An Investigation of the Impact of Frequency on the Development of Latin to Spanish

by

Fiona M. Wilson

B.A., Simon Fraser University, 2012

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Linguistics
Faculty of Arts and Social Sciences

© Fiona M. Wilson 2015

SIMON FRASER UNIVERSITY

Spring 2015

All rights reserved.
However, in accordance with the *Copyright Act of Canada*, this work may be reproduced, without authorization, under the conditions for "Fair Dealing." Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review and news reporting is likely to be in accordance with the law, particularly if cited appropriately.

Approval

Name: **Fiona Wilson**

Degree: **Master of Arts**

Title: ***An Investigation of the Impact of Frequency on the Development of Latin to Spanish***

Examining Committee: **Chair:** Chung-Hye Han
Professor

Panayiotis Pappas
Senior Supervisor
Associate Professor

Arne Mooers
Supervisor
Professor

Maite Taboada
Supervisor
Professor

Alexandra D'Arcy
External Examiner
Associate Professor
Department of Linguistics
University of Victoria

Date Defended/Approved: February 27, 2015

Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the non-exclusive, royalty-free right to include a digital copy of this thesis, project or extended essay[s] and associated supplemental files ("Work") (title[s] below) in Summit, the Institutional Research Repository at SFU. SFU may also make copies of the Work for purposes of a scholarly or research nature; for users of the SFU Library; or in response to a request from another library, or educational institution, on SFU's own behalf or for one of its users. Distribution may be in any form.

The author has further agreed that SFU may keep more than one copy of the Work for purposes of back-up and security; and that SFU may, without changing the content, translate, if technically possible, the Work to any medium or format for the purpose of preserving the Work and facilitating the exercise of SFU's rights under this licence.

It is understood that copying, publication, or public performance of the Work for commercial purposes shall not be allowed without the author's written permission.

While granting the above uses to SFU, the author retains copyright ownership and moral rights in the Work, and may deal with the copyright in the Work in any way consistent with the terms of this licence, including the right to change the Work for subsequent purposes, including editing and publishing the Work in whole or in part, and licensing the content to other parties as the author may desire.

The author represents and warrants that he/she has the right to grant the rights contained in this licence and that the Work does not, to the best of the author's knowledge, infringe upon anyone's copyright. The author has obtained written copyright permission, where required, for the use of any third-party copyrighted material contained in the Work. The author represents and warrants that the Work is his/her own original work and that he/she has not previously assigned or relinquished the rights conferred in this licence.

Simon Fraser University Library
Burnaby, British Columbia, Canada

revised Fall 2013

Abstract

Previous research has suggested a relationship between frequency of use (FoU) and language change (Pagel, Atkinson, & Meade, 2007), but its nature remains unclear. Two research questions were raised in this thesis: 1) whether FoU remains stable over time, 2) whether amount of language change over time can be predicted using FoU. A 1147-word subset of the IDS wordlist (Key & Comrie, 2007) was used to test these questions. The FoU of both Latin and Spanish, and amount of change for each word was measured. There was a lower correlation across time than cross-linguistically, but the effect of genre could not be removed. A weak, highly significant negative relationship between FoU and amount of change was identified, supporting the claim that high frequency words change less than low frequency words. There is an intriguing correlation between FoU and lexical change, but the causal mechanism is not yet understood.

Acknowledgements

This research was made possible due to support from the Joseph-Armand Bombardier Canada Graduate Scholarship provided by the Social Sciences and Humanities Research Council (Award Ref. No.: 766-2012-0950). As well as support from Dr. Mark Collard and the SFU Human Evolutionary Studies Program.

During the course of this project, a number of people have allowed me to make use of their considerable expertise; without their guidance, this thesis would never have been completed. I would therefore like to thank my senior supervisor, Dr. Panayiotis Pappas, as well Dr. Arne Mooers and Dr. Maite Taboada for their assistance. I would also like to express my appreciation to the members of Dr. Mooers' lab for their feedback and suggestions, in particular Gordon Smith and Karen Gordon for their assistance with the statistical analysis and data extraction. I am also grateful for the suggestions of Brian Corrie and Chris Carleton with regard to database functionality, and to Dr. Nancy Hedberg for input on semantic categorization.

Finally, thank you to my friends and colleagues in the Linguistics department, as well as my family and partner, whose unwavering support is incalculably valuable to me.

Table of Contents

List of Tables

List of Figures

List of Acronyms

CdE	Corpus del Español
FoU	Frequency of Use
IDS	Intercontintental Dictionary Series
IE	Indo-European
PDLP	Perseus Digital Library Project

Chapter 1.

Introduction

1.1. Overview

Previous research (Pagel, Atkinson, & Meade, 2007) has provided evidence for a negative relationship between frequency of use and lexical change over time. However, the nature and directionality of that relationship has yet to be definitely established, as argued in Pappas and Mooers (2011). The present thesis further investigates this issue by raising two primary research questions: firstly, whether frequency of use remains stable over time, and secondly, whether amount of lexical change over time can be predicted using frequency of use.

The current research attempted to test both of these questions using a 1147-word subset of the Intercontinental Dictionary Series wordlist (Key & Comrie, 2007), for Latin and its daughter language Spanish. The frequency of use (hereafter FoU) for words for both languages was derived from two corpora: the Perseus Digital Library Project corpus of Latin (Crane, 2010), and the Corpus del Español for Spanish (Davies, 2002). The amount of change that each lexical item has undergone between the two stages was coded according to a scheme adapted from Pappas and Mooers (2011). This scheme comprised six levels, including sound change, paradigm leveling, general analogy, syntactic reanalysis, semantic change, and lexical borrowing. The methods used to derive the data are further described in Chapter 2.

The first research question will be approached in Chapter 3 of this thesis. A linear model predicting Latin frequency from Spanish frequency was run, and these results were compared to prior research on the correlation between modern Indo-European languages in frequency of use (Pagel, Atkinson, & Meade, 2007). The results on the

stability of frequency of use over time were inconclusive. While the correlation between Spanish and Latin was considerably lower than previous correlations performed between between modern languages (suggesting that frequency of use is not stable over time), any effect of change to frequency over time could not be separated from the effect of genre differences among the corpora. So, direct comparisons between different genres of the same language (Spanish) found that some genre comparisons had similar strengths of correlations as did Latin and Spanish.

Chapter 4 describes the analysis of the second research question. The analysis did identify a weak but highly significant negative relationship between frequency of use and amount of change over time, this relationship is weaker than the one previously identified by other researchers (Pagel, Atkinson, & Meade, 2007). This result offers tentative support to the claim that high frequency words change less over time than their low frequency counterparts. Even when the covariates of part of speech and semantic category were controlled for, an independent relationship between frequency of use and change remained. However, post-hoc pairwise comparisons determined that the relationship appeared to be primarily driven by the relationship between the most extreme change categories (namely, the difference between lexical borrowing, no change, and all other categories). Given that previous research attempted to exclude lexical borrowing (one of the primary drivers) from the investigation, the dependence is unexpected. Chapter 5 will further discuss these results, and attempt to place them in the general context of conducting research in historical linguistics using methods from evolutionary biology.

1.2. Parallels between Evolutionary Biology and Historical linguistics

Researchers have observed parallels between the fields of biology and historical linguistics for over a hundred years, as early as Darwin's *Origin of Species by Means of Natural Selection* in 1859 (Atkinson & Gray, 2005). The general observation has been an apparent similarity between the phylogenetic trees representing speciation in biology, and the linguistic trees representing historical divergence between languages. More recently, specific parallels have been drawn between the processes of biological

evolution and linguistic change over time (Pappas & Mooers, 2011). The parallels include genetic drift and linguistic drift, biological extinction and language death, as well as horizontal gene transfer and borrowing. Some of the proposed parallels describe highly similar processes, while others are more tenuous comparisons (e.g. while the word used is the same, *drift* in evolution and in linguistics refers to very different concepts) (Pappas & Mooers, 2011).

Comparisons between the fields of evolutionary biology and historical linguistics have several potential ramifications. First, it may be that the parallels reflect a more general process of evolution that is not constrained to the biological sphere, but instead applies to a wider range of fields, including, e.g., anthropology and linguistics. In this view, historical linguistic change would be an expression of such a general evolutionary process. Alternatively, even superficial similarities between historical linguistics and biology (i.e. similarities which do not reflect underlying shared processes or causes) may give insight into what methods may be useful in approaching historical linguistic systems, based on insights gleaned from biology on how best to analyze or approach structurally similar systems. Whether the observed parallels between biology and historical linguistics are representative of a general process of evolution or merely analogical in nature, they still represent a potentially rich avenue of research in historical linguistics, by suggesting the transfer of methods already applied in the field of phylogenetics. For example, phylogenetic methods have been applied in linguistics to propose internal subgroupings for language families (Bowern & Atkinson, 2012), as well as systematically assessing prior untested hypotheses on the differences between hunter-gatherer and cultivator languages around the world (Bowern, Epps, Gray, Hill, Hunley, McConvell, & Zentz, 2011).

Such proposals have piqued the interest of multiple researchers, including those whose primary area of study is biology and the application of phylogenetic methods. This is the case for the Reading Evolutionary Biology Group, a research group headed by Mark Pagel. The group's research goals include phylogenetic trees of language, rates of word evolution, linguistic half-lives, and deep reconstruction of language (<http://www.evolution.reading.ac.uk/LingCultEvo.html>). Members of this group have done research on the potential relationship between the frequency of use of lexical items

and linguistic change over time, applying methods culled from evolutionary biology, as well as molecular and cultural evolution (Pagel, Atkinson, & Meade, 2007).

In the group's major contribution to date, Pagel et al. (2007) explored the role of frequency in change through the number of cognates per word meaning across Indo-European languages. In 2011, Pappas and Mooers used a case study of a single language with a great deal of historical data (Greek) in order to explore the role of frequency in change across a single language lineage. This thesis seeks to address some of the methodological concerns raised in the previous research, as well as make use of a different language lineage (Italic) and a larger sample size in order to add to this body of research, as discussed in greater detail in the following sections of this chapter.

1.2.1. The role of frequency of use in lexical change

In 2007, a group of researchers from the Reading Evolutionary Biology Group published research using phylogenetic methods to investigate the relationship between the frequency of use of lexical items and linguistic change over time. Asserting that frequency affects change monotonically, Pagel et al. (2007) argue that low frequency items will change more over time than high frequency items. In particular, Pagel et al. claim that frequency of use has a direct and "lawlike" impact on the rates of lexical evolution across Indo-European languages. To test this, they examine the number of cognates across the Indo-European language family for each item on the 200-word Swadesh list (see section 2.2 for more detail), and compare this measurement to the cognate's current frequency of use across 4 languages.

The number of shared cognates across 87 Indo-European languages is used to derive a rate of lexical evolution for each meaning on the Swadesh list (i.e. the more independent, non-cognates represented for a word meaning across the 87 languages, the more change has occurred for that meaning, and the higher its rate of lexical evolution). The number of different cognates is a representation of how likely a given word meaning is to be replaced over the course of Indo-European history (which Pagel et al. approximate as 10,000 years), and consequently a representation of how quickly a given word is replaced and/or changes sufficiently to be unrecognizable as a cognate. A

significant relationship was found between the modern frequency of use of a word meaning on the Swadesh list and this measure of its rate of lexical evolution. The authors take this as evidence that the frequency of use of words “exerts a general and law-like influence on their rates of evolution” (Pagel et al., 2007, p. 717). They further claim that frequency is “a general mechanism of linguistic evolution” (Pagel et al., 2007, p. 719), where high frequency of use impedes language change (i.e. change preferentially occurs in low frequency lexical items).

1.2.2. Lexical change within a single lineage

Pappas and Mooers (2011) test Pagel et al.’s (2007) claims about the role of frequency using Greek as a case study. They argue that any impact of frequency on change over time at the level of Indo-European as a family should also be observable within a single language lineage over time. If, as Pagel et al. (2007) assert, frequency has a lawlike influence on change, then the amount of change to a word over time should be inversely proportional to its frequency of use.

The authors found weak support of a relationship between frequency of use and change within Greek, looking at the 200-words on the Swadesh list. A multi-level logistic regression performed between amount of change and frequency of use was marginally significant (p = 0.066, t = -1.513). However, they also found that a model including amount of change in Greek and word frequency had much greater predictive power regarding the number of cognates across Indo-European than did a model including frequency alone. This suggests that the predictors themselves may be poorly understood.

Additionally, Pappas and Mooers (2011) raise two concerns about the study by Pagel et al. (2007). The first is that linguists have previously found that frequency effects both impede and encourage language change depending on the type of word. The second concern is that the modern frequency of use values used may not accurately reflect the historical frequencies of use for the same words for at least some of the lexical items on the Swadesh list. These two concerns will be addressed in the following sections.

1.2.3. Issues raised

The research by Pagel et al. (2007) raises two primary contextual concerns.

Issue 1. The role of frequency of use in language change

While the existence of some sort of relationship between frequency of use and language change has long been suggested in the field of linguistics, research on the nature of that relationship has been inconclusive. If the proposed relationship between frequency of use and change over time exists, why should the nature of this relationship be expected to take the form hypothesized by Pagel et al. (2007), rather than some other form?

Pagel et al.'s (2007) claims about frequency impeding language change, rather than enabling it, are not unprecedented, but they do contradict a substantial amount of prior research, which has predominantly focused on the ways in which frequency makes change *more* likely. It is perhaps indicative of the direction of prior research that when Phillips (1984) wrote suggesting that frequency both enables and impedes change, it was the proposal that some changes primarily affect low frequency words which was taken to be unexpected, and in need of justification by the author.

For this reason, section 1.3 will be dedicated to summarizing some of the previous research in linguistics on frequency of use and language change, providing some context for the current research within the field. This may shed some light on how frequency functions in general, and how we might expect frequency of use to interact with change over time in particular.

Issue 2. Stability of frequency of use over time

The research conducted by Pagel et al. (2007) makes use of modern frequency of use measurements to describe a hypothesized historical relationship between frequency and change. If frequency changes over time, however, this would not be appropriate; the frequency values being used in the research would not reflect the frequency of a particular word when it either was replaced or failed to be replaced by another lexical item. It has yet to be established, therefore, whether the frequency values

made use of by Pagel et al. (2007) are the right measurements to be using for this kind of research.

Section 1.4 of this chapter will therefore summarize the previous research on the stability of frequency of use over time, describing some of the attempts that have been made to test stability and to determine whether these tests are sufficiently conclusive and generalizable to accept the use of modern frequency of use values in research of this type.

1.3. Role of Frequency of use in Language

1.3.1. Frequency of use in synchronic linguistics

There is a great deal of evidence from linguistics and related fields that frequency of use has a relationship with various components of language processing and use. The nature and degree of this relationship is less well understood.[1]

Some research from the field of psychology has indicated that word frequency facilitates word recognition. Howes and Solomon (1951) found that in list-based word-recognition tasks, high frequency words required shorter periods of exposure in order to be correctly identified. Deese (1960) found that when inter-word association was controlled, an apparent relationship between frequency and word recall disappeared, suggesting to the researcher that frequency has no *intrinsic* effect on recall. Both of these studies are contradicted by later research by Mandler, Goodman, and Wilkes-Gibbs (1982), which describes an apparently paradoxical effect of word frequency on memory. According to their experiments, frequency does have a positive effect on word recall; however, frequency was also found to have a negative effect on word recognition. However, Mandler et al. used a different task to test word recognition than did Howes and Solomon (1951), requiring subjects to recognize whether a word presented had

[1] A huge body of literature exists which investigates the nature of frequency of use in general. Material on this topic was included in the following literature review based on its applicability to the discussion of frequency as it relates to language change. Bybee and Hopper (2001) represents a good overview of this subfield.

been part of a previous list. It is therefore possible that different cognitive measures were being measured.

Free association tasks by Nelson and McEvoy (2000) found that high frequency words produce more associates and are produced as associates of more words than low frequency words are. This indicates that they have more connections to and from other words. They attempt to explain the observation that frequency has a negative effect on recognition (as documented in Mandler et al., 1982) by claiming that high frequency words have stronger connections to other words, but because they also have more connections, they are less likely to elicit any individual connected word in association. For example, the association from 'horse' (high frequency) to 'stirrup' (low frequency) is stronger than the association of 'stirrup' to 'horse', but the latter is more likely to be observed in an association task because there are less possible associations for 'stirrup'. The authors claim that this would explain why low frequency words are more likely to be recognized than high frequency words. In testing this hypothesis, however, they found no evidence that high frequency words had more connections *to* other words (although there is some evidence that they have more connections *from* other words). The authors therefore conclude that the effects of word frequency on cognitive task performance are the result of differences in memory accessibility, not connectedness (Nelson & McEvoy, 2000).

While frequency of use is often assumed to be some measure of how often a speaker/reader is exposed to a particular lexical item, it has been suggested that word frequency may be measuring something else. Other possible factors which have been suggested as being measured (or as being conflated with the measurement) include connectivity between words (Deese, 1960; Nelson & McEvoy, 2000), the probability of a word's occurrence (Howes & Solomon, 1951), accessibility of the word in the lexicon (Nelson & McEvoy, 2000), markedness (Krug, 1998), or concreteness (Nelson & McEvoy, 2000).

Regarding the topic of this thesis, any effect of frequency on diachronic language change should be reflected in the synchronic realities of language use and processing. If frequency affects the course of change in a language by altering the likelihood of errors,

the chance of a sound being produced non-canonically, or the likelihood that speakers will learn and retain a form, this should be observable in synchronic investigation. Unfortunately the research on the relationship between frequency, memory, and language processing in general seems to be inconclusive.

1.3.2. Frequency of use in diachronic linguistics

In an early investigation of the relationship between frequency of use and diachronic change, George Zipf proposed the *Principle of Frequency* (Zipf, 1929). This principle states that "The accent, or degree of conspicuousness, of any word, syllable or sound, is inversely proportionate to the relative frequency of that word, syllable, or sound, among its fellow words, syllables, or sounds, in the stream of spoken language. As usage becomes more frequent, the form becomes less accented, or more easily pronounceable, and *vice versa*." (Zipf, 1929, p. 4). This principle rests on the rationale that greater effort of production coincides with greater salience for the listener (e.g. a syllable with primary stress requires more force to produce, but is louder). Speakers will attempt to minimize effort by emphasizing elements of the speech stream that are most useful for interpretation. Sounds, syllables, and words that are less frequent are more diagnostic of their content, and so will be emphasized (undergoing fortition); high frequency sounds, syllables, and words are less useful for interpretation, and so will be de-emphasized (undergoing lenition).

Zipf supports this claim with a large number of case studies. The location of accent on forms within noun paradigms of Sanskrit (i.e. on the root or on the ending) appears to demonstrate that grammatical cases where the accent falls on the ending (the element of the form which enables the listener to identify the case) occur much less frequently than cases where the accent falls on the root. According to Zipf, the ending is more emphasized in forms where it is unexpected, by virtue of being low-frequency. Zipf also investigated the relative frequency of phonemes for a number of Indo-European languages (as well as Hungarian). For the thirteen languages described, there appears to be a strong relationship between the proportional occurrence of a phoneme within a language, and its relative acoustic salience (e.g. overall the more salient voiceless stops are roughly twice as common as the voiced stops at the same place of articulation).

While the focus is on sound change, the broadness of the *Principle of Frequency* leaves open the possibility of frequency applying to other forms of language change as well (Zipf, 1929).

Later researchers have continued to expand on some of the ideas presented by Zipf (1929), including applying them to levels of language other than phonemes. However, much of this research has been undertaken using the theoretical perspective of lexical diffusion (Hock, 1991, p. 649). Lexical diffusion holds that sound changes begin in a small portion of a lexicon and spread throughout the language as a whole over time. This contrasts with the Neogrammarian hypothesis (Hock, 1991, p. 34) that sound changes are exceptionless (the perspective taken by Zipf in his original research), and affect all the contexts to which they apply at the same time. Therefore the interpretation of the observations made in these studies will differ depending on what theoretical paradigm is followed. This research makes assertions about the impact of frequency on “changes in progress”, which according to other theoretical perspectives are not changes in progress, but potentially sporadic changes instead. Even if one is following the latter perspective, the observations from lexical diffusion still have potentially meaningful implications for the role of frequency in language change. If it can be demonstrated that “changes in progress” are found to have affected words of significantly different frequency from words they have not affected, then both the lexical diffusion interpretation (words of X frequency are affected by diffuse changes first) and the Neogrammarian interpretation (sporadic changes preferentially affect words of X frequency) imply the same general conclusion: there is a significant relationship between frequency of use and historical language change.

Hooper (1976) claims that “the relation between sound change and word frequency is just the reverse of the relation between analogy and word frequency” (Hooper, 1976, p. 95). In order to test this claim, schwa-deletion in American English was examined as a case study. Hooper gathered self-reported data on deletion from 8 speakers on 112 words that contained an appropriate context for deletion. The average frequency of words judged by speakers to be most likely to undergo deletion was between three and six times greater than the average frequency of the words judged by the same speaker to be least likely to undergo deletion. This is taken as confirmation

that high frequency items are more likely to undergo reduction, and is completely consistent with the claims of Zipf (1929), although Hooper (1976) argues that the source of these sound changes is errors in casual speech, rather than frequency itself. If schwa deletion is in fact a sporadic change, then the high frequency items are more likely to have undergone this sporadic change. In the same paper, analogical leveling was tested using the case study of verb regularization in modern English. Comparing the frequencies of verbs that have undergone regularization with those which have remained irregular, verbs in the latter category are more than 10 times more frequent than those in the former. This supports the assertion that low frequency words are more prone to analogical leveling than high frequency words (Hooper, 1976). Phillips (1984) provides additional support for Hooper's (1976) claims about frequency. This author states that changes such as vowel reduction, deletion, and assimilations preferentially affect high frequency words, while analogical leveling is impeded by frequency, and so preferentially affects low frequency words. The phenomenon of Southern American glide deletion is presented as a case of the latter kind of change (Phillips, 1984). Southern American glide deletion refers to the variable deletion of the glide in the consonant-vowel cluster [ju] (e.g. *news, tuber*) in some dialects of English, a change which Phillips claims is more likely to occur in low frequency words.

In agreement with this concept that frequency of use has a paradoxical effect on change (both enabling and impeding it), Bybee and Thompson (1997) describe what they refer to as The Reduction Effect and The Conserving Effect. They refer to the 'reduction effect' as phonological reduction, loss of internal structure, and semantic bleaching that are promoted by high frequency. Other research by Bybee alone has found more specific evidence for this effect in the form of a positive relationship between high frequency of use and word-final /t/ and /d/ deletion in American English (Bybee, 2002). Regarding the 'conserving effect': Bybee and Thompson (1997) claim that high frequency words are resistant to analogical change. This effect is claimed as an explanation for why English pronouns have maintained distinct case forms, even as English lexical nouns have lost their case forms: as high-frequency words, the pronouns

are resistant to analogy. The use of this anecdote as evidence is extremely limited, however.[2]

Fosler-Lussier and Morgan (1998) investigated the relationship between both speaking rate and word frequency on pronunciation. However, frequency here was used as a component of a compound variable 'word predictability', which includes both prior probability (the frequency of a word in the language as a whole) and the probability of a word in context (which relies on the surrounding discourse). For this reason, conclusions drawn from this research about frequency as an independent measurement are not directly comparable to others. In the spoken corpus investigated (the Switchboard corpus), words with greater predictability were less likely to be pronounced canonically ($p < .05$). There was an interaction with speaking rate such that speaking rate had a greater effect on the pronunciation of high frequency words. The authors claim that the greater predictability of high frequency words means that listeners are better able to understand their meaning from context, and therefore that they can be permitted to be more variable in their pronunciation. Only 33% of words in the corpus were produced canonically, and words differed more from canonical pronunciation when speaking rate and predictability were higher. The information-theoretic viewpoint espoused by the authors makes the prediction that since high-frequency words are more predictable overall, greater levels of variability in their pronunciation should be tolerated. In the context of this thesis, this might in turn predict that such words might also have greater rates of change (Fosler-Lussier & Morgan, 1998).

Krug (1998) describes frequency of use as "a central cognitive motivation" in language change, indicating that there is evidence found in multiple languages that the words in the language that are oldest (and the shortest) are generally also the most frequent. Krug is focused, however, on the frequency of strings, rather than of individual lexical items. He argues that two forces are at work: a tendency for high frequency words to remain irregular, and a tendency for high frequency forms to be contracted or

[2] Note that even within the same change (i.e. loss of grammatical case from Old English to Modern English), the definite article 'the' (one of the highest frequency words in English) has completely lost its case system, which once encompassed 14 morphological forms (Hasenfratz & Jambeck, 2005).

shortened through grammaticalization. Krug finds that the likelihood of contraction between two words is significantly increased the higher their string frequency. Logistic regression indicated that a higher ratio of contracted to uncontracted forms (for a string of two words) is predicted by a higher string frequency. This is tested across two corpora separated in time, and while the likelihood of contraction overall goes up over time (i.e., the probability of contraction is higher for the later corpus), the relationship between string frequency and contraction is the same for both corpora. Examples from a number of other languages are given by way of demonstration, indicating that high frequency strings are generally more likely to be contracted and merge. Krug goes so far as to state that "this article proposes that language change in general may be due to frequency constraints" (Krug, 1998, p. 309). Particularly, that string frequency is the most important motivation for coalescence, but that it also impacts the regularization of paradigms, either directly, or through analogy (i.e. the high frequency of a productive pattern might increase the likelihood of the pattern being extended via analogy, regularizing the paradigm).

Gregory, Raymond, Bell, Fosler-Lussier and Jurafsky (1999) argue that frequency and predictability are really measurements of the same phenomenon: probability of occurrence. Logistic and linear regression were used to model three variables (t/d deletion, t/d flapping, and duration shortening) and factors irrelevant to the research but known to be associated with these variables were controlled for (e.g. rate of speech, quality of the following vowel) in a subset of the English Switchboard voice recording corpus (Gregory et al., 1999). Three components of probability were used as independent variables: prior probability (i.e. frequency of use), collocational probability, and discourse probability. While frequency was not a factor in tapping, it was found to be a significant predictor for deletion ($p < .00005$), as well as for duration ($p < .00005$). That is, the highest frequency words were more likely to undergo deletion, and were on average 22% shorter than words of the lowest frequency.

Bybee and Scheibman (1999) grouped instances of the contraction "don't" into four categories based on the level of reduction (i.e. full/reduced vowel, full stop/flap) and tested whether the level of reduction was related to the frequency of the context in which the contraction was found. Frequency here refers to the frequency with which the

contraction “don’t” occurs with a particular surrounding context, out of instances of “don’t” in general, rather than the frequency of occurrence of the context in general. The reduced cases of “don’t” only ever occurred with pronouns (never lexical nouns), and the two most reduced categories occurred almost exclusively with the first person singular pronoun. These form the most frequent contexts in which “don’t” occurs. The authors hypothesize that the pronouns (and in particular the first person singular pronoun) coupled with “don’t” occur often enough in speech that they form a unit in the memory of speakers, leading them to be treated as a phonological unit. The reduction of the initial stop to a flap indicates that the contraction is being treated as a unit with the preceding word, as flapping is generally conditioned by surrounding vowels within the same phonological unit. Frequency of use is therefore leading “don’t” and the preceding pronouns to be used as a unit, facilitating their reduction.

Sequences of phones adjacent at word boundaries palatalize variably in English (compare “would you” with “good, you”). Bush (2001) examines what factors might influence the presence or absence of palatalization in these contexts, including the text frequency of the adjacent elements (i.e. how often the two words are found together in a string). A chi-square test was performed on data extracted from a subset of the CHILDES natural speech corpus (Bush, 2001), and confirmed that the string frequency and likelihood of palatalization are highly interdependent ($p < .0001$). Word pairs that produce the appropriate context for palatalization are more likely to palatalize the greater the frequency with which they occur together.

Pluymaekers, Ernestus, and Baayen (2005) investigated the relationship between frequency and acoustic reduction in Dutch for morphologically complex words. The authors tested whether the frequency of the root word had an impact on the duration of the affix. Their hypothesis was that an affix forming part of a high frequency word would be shorter in duration than the same affix forming part of a low frequency word. For three out of the four affixes investigated, there was a significant effect of frequency on either the pronunciation of the affix as a whole or some segments thereof. This research contributes additional support to the claim that high frequency is correlated with acoustic reduction.

In 2007, Lieberman, Michel, Jackson, Tang, and Nowak repeated Hooper's (1976) previous test of the impact of frequency on the regularization of Old English irregular verbs using more modern statistical methods. A total of 177 Old English irregular verbs (of which 98 remain irregular in modern English), were put into 6 logarithmic bins according to frequency of use (as ascertained by a modern corpus of English). The likelihood of regularization for each bin was plotted against frequency. The results indicate that frequency is negatively correlated with change over time, with the half-life of a verb being roughly proportional to the square root of its frequency. While the results are suggestive, the study is restricted to the observation of a single change, affecting a single part of speech, for a single language, over a period of approximately 1,200 years. The authors do not claim that their results should be broadly generalized without further investigation.

