." *Bilingualism: Language and Cognition* 15(33): 630–643.

Tagliavini, Carlo. 1949. *Introduzione alla glottologia.* Prof. Riccardo Pàtron.

Trask, Larry. 2009. *Why Do Languages Change?* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511841194

Varela Merino, Elena. 2009. *Los galicismos en el español de los siglos XVI y XVII* 1. Editorial CSIC-CSIC Press.

Villarreal, Antonio. 2014. "Cuando el inglés usurpa la riqueza léxica del español". *ABC*, April 28. Accessed December 11, 2016. http://www.abc.es/cultura/20140427/abci-anglicismos-201404261644.html.

Young, Robert E. 1996. *Intercultural communication: Pragmatics, genealogy, deconstruction.* Philadelphia: Multilingual Matters.

Zaharia, Matei, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and **Ion Stoica.** 2010. "Spark: Cluster Computing with Working Sets." *HotCloud* 10: 10–10.

Zenner, Eline, and **Gitte Kristiansen,** (eds.). 2014. *New Perspectives on Lexical Borrowing: Onomasiological, Methodological and Phraseological Innovations* 7. Walter de Gruyter.