The results above collectively seem to suggest a relationship between frequency and linguistic change. Other variables, such as memory and cognitive processing, appear to be involved as well. It remains unclear what the precise nature or directionality of the relationship may be. While this topic has been of interest to linguists for a substantial period of time, there does not appear to be any single explanation which straightforwardly accounts for all the evidence, but rather a slew of varying suggestions, all of which are supported by a handful of individual results. The relationship between frequency and change over time is by no means settled.

1.4. The Stability of Frequency of Use

1.4.1. Implications of stability of frequency of use

If a general relationship between language change over time and frequency of use exists, the historical frequency would be of importance, rather than the modern frequency. Historically high frequency of use might be hypothesized to result in resistance to change in lexical items, but the word would have to be of a particular frequency at the time the change was triggered, or failed to be triggered. Accurate frequency of use values for historical stages of languages are extremely difficult to come by; very few languages have a long written history. For those languages with such a

history (such as Old English, Greek, Latin, and Sanskrit), the material that remains available in the modern day is almost exclusively written material (rather than spoken language), and there is a strong selection bias. For example, much of the written material in Latin that has been preserved over time comes from monastery collections. Monks copied Classical Latin texts over centuries; the process of copying by hand is very labour-intensive, and so generally only texts that were considered valuable or worthwhile in some way would be copied. We would not expect to find a large amount of pagan religious texts to be kept by Christian monks.

The lack of good historical corpus data means that researchers studying change over time almost exclusively use modern frequency values in their research. This assumes that the modern frequency values are the same as or very similar to the historical frequency values; therefore this practice assumes that frequency of use is relatively stable over time. If this assumption is not true, and frequency of use varies greatly over time, then it is highly problematic to use modern frequency to investigate a relationship between change and historical frequency.

1.4.2. Previous attempts to address the stability of frequency of use

Several researchers have previously recognized the importance of establishing whether frequency of use remains stable over time. Some of these studies have attempted to address the issue directly, by performing small-scale tests on their data.

Hooper (1976), in his work on frequency of use and language change (see above) raises the issue of frequency change as a potential area of concern. As a test, Hooper identifies six verbs with a suppletive past-tense form (*keep*, *leave*, *sleep*, *creep*, *leap*, and *weep*), and cross-references whether they can variably occur without schwa (i.e., have participated in the sound change) with their modern frequency. Three of the words occur variably without schwa (*creep*, *leap*, and *weep*) and three do not (*keep*, *leave*, and *sleep*)[3]. The average frequencies of use of the two groups are extremely

[3] The past tense of the verb “creep”, for example, may be pronounced with a schwa (i.e. -[pəd]) or withot a schwa (i.e. -[pt]) (Hooper, 1976).

different, with the former group having an average frequency of 37 compared and the latter group exhibiting an average frequency of 485. Despite the small sample size, Hooper claims that the differences in frequencies are sufficiently large to be able to suggest a real effect. Hooper further claims that differences in frequency over time would have to be very large in order to affect the main results meaningfully.

Woods (2001) conducted research on a historical Spanish corpus consisting of over a million words in order to determine whether the rank orders of the three most common words in Spanish have changed over time. No further statistical tests were done, and only a small number of the most common words were included in the research. The findings indicate, however, that the order of the three most frequent words in written Spanish has indeed changed over time. While the three most common words in modern Spanish are consistently *'de', 'la'* and *'que'* (in descending order of frequency), in the historical Spanish corpus, *'de'* is only the third most frequent word overall, behind *'la'* and *'que'.* What's more, the relative ordering of these three words is a great deal less consistent, varying across texts and authors. The author suggests that there has been a major change in Spanish frequency of use in recent centuries, at least regarding the most common words. This would seem to suggest that frequency of use is not stable over time.

The previously cited research on frequency of use and change over time by Pagel et al. (2007) relies on the assumption that frequency of use is stable over time. The authors state outright that "*If frequency of meaning-use is a shared and stable feature of human languages,* then this could provide a general mechanism to explain the large differences across meanings in observed rates of lexical replacement" (emphasis added) (Pagel et al., 2007, p. 717). The authors are aware of the importance of this assumption, and attempt to test it by means of a cross-linguistic analysis of the frequencies of use of four modern Indo-European languages (English, Spanish, Russian, and Greek). For the 200 word-meanings of the Swadesh list, the frequencies of use between these four languages (as established through modern corpora) are found to be highly correlated. Pagel et al. (2007) take the similarities between these languages to be evidence that frequency of use has remained stable over time for the Indo-European family. This conclusion relies on the unsupported assumption that similarities in the

frequency of use values of related modern languages are evidence that those frequency of use values are inherited from a shared ancestral language, and therefore that those values have remained stable over time. However, it is possible that the similarities observed by Pagel et al. (2007) are the result of convergence of use or borrowing. The cultural context in which speakers of modern English live (with regards to daily life, technology, methods of communication) is much more similar to the cultural context of speakers of modern Spanish than either of them are to that of a speaker of Proto-Indo-European living approximately 10,000 years ago. Lexical frequency of use may be partially a reflection of this context.

Furthermore, populations from different branches of the Indo-European language family have not existed in isolation over the course of their languages' development; the impact of modern languages on each other through mechanisms such as borrowing might also contribute towards similarities between modern languages. While the similarities in frequency of use between the four modern Indo-European languages merit investigation, they do not necessarily tell us anything about the changes in frequency of use that have occurred (or failed to occur) over the course of Indo-European history. As such, this test does not directly address the research question.

Calude and Pagel (2011) partly addresses this concern, using a sample of 16 languages[4], rather than the four examined by Pagel et al. (2007). This paper determines the frequencies of use for the 200 words of the Swadesh list for each of the languages using corpora that range in size from less than 1 million words to over 450 million words. They report that the average inter-correlation found overall is 0.73; the average within-IE correlation is 0.82. Like the paper by Pagel et al. (2007), this work assumes that correlations in frequencies of use between modern languages reflect frequencies of use inherited from an earlier, shared ancestral language. Without this assumption, the test conducted in this paper does not address the research question any more successfully than Pagel et al. (2007) did. These are all still *modern* languages, and without the

[4] The paper claims that 17 languages are used, but this number includes both Spanish and Chilean Spanish, which are dialects of the same language, not different languages. There seems to be no explanation for this decision, as the source of the paper's language classification is the online Ethnologue, which does not consider these dialects to be separate languages (Lewis, Simons, & Fennig, 2014).

assumption that they reflect the inherited frequency of previous ancestral languages, they cannot determine whether frequency of use remains stable over time. At best, they might imply that there is a shared feature of human cognition, and so that the correlations between the frequencies of all languages will remain this high. The authors take these results to indicate that they have confirmed that frequency of use is a "shared feature of human languages" (Calude & Pagel, 2011, p. 1106).

For 2% of the words in Calude and Pagel's (2011) dataset, FoU data were not available. Depending on the source of the data, this could mean that the words are extremely low frequency (i.e. don't appear in the corpora), or alternatively, if the source of the frequency of use data is a list of words, rather than a direct investigation of the corpus, it may be that the word in question was not included in the list, despite appearing in the corpus. Calude and Pagel (2011) dealt with these cases by replacing the missing data with the mean frequency from either all the other languages or the mean IE frequency (depending on whether the missing data was from an IE language itself). While the presence of zeroes is often a problem in this kind of research, when testing the level of similarity across languages, replacing missing data with the mean values is likely to increase the degree of homogeneity in the data, potentially inflating the correlation (although given the small percentage, this is not a huge concern). Calude and Pagel do not dedicate a great deal of space in the paper to this component of the research, which is unfortunate, given the large quantity of cross-linguistic frequency of use information available.

Lieberman et al. (2007), in their investigation of the effects of frequency using the test case of English verb regularization (described above), also raise the issue of changing frequency of use over time. They attempt to directly test changes in the frequency of the verbs under examination using a small corpus of Middle English, an intermediate point in the time period they are studying. They found that only 5 of the 50 verbs investigated had changes greater than a factor of 10 in their frequency. Due to the fact that this research makes use of frequency bins, Lieberman et al. (2007) state that a large number of words would have to change bins (i.e. experience a very large change in frequency of use) in order to alter their results significantly, a level of change which they do not think is supported by the test of the Middle English corpus. This paper

includes a direct test of how frequency has changed or failed to change over time using historical texts, and the results seem to suggest that it remains sufficiently stable over time for use of modern frequency of use values to be appropriate for this kind of research. However, the scope of the Lieberman et al. (2007) paper is limited to those English verbs that were also strong verbs in Old English. While it may be useful for drawing conclusions about the research in that specific subject, it would be overly simplistic to extend this finding to other languages and parts of speech without further investigation.

Several researchers discussed above identify the stability of frequency as being an issue of concern. Small-scale quantitative tests performed by researchers have been inconclusive (Hooper, 1976; Woods, 2001). Corpus-based comparisons of modern languages (both within and across language families) have found a high degree of correlation, but without evidence that similarities in modern languages reflect inherited historical frequencies, these comparisons do not address the research question (Pagel et al., 2007; Calude & Pagel, 2011). The only paper described herein which directly tests the issue of frequency change over time using a large number of words is Lieberman et al. (2007), with the direct comparisons of modern English frequencies with Middle English corpus-derived frequencies. While Lieberman et al.'s conclusions were that the historical frequencies were not sufficiently different from the modern frequencies to create doubt about their conclusions, this may be in part due to the methodology employed, namely binning words according to frequency. Lieberman et al. claim that, due to the use of bins, the frequency of a substantial number of words would have to have changed a great deal in order to invalidate their results. This does not automatically determine, however, that research which does not make use of bins, but instead refers to rank order or number of occurrences per million (such as the current research, and the methodologies of Pagel et al., 2007, and Calude & Pagel, 2011) can draw similar conclusions about the irrelevance of possible changes to frequency of use over time; it is not known if analyses that use correlations of log-transformed frequencies would be more sensitive to changes through time than those, like Lieberman et al.'s (2007), that first bin words into frequency classes.. It is also important to note that Lieberman et al.'s (2007) research was specifically targeted at English verbs (more specifically, a subclass

thereof), and caution should be used when applying their results to other parts of speech or languages.

Pappas and Mooers (2011) raise the concern of frequency of use stability as a potential methodological problem. The authors do not accept Pagel et al.'s (2007) claim that the high degree of correlation between the frequencies of the four modern languages is evidence that the 200 word meanings in question have had the same frequency of use over time. Several individual words are raised as being intuitively problematic in this framing (e.g. it would be reasonable to suggest that the frequency of use of "bark" (as a noun) has changed over the last several thousand years), but they offer no solutions.

1.4.3. Basis for judging similarity

An important question for this component of the current research is how to judge the results. What strength of correlation would be sufficient to determine that it *is* methodologically appropriate to use modern frequency of use values in lieu of historical values? The most straightforward solution would seem to be judging based on comparison with existing research.

Research by Pagel et al. (2007) constitutes the most extensive directly applicable comparison. The frequency of use comparisons between four Indo-European languages found that the languages were highly correlated, with the value of r ranging from a minimum of 0.78 to a maximum of 0.89, with a mean value of $r = 0.84$. If, as the authors suggest, these high correlations are due to shared ancestry and stability, then we would expect a correlation between one of the modern Indo-European languages and an older form of the same language to be equally strong or even stronger than the correlation between the modern languages.

The later work conducted by Calude and Pagel (2011), which also examined the correlation in frequency of use values between languages, used a much larger number of modern languages. The average inter-correlation reported among the Indo-European languages in the study was $r = 0.82$. The average inter-correlation reported for all the languages (with all Indo-European languages being represented by a single

measurement to avoid over-biasing the sample), found a correlation of r = 0.73. The difference in the strength of these two correlations might appear to support Pagel et al.'s (2007) claim that correlations between modern languages are inherited (i.e. the more distantly related languages are less well correlated than the more closely related languages). However, there was a large difference in the relative sizes of the corpora used for the Indo-European sample compared with the non-Indo-European sample: the Indo-European corpora contained on average 150 million words, while the non-Indo-European corpora contained on average less than 5 million words. Given that the smaller corpora may have more variable frequency values (i.e., each individual contribution to the corpus, which may be highly biased in favour of particular low-frequency words, will have a greater impact on a smaller corpus than it would on a larger corpus), a lower correlation when they are used is to be expected, and does not necessarily support Pagel et al.'s claims.

Synchronic research on the correlations between frequency of use measurements might also provide a point of comparison for the results. Alonso, Fernandez and Diez (2011) measured oral frequency norms for Spanish for a total of almost 70,000 word forms. As part of their research, they assessed the correlation between their oral frequency norms and three other sets of norms that had been gathered by other researchers. The correlation with subjective frequency norms was r = 0.68, with written frequency the correlation was r = 0.79, and with subtitle-based frequency r = 0.67 (all significant $p < .001$). These reflect the correlations within a single language between different media of presentation.

1.4.4. Research questions

As the discussion of the relevant literature in this chapter has shown, the study by Pagel et al. (2007) and the response by Pappas and Mooers (2011) raise two important research questions about the role of frequency of use in lexical change:

1. Does frequency of use remain stable over time so that we can confidently use rates from modern corpora in our investigations of diachronic change?

2. Does the negative correlation between frequency of use and lexical change that Pagel et al. (2007) discovered in change across Indo-European also hold within a single language lineage?

In the next chapter, I discuss in detail the methods I employ in this thesis to pursue these questions.

Chapter 2.

Methodology

2.1. Introduction

Chapter 1 presented some of the questions that have been raised by previous studies. This chapter will seek to describe in detail the methodology used to acquire the data that will be used to test the relevant hypotheses in chapters 3 and 4. Section 2.2 will describe the wordlist used, as well as the reasons why this list was chosen above the more standard Swadesh list. The two corpora used (the Perseus Digital Library Project and the Corpus del Español) will be discussed in section 2.3, as well as the search terms and extraction methods used with them. Finally, section 2.4 will detail the coding process, as well as the six-point coding scheme used (taken after Pappas & Mooers, 2011), and what types of language change are indicated by each of the six categories in the scheme.

2.2. Wordlist

The focal research on the relationship between frequency of use and amount of change (Pagel, Atkinson, & Meade, 2007; Pappas & Mooers, 2011) has relied on the use of the 200-word Swadesh list (Swadesh, 1952). Use of this list enables straightforward comparisons between analyses and languages, due to its widespread use, but the list also has disadvantages. The 200-word Swadesh list was specifically designed to include only “universal everyday vocabulary” (Swadesh, 1952, p. 455), so that it could be compared across languages easily. The collection of words was also intended to avoid borrowing, based on the idea that ‘cultural’ vocabulary is more prone

to borrowing than ‘intimate’ vocabulary. By limiting the list of words to intimate vocabulary, the words should be resistant to borrowing (Swadesh, 1952).

For the purposes of the current research, however, the Swadesh list is non-ideal. The present research seeks to investigate the phenomenon of borrowing as a level of lexical change (see Coding, below), and a list that avoids borrowing would hamper this investigation. Note that in Pappas and Mooers (2011), the category of borrowing had to be eliminated due to a small number of words. Further, 200 words is too small for the purposes of some of the statistics performed herein; the fewer data points that are available, the fewer variables and levels thereof that can be investigated. For these reasons, the Intercontinental Dictionary Series list was used for the present research.

The Intercontinental Dictionary Series (IDS) (Key & Comrie, 2007) is a database project that seeks to collect lexical information on a large number of languages for comparative purposes. A list of approximately 1,310 word meanings is available online, and the database currently contains this list of word meanings translated into 241 different languages. The IDS list was used for this project due to its large size and accessibility. This list is also organized by semantic domain into 22 categories. The research here could be more easily replicated in another language using the IDS than would be possible with a less accessible list.

The IDS list for Latin was retrieved from the website directly (Key & Comrie, 2007, <http://lingweb.eva.mpg.de/ids/>). I translated the Spanish words using a variety of dictionary sources, and the resulting Spanish wordlist was checked by a native Spanish speaker. I further refined and altered the Latin translations, while the word meanings (available in English at this time) have been maintained. In places this alteration was to correct errors identified in the IDS translations, as well as keep the Latin terms consistent with the meaning of the Spanish terms. Many of the IDS word meanings are partly ambiguous, and some of the Latin translations are missing from the downloaded list. Some word meanings are also ascribed to multiple Latin words as translations; this is not appropriate for the current research, as the analysis requires that each word meaning be associated with one code and one set of frequency of use values.

Words were removed from the list when the meaning could not be encapsulated by a single lexical item in one of the languages; phrasal translations were not accepted.[5] Repetitions of lexical items in one of the languages were kept in the list, provided that at least one of the languages used different words for the meanings. Semantic distinctions not maintained by either Latin or Spanish were removed to avoid hollow repetition. To demonstrate, see examples 1 and 2.

1. 'nose' LA: *nasus* SP: *nariz*

2. 'nostril' LA: *naris* SP: *nariz*

Latin has separate words for the meanings 'nose' and 'nostril' (*nasus* and *naris*, respectively), while Spanish uses the same word, *nariz*, to refer to both of these meanings, and does not distinguish between them with two individual lexical items. Repetitions of this type were included. On the other hand, the IDS list includes separate meanings for 'we', 'we (inclusive)' and 'we (exclusive)'. These would be necessary to capture the full pronoun system of a language which distinguishes between inclusive and exclusive first person plural pronouns, but neither Latin nor Spanish makes this grammatical distinction, so functionally all three meanings are the same (see examples 3, 4, and 5).

3. 'we' LA: *nos* SP: *nosotros*

4. 'we (inclusive)' LA: *nos* SP: *nosotros*

5. 'we (exclusive)' LA: *nos* SP: *nosotros*

Repetitions of this type were excluded, so that only 'we' in the example above remains on the edited list. The resulting edited IDS word list is comprised of a total of 1147 words. Of these, 696 are nouns, 290 are verbs, 106 are adjectives, 22 adverbs, and the remaining 33 are assorted function words.

[5] The script used to extract frequency values from the Latin results is only capable of identifying single words, therefore any phrasal translation would result in a frequency value of '0'.

2.3. Corpora

2.3.1. Perseus Corpus

Description of the corpus

The Latin texts examined came from the Greek and Roman collection of the Perseus Digital Library Project (PDLP) (Crane, 2010), which is available online. This corpus was assembled largely as an experiment in digital library management. It was used for this research due to the large number of texts it contains, and also due to the very useful vocabulary and lemmatization search functions, which make it possible to gather large amounts of data from the material with relative ease. The PDLP contains 10.5 million words of Latin, but only a subset of 4.2 million words was used. The material from the corpus not used in the present study (approximately 6.3 million words of Latin) was excluded based on a number of considerations. Several texts in the corpus were repeats, and in such cases only one version of the text was used. A small number of texts were excluded due to technical difficulties: the vocabulary tool was unable to gather data from the texts in question.

Texts were further restricted based on approximate time of writing. This was done so that the Latin under examination would be constrained to a particular time period, and therefore remain as internally consistent as possible. Later time periods were excluded as they represented Church Latin, Academic Latin no longer in use as a living language, and Latin already in the process of splitting into the Italic languages. Latin as a language is generally divided into time periods as shown in table 2.1.

Table 2.1. Time periods for Latin Literature (Wheelock & LaFleur, 2005)

Designation	Time period	Estimated dates
A	Archaic through Early republican period	pre 80 B.C.E.
B	Late republican and Augustan period (the "Golden age")	80 B.C.E. - 14 C.E.
C	Post-Augustan period (the "Silver age")	14 C.E. - 138 C.E.
D	Patristic period	Late 2nd c. C.E. - 5th c. C.E.
E	Medieval period	6th c. C.E. - 15th c. C.E.
F	Renaissance	post 15th c. C.E.

This research makes use of texts for which the author is known or suspected to have been alive and working during time periods B and C. This is not intended to be completely certain and precise, as often little is known about the authors of the works in question, but rather an attempt to try and restrict the texts to a general time period.

It may be worth noting that the precise dates used to constrain the time periods are primarily historical, rather than linguistic. For example, the divide between Golden age Latin and Silver age Latin is 14 C.E.; the reason this particular date is used is because it represents the year of Augustus Caesar's death, a major turning point in the history of the Roman Empire. However, the choice of which time periods to include in the current study was due to both the greater availability of texts from these periods and the fact that the end date for the second time period (the second century C.E.) is early enough to ensure that the Latin texts in use will not overlap with early Spanish texts (Wheelock & LaFleur, 2005).

The subset of the corpus used (4.2 million words in total, as described above) therefore consists of exclusively written material from 117 texts attributed to a total of 35 authors, and includes both poetry and prose. Some authors contributed a large number of texts (for example, 31 texts are by M. Tullius Cicero), while others contributed only a single work. A detailed list of the material represented in the Latin corpus can be found in appendix B.

Extraction procedures

The texts used were broken into groups of 1 - 5 texts each, according to the technical limitations of the corpus interface. These groups were each put through the Perseus vocabulary tool, and the source code of the vocabulary tool search results was downloaded into a file. An awk script (available in appendix C) was run through terminal using a list of the relevant keywords, extracting the relevant weighted frequency from the source code of the downloaded files and associating them with the words on the list. These group results (in the form of raw weighted frequencies, rather than normalized frequencies) were tabulated for all texts in the corpus and normalized based on the combined word count of the texts used in the corpus, as determined by the Perseus search function.

The PDLP vocabulary tool takes as an input a number of texts requested by the researcher, it then tabulates the frequencies (including the weighted frequency) for every lemma present in those combined texts and reports them in a table. The source code of this table is then saved; the awk script is then run on this file using a reference file that contains a list of target Latin lemmas. When the target word is found, the script extracts weighted frequency that accompanies it. The combined values of all the extracted frequencies for a particular lemma (for each group of texts the script is run on) represent the raw frequency that is used as a data point.

Use of weighted frequency

Words in the PDLP are not independently tagged as being part of a particular lemma. This means that sometimes a particular token cannot be unambiguously assigned to one lemma or another. The frequency calculations are performed such that the minimum frequency excludes all ambiguous cases, while the maximum frequency includes all ambiguous cases. These measurements of frequency are both problematic for a morphologically rich language like Latin, where the individual forms of a particular lemma may overlap with one or more other lemmas. For example, the Latin adjective *bellus* translates as "beautiful" (see example 6). As an adjective, it agrees in number, gender, and case with the noun it describes. This agreement leads to a large number of different morphological forms.

6. 'beautiful' LA: *bellus* SP: *bello*
   Latin morphological forms: *bellus,* ***belli****,* ***bello****,* ***bellum****, belle,* ***bella****, bellae, bellam,* ***bellorum****,* ***bellis****, bellos, bellarum, bellas*

The Latin noun meaning "war", *bellum*, and its morphological forms, can be seen in example 7.

7. 'war' LA: *bellum* SP: *guerra*
   Latin morphological forms: ***bellum****,* ***belli****,* ***bello****,* ***bella****,* ***bellorum****,* ***bellis***

The forms which overlap between the two lemmas are in bold. It is apparent from this surface comparison that six forms are shared between the two, including all forms of

the noun “war”. In reading, these words are easily disambiguated by a fluent Latin user, due primarily to contextual information. Since the PDLP is not tagged for part of speech or other syntactic information, however, there is no way for the search function to establish which lemma one of these six forms belongs to when it occurs in a text. These are certainly not exceptional cases, indeed, there is further overlap for these six forms: *bellum* may also mean “white daisy”, while *bello* may be a form of the verb “to fight”. For these examples alone, using maximum frequency counts would ensure that every instance of the word “war” would be counted at least twice, with several being counted three or more times, while minimum frequency would fail to count *any* instances of the word “war”. In the absence of a tagged corpus, therefore, a more appropriate measurement is a weighted frequency, which the PDLP provides.

When a weighted frequency is calculated, each individual token in the corpus is assigned a score of 1. For unambiguous words, which could only belong to a single lemma, a frequency of 1 is added to the count for that lemma. When the token is ambiguous, however, the score of 1 is divided by the number of possible lemmas it could be a form of, and this quantity is added to the weighted frequency counts of each possible lemma. For example, if a token is ambiguous between four different lemmas, the total raw frequency score for each of those lemmas will be increased by 0.25 (1/4). In this way unambiguous tokens are weighted more heavily than highly ambiguous tokens. Additionally, each token is only counted once, so that the overall number of words in the corpus is equal to the collective weighted frequencies of all lemmas that are found in it (Crane, 2010, <http://www.perseus.tufts.edu/hopper/help/vocab#wft>).

Problematic cases

Within the PDLP, the pronouns varied in how they were lemmatized. The first person pronoun was divided into singular and plural forms, which were considered separate lemmas. The second and third person pronouns, however, were respectively lemmatized with both singular and plural forms as the same lemma. In order to maintain consistency with this scheme, I merged the second and third person pronouns in Spanish in a similar way so that they would be directly comparable with the Latin.

2.3.2. Corpus del Español

Description of the corpus

The Corpus del Español (CdE) (Davies, 2002) contains over 100 million words of Spanish. It includes texts written from the 1200s to the 1900s. Texts from the 1900s are organized into four genres: News, Fiction, Academic, and Oral. These four genres collectively make up approximately 20 million words of Spanish. For the purposes of this research, the oral material was excluded, as the Latin corpus is exclusively composed of written material. Only texts from the 1900s were used in order to ensure that the sample would be internally consistent producing a set of 15 million words across three genres. This set includes material from a variety of sources across multiple countries, and is not composed solely of a single dialect of Spanish.

No attempt was made to match the genres of the CdE with the Latin corpus. In part, this was due to the small size of the Latin corpus, as separating the PDLP into subsets would result in very small samples that would not be appropriate to generalize. Additionally, these modern Spanish genres do not correspond precisely with anything present in the Latin texts. For example, one of the genres represented in the CdE is 'News', but newspapers and magazines did not exist in their current form two thousand years ago, and consequently the PDLP does not contain any examples of these types of texts.

Extraction procedures

Both raw and normalized frequencies were extracted from the Spanish corpus by individually searching for each item on the list. The values were recorded individually for each of the three genres examined (Academic, Fiction, and News), and the raw frequencies from all three of these were combined to produce the normalized frequency of the word for the entire subset of the corpus. The search terms used specified that the lemma was being searched for, rather than the individual form, and the intended part of speech was also specified.

While the overwhelming majority of words were searched for using the format above, there were a small number of problematic cases for which this was not

appropriate. For example, some interrogative pronouns did not return results when their part of speech was specified; some of the pronouns had to be searched for as multiple precise strings to ensure that the lemmatization of these forms reflected that of the Latin.

Use of maximum frequency

The normalized frequencies for the Spanish corpus are maximum frequencies. This is appropriate for the Spanish results for several reasons. Primarily, the Corpus del Español includes tagging for part of speech; this means that it is possible to specify that only the strings identified as a particular part of speech should be returned. Additionally, the typological character of Spanish (namely, that it is not a case language), means that there are considerably fewer forms for each individual Spanish lemma than for an equivalent Latin lemma; consequently, there is much less direct overlap between the morphological forms of the language, and less ambiguity is expected when returning search results.

2.4. Coding

In order to investigate the relationship between frequency and change, some quantitative measurement of change was required. I coded each lemma pair (Latin and Spanish) using one of the coding schemes outlined in Pappas and Mooers (2011). The more detailed six-point scheme (rather than the three-point scheme also described in the same paper) was used, as the larger word list makes more fine-grained coding feasible without excessively reducing the number of words in each code. This includes (in order of least to greatest degree of change): sound change, paradigm leveling, general analogy (referred to as “four-part analogy” by Pappas & Mooers), syntactic reanalysis, semantic change, and lexical borrowing. After coding was completed, however, it became clear that very few of the words on the IDS list were could be coded as syntactic reanalysis (16 in total). This category was therefore merged with general analogy for the data analysis, so that the coding is effectively on a five-point scale. The individual codes are described in more detail below, including several examples of each of the six original codes.

In cases where more than one form of change has occurred, the item has been coded with the greater form of change (i.e. the higher code). For instance, a word which has undergone paradigm leveling (1) as well as general analogy (2), would be coded as (2); while a word whose origin is a lexical item of a different meaning (4) from a different language (5), would be coded as lexical borrowing (5).

The primary resource for coding used in this research is Elsevier's Concise Spanish Etymological Dictionary (De Silva, 1985). This resource established etymological origins of the majority of words on the list, and the words were then coded according to the criteria set out below. Most often, De Silva (1985) established the etymological connection between the Spanish word and the word from which it was derived, but gave little in the way of specific details on the changes that the lexical item had undergone. In such cases, two resources on the phonological changes between Latin and Spanish were used: Boyd-Bowman (1954) and Mendeloff (1969). When the Spanish word was derived from the Latin word of the same meaning, therefore, the amount of change could plausibly be coded as (0), (1), or (2). The detailed accounts of the changes between Latin and Spanish given in Boyd-Bowman and Mendeloff allowed to coder to determine one of the following scenarios: all changes between a Latin and Spanish word could be accounted for by established regular sound changes (code 0); all changes could be accounted for provided the Spanish word was derived from a specific morphological form in the Latin paradigm (code 1); or some of the changes that the word had undergone were sporadic, irregular, or analogical in nature (code 2).

When De Silva (1985) did not include an entry for a particular Spanish word, its etymological relationship with the Latin was established either by reference to the extensive lists of examples given in Boyd-Bowman (1954), or by a plausible semantic and phonological connection (e.g. a Spanish word is completely phonologically predictable from its Latin equivalent, save for the presence of an unexplained 'll' before the final vowel: this 'll' is characteristic of the Latin diminutive suffix and is found in a large number of Spanish words attested as having been derived from a Latin diminutive form, meaning the word can be coded as (4)).

2.4.1. Code 0: Sound Change

Code (0) describes tokens where the only changes that have taken place are the regular sound changes from Latin to Spanish. For these words, nothing has been removed or added to the original stem, and the meaning and part of speech of the word has remained the same (Hock, 1991, p. 34). Example 8 shows the word for 'salt', where no changes have affected the root form of this word, and the relationship between the terms is completely transparent.

8. 'salt' LA: *sal* SP: *sal*

Only rarely are the Latin and Spanish forms completely identical, however. Another example coded as (0) is the meaning 'storm', example 9.

9. 'storm' LA: *tempestas* SP: *tempestad*

As there is no *s* > *d* development from Latin to Spanish, it would appear that the word has developed irregularly, but this is misleading. The root form of the Latin word is in fact *tempestat-*, with endings added according to case and number. It is only in the Latin nominative singular (which has a null ending) that the *'t'* of the root is realized as *'s'*. The third *'t'* of *tempestat-* is root-final, but word-medial, and in intervocalic contexts the development of *t* > *d* is completely regular (Boyd-Bowman, 1954).

2.4.2. Code 1: Paradigm Leveling

Leveling is a subtype of analogy that occurs at the level of the paradigm. It is an irregular process that serves to reduce morphophonemic alternations within paradigms, resulting in a more regular paradigm (Hock, 1991, p. 168). Code (1) differs from code (0) in that an irregular form of change has occurred, rather than purely regular exceptionless sound change. The source of this change is within the paradigm of the lexical item, rather than from another source in the language, and in this way it is distinguished from code (2), general analogy.

Cases where part of the original Latin ending has been incorporated into the stem of the Spanish word are also coded as (1). The word for ‘brain’ is included in this category (example 10).

10. ‘brain’ LA: *cerebrum* SP: *cerebro*

The Latin paradigm begins *cerebrum (nom. sg.), cerebri (gen. sg.)* demonstrating that the stem of the noun is *cerebr-*. The Spanish paradigm is *cerebro* in the singular, and *cerebros* in the plural, demonstrating that the stem of the noun is *cerebro-*. Here, the final ‘o’ of the Spanish stem is derived from one of the Latin endings, indicating that the Spanish word is not derived from the Latin stem, but a specific Latin form within the paradigm (likely *cerebrum*). This is a case of paradigm leveling, coded as (1). Given that Latin noun roots are most commonly consonant-final, while Spanish noun roots are most commonly vowel-final, a high proportion of the nouns on the list fall into this category. Paradigm-internal change does not necessarily result in completely regular paradigms in which all the forms have a predictable relationship to one another.

2.4.3. Code 2: General analogy

Code (2) is labelled as “four-part analogy” in the original presentation of the coding scheme given by Pappas and Mooers (2011). While four-part analogy is included in code (2) here, it does not fully describe the category. Code (2) describes irregular changes that are cognitive, rather than mechanical, in nature. A cognitive change requires the addition of an unpredictable cognitive association on the part of speakers that is not limited to the specific phonemes found in the word.

An example of general analogy is the development of the word for ‘birch’. In Latin this word is *betula*, while in Spanish ‘birch’ is *abedul*, example 11.

11. ‘birch’ LA: *betula* SP: *abedul*

12. ‘fir’ LA: *abies* SP: *abeto*

The appearance of the word-initial '*a'* in 'birch' is not based on the sequence of sounds in *betula*, and other Latin words beginning with '*b'* do not show this same development. Instead, the source of the '*a'* is the word *abeto*, meaning 'fir' (example 12). The word for 'birch' is altered to be more like the word for 'fir', with which it already shares a number of sounds. The analogy is not based on phonetic similarity, however, but most likely on the fact that both words describe a type of tree. But the relationship between these two words is based on speakers' association of different species of tree with one another, an association that cannot be found directly in the forms of the words in question. Similarly, the Latin *noverca*, meaning 'stepmother' (example 13), developed into the Spanish *madrastra* based on analogy with *padrastro*, 'stepfather' (example 14) (which itself is derived more regularly from the Latin *patraster*); the relationship between stepmothers and stepfathers is based on non-linguistic associations, rather than any mechanical relationship between *noverca* and *patraster*.

13. 'stepmother' LA: *noverca* SP: *madrastra*

14. 'stepfather' LA: *patraster* SP: *padrastro*

Another example is that of the word 'prison'. The Latin word is *carcer*, while the Spanish for the same meaning is *cárcel* (see example 15). Here, the second of two 'r's has changed to 'l' in a form of dissimilation. This process is common, but sporadic. When two 'r's are found in a word, it is often the case that one will disappear (as in 'plough' (example 16), in Latin *aratro*, but in Spanish *arado* (Boyd-Bowman, 1954, p. 112)) or change to 'l' (as is the case for 'prison').

15. 'prison' LA: *carcer* SP: *cárcel*

16. 'plough' LA: *aratro* SP: *arado*

Which change will occur, however, and which 'r' will change is unpredictable. Other sporadic changes include metathesis (reversal) of 'l' and 'r' in a word, or in some cases a single 'r' changing to 'l'. This is therefore not a case of a regular, exceptionless sound change. Dissimilation in general is not a case of regular sound change (Hock,

1991). This is a cognitive, rather than a mechanical change. This example and other similarly sporadic examples are coded as (2).

2.4.4. Code 3: Syntactic reanalysis

Syntactic reanalysis is coded as (3) according to this scheme. This does not refer to historical change in syntactic rules, but rather reinterpretation of a phrase or utterance to include different elements in the final lexical item (Pappas & Mooers, 2011). This code most often reflects one of two developments: the inclusion of new material, or the transfer of meaning to a different part of the phrase. An example of the former is the meaning 'inside, in': in Latin *intra*, but in Spanish *dentro* (see example 17).

17. 'inside, in' LA: *intra* SP: *dentro*

The Spanish term is not derived solely from the Latin term, but rather the larger Latin phrase *de intro*, meaning "from inside". *de*, originally a preposition meaning 'from', has been reinterpreted as part of the following adverb *intro* meaning 'inside'. The boundaries of the lexical item meaning 'inside' have therefore been reanalysed.

Other examples of syntactic reanalysis are wholesale transfers of meaning from one part of a phrase to another. The meaning 'zero, nothing', is expressed in Latin by *nihil*, and in Spanish by *nada* (example 18). The Spanish term is not derived from *nihil*, but instead from the phrase *nulla res nada*, literally translated as "no thing born". This phrase is an emphatic negative, with the negation originating in the word *nulla*, meaning 'no, none', but the phrase has been reanalysed so that instead the negation is ascribed to the word *nada*, which is a Latin word meaning 'born'.

18. 'zero, nothing' LA: *nihil* SP: *nada*

2.4.5. Code 4: Semantic change

Semantic change describes cases where the word in the modern language is derived from a word of another meaning, but which nevertheless was a part of the ancestral language in some form (Hock, 1991, p. 296). The source for this type of

change is therefore language-internal, rather than language-external. For example, the Latin word for ‘week’ is *hebdomas* (itself a Latin borrowing from Greek); however, the Spanish word for the same meaning, *semana*, is not derived from *hebdomas*, as shown in example 19.

19. ‘week’ LA: *hebdomas* SP: *semana*

Instead, the modern Spanish word is derived from the Latin word *septem*, meaning ‘seven’. The semantic association between the concepts is in this case fairly transparent: there are seven days in a week. What is important in coding this example as a (4) is that the Spanish word is derived from a Latin word of a different meaning, rather than that of another language.

One of the most common types of semantic change found in the data is generalization. In generalization, a very specific term in Latin changes to encompass a much broader meaning in Spanish. An example of generalization is the Spanish word for ‘to blow’, *soplar* (example 20). The Latin word for ‘to blow’ is *flare*, and the Spanish is not derived from this, but from the more specific verb *sufflare*, which means ‘to blow up, inflate’. The meaning of the specific term has broadened to supplant the original general term. Other word meaning pairs coded as (4) include cases where the Spanish word in question has been developed from the diminutive form of the Latin term. See example 21: the word for ‘fortress’ is *castrum* in Latin (literally ‘fortified camp’); in Spanish, this meaning is described by *castillo*. The Spanish form is derived not from the base Latin *castrum*, but from *castellum*, the diminutive form, meaning something like ‘little fortified camp’ (De Silva, 1985). This is a specific case of generalization.

20. ‘to blow’ LA: *flare* SP: *soplar*

21. ‘fortress’ LA: *castrum* SP: *castillo*

Small-scale changes in word meaning and connotations, however, are not considered instances of semantic change. Particularly with words that display great polysemy, the likelihood of there being words in both Latin and Spanish that describe exactly the same senses of meaning is low. The cut off point in this coding scheme for

whether a difference in the senses of two words was significant enough to be considered a “semantic change” is the following question: “Could a speaker of Latin and a speaker of Spanish both use these words with the intended meaning described in the IDS wordlist?” If not, semantic change has occurred, and a code of (4) is appropriate. Take as an example, the word ‘court’, as in a court of law. The Spanish term for this meaning is *tribunal,* which is derived from the Latin word *tribunal* (example 22).

22. ‘court’ LA: *tribunal* SP: *tribunal*

This Latin term primarily refers to the platform where magistrates sat, but has ‘court’ as a secondary, less literal meaning. The Latin word here has a meaning that is not included in the Spanish word’s meaning, but speakers of both languages would be able to use the word *tribunal* to refer to a court of law. This is therefore *not* a case of semantic change (as it exhibits no other form of change either, this particular item is coded (0)).

Words of onomatopoeic origin were also coded as a form of semantic change. Although they are not derived from a particular word in the older language, they are words that originate language-internally and take on new meaning. It is therefore not appropriate to describe them as cases of borrowing, since they do not come from an external linguistic source. An example of onomatopoeic origin is the Spanish word for ‘dog’, *perro*. This term is hypothesized to come from the noises shepherds make when they are encouraging sheep dogs (De Silva, 1985, p. 411).

23. ‘dog’ LA: *canis* SP: *perro*

2.4.6. Code 5: Lexical Borrowing

The term lexical borrowing describes the adoption into one language of a lexical item from another language (Hock, 1991, p. 380). Borrowing may occur as a result of need, i.e. when a language does not have a word for a particular concept or item it may borrow a word from another language, or it may replace an existing word in the language. For the purposes of this study, meanings on the IDS list were dropped from consideration if they could not be expressed using a single word in both Latin and

Spanish. Many terms for plants and animals native to the Americas or Oceania, such as ‘potato’ and ‘opossum’, do not have a word in Latin. Due to this filtering, borrowing resulting from need is not expected to occur very often, if at all. Lexical borrowing herein, therefore, primarily describes the replacement of a native Latin-derived term with a word from another language, in which it may or may not have the same meaning.

Spanish includes borrowed words from a wide variety of other languages. Due to historical influences, a large number of Arabic borrowings can be found. For example, the word *almohada*, meaning ‘pillow’; in Latin the word for this meaning is *cervical*, but the Spanish term is derived from the Arabic phrase *al-mukhadda*, meaning ‘the pillow’ (example 24). Other borrowings come from Germanic, including *blanco* ‘white’, which is derived from the Germanic word *blank-* (meaning ‘to shine’) rather than the Latin word for ‘white’: *albus* (example 25).

24. ‘pillow’ LA: *cervical* SP: *almohada*

25. ‘white’ LA: *albus* SP: *blanco*

A change is still coded as borrowing even if the language that the word is borrowed from is an Italic language, such as French or Italian. In such cases, it may be that the borrowed word is a cognate of the word being replaced. These cognates are still language-external in their source, and speakers are not expected to treat languages differently based on their opaque historical relationship to their own language. While these borrowings may appear superficially similar to the original Latin term, the influence of their intermediate Italic language makes it possible to identify them. For example, the word for ‘chimney’ is *caminus* in Latin, and *chimenea* in Spanish (example 26). The Spanish word is derived from the Old French *cheminee* (also the source of the modern French *cheminée*) which has the same meaning; this Old French term is itself derived from the Latin *caminus*. This is still a borrowing, however, due to the language-external source for the modern Spanish word. In this case, the development of Latin word initial [k] to [tʃ] is the relic of a French sound change.

26. ‘chimney’ LA: *caminus* SP: *chimenea*

Another group of words coded as (5) are Latinisms and semi-learned words. These are terms borrowed directly from Latin during a later stage of Spanish. This phenomenon is due to the fact that even after the appearance of its daughter Italic languages in everyday speech, Latin continued to be learned and used for specific purposes. The use of Latin in some church services continues to this day. These forms of the language are known as Late Latin and Church Latin, respectively. Even though Spanish is the daughter language of Latin, modern Spanish and any currently extant forms of Latin are now separate languages, and so words from Late or Church Latin which are borrowed into Spanish are classified as external linguistic influence, and coded as (5). Semi-learned words were borrowed into Spanish at an earlier stage than Latinisms, but they have in common that the words have not participated in some or all of the historical sound changes affecting the development of Latin to Spanish. This absence of expected changes is often what makes it possible to identify them as later borrowings.

An example of an unambiguous Latinism is the word for ‘to boast’. In Latin this verb is *iactare*, and in Spanish the same meaning is expressed by the word *jactarse*, which can be seen in example 27.

27. ‘to boast’ LA: *iactare* SP: *jactarse*

The words are overwhelmingly similar to one another, and express the same meaning, but rather than being coded as (0) (sound change), they are instead instances of (5) (borrowing). The key feature is the consonant cluster shared by both terms: ‘ct’ [kt]. The development of ‘ct’ from Latin to Spanish does not result in ‘ct’. Rather, regular sound changes derive Spanish ‘ch’ [tʃ] from Latin ‘ct’ [kt]. This development is well established, and supported by a large number of examples, including: ‘milk’ *lacte* > *leche*, ‘eight’ *octo* > *ocho*, and ‘night’ *nocte* > *noche* (examples 28-30), among others. Given that this very regular sound change is not apparent in ‘to boast’, it can be established that the word was not a part of the Spanish language at the time that this sound change took place, and was instead borrowed directly from Latin at a later date. Other Latinisms also display this pattern (Boyd-Bowman, 1954, p. 40).

28. ‘milk’ LA: *lacte* SP: *leche*

29. ‘eight’ LA: *octo* SP: *ocho*

30. ‘night’ LA: *nocte* SP: *noche*

The following chapters will detail the analysis of the data derived using the above methodology. Chapter 3 will investigate the question of the stability of frequency of use over time by comparing the frequency of use values derived from the PDLP and the CdE with each other, as well as with the results of prior research. Chapter 4 will use the results of chapter 3’s investigation to inform an analysis of the relationship between frequency of use (as derived from the two corpora) and change over time (as coded according to the above six-point scheme).

Chapter 3.

Stability of Frequency of use from Latin to Spanish

3.1. Testing Pagel et al.'s Results

Pagel, Atkinson and Meade (2007) find a high degree of correlation between the frequency of use values for the 200 Swadesh list words across all pairs from four Indo-European languages (English, Spanish, Greek, and Russian). Given the well-established historical relationship between these languages, Pagel et al. (2007) assert that this correlation in frequency of use must be due to a high degree of consistency in frequency over time. The authors hypothesize that the similar frequencies of Spanish and English, for example, are both inherited from their shared linguistic ancestor, proto-Indo-European. The claim is that the similarities between modern languages in frequency of use are not due to convergence, but are instead historical legacies. If this is indeed the case, and differences in use accrue over time, then we would predict that the strength of correlation between Spanish and its recent ancestor Latin should be greater than the strength of correlation between Spanish and the modern languages with which it shares more distant ancestry.

The use of a Latin corpus in the current research allows the estimation of frequency of use for words in that language. Using these estimates, the Pagel et al. (2007) tests of correlation in the frequency of use across languages can be replicated by comparing Latin with Spanish. Due to the nature of the Latin corpus, as well as the typological characteristics of the language, the frequency of use measurement is slightly different from that used for the Spanish corpus (see section 2.3.1). In order to remove this as a potential source of error, Pagel et al.'s original tests, as well as the current comparisons, will be conducted using non-parametric statistics. The use of Spearman's ρ, rather than Pearson's r, means that small systematic differences in the absolute

frequency values caused by the differences in measurements will be ignored, and only the rank order of the frequency items will be considered (Schwarz, 2014c).

Although the current word list (see section 2.2) is much larger than the one used by Pagel et al., in section 3.1 I first consider the same word list as Pagel et al. Of the 200 words used by Pagel et al., 27 are not in the IDS list, leaving 173 words as the base dataset. Pagel et al.'s (2007) original frequency of use data is available in their supplementary material, and numbers from this source are used when indicated. In section 3.2 I return to the full word list.

3.1.1. Reinterpreting Pagel et al.'s frequency of use results

The original frequency-of-use comparisons conducted by Pagel et al. (2007) made use of parametric statistics (Pearson's *r*), and the full 200 words available on the Swadesh list. The reported correlations ranged from an *r* of 0.78 (Greek vs. Russian) to an *r* of 0.89 (English vs. Spanish) (Pagel, Atkinson, & Meade, 2007, Supplementary information table 2).

Table 3.1 reports the non-parametric correlations among the 4 modern languages using the 173 word subset of the Swadesh list. The non-parametric results remained similar to those reported by Pagel et al., with the lowest value of ρ being 0.76 (Greek vs. Russian), and the highest value being $\rho = 0.88$ (English vs. Spanish).

Table 3.1. Pagel et al. frequency comparison of ρ and r

Language Comparisons	Pearson's *r* (n = 200)	Spearman's ρ (n = 173)
English - Spanish	0.89	0.88
Russian - Spanish	0.84	0.82
Greek - Spanish	0.87	0.86
English - Russian	0.87	0.82
Greek - Russian	0.78	0.76
English - Greek	0.85	0.84

The use of weighted, rather than maximum frequency as the measurement for the Latin corpus motivates the use of non-parametric statistics in this research (Crane, 2014). The replication of Pagel et al.'s (2007) results using non-parametric statistics and

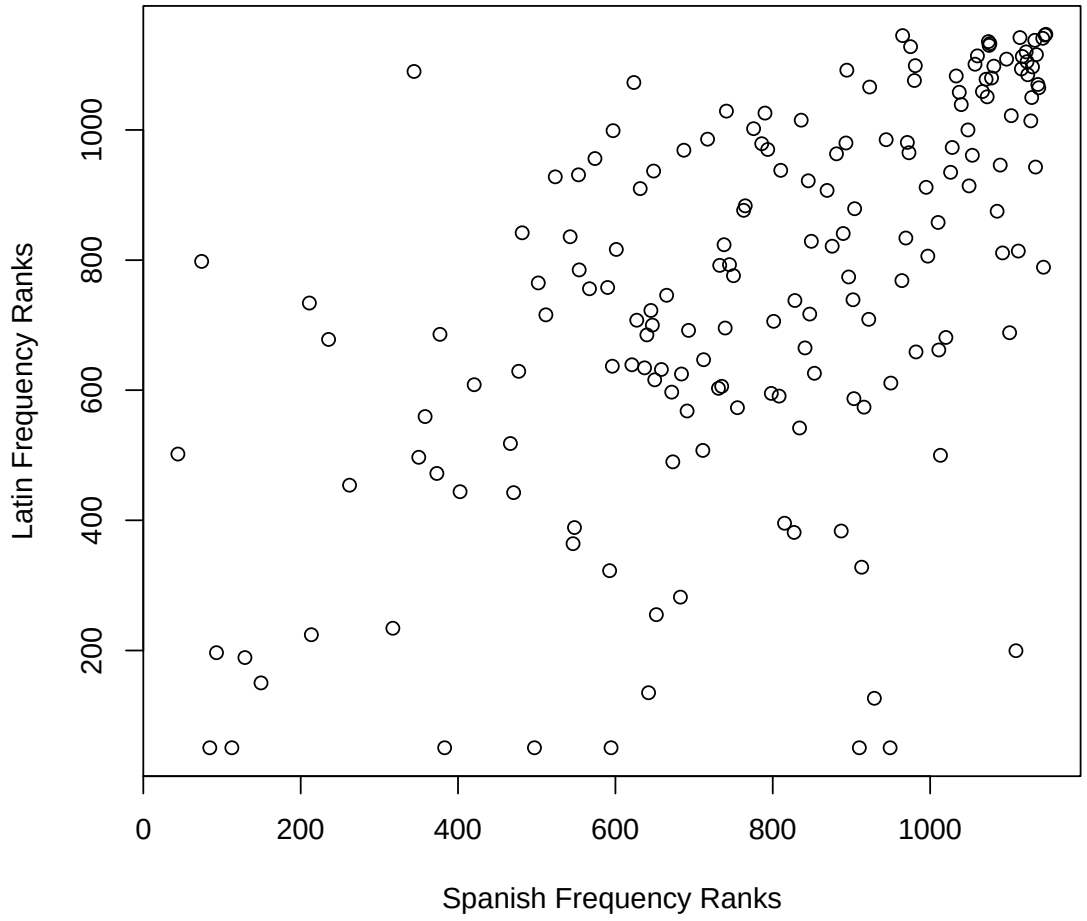
a slightly smaller wordlist, as seen in table 3.1, indicates that using non-parametric statistics does not greatly distort these results. Further comparisons between the current research and Pagel et al.'s correlations will therefore make use of the replicated non-parametric results for ease of interpretation.

3.1.2. Swadesh list comparison of Latin and Spanish

If, as Pagel et al. (2007) claim, the strong degree of correlation in frequency of use among the four modern languages is the result of the frequency having been inherited from proto-Indo-European, then we would expect that the correlation within these language lineages should be similar. Therefore, if the high correlation between modern languages is the result of frequency of use being stable over time, then Spanish should have an equally strong (or perhaps even stronger) correlation in frequency of use with its more recent ancestor Latin.

Using the same 173 words as in the previous tests, a test of the correlation between Latin frequency of use (as derived from the Perseus Digital Library Project) and Spanish frequency of use (as derived for this research from the Corpus del Español) resulted in $\rho = 0.60$. This correlation is considerably weaker than even the lowest modern correlation (Greek vs. Russian $\rho = 0.76$). At first glance, this result is not consistent with Pagel et al.'s (2007) claim about the stability of frequency of use over time. A plot of the ranks of Latin and Spanish can be seen in Figure 3.1.

Figure 3.1. Latin FoU vs. Spanish FoU, n = 173



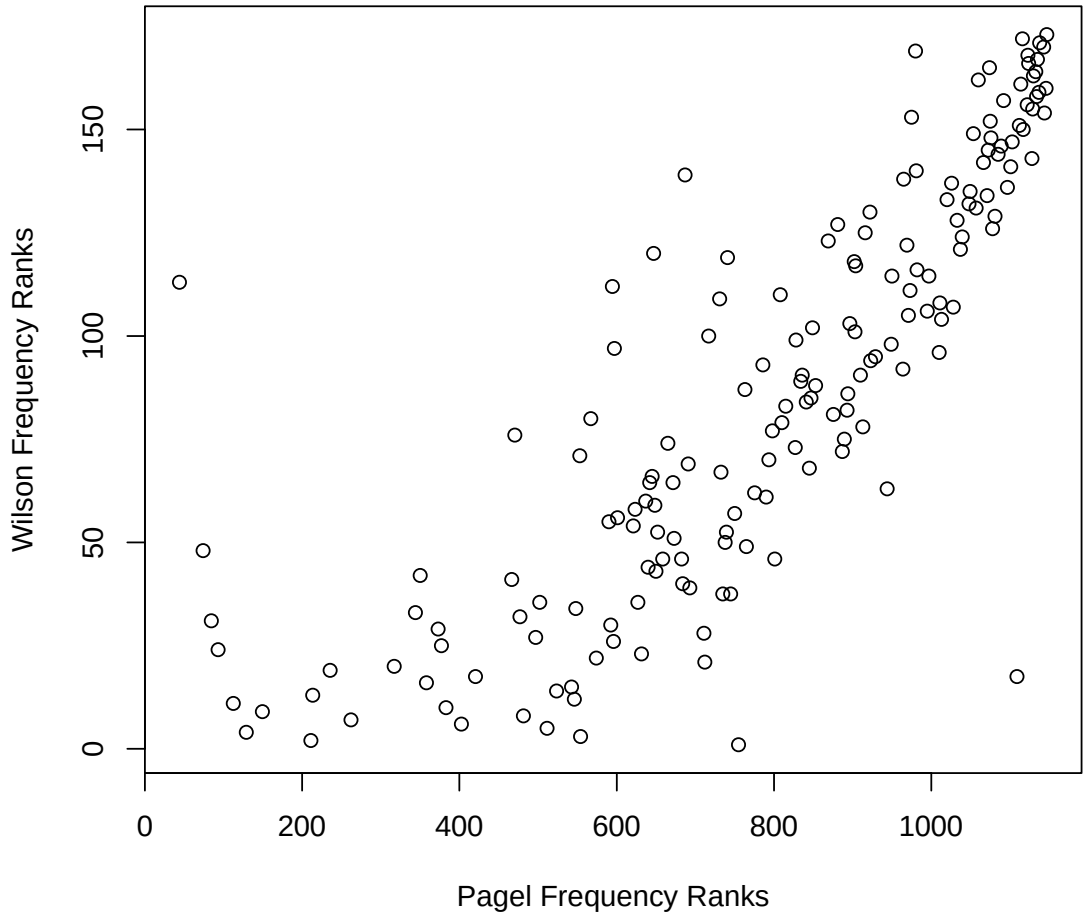
3.1.3. Comparing two sets of frequencies from Modern Spanish

To establish whether the source of the low correlation between Latin and Spanish (relative to the correlation between modern languages as reported by Pagel et al.) might be from the different dataset used[6], I compared the 173 Spanish word frequencies used by Pagel et al. (2007) (a subset of the originally used 200 words) with the 173 Spanish word frequencies collected for this research. This resulted in a ρ of 0.87. Both Pagel et al. (2007) and the current research made use of the Corpus del Español, and so this is surprisingly low. The difference between the datasets may be

[6] i.e., Spanish frequency of use values for this research were taken directly from the Corpus del Español, rather than from Pagel et al.'s supplementary information, unless otherwise stated.

due to a somewhat different subset of the corpus being used (Pagel et al. do not specify whether the oral section of the corpus was included nor whether any historical texts were included), or somewhat different search terms being used (again, these terms are not specified in the Pagel et al. paper). Changes made to the Spanish lemmatization for the current research (as described in section 2.3.1) in order to maintain consistency with the Latin corpus may also have reduced the correlation. A plot of the ranks of the two Spanish datasets can be seen below in Figure 3.2.[7]

Figure 3.2. Wilson Spanish FoU vs. Pagel Spanish FoU, n = 173



[7] Two fairly dramatic outliers are apparent in figure 3-2; these reflect differences in the lemmatization of Spanish pronouns, which the current research grouped in such a way as to be directly comparable with the Latin corpus (a concern which Pagel et al. (2007) did not have).

In addition to the direct comparison of the two Spanish frequency datasets, for the sake of completeness, my Latin FoU values were compared to the Spanish FoU values reported by Pagel et al. (2007). This resulted in a correlation of $\rho = 0.62$. This is only marginally stronger than the correlation between Latin FoU and my Spanish FoU values reported above ($\rho = 0.60$).

3.1.4. Genre comparison

Pagel et al.'s (2007) interpretation of the high correlation in FoU values between modern Indo-European languages ($0.76 < \rho < 0.88$) is that the FoU values of these related languages is inherited from their shared ancestral language (in this case: proto-Indo-European). If this interpretation is correct, then the correlation between a modern language, Spanish, and its close ancestral language, Classical Latin, should be at least as strong. It does not seem to be.

However, before interpreting this weak correlation in the light of inheritance of FoU vs., for example, convergence, we must consider confounding variables. One major potential confound is the make-up of the corpora themselves, specifically genre. Due to the small amount of written material available in Classical Latin, it was not possible to control the Latin corpus for genre. If it is the case that the modern corpora used by Pagel et al. (2007) *were* balanced for genre (such that the corpora were similar to each other in this respect), then differences in the genres of the Classical Latin and Spanish corpora might result in the observed weaker correlation (Biber, 1993).

In order to explore this possibility, the FoU values for the 173 words were measured for the three individual subcorpora in the CdE for modern Spanish. This corpus is subdivided into Academic Writing, Fiction, and News; each of the three subcorpora are approximately 5 million words in size. If the correlations in FoU among these three subcorpora are strong, this would support the assertion that the differences between the Classical Latin and Spanish corpora cannot be due solely to differences in genre. Table 3.2 summarizes the results of comparing the three Spanish genre-based subcorpora to each other.

Table 3.2. Genre comparison, non-parametric correlations, n = 173

Comparison	Spearman's ρ, n = 173
Spanish Fiction - Spanish Academic	0.67
Spanish News - Spanish Academic	0.82
Spanish News - Spanish Fiction	0.88
Latin - Spanish Academic	0.54
Latin - Spanish Fiction	0.58
Latin - Spanish News	0.63

The correlations between News and both Academic Writing and Fiction are within the ranges of values found for the modern languages in Pagel et al. (2007) ($0.82 < \rho < 0.88$). The correlation between Academic Writing and Fiction, however, is $\rho = 0.67$. The highest correlation between Latin FoU and modern Spanish FoU (using Pagel et al.'s (2007) data) is $\rho = 0.62$. Given the similar correlation strengths between these two comparisons, we cannot exclude the possibility that the relatively weak correlation between Classical Latin and Spanish is due to the differences in genre between the materials of the two corpora.

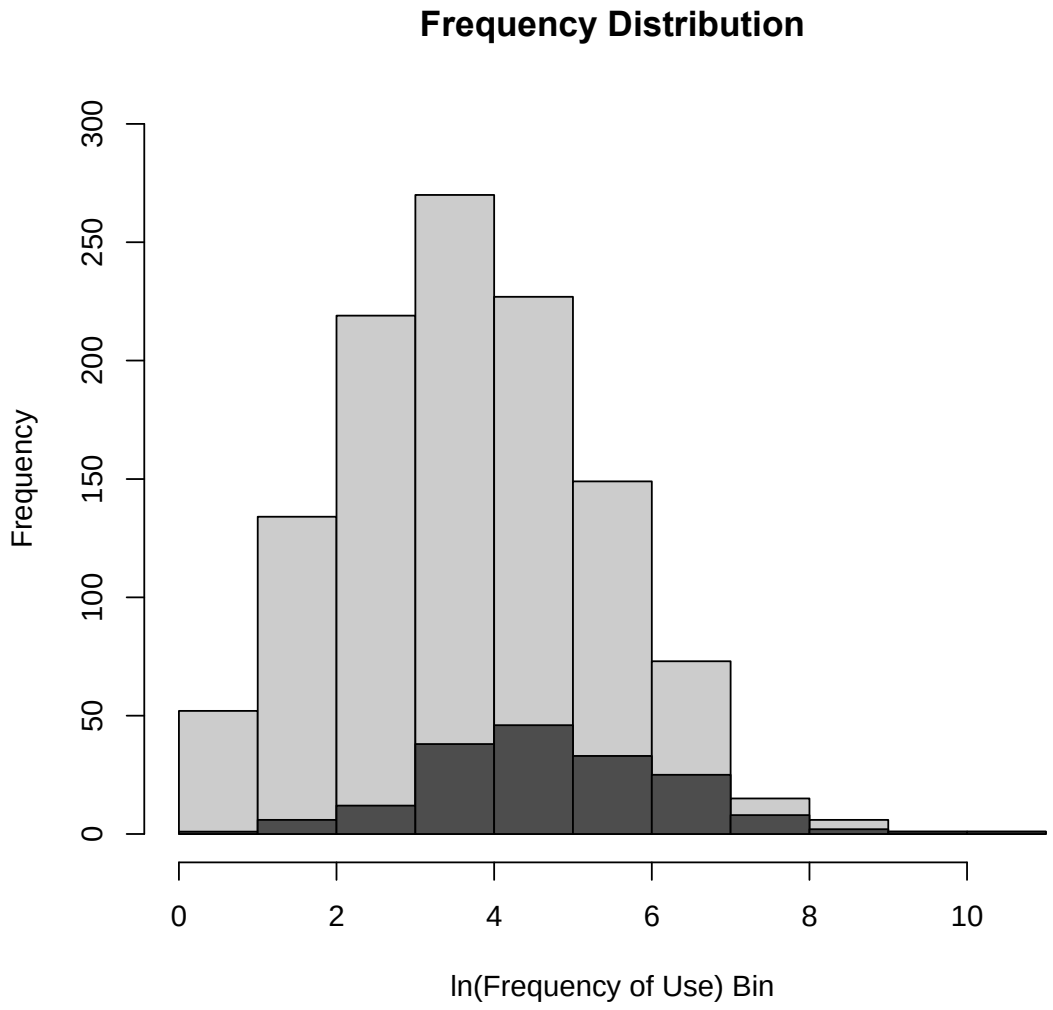
In order to determine whether the Latin corpus might be closer in frequency values to any one of the three available Spanish genres, I performed correlations between the three Spanish subcorpora and the Latin corpus (table 3.2). These comparisons range from ($0.54 < \rho < 0.63$), all lower than the Spanish cross-genre correlations and lower than or comparable to the overall Latin and Spanish correlations. From these results, it does not appear to be possible to match the CdE with the PDLP in such a way as to eliminate the possible confounding influence of genre differences.

Based on these results, it appears to be possible that the lower correlation observed between Latin and Spanish (relative to the correlations observed between modern languages in Pagel et al. (2007)) may be due to differences in genre between the modern and historical corpora used. Due to this confound, the analysis of these data does not present a direct challenge to the assumption of Pagel et al. (2007).

3.2. Full list results

Comparisons in the previous section made use of only the smaller number of words shared by both the IDS list and the Swadesh list. The following section will make use of the full IDS list to determine the level of correlation between Spanish and Latin for the larger set of words. The IDS list is large enough that it becomes feasible to introduce additional variables into the model, such as part of speech and semantic category, to determine whether these characteristics improve the explanatory power of the model predicting Latin frequency of use with Spanish frequency of use (Schwarz, 2014b). Figure 3.3 shows the distribution of the natural log-transformed frequency of use for the overall Spanish corpus for both the IDS and Swadesh lists (Schwarz, 2014a).

Figure 3.3. Histogram comparing Spanish FoU for the Swadesh and IDS lists



Even though the distribution is non-normal (most likely due to the presence of some extreme high outliers), the unimodal distribution gives us some confidence in our ability to run statistical analyses on it (Schwarz, 2014b). The primary apparent difference between the full list and the Swadesh subset is that the Swadesh list appears to skew slightly towards higher frequency words than does the IDS list; given that the Swadesh list is intentionally biased towards common words that are likely to be present in any language, and the overwhelming majority of words in a language are low frequency, this decrease in average frequency for a larger list is to be expected (Sichel, 1975).

3.2.1. Full list comparison of Latin and Spanish

A general linear model was run on the rank frequency data, using modern Spanish FoU to predict the Latin FoU for the full list of 1147 words. This resulted in a value of $\rho = 0.55$ ($p < 0.0001$)[8]. This value is comparable with the Swadesh list correlation ($\rho = 0.60$). Figure 3-4 shows a plot of the Latin ranks against the Spanish ranks for the full list. From a visual observation of the plot, there appears to be a concentration of strongly correlated high rank values. If it is the case that the ranks of high frequency words are more strongly correlated, then the small bias in favour of more frequent words observed in the Swadesh list might account for the stronger correlation reported in 3.1.3 for the Swadesh list words relative to the full list.

There are a relatively high number of zero values within the IDS dataset (a total of 107 out of 1147 total words between the two languages). The majority of these zeroes were Latin FoU values, possibly due to the relatively small size of the Latin corpus relative to the Spanish corpus. In order to ensure that these zeroes were not causing problems for the analysis, the full list results were replicated without the zeroes included.[9] A general linear model predicting Latin FoU using Spanish FoU for the 1040

[8] The general linear model run presents the results in the form of $r^2 = 0.30$. From this we can derive $r = 0.55$. Since the model was run on the rank data, rather than the raw data, Pearson's r is here equivalent to Spearman's ρ. Results from the linear models hereafter will therefore be reported as ρ for ease of interpretation.

[9] The use of zero-inflated statistics to account for the high number of zeroes was also attempted, but was abandoned as the data does not conform to any of the appropriate statistical distributions (Loeys, Moerkerke, De Smet, & Buysse, 2012).

word list (exluding zeros) gave $\rho = 0.57$ ($p < 0.0001$); the zero values do not seem to be exerting undue influence on the patterns reported here.

3.2.2. Part of speech comparison

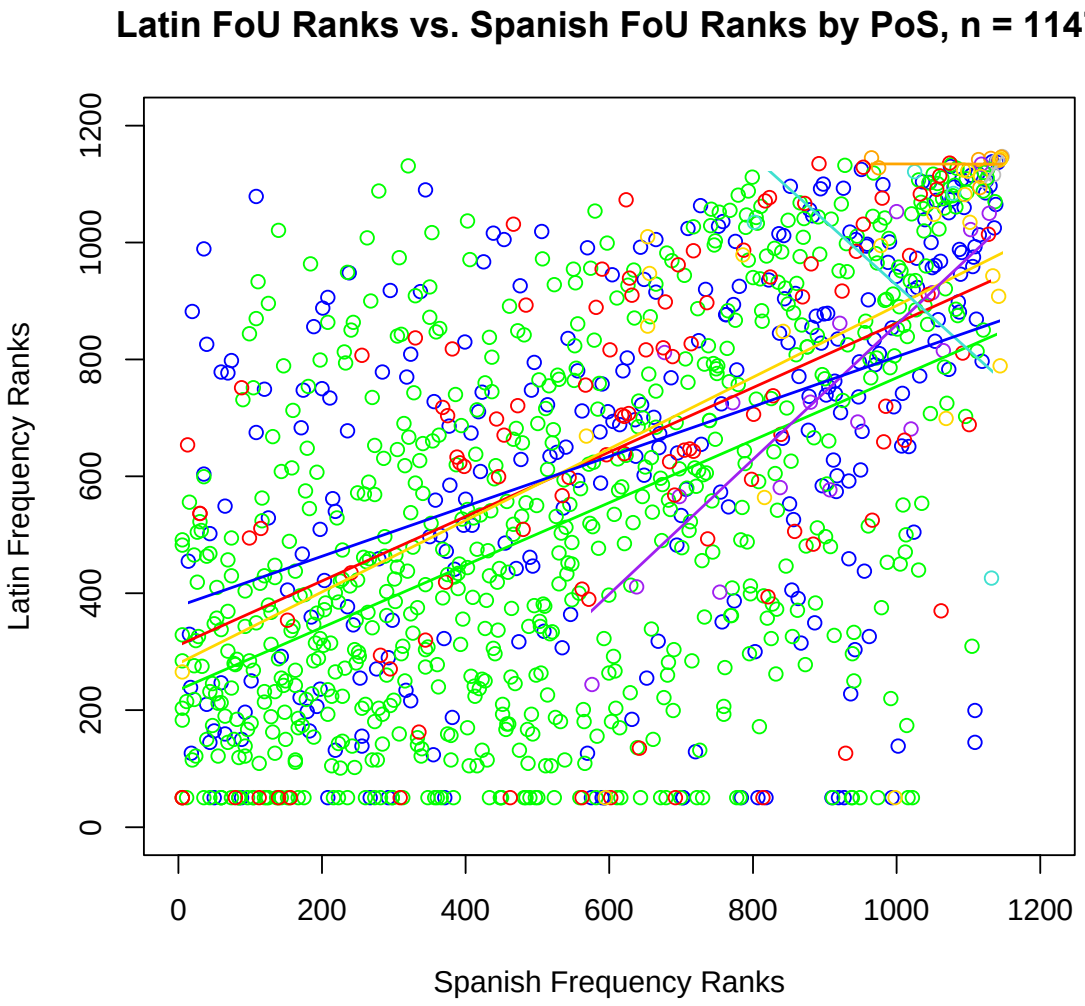
A linear model predicting Latin FoU using Spanish FoU and including part of speech as a covariate was also run on the Latin and Spanish frequencies. This was to determine whether the addition of the variable part of speech would increase the explanatory power of the model, accounting for additional variation in frequency of use (Schwarz, 2014d). Eight part of speech categories were used: adjectives, adverbs, conjunctions, nouns, numbers, prepositions, pronouns, and verbs. The initial test showed no significant interactions. The model without interactions found a significant negative main effect of part of speech, with lower FoU for nouns in Latin than in Spanish ($p < .002$), and a higher FoU of pronouns in Latin than in Spanish ($p < .03$). The model predicting Latin FoU rank using both Spanish FoU rank and part of speech resulted in an ρ of 0.57. Given that ρ for the model without part of speech is 0.55, this variable does not appear to add explanatory power to the model, i.e. the relationship between Latin FoU and Spanish FoU does not vary by part of speech.

Figure 3.4 shows the correlation between Latin FoU ranks and Spanish FoU ranks, with the partial slopes for parts of speech colour-coded for illustrative purposes. Several of the categories exhibit very different slopes from the general trend, particularly prepositions (turquoise), which have a negative slope, and pronouns (orange), where there appears to be no correlation at all. In the linear model, however, there are no significant interactions, and so none of these visual differences in slopes are significant. Both of these unusual categories are represented by a very small number of words in the dataset, however (7 pronouns and 4 prepositions), and so the non-significant different slopes may be due to small sample size.

The functional part of speech categories included are prepositions, pronouns, numbers, and conjunctions. Each of these categories is represented by fewer than 20 words in the full IDS list, and all four appear to be constrained towards the more frequent end of the range for both languages, while the lexical part of speech categories are

spread across the entirety of the range. The category of conjunctions, in particular, includes only extremely high-frequency words.

Figure 3.4. **Latin FoU ranks and Spanish FoU ranks with Part of Speech lines. Adjectives (red), adverbs (yellow), conjunctions (gray), nouns (green), numbers (purple), prepositions (turquoise), pronouns (orange), verbs (blue).**



3.2.3. Semantic category comparison

The IDS list of word meanings is coded according to semantic category; there are 22 of these semantic categories in total. This number of levels for the variable of semantic category is too large to be appropriate for a dataset of only 1147 items. In order to facilitate statistical analyses for this research, therefore, these 22 categories

were further grouped into five sets. The grouping of the categories was checked by a semanticist (N. Hedberg, personal communication, October 8, 2014). The original IDS semantic categories and their groupings can be seen in table 3.3.[10]

Table 3.3. Grouping of the 22 IDS Semantic categories.

Group	Name	IDS Categories	Number of words
A	Society and Governance	Law Religion & belief Social & political relations Kinship	127
B	Hunting or "male" sphere	The physical world Animals Warfare & hunting The body	317
C	Abstract relations in the physical world	Spatial relations Quantity Time Sense perception Motion	253
D	Domestic or "female" sphere	Food & drink Agriculture & vegetation Basic actions & technology The house Possession Clothing & grooming	317
E	Mental domain	Cognition Emotions & values Speech & language	133

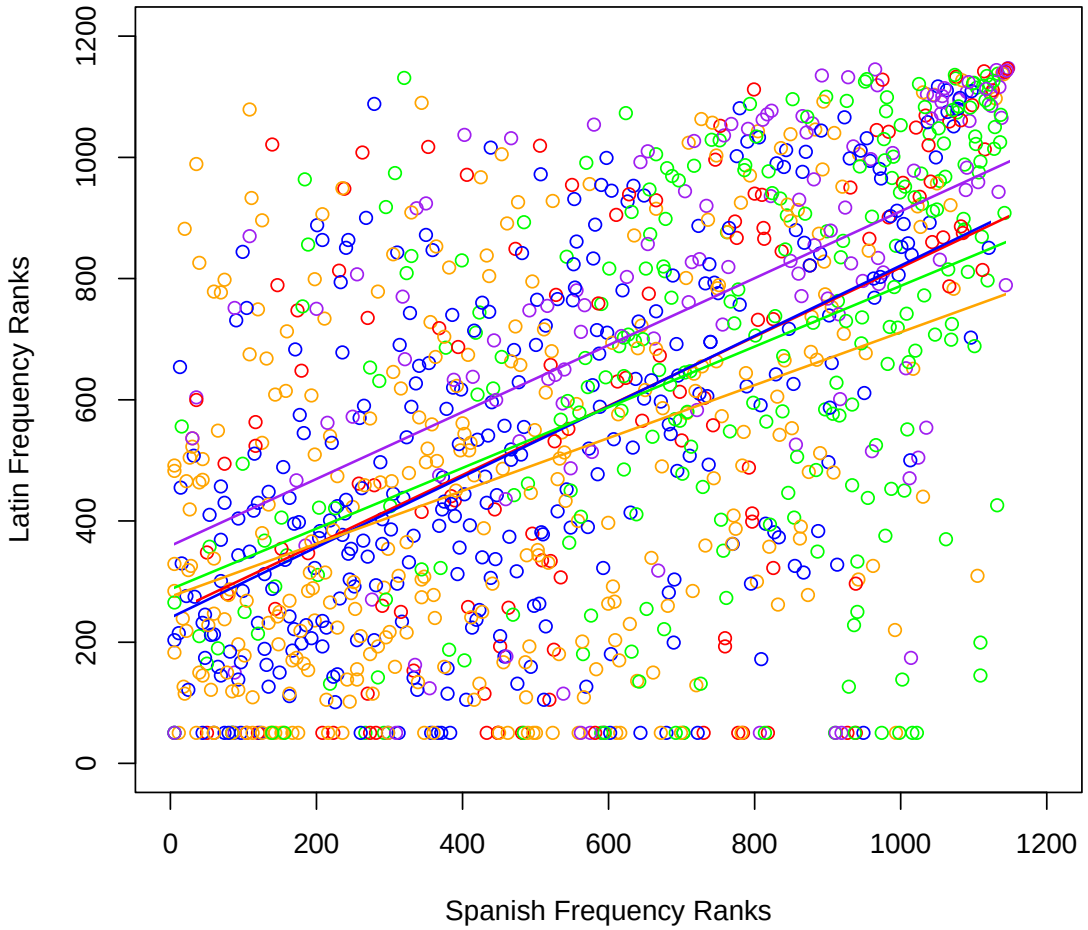
A linear model predicting Latin FoU ranks using Spanish FoU ranks with Semantic category group as an additional variable resulted in $\rho^2 = 0.32$ ($\rho = 0.57$). Compared to $\rho^2 = 0.30$ ($\rho = 0.55$), for a model without Semantic category group, this variable also does not appear to contribute a great deal of explanatory power to the model. As with part of speech, there are no interaction effects, indicating that the relationship between Latin frequency of use and Spanish frequency of use is not

[10] These groupings are intended as a preliminary investigation and have not been independently tested.

significantly different depending on Semantic category group. The model without interaction did return a significant positive main effect of Semantic Category E (Mental Domain) ($p < .002$); this category includes words related to cognition, emotions, values, and language, and these abstract terms are significantly more frequent in Latin than in Spanish.

All of the groups are represented by more than 100 words, and all five appear to be spread across the entirety of the range of possible ranks. Unlike for part of speech, the slopes representing each group appear to be roughly parallel, which is consistent with the lack of significant interaction effects in the model.

Figure 3.5. Latin FoU ranks and Spanish FoU ranks with Semantic Category Group lines. Society and governance (red), Hunting or "male" sphere (blue), Abstract relations and the physical world (green), Domestic or "female" sphere (orange), Mental domain (purple).



3.3. Discussion

In this chapter, frequency of use was evaluated first using a list of 173 terms shared between the Swadesh list and the IDS list. Frequency was then evaluated using the full 1147-word IDS list. The investigation of the smaller list was focused on testing the relationship between Latin and Spanish frequencies of use in light of the previous research conducted by Pagel et al. (2007). These earlier results were redone using non-parametric statistics with their original frequency of use data for four modern languages (Spanish, English, Greek, and Russian). The correlation between Latin and Spanish was tested for the smaller list, and found to be considerably weaker than the correlations between the modern languages. Comparisons were run between the genre-controlled subcorpora of the CdE, to determine whether the weaker correlation between Latin and Spanish might be due to differences in genre between the Latin and Spanish corpora. The lowest between-genre correlation was found to be comparable to the Latin and Spanish correlation. Based on these tests, we cannot reject the possibility that the lower correlation between Spanish and Latin is due to genre differences alone. There is therefore no definitive evidence that frequency of use is not stable over time. Direct comparison of the Spanish genre-controlled subcorpora with the Latin corpus did not result in stronger correlations, and so it does not appear to be possible to control for the genre of the Latin corpus using these data.

Investigation of the relationship between Latin and Spanish frequency of use for the full IDS list found the relationship to be somewhat weaker than for the Swadesh list. Models predicting Latin frequency of use from Spanish frequency of use that included the additional variables of part of speech and semantic category group, respectively, found that these variables increased the explanatory power of the model only marginally. There was no evidence of interaction, which suggests that the relationship between Spanish and Latin frequency of use does not differ significantly by part of speech or semantic category group.

Although the characteristics of part of speech and semantic category did not add to the explanatory power of the model, it is possible that there are other variables which could be investigated that do add explanatory power. Zipf's research on the relationship

between frequency and length of words suggests that word length might be a variable of interest (Zipf, 1935), however this raises the question of how to operationalize “word length”. Number of phonemes, number of morphemes, or even average length of the word in casual speech are all possible ways of measuring length (although the latter presents a problem when dealing with a dead language like Classical Latin, for which there are no longer any native speakers). Some research also suggests that there is a relationship between the concreteness or abstractness of words and their frequency of use (Nelson & McEvoy, 2000). Finally, measures of word categorization with fewer levels than part of speech might resolve any issues arising from the small number of items in some categories (such as prepositions and pronouns). Such measures might include testing closed vs. open class words, or possibly lexical vs. functional words.

The following chapter will investigate the relationship between frequency of use and amount of change over time. Based on the above results, the assumption of stable frequency of use over time cannot be rejected. For this reason amount of change will be investigated in relation to modern Spanish frequency of use, Latin frequency of use, and additionally the mean of these two frequency values; this will establish which, if any, of the three frequency measurements best predicts amount of change.

Chapter 4.

Frequency of use and Lexical Change

The previous chapter dealt with the stability of frequency of use over time. It determined that the analysis of the current data does not present a direct challenge to the assumption of stability of frequency over time, due to the confounding influence of genre. The relationship between Latin and Spanish frequency of use measurements is weaker than the relationships between modern language frequency of use measurements (as measured by Pagel, Atkinson and Meade, 2007), but comparisons of frequency of use across genres conducted for the current research indicates that the weaker relationship between Latin and Spanish may simply be due to difference in genres between the two corpora used.

This chapter will investigate the relationship between frequency of use and change over time (a variable coded on a 5-point ordinal scale of change from no change to complete lexical replacement, following Pappas and Mooers, 2011). The research question is whether there is a relationship between frequency of use and amount of lexical change over time (whether the amount of change a word has undergone over a period of time can be predicted from its frequency of use).

The primary hypothesis (following Pagel et al. 2007) is therefore: a greater degree of change over time will be predicted by a low frequency of use (i.e., common words are more resistant to change than rare words are).

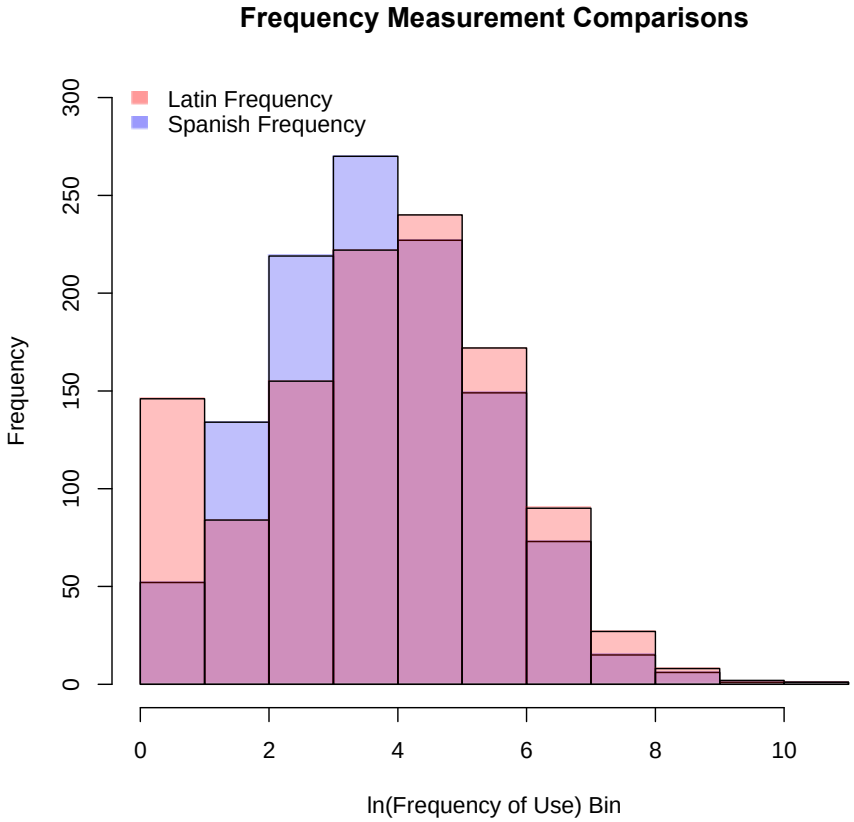
4.1. Relationship between frequency of use and lexical change

Given the results of the previous chapter, we are left with three possible measurements of frequency of use that can be used to predict amount of change. These are: 1) The modern Spanish frequency of use values (taken from the Corpus del Español), 2) The historical Latin frequency of use values (taken from the Perseus Digital Library Project), or 3) The mean frequency of use of modern Spanish and historical Latin (derived from measurements from both corpora). These three possible measurements of frequency of use will be investigated in the following section.

4.1.1. Frequency of use as a predictor variable

The distributions of frequency of use for both Latin and Spanish can be seen in figure 4.1.

Figure 4.1. Histogram of Spanish frequency



All measurements are unimodal, and visually appear to be relatively normally distributed. Both Spanish and Latin measurements (as well as the mean of both) fail formal tests of normality, most likely because of the right-ward skew caused by several high outliers. Given these observations, any one of the three measurements could be used to perform the statistical analyses below. Due to the large size of the Corpus del Español, however, as well as the part of speech tagging in that corpus, the modern Spanish frequency of use values are most likely the best measurements (as far as being closest to the "real" frequency of use), for this reason, modern Spanish frequency of use values will be the main focus for the remaining tests in this chapter.

4.1.2. Amount of change as a variable

The original coding scheme followed was based on a six-point scheme outlined in Pappas and Mooers (2011). As previously mentioned in Chapter 2, not all codes were well represented. Of particular concern was code (3), syntactic reanalysis, of which there were only 16 examples. The distribution of words according to code can be found in table 4.1. It is apparent from this breakdown that code (3) represents less than 2% of the overall total. For this reason, code (3) was binned together with code (2) for the purpose of the analyses below.

Table 4.1. Distribution of word meanings across codes

Type of Change	Code	Number of words	Percentage of total
Sound change	0	199	17.35%
Paradigm leveling	1	317	27.64%
General analogy	2	129	11.25%
Syntactic reanalysis	3	16	1.39%
Semantic change	4	320	27.90%
Lexical borrowing	5	166	14.47%

Code (3) was binned with the adjacent code (2) (general analogy) for theoretical reasons. It is more appropriate to bin (3) in this way as it also represents a form of cognitive association, and is therefore more akin to analogical change (code (2)) than to semantic change (code (4)), which represents a replacement of one word with another from the same language.

For the purpose of the following analyses, therefore, codes (2) and (3) were both designated as (2), with the numbers of the larger two codes (4 and 5) also being decreased in order to avoid gaps. A comparison of the code designations used in the original Pappas and Mooers (2011) paper and those used in the following results can be found in table 4.2.

Table 4.2. Code designations

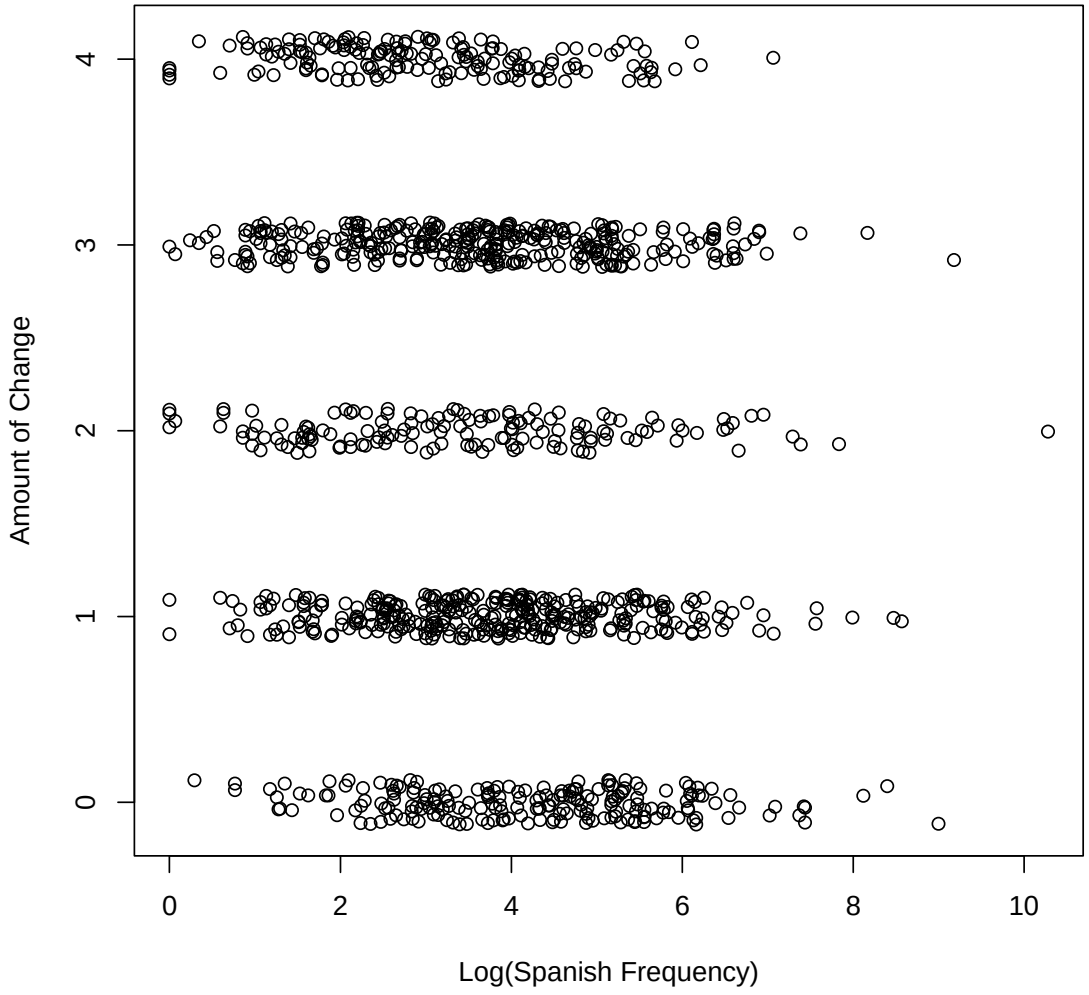
Code value	Designation in Pappas & Mooers 2011	Designation in the following results
Sound change	0	0
Paradigm leveling	1	1
General analogy	2	2
Syntactic reanalysis	3	2
Semantic change	4	3
Lexical borrowing	5	4

4.1.3. Relating frequency of use to lexical change

In a first step, I treated amount of change as a pseudo-continuous variable, rather than an ordinal variable. This assumes equidistance between the ordered codes (i.e. the assumption being that the difference between code 0 and code 1 is the same as the difference between code 1 and code 2).

A scatterplot of the amount of change by the log frequency can be found in figure 4.2. This scatterplot appears to show an overall negative trend (i.e. code 4 appears to have a lower mean than code 0). This trend is more apparent in figure 4.3, which is the same data represented as a box-and-dot plot.

Figure 4.2. Scatterplot of frequency of use and amount of change, n = 1147



The linear model predicting amount of change from log Spanish frequency of use results in an estimated r^2 of 0.038 ($p < 0.00001$). The estimate is -0.161. This is a very weak, but highly significant negative correlation between amount of change and frequency of use, as predicted from the work by Pagel et al. (2007): less than 4% of the variability in amount of change can be explained by the frequency of use of the words. This indicates that there is a relationship between frequency of use and amount of change, but not a very strong one.

Figure 4.3. Box-and-dot plot of frequency of use and change

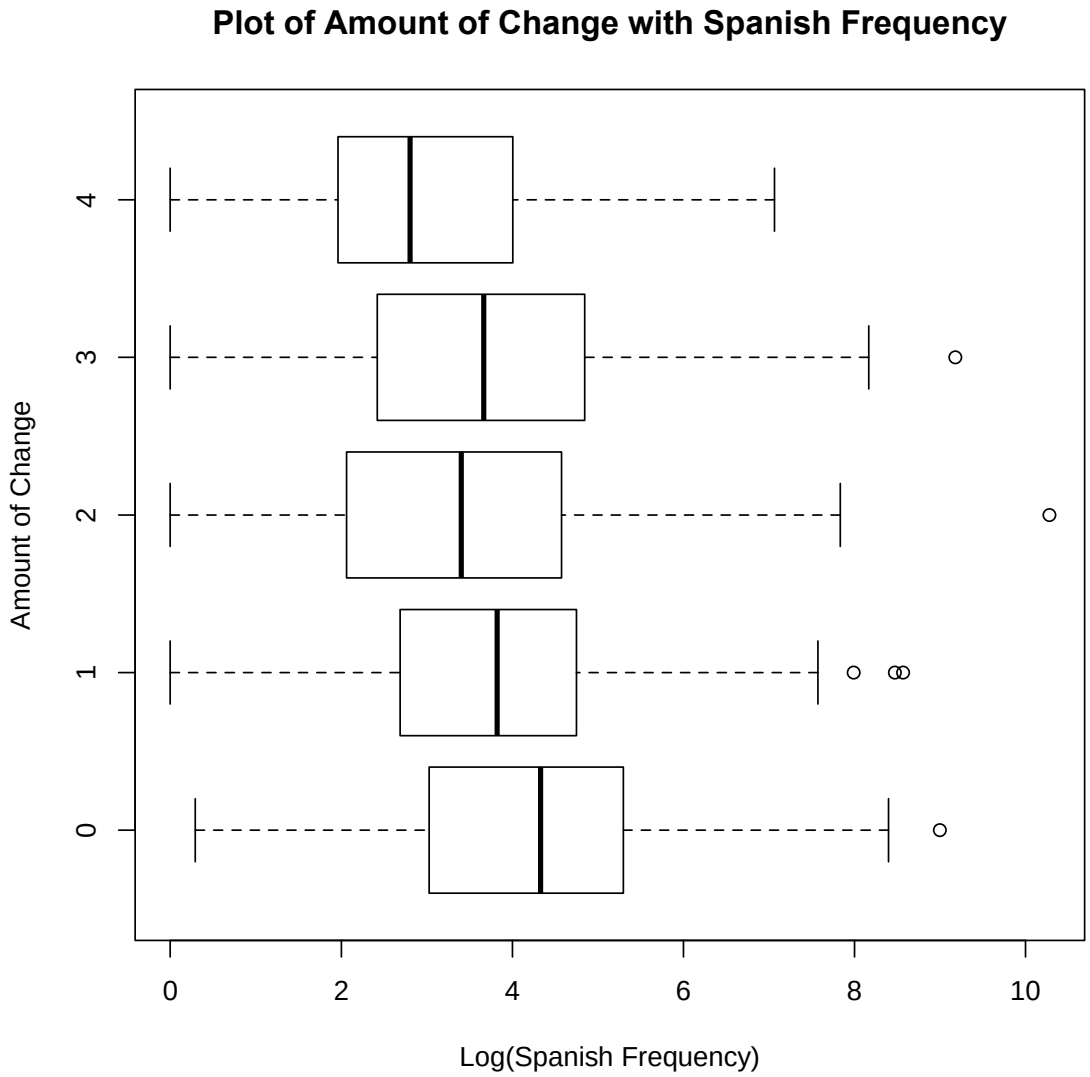


Figure 4.3 shows a box-and-dot plot of amount of change against log-Spanish frequency. This plot allows us to visually compare the means of different categories. By inspection, there would appear to be a general negative trend such that the higher amounts of change have lower mean frequencies. The exception to this is the mean of code 2, which is out of line with the rest of the means.

4.2. Pairwise comparisons

The data support a very weak but significant negative relationship between amount of change (as classified on the ordinal scale used herein) and frequency of use. Post-hoc tests of the different classes of change may shed light on the drivers of this relationship.

We did these tests in an ANOVA framework, treating amount of change as a nominal variable, rather than as a pseudo-continuous variable as done in the previous section. Tukey's HSD was used to correct for the multiple testing problem (i.e. an increase in the number of pairwise comparisons being performed increases the likelihood of a false-positive if the threshold for determining a significant p-value is kept constant (Schwarz, 2014a)). The results of the pairwise comparisons (as adjusted by Tukey's test) can be seen in table 4.3.

Because the p-values here have already been adjusted for multiple testing, they can be read with the same .05 cut-off for significance as would normally apply. The pairwise tests compare frequency distributions of two codes (for example, 1 and 0) (essentially the same as running a t-test here would be), and so test whether two codes have different means. If the p-value is significant, there is evidence that the two codes have different population means. Significant p-values are marked with an asterisk.

Table 4.3. Tukey's HSD pairwise tests on the ANOVA

Comparison	Difference	Lower 95%	Upper 95%	p (adjusted)
1-0	-0.444	-0.840	-0.049	< 0.02*
2-0	-0.782	-1.260	-0.304	< 0.0001*
3-0	-0.603	-0.999	-0.208	< 0.001*
4-0	-1.233	-1.693	-0.772	< 0.0000001*
2-1	-0.337	-0.776	-0.102	< 0.23
3-1	-0.159	-0.506	0.188	< 0.73
4-1	-0.788	-1.207	-0.369	< 0.00001*
3-2	0.179	-0.260	0.617	< 0.80
4-2	-0.451	-0.948	0.047	< 0.10
4-3	-0.629	-1.048	-0.210	< 0.001*

By examining the table, we can see that six of the ten comparisons are significant at alpha = 0.05. Code 0 is significantly different from all other codes. Code 4 is significantly different from all codes at alpha = 0.1, and different from all but code 2 at alpha = 0.05. Codes 1, 2 and 3, however, are not significantly different from each other. Note that the Tukey's test pairwise comparisons always compare a higher code to a lower code, and that with one exception (the comparison of code 3 and code 2) the difference between the codes is always negative. This is consistent with the apparent negative trend seen in figures 4 and 5, with Code 2 being the one not following this trend.

4.3. Covariates

As presented above, the model predicting amount of change based on log frequency alone resulted in a r^2 of 0.038 ($p < 0.0001$) with an estimated slope of -0.16. Some of the variation in amount of change explained by frequency of use may in fact be due to the covariates of part of speech and semantic category group that may differ in both frequency of use and probability of change. In order to test this question, I ran additional models that include the covariates of part of speech and semantic category group (individually and together), to determine whether this affects the slope (i.e. the estimate of the relationship between amount of change and frequency).

A covariate is a variable whose effect interests us only insofar as it mediates a relationship of interest. If we introduce the covariates and the estimate between our focal measures changes, then some of the relationship we were observing in the base model can also be explained by the covariates, and including these allows us to isolate the independent effects of our focal variable. As in Chapter 3, I considered two possible covariates: part of speech and semantic category.

4.3.1. Part of speech

There are two possible ways of including part of speech, either as an 8-category factor (i.e. all part of speech categories identified previously included in the estimate) or as a 4-factor (i.e. where all the functional part of speech categories have been binned

together as a the single “function” part of speech, while the lexical categories of noun, adjective and verb are left independent). Binning the functional categories together is better statistically, as some of the functional categories include very few words, a situation that can produce unstable estimates. The 8-category tests approach has been used in previous research (i.e. Pagel, Atkinson and Meade, 2007) as well as in Chapter 3, so I include it here with this caveat.

When all eight part of speech categories are included, the model predicting amount of change with frequency of use gives a full model r^2 of 0.057 ($p < 0.0001$), with the estimate of the slope for FoU -0.14. When part of speech is rescored with only four categories (binning functional categories) the results are similar, with an r^2 of 0.053 ($p < 0.0001$) and an estimate of -0.13. In both models, the effect of FoU on amount of change remains highly significant.

Both of these models give a slightly weaker estimate of the relationship between amount of change and frequency of use than the base model (slope = -0.16). This would seem to indicate that the relationship between FoU and amount of change in the base model is affected by, but not wholly due to the confounding effects of part of speech.

4.3.2. Semantic category group

All five levels for semantic category group are well-represented, and so no further grouping is necessary. A model of amount of change as predicted by frequency of use that includes semantic category group as a covariate results in an r^2 of 0.053 ($p < 0.0001$), with a highly significant estimate of slope for FoU of -0.13. This gives the same results as the part of speech model, in that the estimate is weaker but still significant when semantic category group is included. Some of the variation accounted for in the base model appears to be due to semantic category group, but not all of it.

4.3.3. Part of speech and semantic category group

The model predicting amount of change with frequency of use and including both part of speech (the binned variable) and semantic category group similarly results in a

full model r^2 of 0.065 ($p < 0.0001$). The estimated relationship between amount of change and frequency of use in this more complex model is -0.12 ($p < 0.0001$).

The estimate of the relationship between amount of change and frequency of use for the model including both part of speech and semantic category group is the weakest of all, which is consistent with both of the above models (models including only one of the two covariates). Even with both covariates included, however, there is still a relationship. So the weak but highly significant negative correlation between amount of change and frequency of use (observed above) is not due solely to the influence of part of speech and semantic category group. There is a relationship between amount of change and frequency of use that is *independent* of these two variables.

4.4. Summary

The current chapter tested the relationship between amount of change from Latin to Spanish (using a 5-point ordinal scale of change) and modern Spanish frequency of use values for a set of 1147 words. The results indicate that there is a weak but highly significant negative correlation between these two variables. Post-hoc pairwise comparisons indicated that the relationship is primarily driven by the most extreme levels of change (code 0 and code 4), both of which are significantly different from almost all the other codes. The three middle subsets (groups coded as 1, 2, and 3) were not found to be significantly different from each other.

Tests of models which included the additional covariates of part of speech and semantic category group found that these variables were contributing to the relationship between amount of change and frequency of use. Even in models controlling for both of these covariates, however, there remained an independent relationship between amount of change and frequency of use.

Chapter 5.

Discussion

The research in this thesis was motivated by previous research that provided evidence for a negative relationship between frequency of use and language change over time (Pagel, Atkinson, & Meade, 2007). The nature and directionality of that relationship has yet to be definitely established, however. Two primary research questions were raised: firstly, whether frequency of use remains stable over time, and secondly, whether amount of language change over time can be predicted using frequency of use. The current research attempted to test both of these questions using a 1147-word subset of the Intercontinental Dictionary Series wordlist (Key & Comrie, 2007), for Latin and its daughter language Spanish. This research was able to extend Pappas and Mooers' (2011) replication of Pagel et al.'s (2007), similarly within a single language lineage (Italic, rather than Greek in this case), and also attempted to address two major concerns raised by Pappas and Mooers: the restricted nature of the 200-word Swadesh list, as well as the issue of the stability of frequency of use over time.

The question of the stability of frequency of use over time was tested using two corpora, one of Classical Latin (the PDLP) and another of modern Spanish (the CdE). The goal of this test was to establish whether it is appropriate to use modern frequency of use values in lieu of historical frequency of use values when the latter is unavailable. Due to the paucity of written material in many languages, research on the historical impact of frequency on language is almost entirely dependent on these modern frequencies, and so their appropriateness is a matter of concern. It was the determined that the correlation between the frequencies of two different points in time for a single language lineage (Latin and Spanish) was lower than the correlation between the frequencies of modern Indo-European languages (compare ρ = 0.60 for Latin vs. Spanish with ρ = 0.76 for Russian vs. Greek). However, within-language tests of

different genres found that a similarly low correlation could be achieved by testing the correlation of Fiction with Academic Writing in Spanish ($\rho = 0.67$).

Given the opportunistic nature of the Latin corpus, and the large difference in time period between the two corpora, it does not appear to be possible to control the two corpora for genre in such a way as to remove its influence (i.e. the PDLP and the CdE cannot be matched for genre). Therefore, the lower correlation between different Spanish genres indicates that the observed lower correlation between two time-stages of the same language lineage (the Latin vs. Spanish correlation) may simply be due to differences in genre. The results of this test are therefore inconclusive. Because genre cannot be rejected as a source of difference, it is not possible to determine conclusively that the relatively low correlation between Latin and Spanish frequency of use is due to deviations in frequency of use over time. Therefore, the null hypothesis, which is that frequency of use does not change over time, cannot be rejected.

I then investigated the second research question, whether there is a relationship between frequency of use and amount of change over time. The question might be alternatively phrased as to whether current research addressing change within a single lineage with a much larger data set of words could reproduce the relationship between frequency of use and number of cognates across Indo-European languages identified by previous research (Pagel, Atkinson & Meade, 2007). Amount of change between Latin and Spanish was coded on a five-point scale (after the scale used in Pappas & Mooers, 2011). In the absence of conclusive evidence against the use of modern frequency of use values, the modern Spanish frequency data was used (given that the CdE is both tagged for part of speech and three times the size of the PDLP, these values are expected to be more accurate).

Tests of the relationship between amount of change and frequency of use identified a weak but highly significant negative correlation (slope = -0.16, $r^2 = 0.038$, $p < .00001$). This result indicates that the greater the frequency of use of a word, the less change it has undergone. Investigation of the results in more detail indicate that this relationship is primarily driven by the difference between the most extreme categories. Language-external borrowing (code 4) seems most common in low frequency of words

and no change (code 0) is most common in higher frequency of use words; there does not appear to be a significant difference in the mean frequency of use among the three intermediate categories of change, however.

These results add to the growing body of evidence of a relationship between frequency of use and language change over time. Further research examining this relationship using other languages and other language lineages is needed. With regard to the question of the stability of frequency of use over time, the results of the current research were inconclusive. In order to establish definitively whether frequency of use is a stable characteristic of language, it will be necessary to test corpora that are better controlled for genre. Additional potential confounds (which were not addressed in the current research due to the overwhelming genre issue) include different sized corpora (i.e. a historical corpus is likely to be a great deal smaller than an equivalent modern corpus due to preservation issues) and different measurements (such as the difference between maximum frequency and weighted frequency described in chapter 2). One potential way of addressing this question might be to test more than two time-periods. Testing an intermediate time stage (for instance, medieval Spanish in addition to Classical Latin and modern Spanish), might better establish the presence or absence of a trend in frequency of use. If it is the case that frequency of use changes over time, we would expect the frequencies of an intermediate language stage to also be intermediate in frequency (e.g. if a word has a frequency of 10 in Latin and 100 in modern Spanish, its frequency in medieval Spanish should be greater than 10 but less than 100).

Regarding amount of change and frequency of use, the current research has been largely preliminary, and the hypothesized relationship between these two variables is not well understood. Models including part of speech and semantic category as covariates indicate that there is a relationship between change and frequency which is independent of these additional two covariates, but the fact that the relationship is altered by the presence of the covariates in the models does indicate that the intersection of all variables involved should be investigated more thoroughly. Additional covariates might also be investigated as well.

The pairwise comparisons of the mean frequencies between codes also suggest that the coding scheme in use requires more investigation. The observation that language-external borrowing and absence of change were significantly different from the other codes, which were themselves not significantly different from each other, suggests that the relationship between frequency and change may not be a simple linear one of least to greatest amount of change. There appears to be a relationship, but the specific nature of that relationship, and the mechanism through which it appears remains unclear.

The finding of a statistically significant correlation between amount of change and frequency of use indicates the presence of a relationship between the two variables. The weakness of the correlation, however, indicates that this relationship may not be particularly meaningful or important (Schwarz, 2014b). Although attempts were made to control for two covariates (part of speech and semantic category) which were suggested to be involved in the relationship, the possibility remains that some additional unknown variable is participating in the relationship. If either frequency of use or amount of change is being conflated with a third variable, the statistical relationship that we observe may not be direct. It is important to avoid the assumption of a causal relationship between frequency of use and amount of change, when all that has been observed is a correlation.

There remains the additional concern regarding the appropriateness of our measurements. Both of our variables are proxy measurements for a hypothetical linguistic characteristic. “Frequency” for the purpose of this research is a measurement of the number of times a particular word appears in a written corpus, controlled for the overall size of that corpus. The words in this corpus are not what is of interest to us, however, but instead how regularly a given word or concept appears in spoken language, or perhaps how accessible said word is within the memory of a typical speaker. The variable of amount of change is also potentially problematic, as it represents an attempt to quantify the degree to which a word has changed over time on a relatively new ordinal scale. In interpreting the present results, we must leave room for the possibility that we are not measuring precisely what we think we are measuring, and therefore that the relationship that we observe is not what we first assume it to be.

The current results raise some questions about the original Pagel et al. (2007) results with regards to number of forms in Indo-European. The Pagel et al. research made use of the Swadesh list, which is intentionally designed to minimize language-external borrowing, and includes only words which are deemed unlikely to be borrowed (Swadesh, 1952). While it is the case that over 87 languages, the 200 words in question undoubtedly include *some* instances of borrowing, it is likely that the intentional exclusion of this type of change means that the list has proportionally less borrowing included than a list not intended for this purpose (like the IDS list). Note that in Pappas and Mooers (2011), which also made use of the Swadesh list, the category of borrowing had to be dropped, due to a low number of items.

The negative correlation observed between amount of change and frequency of use in the current research, however, was deemed to be driven primarily by the relationship between language-external borrowing and the other categories. Given that this driving category is not supposed to be strongly represented in the Swadesh list, we would expect that the same correlation using the Swadesh list (i.e. comparing the Pagel et al. 2007 research to the current results) would be weaker. Instead, we find that the correlation for Pagel et al. was stronger than the current correlation (compare Pagel et al.'s $r = 0.35$, with the current research $r = 0.19$).

The overarching question motivating this research is Pagel et al.'s (2007) initial claim to have uncovered a new mechanism of language change in frequency of use. Based on the current results, frequency does not have a strong enough impact, or an unambiguous enough impact to conclusively identify it as a mechanism of change. The primary driver of the correlation herein appears to be borrowing, which is already known to be an independent mechanism of language change. The previous research has instead suggested that frequency mediates, exaggerates, or otherwise interacts with mechanisms of change, rather than being a mechanism of change itself (Bybee, 2002). It is therefore premature to describe frequency of use as a mechanism of language change that operates in a monotonic fashion.

The conclusion that frequency of use is a new mechanism of lexical change, or at least, the conclusion that frequency has a linear, negative impact on lexical change,

continues to appear in research as an established truth (Altmann, 2013; Keller, 2013; Alonso, Fernandez, & Diez, 2011), and this can have unfortunate consequences. When this claim appears, the sources given are the work Pagel, Atkinson and Meade (2007), and additionally the research on English irregular verbs by Lieberman et al. (2007) (a suggestive, but highly specific result which the authors do not claim should be generalized to all languages and parts of speech). More recent research by members of the Pagel group has been based on the 2007 results, including Calude and Pagel (2011), which makes use of the rate of lexical replacement data from Pagel et al. (2007), and Pagel, Atkinson, Calude and Meade (2013), which goes so far as to utilize the results of Pagel et al. (2007) to identify "ultra-conserved words", which are determined to have an extremely slow rate of lexical replacement, as a basis for identifying linguistic 'superfamilies', like the proposed Nostratic superfamily, which includes at least Indo-European, Uralic, Altaic, and the Kartvelian language families. A common criticism of the original Nostratic hypothesis (Renfrew & Nettle, 1999) is that it is the result of applying historical reconstruction on datasets that are themselves the results of historical reconstruction, and therefore, highly speculative (Campbell, 2004). Unfortunately, these newer claims must also be seen as speculative because the basic assumption about the role of frequency cannot be taken at face value. The most we can say, as this thesis has shown, is that there is an intriguing correlation between frequency and lexical change, but the causal mechanism behind it is not yet understood.

References

Alonso, M. A., Fernandez, A., & Díez, E. (2011). Oral frequency norms for 67,979 Spanish words. *Behavior Research*, (43), 449–458.

Altmann, E. G., Whichard, Z. L., & Motter, A. E. (2013). Identifying trends in word frequency dynamics. *Journal of Statistical Physics, 151,* 277-288.

Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections - Phylogenetic thinking in biology and historical linguistics. *Systematic Biology, 54*(4), 513-526.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133,* 283-316.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257.

Bowern, C., & Atkinson, Q. (2012). Computational phylogenetics and the internal structure of Pama-Nyungan. *Language 88*(4), 817-845.

Bowern, C., Epps, P., Gray, R., Hill, J., Hunley, K., McConvell, P., & Zentz, J. (2011). Does lateral transmission obscure inheritance in hunter-gatherer languages? *PloS one, 6*(9), e25195.

Boyd-Bowman, P. (1954). *From Latin to Romance in Sound Charts*. Kalamazoo College Press.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change,* (14), 261-290.

Bybee, J., & Hopper, P. (Eds.). (2001). *Frequency and the emergence of linguistic structure* (Vol. 45). John Benjamins Publishing.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of *don't* in English. *Linguistics 27*(4), 575-596.

Bybee, J., & Thompson, S. (1997). Three Frequency Effects in Syntax. *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure*, 378–388.

Bush, N. (2001). Frequency effects and word-boundary palatalization in English. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 255-280). Amsterdam: Benjamins.

Calude, A. S., & Pagel, M. (2011). How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B*, *366*, 1101–1107.

Campbell, L. (2004). *Historical linguistics: An introduction*. MIT press.

Crane, G. R. (2010). Perseus digital library project. *Tufts University, Retrieved from <http://www.perseus.tufts.edu/hopper/>*

Davies, M. (2002). Corpus del Español: 100 million words, 1200s-1900s. *URL <http://www.corpusdelespanol.org>*

Deese, J. (1960). Frequency of usage and number of words in free recall: The roll of association. *Psychological Reports, 7*, 337-344.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12,* 627-635.

Fosler-Lussier, E., & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciation in conversational speech. *Speech Communication 29*, 137-158.

Gómez de Silva, G. (1985). *Elsevier's concise Spanish etymological dictionary.* Elsevier Scientific Publishing Company*.*

Grainger, J. (1990). Word frequency and neighbourhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29,* 228-244.

Gregg, V. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), *Recall and recognition* (pp. 183-216). New York: Wiley.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. *Proceedings of the 35th meeting of the Chicago Linguistic Society* (pp. 151-166). Chicago: Chicago Linguistic Society.

Hasenfratz, R. J., & Jambeck, T. (2005). *Reading Old English: A primer and first reader.* WVU Press.

Hock, H. H. (1991). *Principles of historical linguistics*. Walter de Gruyter.

Hooper, J. (1976) Word frequency in lexical diffusion and the source of morphophonological change. In W. M. Christie, Jr. (Ed.), *Current progress in historical linguistics*. Amsterdam: North Holland.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, *41*, 401-410.

Keller, D. B., & Schultz, J. (2013). Connectivity, not frequency, determines the fate of a morpheme. *PloS ONE 8*(7): e69945.

Key, M. R., & Comrie, B. (2007). Intercontinental dictionary series.

Krug, M. (1998). String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics 26*, 286-320.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (eds.). (2014). *Ethnologue: Languages of the World, Seventeenth edition*. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com>

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*, 713–716.

Loeys, Moerkerke, De Smet & Buysse, 2012 The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. British Journal of Mathematical and Statistical Psychology.

Mandler, G., Goodman, G. O., & Wilkes-Gibbs, D. L. (1982). The word-frequency paradox in recognition. *Memory and Cognition, 10*, 33-42.

Mendeloff, H. (1969). *A manual of comparative Romance linguistics: phonology and morphology.* Catholic University of America Press.

Nelson, D. L., & McEvoy, C. L. (2000). What is this thing called frequency? *Memory & Cognition, 28*, 509-522.

Pagel, M., Atkinson, Q. D., Calude, A. S., & Meade, A. (2013). Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences, 110*(21), 8471-8476.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*, 717–720.

Pappas, P. A., & Mooers, A. O. (2011). Phylogenetic methods in historical linguistics: Greek as a case study. *Journal of Greek Linguistics*, *11*(2), 198–220.

Phillips, B. S. (1984). Word frequency and the actuation of sound change. *Language 60*(2), 320-342.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Acoustical Society of America*, 2561-2569.\

Renfrew, C. & Nettle, D. (1999). *Nostratic: examining a linguistic macrofamily*. The McDonald Institute for Archaeological Research.

Schwarz, C. J. (2014a). Chapter 2: Introduction to Statistics. In *Course Notes for Beginning and Intermediate Statistics.* Available at http://www.stat.sfu.ca/~cschwarz/CourseNotes. Retrieved 2014-12-15.

Schwarz, C. J. (2014b). Chapter 14: Correlation and simple linear regression. In *Course Notes for Beginning and Intermediate Statistics.* Available at http://www.stat.sfu.ca/~cschwarz/CourseNotes. Retrieved 2014-12-19.

Schwarz, C. J. (2014c). Chapter 15: Detecting trends over time. In *Course Notes for Beginning and Intermediate Statistics.* Available at http://www.stat.sfu.ca/~cschwarz/CourseNotes. Retrieved 2014-12-15.

Schwarz, C. J. (2014d). Chapter 19: Multiple linear regression. In *Course Notes for Beginning and Intermediate Statistics.* Available at http://www.stat.sfu.ca/~cschwarz/CourseNotes. Retrieved 2014-12-15.

Schwarz, C. J. (2014e). Chapter 23: A short primer on residual plots. In *Course Notes for Beginning and Intermediate Statistics.* Available at http://www.stat.sfu.ca/~cschwarz/CourseNotes. Retrieved 2014-12-15.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association, 70*(351), 542-547.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society, 96*(4), 452-463.

Wheelock, F. M., & LaFleur, R. A. (2005). *Wheelock's Latin (6th ed. revised).* HarperResource.

Woods, M. J. (2001). Spanish word frequency: A historical surprise. *Computers and the Humanities*, *35*(2), 231–236.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology 40*, 1-95.

Zipf, G. K. (1935). *The psycho-biology of language.* Cambridge MA: MIT Press.

Appendix A.

Wordlist and coding justifications

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
2nd person (sg. & pl.)	tu	tú	0	No change to stem.	pronoun	2	5609.24	134.80	469.80	3808.49	783.45
3rd person (sg. & pl)	is	él	4	From *ille* "that" (demonstrative)	pronoun	2	9431.86	18194.60	19274.25	32047.39	9703.13
account, reckoning	ratio	cuenta	4	From the verb *computare* "to reckon, compute, sum up"	noun	11	1044.35	123.40	334.42	511.96	320.29
accuse	accuso	acusar	0	Geminate simplification	verb	21	182.03	51.00	141.42	42.56	78.73
acid, sour	acidus	ácido	0	No change to stem.	adjective	15	12.25	27.60	7.25	6.08	13.78
acorn	glans	bellota	5	Arabic origin	noun	8	30.24	1.20	0.20	3.14	1.49
acquit	absolvo	absolver	0	No change to stem.	verb	21	103.02	3.60	8.26	4.61	5.50
adultery	adulterium	adulterio	1	LA ending vowel added to SP stem.	noun	21	44.34	3.60	2.01	5.03	3.53
after	post	después	3	From *de post* "from after"	adverb	12	458.35	639.01	479.07	937.34	681.71
afternoon	postmeridianum	tarde	4	From the adverb *tarde* 'slowly, late'	noun	14	0.00	8.80	144.85	382.40	175.59
age	aetas	edad	0	Regular ae > e development; intervocalic 'd' voicing; LA stem ends in 't'.	noun	14	450.58	316.80	149.08	171.07	213.12
all	totus	todo	0	Intervocalic 't' voicing	adjective	13	656.52	313.20	179.70	153.88	216.65
alone	solus	solo	0	No change to stem.	adjective	13	196.23	123.20	277.21	644.67	343.91

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
altar	altar	altar	0	No change to stem.	noun	22	8.05	19.20	9.87	42.14	23.48
always	semper	siempre	1	Diphthongization of short stressed 'e' > 'ie'; *	adverb	14	692.25	222.80	508.28	141.96	616.55
ambush	insidiae	emboscada	5	Germanic origin	noun	20	104.22	2.40	4.03	2.73	3.05
ancestors	major	antepasado	3	From *ante pasado* "passed before"	noun	2	0.00	33.80	7.05	21.80	20.90
anchor	ancora	ancla	2	r > l sporadic change	noun	10	15.54	1.00	1.01	3.35	1.76
and	et	y	2	Loss of final -t is regular, probably developed into yod due to being placed in hiatus with following vowels	conjunction	17	26891.67	31112.34	25279.76	31176.51	29168.11
anger	ira	ira	1	LA ending vowel added to SP stem.	noun	16	462.30	9.80	11.28	36.48	18.94
animal	animal	animal	0	No change to stem.	noun	3	112.59	365.60	68.70	164.78	200.56
ankle	talus	tobillo	4	From the diminutive of *tuber* 'hump, bump, swelling'	noun	4	160.75	2.20	3.63	21.17	8.82
announce	annuntio	anunciar	2	Missing nasalization	verb	18	0.00	70.00	294.74	122.02	162.55
answer	respondeo	responder	0	No change to stem.	verb	18	392.38	76.20	191.19	262.27	175.18
ant	formica	hormiga	1	Word-initial 'f' > 'h'; intervocalic [k] voicing; LA ending vowel added to SP stem.	noun	3	15.42	21.20	4.23	29.98	18.33
anvil	incus	yunque	2	From a metathesized form	noun	9	5.74	1.60	0.20	0.63	0.81
anxiety	anxietas	ansiedad	2	Sporadic x > s	noun	16	3.35	13.00	9.87	47.38	23.08
approach	accedo	acercar	2	a + cerca + ar	verb	10	345.29	84.60	147.27	477.58	232.94
arch	arcus	arco	1	LA ending vowel added to SP stem.	noun	7	74.70	27.40	8.06	10.48	15.41

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
armor (defensive)	armatura	armadura	1	Intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	20	13.51	26.60	0.60	11.74	13.03
arrow	sagitta	flecha	5	French origin	noun	20	63.46	14.80	4.03	21.59	13.37
ashes	cinis	ceniza	2	No reason for short-i > e	noun	1	145.10	16.00	12.89	48.22	25.38
ask, request	peto	pedir	0	Intervocalic [t] voicing	verb	18	966.39	47.80	396.27	538.38	324.02
ass, donkey	asinus	burro	4	From *burricus* 'small horse'	noun	3	16.85	2.80	4.03	23.90	10.05
attack	impetus	ataque	5	Old Italian origin	noun	20	298.32	142.60	99.12	60.80	101.47
aunt	amita	tía	5	Greek origin	noun	2	4.90	2.60	7.66	218.04	74.05
avaricious, stingy (greedy)	avarus	avaro	0	No change to stem.	adjective	11	40.99	0.40	0.81	2.31	1.15
awl	subula	punzón	4	From *punctio* "puncture, pricking pain"	noun	6	0.72	2.00	0.40	3.14	1.83
ax	securis	hacha	5	French origin	noun	9	53.31	8.80	1.61	9.85	6.72
back	dorsum	espalda	4	From the dimiutive of *spatha* 'straight sword'	noun	4	18.29	20.60	40.49	246.55	100.45
bad	malus	malo	0	No change to stem.	adjective	16	237.32	45.00	121.88	233.55	131.94
bait	esca	cebo	4	From *cibus* "food, fodder"	noun	20	9.56	4.00	1.21	3.35	2.85
bake	coquo	hornear	4	From the noun *furnus* "oven"	verb	5	115.45	1.00	0.81	3.98	1.90
bald	calvus	calvo	0	No change to stem.	adjective	4	10.43	3.80	13.90	15.09	10.86
barley	hordeum	cebada	4	From *cibus* "food, fodder"	noun	8	30.72	19.80	0.60	3.56	8.08
basket	cista	cesto	2	No reason for short-stress-i > e	noun	9	3.47	2.20	1.21	5.66	2.99
bat	vespertilio	murciélago	4	From *mur ciego* "blind mouse"	noun	3	4.78	18.80	2.62	10.90	10.79

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
bathe	lavo	bañar	4	From the noun *balneum* "bath"	verb	4	91.21	6.00	11.68	72.33	29.39
bay	sinus	bahía	5	French origin	noun	1	262.07	46.40	14.91	24.11	28.57
bead	baca	cuenta	4	From the verb *contar* "to count"	noun	6	27.48	123.40	334.42	511.96	320.29
bean	faba	judía	4	From the feminine form of *Judaeus* "Jew, Jewish"	noun	5	38.01	2.60	0.60	0.21	1.15
beard	barba	barba	1	LA ending vowel added to SP stem.	noun	4	23.70	10.20	14.30	70.65	31.15
beautiful	bellus	bello	0	Geminate 'll' palatalizes regularly.	adjective	16	657.01	63.00	103.95	132.29	99.23
because	quia	porque	3	From *por que* "for that"	conjunction	17	1011.37	225.80	1083.85	1858.54	1043.47
bed	lectus	cama	5	Celtic/Iberian origin (from LL)	noun	7	31.18	12.40	20.75	365.42	129.50
bee	apis	abeja	4	From *apicula*, diminutive of *apis* "bee"	noun	3	25.97	30.20	6.25	14.68	17.10
beehive	alveus	colmena	5	Celtic origin	noun	3	32.51	5.60	2.62	5.24	4.48
beer	cervesia	cerveza	2	No reason for si > z	noun	5	0.00	19.20	16.12	70.65	34.82
before	ante	antes	2	Addition of the 's' unexplained	preposition	12	1297.25	246.20	255.05	409.86	302.17
beget (of father)	genero	engendrar	4	From *ingenerare* "to produce, cause"	verb	4	177.37	3.80	5.24	12.16	6.99
beggar	mendicus	mendigo	1	Intervocalic [k] voicing; LA ending vowel added to SP stem.	noun	11	5.26	3.60	4.43	21.59	9.71
begin	incipio	comenzar	4	Generalization: From *com* "thoroughly" + *initiare* "to initiate"	verb	14	246.36	498.81	344.90	479.89	440.83
behind	post	detrás	2	de' + 'tras'	preposition	12	458.35	25.20	48.35	196.65	88.51
believe	credo	creer	0	Intervocalic loss of 'd'.	verb	17	837.50	171.40	469.80	858.93	509.52
belt	cingulum	cinturón	2	"cintura" + "on"	noun	6	2.15	20.80	6.25	21.80	16.22
bend	curvo	doblar	4	From *duplus* "double"	verb	9	42.07	13.80	13.70	68.56	31.49

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
betray	prodo	traicionar	4	From *traditio* "surrender, hand over"	verb	17	201.35	3.20	10.88	26.84	13.44
beverage, drink	potus	bebida	4	From the verb *bibere* "to drink"	noun	5	142.82	12.00	14.71	36.69	20.90
birch	betula	abedul	2	Suspected analogy with "abeto"	noun	8	0.72	6.00	1.61	0.42	2.71
bird	avis	ave	1	LA ending vowel added to SP stem.	noun	3	82.03	133.20	19.34	46.96	66.92
bite	mordeo	morder	0	No change to stem.	verb	4	64.50	4.00	5.64	66.88	24.91
bitter	amarus	amargo	2	Influence of "amargar"	adjective	15	74.70	9.20	14.91	43.61	22.26
blacksmith	ferrarius	herrero	1	a > e under influence of following r + yod	noun	9	9.56	1.60	3.02	15.30	6.52
blame	culpa	culpa	1	LA ending vowel added to SP stem.	noun	16	107.45	9.20	36.06	137.74	59.86
blanket	lodix	manta	5	Gaulish origin	noun	7	1.20	3.60	4.03	34.17	13.64
blind	caecus	ciego	0	CL 'ae' > VL open 'e' > diphthongizes in SP to 'ie'; intervocalic [k] voicing.	adjective	4	60.27	9.00	7.45	40.46	18.66
blister	pustula	ampolla	4	From *ampulla* "flask, bottle"	noun	4	3.11	1.20	0.81	2.73	1.56
blood	sanguis	sangre	2	Sporadic additional syllable from VL	noun	4	569.03	133.00	72.93	297.70	166.08
blow	flo	soplar	4	Generalization: From *sufflare* "to blow up, inflate"	verb	10	55.71	15.60	4.43	41.93	20.36
blue	caeruleus	azul	5	Arabic origin	adjective	15	20.56	47.80	57.62	193.93	98.41
boar	verres	jabalí	5	Arabic origin	noun	3	41.95	14.00	0.81	3.35	6.11
boast	iacto	jactarse	5	Semi-learned? No ct > ch	verb	18	0.00	0.20	3.42	4.19	2.58
boat	linter	barca	5	Greek origin (from LL)	noun	10	5.50	2.00	1.61	7.34	3.60
body hair	pilus	vello	4	From *villus* "shaggy hair, tuft of hair, wool"	noun	4	20.57	1.40	0.40	13.84	5.09
boil	ferveo	hervir	0	Word initial 'f' > 'h' (now silent in	verb	5	34.06	8.00	4.03	36.06	15.75

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				SP)							
bone	os	hueso	1	From 'ossum' a variant of 'bone'	noun	4	135.89	71.80	26.39	103.57	66.79
boot	caliga	bota	5	Greek origin (from LL)	noun	6	1.08	7.20	7.86	47.38	20.43
booty, spoils	praeda	botín	5	Provencal origin	noun	20	186.25	7.20	8.46	15.72	10.38
bore	perforo	perforar	0	No change to stem.	verb	9	15.30	16.00	4.63	13.84	11.47
born	nascor	nacer	2	Deponent verb becoming active	verb	4	309.73	234.80	150.09	197.91	194.32
boy	puer	niño	4	A nursery word (De Silva)	noun	2	403.73	160.20	266.33	579.05	331.56
bracelet	armillae	pulsera	2	"pulso" + "era"	noun	6	0.00	3.60	1.01	10.69	5.02
braid	spira	trenza	4	From *tres "three"*	noun	6	8.49	0.60	0.60	22.85	7.81
brain	cerebrum	cerebro	1	LA ending vowel added to SP stem.	noun	4	41.83	50.40	58.62	55.98	54.98
branch	ramus	rama	1	LA ending vowel added to SP stem.	noun	8	121.67	100.20	29.61	85.75	71.74
brave	fortis	bravo	4	From *barbarus* "foreign, barbarous"	adjective	16	453.45	13.20	32.03	26.42	23.82
bread	panis	pan	0	No change to stem.	noun	5	46.45	15.60	26.79	89.31	43.23
break	rumpo	romper	0	CL short stressed 'u' > VL closed 'o', which remained 'o' in SP.	verb	9	197.92	81.20	86.43	190.99	118.50
break wind	pedo	peer	0	Intervocalic loss of 'd'.	verb	4	29.12	0.80	0.20	0.00	0.34
breakfast	ientaculum	desayuno	4	From *jejunus* "fasting, hungry"	noun	5	0.00	0.60	8.86	33.33	13.98
breast (of woman)	mamma	teta	5	Old French origin	noun	4	26.06	0.40	0.40	10.27	3.60
breathe	spiro	respirar	4	Generalization: From *respirare* "to breathe again"	verb	4	65.38	12.80	22.76	106.71	46.56
brick	later	ladrillo	4	From the diminutive of *later* "brick"	noun	7	178.72	24.60	8.06	37.53	23.21
bring	fero	traer	4	From *trahere* "to pull, drag"	verb	10	1473.02	45.40	116.85	408.82	187.12

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
broken	perditus	roto	1	From the past participle	adjective	9	0.00	10.20	9.67	51.15	23.28
bronze	aes	bronce	5	Italian origin	noun	9	0.00	75.80	15.31	37.32	42.96
brother	frater	hermano	4	From *germanus* "having the same parents"	noun	2	481.90	121.00	154.12	411.33	226.15
brush	peniculus	cepillo	4	From *cippus* "stake, post, pillar"	noun	6	0.00	3.80	1.41	8.18	4.41
build	construo	construir	0	No change to stem.	verb	9	12.67	343.40	160.97	81.76	197.24
bunch	fasciculus	manojo	4	From *manipulus* "a handful"	noun	5	2.87	0.60	2.82	11.95	5.02
burn (intrans)	ardeo	quemar	4	From *cremare* "to burn up, consume by fire"	verb	1	219.20	32.60	34.05	82.60	49.28
burn (trans)	cremo	quemar	2	No reason for cr > qu	verb	1	52.35	32.60	34.05	82.60	49.28
bury (the dead)	sepelio	enterrar	4	From *in* "in", *terra* "earth" combined with a verbalizing suffix	verb	4	49.00	29.40	18.53	74.43	40.32
butterfly	papilio	mariposa	4	From the name *María*	noun	3	3.35	16.40	5.24	48.22	22.94
buttocks	natis	nalga	2	From an adjectival form derived from LA noun	noun	4	28.09	1.00	0.81	34.59	11.81
button	globulus	botón	5	French origin	noun	6	1.43	1.20	4.23	21.38	8.76
buy	emo	comprar	4	From *comparare* "to prepare, provide, match, obtain"	verb	11	168.78	41.00	143.84	228.73	136.42
calf	vitulus	ternero	4	From *ternerum* "tender, delicate, young"	noun	3	29.32	0.20	0.60	5.66	2.10
call (=summon)	voco	llamar	4	Generalization: From *clamare* "to cry out, shout"	verb	18	1078.89	601.61	370.28	788.91	584.31
call (a name)	nomino	llamar	4	From *clamare* "to cry out, shout"	verb	18	279.08	601.61	370.28	788.91	584.31
calm (of sea)	tranquillus	tranquilo	2	No reason for ll > l development	adjective	1	33.58	5.00	37.87	159.12	65.97

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
camel	camelus	camello	5	Greek origin (Presence of 'll' in Spanish indicates semi-learned or borrowed origin)	noun	3	8.77	9.80	7.86	6.92	8.21
candle	candela	vela	4	From the verb *vigilare* "to be awake, watch"	noun	7	2.87	39.60	13.50	79.25	43.64
captive, prisoner	captivus	cautivo	0	CL 'pt' cluster > SP 'ut'.	noun	20	94.58	1.60	2.22	4.19	2.65
carpenter	faber	carpintero	4	Generalization: From *carpentarius* "carriage-maker"	noun	9	11.71	8.40	2.82	11.53	7.53
carry (bear)	porto	llevar	4	From *levare* "to raise, lift"	verb	10	164.22	590.21	639.03	1008.41	742.05
carve	sculpo	esculpir	0	Regular 'e' insertion before word initial 's' + consonant clusters.	verb	9	2.03	13.00	2.22	13.42	9.50
cast (metals)	fundo	fundir	5	Semi-learned (late borrowing)	verb	9	37.01	32.20	6.25	19.50	19.34
cat	feles	gato	5	Afro-Asiatic origin	noun	3	4.66	41.40	25.38	103.57	56.13
catch (ball)	excipio	atrapar	5	French origin	verb	10	302.14	24.80	26.79	61.01	37.19
cattle	boves	ganado	5	Germanic origin	noun	3	0.00	0.00	0.00	0.00	0.00
cause	causa	causa	1	LA ending vowel added to SP stem.	noun	17	1920.66	222.80	185.34	106.50	172.53
cave	caverna	caverna	1	LA ending vowel added to SP stem.	noun	1	18.53	9.60	2.22	9.85	7.19
cease, stop	desisto	desistir	0	No change to stem.	verb	14	24.14	1.00	7.86	12.79	7.13
centipede	centipeda	ciempiés	2	Influence of "cien" with the loss of final 't'	noun	3	0.72	0.60	0.00	2.10	0.88
chain	catena	cadena	1	Intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	9	42.19	113.80	85.02	49.27	83.21
chair	sella	silla	1	Influence of the palatalization of [ll] closes e > i	noun	7	30.36	24.60	24.58	161.64	68.96

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
change	muto	cambiar	5	Celtic origin	verb	12	153.89	153.80	221.20	334.81	235.11
charcoal	carbo	carbón	0	The 'n' is part of the CL stem (*carbo, carbonis*), which does not appear in the nominative form.	noun	1	23.55	108.20	9.67	16.56	45.34
cheap	vilis	barato	5	French origin	adjective	11	90.83	25.80	19.14	28.51	24.43
cheek	gena	mejilla	4	From *maxilla* "jaw"	noun	4	28.85	1.20	2.22	96.23	32.31
cheese	caseus	queso	2	No reason for c > qu	noun	5	19.60	22.40	6.25	33.96	20.70
chest	pectus	pecho	1	Regular 'ct' > 'ch' development; LA ending vowel added to SP stem.	noun	4	455.60	13.40	18.53	243.40	89.59
chew	manduco	masticar	5	Latin origin (late borrowing)	verb	5	7.89	2.40	0.81	25.58	9.37
chicken	pullus	pollo	2	No reason for short-u > o	noun	3	11.69	5.60	23.37	25.58	18.05
chief, chieftain	princeps	jefe	5	French origin	noun	19	229.52	143.20	264.11	114.05	174.50
chimney	caminus	chimenea	5	French origin	noun	7	7.05	12.80	7.45	25.58	15.14
chin	mentum	barba	4	From *barba* 'beard'	noun	4	111.10	10.20	14.30	70.65	31.15
chisel	scalprum	cincel	5	Old French origin	noun	9	6.21	0.20	0.60	4.40	1.70
choose	eligo	elegir	0	CL short 'i' > VL closed 'e', which remained 'e' in SP.	verb	16	123.68	274.60	143.24	128.72	183.12
chop, hew	dolo	tajar	4	From *talea* "twig, rod, stick, cutting"	verb	9	136.61	0.40	1.61	2.31	1.43
circle	circulus	círculo	1	LA ending vowel added to SP stem.	noun	12	29.28	65.40	46.74	80.71	64.07
citizen, subject	civis	ciudadano	2	Sporadic loss of -v-, derived from "city"	noun	19	525.04	84.20	119.26	11.11	72.35
citrus fruit	citrea	cítrico	2	Unexplained additional syllable	noun	8	0.00	2.40	1.41	0.21	1.36
city, town	civitas	ciudad	2	Sporadic loss of -v-	noun	19	727.39	1063.41	510.50	366.47	651.50

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
claw	unguis	garra	5	Gaulish origin	noun	4	81.03	7.00	6.65	23.48	12.22
clean	mundus	limpio	4	From *limpidus* "clear"	adjective	15	83.89	13.80	31.43	90.15	44.46
clear, plain	clarus	claro	0	No change to stem.	adjective	17	271.22	116.60	294.94	438.80	280.99
clever	intellegens	listo	5	Germanic origin	adjective	16	0.00	7.60	38.48	62.69	35.84
climb	scando	subir	4	From *subire* "to go under, enter"	verb	10	29.64	37.60	122.29	373.59	174.91
cloak	amictus	manto	5	Gaulish origin	noun	6	30.84	22.80	11.28	26.21	20.02
clock, timepiece	horologium	reloj	5	Catalan origin	noun	14	5.98	26.80	25.59	111.53	53.82
cloth	textum	tela	4	From *tela* "web, net"	noun	6	6.37	38.20	27.00	71.49	45.20
clothing, clothes	vestitus	vestido	1	Intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	6	10.40	11.00	9.67	109.23	42.35
club	clava	garrote	5	French origin	noun	20	11.19	1.80	2.01	2.52	2.10
coat	paenula	chaqueta	5	French origin	noun	6	5.02	4.60	2.82	27.46	11.40
cock, rooster	gallus	gallo	1	LA ending vowel added to SP stem.	noun	3	43.68	5.00	10.07	55.14	22.94
cockroach	blatta	cucaracha	4	From *cuca* "caterpillar, moth"	noun	3	1.55	4.40	3.42	16.35	7.94
coin	moneta	moneda	1	Intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	11	0.00	128.20	54.80	62.48	82.19
cold (illness)	gravedo	catarro	5	Greek origin	noun	4	6.45	1.00	0.20	3.35	1.49
collar	torquis	collar	4	From *collum* "neck"	noun	6	11.23	11.40	3.02	27.46	13.78
collect, gather	colligo	coleccionar	1	From the past participle	verb	12	22.63	1.20	4.84	7.34	4.41
color	color	color	0	No change to stem.	noun	15	229.81	287.00	141.02	346.13	256.96
comb	pecten	peine	2	Unexplained loss of 'ct'	noun	6	0.00	2.00	1.21	8.18	3.73

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
come	venio	venir	0	No change to stem.	verb	10	1838.95	104.20	379.55	986.82	482.71
come back	revenio	volver	4	From *volvere* "to roll, turn, turn round"	verb	10	1.67	262.00	450.87	1526.25	734.92
command	iubeo	mandar	4	From *mandare* "to commit to one's charge, entrust"	verb	19	0.00	29.00	57.82	207.76	96.58
conceive	concipio	concebir	0	Intervocalic [p] voicing; CL short 'i' > VL closed 'e', which remained 'e' in SP.	verb	4	76.61	51.80	37.47	30.82	40.18
condemn	condemno	condenar	2	Missing nasal palatalization	verb	21	75.77	46.40	59.43	50.94	52.26
conspiracy, plot	conspiratio	conspiración	0	CL [t] followed by jod results in SP 'c' or 'z' (same pronunciation).	noun	19	7.65	15.40	12.49	8.81	12.28
cook	coquo	cocer	0	CL 'qu' (velar stop with lip rounding) lost labial articulation before any vowel other than 'a'.	verb	5	115.45	15.60	3.22	14.05	10.93
cookhouse	popina	cocina	4	From *coquere* "to cook"	noun	7	9.56	8.40	23.97	154.09	60.81
corner	angulus	ángulo	1	LA ending vowel added to SP stem.	noun	12	47.33	63.40	17.12	26.21	35.77
corpse	cadaver	cadáver	0	No change to stem.	noun	4	40.88	10.80	38.68	80.09	42.62
cotton	gossypium	algodón	5	Arabic origin	noun	6	0.00	61.60	3.63	25.37	30.34
cough	tussio	toser	0	Geminate; CL short 'u' > VL closed 'o', which remained 'o' in SP.	verb	4	14.90	1.40	0.40	25.37	8.82
count	numero	contar	4	From *computare* "to reckon, compute, sum up"	verb	13	235.93	317.80	358.20	551.59	407.10
country	terra	país	5	French origin	noun	19	1127.54	1583.62	1655.39	226.00	1168.28
court	tribunal	tribunal	0	No change to stem.	noun	21	44.70	74.00	60.03	11.11	48.94
cousin	consobrinu	primo	3	From *consobrinus primus* "first	noun	2	4.06	9.00	10.48	73.80	30.47

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
	s			cousin"							
cover	operio	cubrir	4	From *cooperire* "to cover completely"	verb	12	96.01	186.80	117.85	246.34	182.85
crawl	repo	arrastrar	4	From *rastro* "rake"	verb	10	118.32	42.60	35.46	167.51	80.63
crooked	curvus	curvo	0	No change to stem.	adjective	12	33.70	6.00	0.40	2.10	2.85
crop, harvest	messis	cosecha	4	From the verb *colligere* "to gather together"	noun	8	30.27	47.40	23.57	22.01	31.15
crouch	conquinisco	agachar	4	Probably from *cogere* "to collect"	verb	10	0.00	0.40	1.61	32.91	11.33
crush, grind	molo	moler	0	No change to stem.	verb	5	41.69	11.20	5.04	7.13	7.81
cry, weep	ploro	llorar	0	Word-initial 'pl' in CL > 'll' in SP.	verb	16	39.20	5.80	29.82	326.84	117.83
cultivate	colo	cultivar	4	From the adjective *cultus* "cultivated" and the suffix *-ivus* "tending towards"	verb	8	47.00	126.40	24.58	26.21	59.66
cure, heal	curo	curar	0	No change to stem.	verb	4	381.42	11.40	10.27	42.56	21.11
custom	consuetudo	costumbre	2	From a contracted form	noun	19	173.54	87.20	47.75	149.90	94.21
cut	seco	cortar	4	Generalization: From *curtare* "to cut off"	verb	9	73.15	87.40	52.18	201.47	112.46
cut down (to fell a tree)	demitto	talar	5	Germanic origin	verb	9	115.37	3.00	1.41	1.89	2.10
dance	salto	bailar	5	Old Provencal origin	verb	10	0.00	15.40	36.26	128.10	58.91
dare	audeo	atreverse	3	From *tribuere sibi* "to attribute to oneself (the ability to do something)"	verb	16	308.92	1.80	26.39	118.45	47.85

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
dark (in color)	fuscus	oscuro	2	Sporadic addition of a syllable from VL	adjective	15	35.49	67.80	52.18	238.79	117.89
darkness	tenebra	oscuridad	2	From *obscurus* (adj.) "covered"	noun	1	0.00	13.20	8.66	144.87	54.30
dawn	aurora	alba	4	From *alba* "white"	noun	14	25.10	4.40	7.25	23.69	11.61
deaf	surdus	sordo	0	CL short 'u' > VL closed 'o', which remained 'o' in SP.	adjective	4	23.90	7.80	3.83	37.53	16.09
deep	profundus	profundo	0	No change to stem.	adjective	12	31.31	145.00	104.76	151.58	133.57
deer	cervus	ciervo	1	Diphthongization of short stressed 'e' > 'ie'; LA ending vowel added to SP stem.	noun	3	33.82	26.80	2.62	3.56	11.13
defeat	clades	derrota	5	French origin	noun	20	176.89	81.60	77.16	27.67	62.65
defend	defendo	defender	0	No change to stem.	verb	20	371.14	121.60	122.29	96.44	113.69
defendant	reus	acusado	4	From *accusare* "to call to account"	noun	21	354.98	6.80	15.92	7.13	9.98
demon (evil spirit)	daemoniu m	demonio	1	Development of 'ae' > 'e'; LA ending vowel added to SP stem.	noun	22	0.00	13.40	6.85	46.33	21.85
deny	nego	negar	0	No change to stem.	verb	18	533.41	78.20	153.11	160.59	130.11
descendants	progenies	descendient e	4	From *descendere* "to descend, climb down"	noun	2	23.90	43.40	6.65	13.84	21.45
destroy	destruo	destruir	0	No change to stem.	verb	11	11.00	121.00	68.90	75.89	88.84
dew	ros	rocío	4	From the adjective *roscidus* "dewy"	noun	1	0.00	3.60	3.83	29.56	12.08
die	morior	morir	2	Deponent > active	verb	4	339.43	202.80	251.62	689.54	376.83
difficult	difficilis	difícil	0	Geminate simplification.	adjective	17	201.63	103.40	227.25	175.69	168.53
dig	fodio	cavar	4	From *cavare* "to hollow out"	verb	8	94.98	4.80	2.01	15.72	7.40
dinner	cena	comida	4	From the verb *comedere* "to eat up"	noun	5	72.55	35.60	32.44	147.38	70.72

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
dirty, soiled	sordidus	sucio	4	From *sucidus* "wet, juicy"	adjective	15	66.09	3.00	22.56	110.69	44.46
disappear	evanesco	desaparecer	2	"des" + "aparecer"	verb	10	15.06	87.20	99.92	222.23	135.20
ditch	fossa	fosa	5	Retention of word-initial [f] indicates late borrowing from Latin	noun	8	25.06	9.20	5.44	12.79	9.09
divide	divido	dividir	0	No change to stem.	verb	12	190.99	256.00	51.78	29.77	113.96
dog	canis	perro	4	Onomatopoeic origin (sound made by shepherds) (De Silva)	noun	3	49.09	49.40	32.64	197.49	91.70
dolphin	delphinus	delfín	5	Semi-learned	noun	3	0.00	13.20	13.70	1.89	9.71
doorpost, jamb	postis	jamba	5	Old French origin	noun	7	30.20	1.60	0.00	0.84	0.81
doubt	dubium	duda	4	From *dubdar* "to hesitate"	noun	17	0.00	51.60	220.60	201.68	157.12
dough	massa	masa	1	Geminate simplification; LA ending vowel added to SP stem.	noun	5	8.29	189.40	60.03	56.40	102.76
down, below	infra	abajo	3	From *a bajo* "to below"	adverb	12	44.58	45.80	43.11	203.57	95.97
dream	somnio	soñar	0	CL [mnj] results in SP 'ñ'.	verb	4	31.19	3.40	34.65	138.16	57.56
drip	stillo	gotear	2	From "drip" + verbalizing suffix	verb	10	10.40	0.60	0.20	11.95	4.14
drop	solvo	soltar	4	From *soltus* "loose"	verb	10	350.79	7.00	17.33	135.22	51.99
drown	demergo	ahogar	4	From *ob-* "against" and *foces* "throat"	verb	4	11.79	2.20	10.07	62.89	24.50
drum	tympanum	tambor	5	Arabic origin	noun	18	27.13	26.80	6.65	36.06	23.01
dry	siccus	seco	0	Geminate simplification; CL short 'i' > VL open 'e', which remained 'e' in SP.	adjective	15	85.18	76.20	27.20	163.74	88.03
duck	anas	pato	5	Arabic origin	noun	3	1.45	14.40	4.63	20.96	13.24
dye	tingo	teñir	0	CL [gn] results in SP 'ñ'.	verb	6	75.97	14.20	7.25	37.11	19.28

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
ear	auris	oreja	4	From *auricula*, diminutive of *auris* "ear"	noun	4	334.42	18.80	7.66	99.16	41.06
early	mature	temprano	4	From *temporaneus* "timely"	adjective	14	0.00	32.00	49.76	91.83	57.35
earn	mereo	ganar	5	Gothic origin	verb	11	203.63	213.60	410.78	248.22	291.24
earring	inaures	pendiente	4	From *pendere* "to hang"	noun	6	0.36	18.00	29.61	29.56	25.66
earth (ground, soil)	solum	suelo	1	Diphthongization of short stressed 'o' > 'ue'; LA ending vowel added to SP stem.	noun	1	240.62	290.40	52.18	282.61	207.62
earth, land	terra	tierra	1	Diphthongization of short stressed 'e' > 'ie'; LA ending vowel added to SP stem.	noun	1	1127.54	663.21	233.09	449.49	449.11
earthquake	concussio	terremoto	5	Italian origin	noun	1	0.24	32.00	9.87	15.09	19.07
east	oriens	este	5	Old English origin (via French)	noun	12	43.62	0.00	0.81	0.42	0.41
easy	facilis	fácil	0	No change to stem.	adjective	17	629.86	50.20	102.74	137.32	96.11
eat	edo	comer	4	Generalization: From *comedere* "to eat up, eat thoroughly"	verb	5	86.61	48.00	59.63	396.87	164.86
edge	ora	orilla	4	From the diminutive of *ora* "edge, border"	noun	12	168.20	46.00	20.95	88.05	51.18
egg	ovum	huevo	1	Diphthongization of short stressed 'o' > 'ue'; LA ending vowel added to SP stem.	noun	5	59.28	80.00	20.35	48.22	49.61
egg yolk	vitellus	yema	4	From *gemma* "bud, gem"	noun	5	6.10	9.60	1.41	13.42	8.08
eight	octo	ocho	0	CL consonant cluster 'ct' results in SP 'ch'.	number	13	47.81	64.40	154.72	103.57	107.51
elephant	elephantus	elefante	5	Semi-learned (late borrowing)	noun	3	60.72	32.40	8.66	10.48	17.31

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
eleven	undecim	once	2	No reason for u > o	number	13	7.17	17.80	44.93	51.78	37.94
embers	favilla	brasa	5	Germanic origin	noun	1	17.21	1.00	2.22	18.66	7.13
embrace	amplector	abrazar	4	From the noun *bracchium* "arm"	verb	16	83.66	6.00	18.74	134.39	51.85
empty	vacuus	vacío	2	Sporadic loss of -v-	adjective	13	135.77	15.80	22.76	124.32	53.28
end	finis	fin	0	No change to stem.	noun	12	411.38	332.20	423.67	542.99	431.26
enough	satis	bastante	5	Greek origin	adverb	13	148.30	98.40	85.02	159.12	113.55
enter	intro	entrar	0	CL short 'i' > VL open 'e', which remained 'e' in SP.	verb	10	113.26	202.60	224.63	547.39	321.65
evening	vespera	atardecer	4	From the adverb *tarde* "slowly, late"	noun	14	2.87	2.80	3.42	51.99	18.94
explain	explico	explicar	0	No change to stem.	verb	17	70.32	121.20	374.51	277.79	257.24
extinguish	extinguo	extinguir	0	No change to stem.	verb	1	0.00	24.00	7.86	16.14	16.02
eyebrow	supercilium	ceja	4	From *cilium* "eyelid"	noun	4	26.29	2.20	4.63	42.98	16.22
face	facies	cara	5	Greek origin (from LL)	noun	4	153.58	52.40	93.48	608.19	246.17
fairy, elf	nympha	hada	4	From *fata* "fate"	noun	22	29.76	13.20	4.23	11.32	9.57
faithful	fidelis	fiel	0	Intervocalic loss of 'd'.	adjective	16	65.26	11.20	17.53	23.90	17.44
fall	cado	caer	0	Intervocalic loss of 'd'.	verb	10	468.52	119.80	213.35	679.89	332.64
family	familia	familia	1	LA ending vowel added to SP stem.	noun	2	158.00	355.20	277.61	385.13	338.75
fan	flabellum	abanico	4	From *evannere* "to winnow"	noun	9	0.36	1.40	4.63	3.98	3.33
fan	ventilo	ventilar	0	No change to stem.	verb	9	5.62	11.20	10.27	17.19	12.83
far	procul	lejos	4	From *laxius*, the comparative of	adverb	12	271.31	18.20	49.76	187.01	83.48

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				laxe "widely, spaciously"							
farmer	agricultor	agricultor	0	No change to stem.	noun	8	0.24	29.20	25.18	5.03	20.02
fast	ieiuno	ayunar	2	Likely dissimilation	verb	22	0.00	2.00	1.41	1.47	1.63
father-in-law	socer	suegro	4	From *socrus* "mother-in-law"	noun	2	69.08	2.00	1.01	15.09	5.90
fear, fright	metus	miedo	1	Diphthongization of short stressed 'e' > 'ie'; intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	16	478.91	17.00	43.72	330.41	127.46
feather	pluma	pluma	1	LA ending vowel added to SP stem.	noun	4	19.72	26.60	14.51	59.54	33.19
feel	sentio	sentir	0	No change to stem.	verb	15	651.55	76.60	213.55	1207.37	488.81
felt	coacta	fieltro	5	Germanic origin	noun	6	0.00	4.00	0.60	7.97	4.14
female	femina	femenino	4	From the adjective *femininus* "feminine"	adjective	3	178.08	75.20	11.28	28.72	38.62
fence	saepes	valla	4	From *valla* "rampart, entrenchment"	noun	8	229.43	9.20	10.27	5.45	8.35
fever	febris	fiebre	1	Diphthongization of short stressed 'e' > 'ie'; LA ending vowel added to SP stem.	noun	4	135.89	35.20	15.92	51.15	33.87
fig	ficus	higo	1	Word-initial 'f' > 'h' development; intervocalic [k] voicing; LA ending vowel added to SP stem.	noun	5	74.10	4.60	0.20	6.92	3.87
fight	pugno	luchar	4	Generalization: From *luctare* "to wrestle" (a gymnastic term)	verb	20	346.96	70.80	66.88	102.10	79.61
fin (dorsal)	pinna	aleta	4	From diminutive of *ala* "wing"	noun	3	0.08	17.40	1.01	5.45	8.01
find	invenio	hallar	4	From *afflare* "to blow on, breathe on"	verb	11	503.05	139.60	113.82	147.59	133.50
fine	multa	multa	1	LA ending vowel added to SP stem.	noun	21	382.60	2.00	33.44	7.34	14.32

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
finger	digitus	dedo	2	No reason for i > e	noun	4	157.64	32.80	37.27	313.43	125.16
finish	finio	acabar	4	From the noun *caput* "head"	verb	14	224.46	163.00	200.05	378.21	245.15
fir	abies	abeto	2	No reason for loss of i	noun	8	18.17	9.20	0.20	1.68	3.73
fire	incendium	incendio	1	LA ending vowel added to SP stem.	noun	1	102.31	25.40	100.93	28.93	51.99
firefly	cicindela	luciérnaga	4	From *lucere* "to shine"	noun	3	0.48	0.20	1.41	13.63	4.95
fireplace, hearth	focus	hogar	2	VL origin for the 'r', diminutive/extra syllable	noun	7	58.80	38.60	70.71	75.05	61.22
first	primus	primero	4	Generalization: From *primarius* "of the first rank"	number	13	2003.49	1284.41	964.59	539.22	935.41
fishhook	hamus	anzuelo	4	From the diminutive of *hamus* "hook, fishhook"	noun	20	14.58	2.60	1.41	4.19	2.71
fishnet	rete	red	0	Intervocalic [t] voicing.	noun	20	17.25	162.20	109.19	35.85	103.44
five	quinque	cinco	2	"o" from analogy with "four"	number	13	77.69	198.40	402.31	246.13	282.55
flat	planus	plano	0	No change to stem.	adjective	12	44.75	55.60	17.73	20.34	31.42
flay, skin	decutio	desollar	4	From *follis* "leather bag"	verb	9	6.97	0.00	0.00	1.68	0.54
float	fluito	flotar	5	French origin	verb	10	16.25	14.60	8.66	77.99	33.12
flour	farina	harina	1	Word-initial 'f' > 'h' development; LA ending vowel added to SP stem.	noun	5	68.13	15.20	7.66	12.58	11.81
flow	fluo	fluir	0	No change to stem.	verb	10	142.11	57.20	12.89	26.42	32.31
flute	tibia	flauta	5	Old Provencal origin	noun	18	30.84	13.20	20.55	13.84	15.88
fly	volo	volar	0	No change to stem.	verb	10	27.81	23.80	10.88	38.99	24.37
foal, colt	pullus	potro	2	Irregular development through VL **pullitrus* "colt"	noun	3	11.69	1.80	2.22	7.55	3.80
foam	spuma	espuma	1	e' insertion before word-initial 's' +	noun	1	41.11	4.20	5.84	29.14	12.83

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				consonant clusters; LA ending vowel added to SP stem.							
follow	sequor	seguir	2	Deponent verb become active	verb	10	1259.64	585.01	655.95	1110.93	779.17
food	alimentum	alimento	1	LA ending vowel added to SP stem.	noun	5	53.07	201.80	61.04	41.09	102.35
footprint	vestigium	huella	4	From *fullo* "fuller (of cloth)" (De Silva)	noun	4	158.48	19.00	28.61	57.23	34.61
forbid	prohibeo	prohibir	0	No change to stem.	verb	18	204.14	77.20	57.42	61.22	65.36
forge	fabrico	forjar	5	French origin	verb	9	10.11	14.60	9.07	10.27	11.33
forget	obliviscor	olvidar	2	From the *oblitus,* the past participle of *oblivisci*; underwent metathesis in VL	verb	17	98.96	22.00	123.90	341.52	159.77
forgive	perdono	perdonar	0	No change to stem.	verb	16	0.00	3.20	22.76	98.54	40.66
fork	furca	tenedor	2	"tener" + "dor"	noun	8	6.45	0.20	1.61	10.27	3.94
fortress	castrum	castillo	4	From *castellum*, diminutive of *castrum* "fortified camp"	noun	20	229.92	42.40	60.24	34.38	45.81
four	quattuor	cuatro	2	Metathesis of the vowel and r	number	13	198.40	342.80	505.87	302.73	384.77
fowl	avis	ave	1	LA ending vowel added to SP stem.	noun	3	82.03	133.20	19.34	46.96	66.92
fox	vulpes	zorro	5	External origin, possibly Basque	noun	3	0.00	32.60	2.62	8.60	14.73
freshwater eel	anguilla	anguila	1	Development of 'll' > 'l' occurs only under influence of a following yod (Mendeloff)	noun	3	3.82	3.60	0.40	3.35	2.44
friend, companion	amicus	amigo	1	Intervocalic [k] voicing; LA ending vowel added to SP stem.	noun	19	233.51	49.00	191.99	530.83	253.16

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
frog	rana	rana	1	LA ending vowel added to SP stem.	noun	3	27.25	22.80	2.22	12.37	12.49
front	frons	frente	2	No reason for o > e	noun	12	64.66	12.80	14.51	5.45	11.00
fruit	fructus	fruto	5	Semi-learned	noun	5	79.24	55.80	46.13	39.83	47.37
full	plenus	lleno	0	Word-initial 'pl' in CL > 'll' in SP.	adjective	13	292.58	26.00	52.38	178.83	84.37
fur	pellis	piel	1	Word final LL > L in O. SP	noun	6	36.25	118.40	36.06	260.80	136.76
garden	hortus	jardín	5	Old French origin	noun	8	67.41	53.80	46.34	157.24	84.77
get, obtain	adipiscor	obtener	4	From *obtinere* "to hold completely"	verb	11	70.28	471.21	290.10	66.25	279.09
ghost, phantom	phantasma	fantasma	5	Greek origin	noun	22	0.24	4.40	21.35	66.04	30.07
gill	branchiae	agalla	4	From *glandula* "glandular swelling"	noun	3	1.43	3.00	1.01	4.19	2.71
girl	puella	niña	4	A nursery word (De Silva)	noun	2	163.26	11.60	47.95	229.15	94.27
give	do	dar	0	No change to stem.	verb	11	860.75	916.21	1453.33	2430.25	1587.33
give back	reddo	devolver	4	From *devolvere* "to roll or tumble off/down"	verb	11	557.91	30.60	64.87	96.65	63.53
glove	chirotheca	guante	5	Catalan origin	noun	6	0.00	8.00	6.25	31.66	15.07
glue	gluten	cola	5	Greek origin	noun	9	0.00	54.40	36.67	66.67	52.40
gnat	culex	jején	5	Arawak origin	noun	3	4.78	0.00	0.00	0.00	0.00
go away, depart	discedo	salir	4	From *salire* "to jump"	verb	10	172.82	178.60	433.94	1150.55	579.29
go down	descendo	bajar	5	Greek origin	verb	10	170.71	20.20	89.45	332.50	144.64
go out	exeo	salir	4	From *salire* "to jump"	verb	10	294.73	178.60	433.94	1150.55	579.29
go up	ascendo	subir	4	From *subeo* "to go under, enter"	verb	10	49.12	37.60	122.29	373.59	174.91
gold	aurum	oro	1	Regular 'au' > 'o' development; LA ending vowel added to SP stem.	noun	9	474.49	207.20	84.81	176.31	155.97

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
good	bonus	buen	0	CL short stressed 'o' diphthongized to SP 'ue'.	adjective	16	1163.38	188.60	506.27	627.48	437.71
good fortune, luck	fortuna	fortuna	1	LA ending vowel added to SP stem.	noun	16	530.78	15.00	39.89	61.85	38.55
goose	anser	ganso	5	Gothic origin	noun	3	15.06	3.80	1.41	4.19	3.12
gourd	cucurbita	calabaza	5	Iberian origin	noun	8	8.61	7.20	1.41	6.29	4.95
grain (barley, oats etc)	granum	grano	1	LA ending vowel added to SP stem.	noun	8	35.26	47.60	20.15	26.00	31.36
grandfather	avus	abuelo	4	From the diminutive form of *avia* "grandmother"	noun	2	94.97	14.60	28.81	264.16	100.18
grandmother	avia	abuela	4	From the diminutive form of *avia* "grandmother"	noun	2	12.43	14.60	28.81	264.16	100.18
grandson	nepos	nieto	4	From *neptis* "granddaughter, niece"	noun	2	71.11	22.20	25.18	43.82	30.20
grape	uva	uva	1	LA ending vowel added to SP stem.	noun	5	45.18	22.60	4.23	14.26	13.71
grass	herba	hierba	1	Diphthongization of short stressed 'e' > 'ie'; LA ending vowel added to SP stem.	noun	8	209.87	44.20	11.28	35.64	30.34
grasshopper	gryllus	saltamontes	4	Possibly from *saltus* "jump" and *montem* accusative of "mountain"	noun	3	2.63	6.00	0.20	2.10	2.78
grave, tomb	tumba	tumba	1	LA ending vowel added to SP stem.	noun	4	0.00	63.40	32.03	48.01	47.85
grease, fat	adeps	grasa	4	From the adjective *crassus* "fat, gross, thick"	noun	5	0.00	50.00	11.68	20.13	27.42
greedy	avarus	avaro	0	No change to stem.	adjective	16	40.99	0.40	0.81	2.31	1.15
groan	gemo	gemir	0	No change to stem.	verb	16	112.71	1.20	1.01	36.06	12.42

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
guard, sentinel	custodia	guardia	5	Gothic origin	noun	20	88.44	10.60	31.83	76.10	38.96
guess	conicio	adivinar	4	From *ad-* "thoroughly" + *divinare* "divine"	verb	17	117.13	1.40	7.66	97.49	34.61
guest	conviva	invitado	4	From the verb *invitare* "to invite"	noun	19	26.73	2.00	22.36	29.56	17.78
guilty	sons	culpable	4	From the noun *culpa* "blame, fault"	adjective	21	15.30	0.60	2.82	10.06	4.41
gums	gingiva	encía	2	Sporadic loss of -v-, among other things	noun	4	22.47	3.00	0.60	4.61	2.71
half	dimidius	medio	5	Semi-learned	adjective	13	39.12	274.40	149.48	181.98	202.39
hammer	malleus	martillo	5	Medieval Latin origin (late borrowing)	noun	9	4.30	10.20	9.47	13.21	10.93
hand	manus	mano	1	LA ending vowel added to SP stem.	noun	4	1084.39	227.60	308.43	1423.10	641.87
handkerchief	sudarium	pañuelo	2	pano + diminutive suffix	noun	6	2.63	4.40	3.22	62.69	22.87
harbor, port	portus	puerto	1	Diphthongization of short stressed 'o' > 'ue'; LA ending vowel added to SP stem.	noun	10	215.13	165.80	63.86	85.12	105.34
hard	durus	duro	0	No change to stem.	adjective	15	224.31	95.00	77.96	126.21	99.36
harm, injure, damage	noceo	dañar	4	From from the noun *damnum* "harm, injury"	verb	11	243.70	25.00	26.79	9.22	20.50
hasten, hurry	propero	precipitar	4	From *praecipitare* "to cast down"	verb	14	100.87	15.00	10.48	36.48	20.43
hate	odium	odio	1	LA ending vowel added to SP stem.	noun	16	104.81	6.20	19.14	64.99	29.59
have	habeo	tener	4	From *tenere* "to hold"	verb	11	3186.72	2435.23	3542.87	4637.44	3521.34

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
hawk	accipiter	halcón	4	Generalization: from *falco* "falcon"	noun	3	0.00	5.20	1.81	3.98	3.67
hay	fenum	heno	1	Word-initial 'f' > 'h' development; LA ending vowel added to SP stem.	noun	8	0.00	4.20	1.21	3.77	3.05
headband, headdress	fascia	tocado	5	Germanic origin	noun	6	12.91	0.00	0.00	0.00	0.00
heart	cor	corazón	4	Generalization: From a VL derivation of *cor*, **coratione* "big heart"	noun	4	48.92	58.20	86.22	305.46	147.69
heavy	gravis	pesado	4	From *pensare,* frequentative of *pendere* "to weigh"	adjective	15	649.46	67.40	21.96	44.45	44.66
heel	talus	talón	2	Influence of VL?	noun	4	160.75	2.60	5.84	13.84	7.33
hell	inferna	infierno	1	Diphthongization of short stressed 'e' > 'ie'; LA ending vowel added to SP stem.	noun	22	0.00	11.80	15.11	50.32	25.38
helmet	cassis	casco	4	From *quassare* "to shake, break"	noun	20	18.13	31.00	19.14	37.11	28.98
hen	gallina	gallina	1	LA ending vowel added to SP stem.	noun	3	13.74	9.80	4.63	51.36	21.52
herdsman	pastor	pastor	0	No change to stem.	noun	3	50.55	24.60	23.77	14.89	21.18
heron	ardea	garza	5	Celtic origin	noun	3	5.34	0.60	2.42	2.52	1.83
hide, conceal	occulto	ocultar	0	Geminate simplification	verb	12	69.20	17.80	38.28	102.52	52.13
hinder, prevent	impedio	impedir	0	No change to stem.	verb	19	145.57	92.60	119.06	108.39	106.63
hip	coxa	cadera	4	From *cathedra* "chair"	noun	4	22.95	8.00	5.64	40.67	17.78
hoe	sarculum	azada	4	From *ascia* "axe, mason's trowel"	noun	8	1.91	1.20	0.00	3.14	1.43
hold	teneo	sostener	4	Generalization: From *sustinere* "to	verb	11	795.03	89.20	194.41	129.56	137.71

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				sustain, hold up"							
hole	foramen	hoyo	4	Generalization: From *fovea* "small pit"	noun	12	70.40	7.00	10.88	13.84	10.52
holy, sacred	sacer	sagrado	4	From *sacrare* "to consecrate, make sacred"	adjective	22	223.50	62.80	25.59	46.75	45.07
honey	mel	miel	0	Diphthongization of short stressed 'e' > 'ie'.	noun	5	127.39	8.00	13.30	42.77	21.04
hook	uncus	gancho	5	Hispano-Celtic origin	noun	12	11.63	3.20	3.42	14.89	7.06
hope	spes	esperanza	2	esperar + anza (nominalizing suffix)	noun	16	557.20	49.40	106.17	127.47	93.80
horn	cornu	cuerno	1	Diphthongization of short stressed 'o' > 'ue'; LA ending vowel added to SP stem.	noun	4	145.27	13.00	3.02	20.34	12.01
horse	equus	caballo	4	From *caballus* "inferior horse, nag"	noun	3	355.73	104.40	36.06	243.82	126.51
host	hospes	anfitrión	5	Greek origin (from name)	noun	19	95.24	3.20	16.52	11.74	10.45
hot	calidus	caliente	4	From the verb *calere* "to be warm"	adjective	15	134.96	43.20	18.94	88.68	49.75
hour	hora	hora	1	LA ending vowel added to SP stem.	noun	14	92.03	127.60	598.74	683.46	466.28
house	casa	casa	1	LA ending vowel added to SP stem.	noun	7	11.59	224.80	473.03	1510.94	724.81
how?	quo	como	1	From the ablative form.	adverb	17	121.67	6004.87	4348.51	5429.29	5260.50
howl	ululo	aullar	3	The addition of the initial *a* is from the adverb *ad* "to, towards"	verb	18	15.90	0.40	0.20	18.87	6.31
hundred	centum	cien	2	Loss of the final 't' is very recent, irregular	number	13	131.71	29.20	67.49	65.83	53.96
hunger	fames	hambre	2	Sporadic addition of a syllable from VL	noun	5	119.76	18.00	61.24	136.27	70.86

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
hunt	venor	cazar	4	From *captare* "to chase, strive after"	verb	20	30.48	19.40	4.63	24.74	16.15
husband	maritus	marido	1	Intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	2	52.79	14.20	20.35	164.16	64.82
hut	casa	choza	5	Galician or Portuguese origin	noun	7	11.59	5.00	1.81	30.19	12.08
ice	glacies	hielo	4	From *gelu* "cold, frost"	noun	1	12.43	65.00	15.71	40.67	40.52
idea, notion	idea	idea	1	LA ending vowel added to SP stem.	noun	17	3.11	276.80	243.97	294.56	271.49
idol	idolum	ídolo	1	LA ending vowel added to SP stem.	noun	22	0.48	5.80	9.07	16.77	10.45
if	si	si	0	No change to stem.	conjunction	17	1236.88	700.21	1488.99	2958.36	1697.01
innocent	innocens	inocente	2	Missing nasal palatalization	adjective	21	89.88	2.00	13.50	32.91	15.88
inquire	inquiro	inquirir	0	No change to stem.	verb	18	30.22	0.60	1.61	14.47	5.43
insane, crazy	insanus	loco	5	Arabic origin	adjective	17	51.39	3.40	6.25	56.81	21.65
insect	bestiola	insecto	4	From *insectum* "segmented"	noun	3	3.59	78.00	11.89	35.01	41.81
inside, in	intra	dentro	3	From *de intro* "of inside"	adverb	12	74.49	16.40	24.78	71.07	36.92
intestines, guts	intestinum	intestino	1	LA ending vowel added to SP stem.	noun	4	29.00	18.60	1.61	9.43	9.91
intoxicated	ebrius	borracho	5	Catalan origin	adjective	4	27.01	0.00	0.60	26.00	8.62
iron	ferrum	hierro	1	Word-initial 'f' > 'h' development; dipthongization of short stressed 'e' > 'ie'; LA ending vowel added to SP stem.	noun	9	488.23	170.20	27.20	69.18	89.32
itch	prurigo	comezón	4	From *comedere* "to eat up"	noun	4	7.41	0.00	0.20	4.61	1.56
jaw	maxilla	mandíbula	4	From *mandere* "to chew"	noun	4	20.32	24.60	5.64	17.40	15.88
jewel	gemma	joya	5	French origin	noun	6	57.01	26.40	25.38	36.69	29.39
join, unite	unio	unir	0	No change to stem.	verb	12	6.21	292.40	143.04	85.96	175.25

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
jump, leap	salio	saltar	4	From *saltare*, the frequentative of *salire* "to jump"	verb	10	15.14	25.40	38.08	159.33	73.03
kettle	lebes	caldera	4	From *caldaria* "warm bath"	noun	5	0.24	15.00	3.02	4.61	7.60
kick	calcitro	patear	5	Germanic origin	verb	10	1.08	0.40	4.03	19.50	7.81
kid	haedus	cabrito	4	From the diminutive of *capra* "she-goat"	noun	3	23.07	0.40	1.01	3.98	1.76
kill	interficio	matar	4	Generalization: From *mactare* "to sacrifice"	verb	4	96.89	36.80	97.91	314.26	147.21
kiss	basio	besar	0	a > e under the influence of following s + yod	verb	16	11.95	1.40	8.46	150.53	52.06
knead	subigo	amasar	4	From *a-* "to cause to be", *masa* "dough", and a verbalizing ending	verb	5	76.25	1.20	1.61	6.71	3.12
knee	genu	rodilla	4	From the diminutive of *rota* "wheel"	noun	4	209.67	11.60	9.67	116.15	44.80
kneel	genuflecto	arrodillar	2	"a" + "rodilla" + "ar" + "se"	verb	10	0.00	0.60	4.23	26.84	10.32
knife	culter	cuchillo	4	From *cultellus*, the diminutive of *culter* "knife"	noun	5	14.58	10.60	8.86	67.09	28.30
knot	nodus	nudo	1	long-o > u sporadically under influence of d + yod	noun	9	39.92	20.60	7.05	31.24	19.48
know	scio	saber	4	From *sapere* "to have taste, be wise"	verb	17	712.49	119.20	643.66	2552.90	1083.78
lagoon	lacuna	laguna	1	Intervocalic [k] voicing; LA ending vowel added to SP stem.	noun	1	11.35	17.40	20.55	25.58	21.11
lake	lacus	lago	1	Intervocalic [k] voicing; LA ending vowel added to SP stem.	noun	1	122.51	140.20	35.26	31.24	69.57
lamb	agnus	cordero	4	From the adjective *chordus* "late	noun	3	10.40	7.60	13.50	6.50	9.23

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				born, produced late in the season"							
lame	claudus	cojo	4	From *coxa* "hip"	adjective	4	70.46	0.00	0.60	0.21	0.27
lamp, torch	lampas	lámpara	2	Influence of "chamber"	noun	7	15.18	18.80	7.05	59.75	28.10
land	expono	desembarcar	2	des + em + barca + ar	verb	10	202.11	19.20	5.04	9.85	11.40
large, big	grandis	grande	1	LA ending vowel added to SP stem.	adjective	12	80.08	934.01	580.61	504.00	675.73
last	ultimus	último	0	No change to stem.	adjective	13	0.00	11.00	0.20	0.84	4.07
last, endure	duro	durar	0	No change to stem.	verb	14	170.43	69.00	59.83	96.44	74.80
late	tarde	tarde	0	No change to stem.	adverb	14	0.00	313.00	112.82	288.90	237.76
lazy	piger	perezoso	2	pereza + oso	adjective	4	36.21	1.40	1.61	7.55	3.46
lead	duco	conducir	4	Generalization: From *conducere* "to lead together"	verb	10	859.94	119.40	115.84	113.21	116.20
lead	plumbum	plomo	2	No reason for short-stress-u > o	noun	9	31.19	0.00	0.00	0.00	0.00
learn	disco	aprender	4	From *apprendere* "to grasp"	verb	17	219.16	63.20	64.87	171.07	98.69
left (side)	sinister	izquierda	5	Basque origin	noun	12	83.54	39.80	47.75	60.17	49.07
leg	crus	pierna	4	From *perna* "haunch, ham"	noun	4	89.52	23.00	39.49	290.57	115.18
lend	commodo	prestar	4	From *praestare* "to be responsible for, perform"	verb	11	19.34	56.80	83.20	98.74	79.27
let, permit	permitto	permitir	0	Geminate simplification	verb	19	163.78	548.21	492.17	240.47	429.70
lie, tell lies	mentior	mentir	2	Deponent > active	verb	16	103.02	1.60	28.61	51.15	26.74
light (in weight)	levis	ligero	5	French origin	adjective	15	237.21	77.40	34.45	63.52	58.44
lightning	fulmen	relámpago	4	From *lampas* "torch, lamp"	noun	1	99.56	4.20	6.45	29.35	13.10
like, similar	similis	similar	2	VL origin for the 'r', diminutive/extra	adjective	12	605.88	235.20	122.69	22.43	128.41

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				syllable							
limp	claudico	cojear	4	From *coxa* "hip"	verb	10	5.26	0.00	0.60	3.77	1.43
line	linea	línea	1	LA ending vowel added to SP stem.	noun	12	28.92	364.80	197.63	123.06	230.22
linen	linum	lino	1	LA ending vowel added to SP stem.	noun	6	44.98	17.80	4.63	9.01	10.52
lip	labium	labio	1	LA ending vowel added to SP stem.	noun	4	3.09	14.60	10.68	259.34	92.51
listen	ausculto	escuchar	2	Confusion with prefix "ex" (Mendeloff pg. 12)	verb	15	1.43	21.40	164.59	579.68	250.38
live	vivo	vivir	0	No change to stem.	verb	4	524.29	328.00	397.68	708.19	474.56
liver	iecur	hígado	4	From *ficus* "fig"	noun	4	0.00	28.40	8.26	14.26	17.04
livestock	pecus	ganado	5	Germanic origin	noun	3	56.00	87.20	64.47	36.69	63.19
lock	sera	cerradura	2	Influence of "to saw"	noun	7	9.37	1.00	1.81	24.74	8.96
long	longus	largo	4	From *largus* "abundant, plentiful"	adjective	12	693.09	310.20	217.37	444.67	322.46
look, look at	aspicio	mirar	4	From *mirus* "wonderful, wonder"	verb	15	185.85	26.60	211.94	1526.25	574.54
loom	tela	telar	2	Sporadic addition of additional syllable from VL	noun	6	125.21	17.40	0.81	2.10	6.86
lose	perdo	perder	0	No change to stem.	verb	11	233.06	199.80	343.49	621.61	384.77
loud	magnus	fuerte	4	From *fortis* "strong"	adjective	15	1781.98	198.20	166.61	202.10	188.82
louse	pediculus	piojo	1	Regular derivation from variant "peduculus"	noun	3	4.10	3.00	1.81	6.71	3.80
love	amor	amor	0	No change to stem.	noun	16	352.82	85.20	134.37	484.94	231.17
low	humilis	bajo	5	Greek origin	adjective	12	90.68	345.20	188.57	144.24	227.37
lunch	prandium	almuerzo	4	From *ad-* "into" and *mordere* "to bite"	noun	5	10.52	0.00	15.31	49.69	21.24

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
magic	magia	magia	1	LA ending vowel added to SP stem.	noun	22	0.48	9.80	16.72	36.48	20.77
male	mas	masculino	4	From *masculinus* "masculine"	adjective	3	229.96	70.00	33.04	17.19	40.45
male	mas	macho	4	From *masculus*, diminutive of *mas* "male"	noun	2	229.96	65.00	13.70	20.96	33.46
man (vs. woman)	vir	hombre	4	From *homo* "human being"	noun	2	134.08	329.40	548.78	1434.42	761.05
market (place)	mercatus	mercado	1	Intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	11	5.97	188.20	457.11	49.90	234.02
marriage, wedding	matrimoniu m	matrimonio	1	LA ending vowel added to SP stem.	noun	2	43.74	79.80	39.49	86.79	68.48
mason	faber	albañil	5	Arabic origin	noun	7	11.71	1.00	3.63	15.72	6.65
mast	malus	mástil	5	French origin	noun	10	237.32	9.00	1.01	8.81	6.24
master	dominus	señor	4	From *senior*, comparative of *senex* "old"	noun	19	235.53	57.60	176.48	794.78	336.31
mat	storea	estera	2	De Silva: "Irregularly"	noun	9	0.60	1.60	0.20	5.87	2.51
match	sulfurata	cerilla	4	From the diminutive form of *cera* 'wax'	noun	1	0.00	0.80	0.20	3.56	1.49
meal	cibus	comida	4	From the verb *comedere* "to eat up"	noun	5	348.28	35.60	32.44	147.38	70.72
measure	metior	medir	2	Deponent verb becoming active	verb	12	108.62	274.60	134.58	65.20	159.64
medicine, drug	medicina	medicina	1	LA ending vowel added to SP stem.	noun	4	62.63	60.00	37.87	16.14	38.35
meet	convenio	encontrar	4	From *in* "in" + *contra* "against"	verb	19	373.42	895.81	684.56	947.82	841.48
merchant	mercator	mercader	5	Catalan origin	noun	11	18.88	17.40	2.22	3.98	7.94
middle, center	medius	medio	5	Semi-learned	noun	12	217.09	225.40	543.94	65.83	281.06
milk	mulgeo	ordeñar	4	From *ordinare* "to put in order"	verb	5	0.96	0.00	0.20	1.89	0.68

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
millstone	molina	molino	1	LA ending vowel added to SP stem.	noun	5	0.12	11.40	4.63	10.69	8.89
misfortune, bad luck	infortunium	desgracia	4	From *gratia* "charm, pleasure, thanks"	noun	16	0.48	9.80	20.15	66.88	31.76
miss (target)	erro	errar	0	No change to stem.	verb	20	78.72	0.80	6.45	8.60	5.23
mistake, error	error	error	0	No change to stem.	noun	16	135.06	47.00	112.82	65.41	75.13
molartooth	molaris	muela	4	From *mola* "millstone"	noun	4	3.94	1.60	4.03	9.64	5.02
mold (clay)	fingo	moldear	5	Catalan origin	verb	9	191.35	12.80	2.82	6.29	7.33
money	pecunia	plata	5	Greek origin	noun	11	468.51	141.80	59.43	146.75	115.65
monkey	simia	mono	5	Arabic origin	noun	3	6.10	24.40	13.50	21.59	19.82
moon	luna	luna	1	LA ending vowel added to SP stem.	noun	1	62.35	75.00	40.90	141.93	85.18
mortar	mortarium	mortero	1	a > e under influence of following r + yod	noun	7	7.65	5.00	2.42	5.66	4.34
mosquito	culex	mosquito	4	From the diminutive of *musca* "fly"	noun	3	4.78	8.80	4.63	15.30	9.50
mother's brother	avunculus	tío	5	Greek origin	noun	2	23.43	14.00	14.30	238.37	86.74
mother's sister	matertera	tía	5	Greek origin	noun	2	4.30	2.60	7.66	218.04	74.05
mountain	mons	montaña	4	From the adjective *montanus* "mountainus"	noun	1	419.75	134.40	37.27	84.28	85.45
mouse, rat	mus	ratón	4	From the verb *radere* "to scrape" (De Silva)	noun	3	78.53	12.20	25.59	35.85	24.37
mouth	oris	boca	4	From *bucca* "distended cheeck, mouth"	noun	4	0.00	46.40	67.29	450.75	184.34
move	moveo	mover	0	No change to stem.	verb	10	804.96	108.20	80.58	324.12	168.80
mow, reap	meto	segar	4	From *secare* "to cut"	verb	8	7.53	1.20	0.81	3.77	1.90

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
much, many	multus	mucho	2	Derivation from muito > mucho has no rationale	adjective	13	2479.29	280.80	450.67	713.44	478.09
mud	lutum	barro	5	Pre-Roman Celtic origin	noun	1	6.11	21.00	7.25	61.22	29.39
mule	mulus	mulo	1	LA ending vowel added to SP stem.	noun	3	8.96	1.20	1.41	1.26	1.29
multitude, crowd	multitudo	muchedumbre	2	No reason for lt't > ch	noun	13	265.09	3.40	9.27	23.48	11.88
mumble	murmuro	murmurar	0	No change to stem.	verb	18	0.00	0.00	2.42	114.47	37.87
murder	caedes	asesinato	5	Arabic origin	noun	21	191.23	36.00	78.17	13.63	42.96
mute	mutus	mudo	0	Intervocalic [t] voicing.	adjective	4	36.09	14.20	6.65	56.40	25.32
nail	clavus	clavo	1	LA ending vowel added to SP stem.	noun	9	31.03	5.40	5.44	28.93	13.03
naked	nudus	desnudo	4	From the verb *desnudare* "to make bare"	adjective	4	128.00	15.60	16.92	138.79	55.93
nape of neck	cervix	cogote	5	Aymara or Quechua origin	noun	4	145.33	0.00	0.00	8.81	2.85
narrow	angustus	estrecho	4	From *strictus* "tight, close"	adjective	12	62.99	83.00	35.05	45.49	54.70
near	circa	cerca	2	No reason for i > e (unless function words are unstressed)	adverb	12	0.00	11.60	35.05	76.94	40.66
neck	collum	cuello	1	Diphthongization of short stressed 'o' > 'ue'; LA ending vowel added to SP stem.	noun	4	114.94	34.20	21.76	161.22	71.13
necklace	monile	collar	4	From *collum* "neck"	noun	6	8.84	11.40	3.02	27.46	13.78
needle	acus	aguja	4	From the diminutive of *acus* "needle"	noun	6	7.09	25.00	5.64	28.93	19.75
nest	nidus	nido	1	LA ending vowel added to SP stem.	noun	3	30.60	30.60	6.04	26.00	20.84

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
nipple	papilla	teta	5	Old French origin	noun	4	6.45	0.40	0.40	10.27	3.60
noon, midday	meridies	mediodía	2	medio + dia	noun	14	28.68	4.00	42.11	54.51	33.19
north	septentrio	norte	5	Old Frankish origin (via French)	noun	12	0.00	635.81	137.40	50.94	278.55
nose	nasus	nariz	4	From *naris* "nostril"	noun	4	9.42	13.00	12.69	141.93	54.64
nostril	naris	nariz	2	The z is the result of the addition of the suffix 'ic'	noun	4	43.86	13.00	12.69	141.93	54.64
now	nunc	ahora	3	From *hac hora* "at this time"	adverb	14	1343.63	87.40	784.69	1118.06	655.98
oak	robur	roble	2	Sporadic r > l change	noun	8	143.18	25.00	4.43	5.66	11.81
oar	remus	remo	1	LA ending vowel added to SP stem.	noun	10	23.96	11.80	2.01	10.27	8.01
oath	iuramentum	juramento	1	LA ending vowel added to SP stem.	noun	21	0.00	8.80	12.49	10.27	10.52
oats	avena	avena	1	LA ending vowel added to SP stem.	noun	8	9.08	6.60	0.20	0.21	2.38
obey	oboedio	obedecer	2	Sporadic addition of syllable from VL	verb	19	11.39	15.20	25.38	54.72	31.42
obscure	obscurus	obscuro	0	No change to stem.	adjective	17	104.22	0.00	0.40	7.55	2.58
ocean	oceanus	océano	1	LA ending vowel added to SP stem.	noun	1	0.00	140.60	17.93	26.00	62.17
oil	oleum	óleo	1	LA ending vowel added to SP stem.	noun	5	153.70	9.60	25.38	11.32	15.47
old	vetus	viejo	4	From *vetulus*, the diminutive of *vetus* "old"	adjective	14	192.78	33.40	52.18	300.22	126.11
old man	vetulus	viejo	4	From the diminutive of *vetus* "old"	noun	2	10.52	31.00	76.35	431.46	175.92
old woman	vetula	viejo	4	From the diminutive of *vetus* "old"	noun	2	0.00	31.00	76.35	431.46	175.92
olive	oliva	aceituna	5	Arabic origin	noun	5	25.82	4.00	1.21	6.71	3.94

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
omen, portent	augurium	augurio	1	LA ending vowel added to SP stem.	noun	22	38.25	1.60	2.22	6.92	3.53
one	unus	uno	1	LA ending vowel added to SP stem.	pronoun	13	828.50	678.41	912.61	1002.96	862.38
open	aperio	abrir	1	The 'e' only drops in pretonic position, which only occurs in some forms of the paradigm	verb	12	205.97	117.00	234.30	620.56	319.54
ornament	ornamentum	ornamento	1	LA ending vowel added to SP stem.	noun	6	87.25	9.40	2.22	2.94	4.89
orphan	orbus	huérfano	5	Greek origin	noun	2	121.67	1.40	2.62	8.81	4.21
outside	foras	fuera	2	Loss of the 's' unexplained	adverb	12	8.49	0.00	0.00	0.00	0.00
owe	debeo	deber	0	No change to stem.	verb	11	1015.43	954.21	1404.17	1002.96	1121.59
owl	noctua	lechuza	2	Influence of *leche* 'milk'	noun	3	5.38	1.60	0.40	3.77	1.90
own, possess	possideo	poseer	1	i > e only when unstressed (only part of the paradigm)	verb	11	77.13	210.40	95.49	80.92	129.77
ox	bos	buey	2	Irregular loss of -v-	noun	3	166.61	11.80	1.81	29.14	14.05
pain	dolor	dolor	0	No change to stem.	noun	16	608.47	48.60	58.83	266.88	122.71
paint	color	pintura	4	From *pictura* "painting"	noun	9	229.81	223.80	185.75	76.52	163.30
paint	pingo	pintar	4	From the verb *pingere* "to paint, embroider, tattoo"	verb	9	70.98	47.40	55.20	90.15	63.87
pair	par	par	0	No change to stem.	noun	13	268.91	66.80	71.92	153.88	96.72
palm of hand	palma	palma	1	LA ending vowel added to SP stem.	noun	4	37.64	19.00	31.23	62.89	37.33
palm tree	palma	palmera	5	Catalan origin	noun	8	37.64	9.40	4.43	29.14	14.12
pan	patina	cazuela	5	Old Provencal origin	noun	5	9.80	0.60	1.01	5.87	2.44
paper	charta	papel	5	Catalan origin	noun	18	30.36	304.00	232.89	257.24	264.90
parrot	psittacus	loro	5	Cariban origin	noun	3	4.30	10.80	2.62	8.60	7.33

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
pasture	pastus	pasto	1	LA ending vowel added to SP stem.	noun	3	6.77	18.00	9.27	34.17	20.29
path	semita	camino	5	Celtic origin	noun	10	17.93	99.60	210.53	341.10	215.15
pay	solvo	pagar	4	From *pacare* "to pacify"	verb	11	350.79	68.80	254.44	208.60	176.60
peel	decortico	pelar	4	From *pilare* "to remove the hair from"	verb	5	1.43	4.80	11.08	12.79	9.50
pen	calamus	pluma	4	From *pluma* "feather"	noun	18	31.91	26.60	14.51	59.54	33.19
penis	penis	pene	1	LA ending vowel added to SP stem.	noun	4	46.13	8.60	2.22	6.29	5.70
pepper	piper	pimienta	4	From *pigmentum* "colour, paint, drug"	noun	5	0.00	3.60	4.23	4.40	4.07
person, human being	homo	persona	4	From *persona* "character played by an actor"	noun	2	1936.44	447.40	676.90	328.31	486.17
perspire	sudo	sudar	0	No change to stem.	verb	4	48.52	0.40	4.43	45.91	16.49
pestle	pistillum	majadero	4	From *malleus* "hammer"	noun	5	0.48	0.40	0.20	1.68	0.75
physician	medicus	médico	1	LA ending vowel added to SP stem.	noun	4	66.52	69.40	110.20	124.11	100.86
pick up	colligo	recoger	4	Generalization: From *recolligere* "to recover again"	verb	12	22.63	101.80	103.35	146.34	116.74
pig	porcus	puerco	1	Diphthongization of short stressed 'o' > 'ue'; LA ending vowel added to SP stem.	noun	3	16.97	0.60	1.61	4.19	2.10
pile up	accumulo	acumular	1	Geminate simplification; *	verb	12	3.82	42.60	50.57	43.19	45.47
pillow	cervical	almohada	5	Arabic origin	noun	7	2.15	1.40	0.60	34.59	11.88
pin	fibula	alfiler	5	Arabic origin	noun	6	10.04	3.40	2.01	15.30	6.79
pinch	vellico	pellizcar	2	No reason for p > v, among other	verb	15	2.39	0.20	0.40	4.40	1.63

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				things							
pine	pinus	pino	1	LA ending vowel added to SP stem.	noun	8	41.83	26.00	12.09	18.66	18.94
pitcher, jug	urceus	jarra	5	Arabic origin	noun	5	3.11	5.20	0.81	10.48	5.43
place	locus	lugar	4	From *localis* "of a place"	noun	12	1369.32	826.21	586.25	601.69	672.68
plain, field	campus	campo	1	LA ending vowel added to SP stem.	noun	1	142.98	362.00	261.09	209.02	278.48
plaintiff	petitor	demandante	4	From *demandare* "to entrust, commit"	noun	21	9.24	0.80	5.24	0.84	2.31
plant	planta	planta	1	LA ending vowel added to SP stem.	noun	8	28.68	10.60	12.89	30.61	17.85
plant	planto	plantar	0	No change to stem.	verb	8	11.23	346.00	86.63	92.25	176.47
play	ludo	jugar	4	From *jocare* "to jest, joke"	verb	16	140.50	77.80	291.92	242.98	203.41
pocket	loculus	bolsillo	5	French origin	noun	6	4.78	2.20	19.74	109.65	42.90
poet	poeta	poeta	1	LA ending vowel added to SP stem.	noun	18	111.75	241.80	109.19	82.60	145.59
poison	venenum	veneno	1	LA ending vowel added to SP stem.	noun	4	147.25	15.60	6.65	22.64	14.86
poncho	pallium	poncho	5	Quechua origin	noun	6	16.73	0.40	1.21	19.08	6.72
post, pole	postis	poste	1	LA ending vowel added to SP stem.	noun	7	30.20	6.60	12.29	16.56	11.74
pot, cooking vessel	olla	olla	1	LA ending vowel added to SP stem.	noun	5	5.73	1.40	6.85	27.04	11.54
potter	figulus	alfarero	5	Arabic origin	noun	9	3.66	0.00	0.00	0.00	0.00
pound with fist	percido	pegar	4	From *picare* "to smear with pitch"	verb	9	1.04	5.60	31.83	151.16	61.56
pour	fundo	verter	4	From *vertere* "to turn around, exchange"	verb	9	37.01	28.60	11.28	15.72	18.60
praise	laus	alabanza	4	Possible from *alapa* "slap"	noun	16	164.30	3.80	1.61	3.98	3.12

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
pray	oro	orar	0	No change to stem.	verb	22	228.52	6.60	6.65	12.16	8.42
precipice	praecipitium	precipicio	2	No reason for p failing to voice	noun	1	10.02	0.80	1.81	9.85	4.07
pregnant	praegnas	embarazado	5	Italian origin	adjective	4	0.00	1.60	2.01	1.68	2.38
preserve, look after	conservo	conservar	0	No change to stem.	verb	11	96.09	209.00	74.94	104.41	129.98
press	premo	prensar	5	Catalan origin	verb	9	297.84	2.80	0.40	0.84	1.36
prison, jail	carcer	cárcel	2	Sporadic r > l change	noun	21	64.54	9.20	77.56	42.98	43.17
prostitute	prostituta	prostituta	1	LA ending vowel added to SP stem.	noun	19	0.00	5.60	10.68	16.56	10.86
proud	superbus	soberbio	1	Intervocalic [p] voicing; LA ending vowel added to SP stem.	adjective	16	128.84	5.20	7.86	15.72	9.50
pull	traho	tirar	5	Germanic origin	verb	9	378.56	42.20	50.77	231.45	106.36
pumpkin, squash	colocynthis	calabaza	5	Iberian origin	noun	8	0.48	7.20	1.41	6.29	4.95
pursue	persequor	perseguir	2	Deponent verb become active	verb	10	112.59	28.80	46.94	90.99	55.04
pus	pus	pus	1	LA ending vowel added to SP stem.	noun	4	0.00	0.40	0.00	6.50	2.24
push, shove	trudo	empujar	4	Generalization: From *impellere* "to push against"	verb	10	7.89	21.20	27.00	104.20	50.02
put	pono	poner	0	No change to stem.	verb	12	1196.14	361.80	606.19	1173.62	706.96
put on (clothes)	vestio	vestir	0	No change to stem.	verb	6	43.27	40.20	47.75	266.25	115.93
raft	ratis	balsa	5	Iberian origin	noun	10	104.54	6.00	3.63	8.39	5.97
rafter	cantherius	viga	4	From *biga* "two-horse chariot"	noun	7	3.59	13.20	4.63	13.84	10.52
rake	rastrum	rastrillo	4	From *rastillo*, the diminutive of *rastrum* "rake"	noun	8	6.93	1.20	0.20	5.45	2.24

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
rape	stuprum	violación	4	From *violare* "to violate, treat with violence"	noun	21	40.04	15.60	66.08	9.64	30.68
raw	crudus	crudo	0	No change to stem.	adjective	5	62.39	19.60	11.48	22.64	17.99
reach, arrive	advenio	llegar	4	From *plicare* "to fold, bend, roll up"	verb	10	123.82	591.21	925.71	1471.11	988.76
ready	paratus	listo	5	Germanic origin	adjective	14	22.79	7.60	38.48	62.69	35.84
rebuke, scold	increpo	reprender	4	From *reprehendere* "to seize, hold back"	verb	18	52.98	0.40	1.01	2.73	1.36
red	russus	rojo	2	No reason for ss > j	adjective	15	0.96	133.20	163.38	224.11	172.80
reef	cautes	arrecife	5	Arabic origin	noun	1	28.45	8.80	1.61	1.26	3.94
refuse	recuso	rehusar	4	From *refundere* "to give back"	verb	18	103.50	4.60	8.86	7.34	6.92
regret	paeniteo	lamentar	4	From the verb *lamentari* "to wail"	verb	16	69.68	3.00	43.31	43.82	31.83
relative, kinsman	cognatus	pariente	4	Generalization: from *parens* "parent"	noun	2	18.16	14.40	15.51	53.25	27.35
release, let go (to free)	libero	soltar	4	From the adjective *soltus* "loose"	verb	11	199.32	7.00	17.33	135.22	51.99
remain, stay	maneo	quedar	4	From *quietus* "quiet"	verb	12	266.50	352.80	565.90	1291.23	728.40
remains, left overs	reliquia	resto	4	From the verb *restare* "to stand back"	noun	12	0.00	263.00	189.57	170.44	208.30
rib	costa	costilla	4	From the diminutive of *costa* "rib"	noun	4	15.30	6.00	2.42	18.87	8.96
rice	oryza	arroz	5	Arabic origin	noun	8	3.59	51.00	20.55	21.59	31.22
rich	dives	rico	5	Gothic origin	adjective	11	182.83	99.40	37.07	60.59	65.84
ride (a horse)	equito	cabalgar	4	From *caballus* "inferior horse, nag"	verb	10	156.01	1.40	1.61	16.98	6.52
right (side)	dexter	derecha	4	From *directus* "straight, direct"	noun	12	251.59	44.80	48.95	73.59	55.52
right, correct	rectus	correcto	4	Generalization: From *com-*	adjective	16	104.96	33.40	45.93	31.03	36.85

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				"thoroughly" + *rectus* "straight"							
ring (for finger)	anulus	anillo	4	From *anellus* "small ring", diminutive of *anus* "ring"	noun	6	26.41	42.80	16.12	37.74	32.17
ripe	maturus	maduro	0	Intervocalic [t] voicing.	adjective	5	84.38	13.00	11.48	24.53	16.22
river, stream, brook	rivus	río	2	Sporadic (ie. irregular) loss of -v-	noun	1	35.02	463.61	148.48	181.56	266.13
road	via	carretera	4	From *carreta* "wagon"	noun	10	426.68	86.40	66.88	47.17	67.13
roast	asso	asar	0	Geminate simplification.	verb	5	6.70	2.00	1.61	17.40	6.86
roll	roto	rodar	0	Intervocalic [t] voicing.	verb	10	29.76	14.20	11.48	55.35	26.61
roof	tectum	techo	1	Regular 'ct' > 'ch' development; LA ending vowel added to SP stem.	noun	7	125.93	21.20	35.66	124.74	59.59
room	conclave	cuarto	4	From *quartus* "quarter, fourth"	noun	7	9.08	51.40	72.73	270.24	129.43
rope, cord	funis	cuerda	4	From *chorda* "catgut, string of a musical instrument"	noun	9	32.99	69.60	28.00	61.01	52.80
rotten	puter	podrido	2	podrir + adjectival suffix	adjective	5	0.00	0.60	0.60	9.01	3.33
rough	asper	áspero	0	No change to stem.	adjective	15	142.47	6.60	4.23	29.35	13.17
round	rotundus	redondo	2	No reason for o > e (the claimed LA alternate "retundus" is not attested in Perseus)	adjective	12	51.27	16.20	21.56	60.59	32.37
row	remigo	remar	4	From the noun *remus* "oar"	verb	10	7.65	1.40	0.81	6.92	2.99
rub, wipe	frico	frotar	5	French origin	verb	9	43.86	4.40	2.42	38.58	14.80
rudder	gubernaculum	timón	4	From *temo* "pole, beam"	noun	10	14.58	5.20	6.04	3.35	4.89

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
rug	stragulum	alfombra	5	Arabic origin	noun	9	2.63	26.20	5.84	45.28	25.52
rule, govern	rego	regir	0	No change to stem.	verb	19	402.01	47.40	28.81	8.39	28.51
sail	velum	vela	1	LA ending vowel added to SP stem.	noun	10	60.43	27.40	15.11	21.38	21.31
salt	sal	sal	0	No change to stem.	noun	5	0.00	2.80	1.01	3.77	2.51
salty	salsus	salado	4	From the noun *sal* "salt"	adjective	15	8.72	18.20	3.02	13.00	11.40
sap	sucus	savia	4	From *sapa* "new wine boiled thick"	noun	8	182.27	1.80	1.41	8.18	3.73
saucer	patella	platillo	5	Greek origin	noun	5	5.74	3.00	3.22	11.32	5.77
sausage	farcimen	salchicha	5	Italian origin	noun	5	0.00	0.40	1.61	3.77	1.90
save, rescue	salvo	salvar	0	No change to stem.	verb	11	20.20	62.80	112.62	192.88	121.69
saw	serra	sierra	1	Diphthongization of short stressed 'e' > 'ie'; LA ending vowel added to SP stem.	noun	9	4.70	53.40	25.99	28.09	35.97
say	dico	decir	2	No reason for long [i] > e, must be analogy	verb	18	564.52	526.21	2485.00	4656.73	2523.35
scissors, shears	forfices	tijera	4	From *tondere* "to shave, clip"	noun	9	0.00	1.40	4.43	17.82	7.74
scrape	rado	raspar	5	Germanic origin	verb	5	39.92	1.40	1.61	8.39	3.73
sculptor	sculptor	escultor	0	Regular 'e' insertion before word initial 's' + consonant clusters.	noun	9	0.00	39.60	21.96	3.98	22.13
scythe	falx	guadaña	5	Germanic origin	noun	8	23.90	0.80	0.00	2.31	1.02
sea	mare	mar	0	No change to stem.	noun	1	157.24	361.60	122.69	290.99	258.25
season	tempus	estación	4	From *statio* "outpost, season"	noun	14	740.54	92.60	84.21	81.55	86.20
second	secundus	segundo	0	Intervocalic [k] voicing.	number	13	134.61	496.41	550.99	298.75	450.81
secret	secretum	secreto	1	LA ending vowel added to SP stem.	noun	17	48.20	16.20	58.62	112.37	61.63

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
seed	semen	semilla	4	From the diminutive of *semen* "seed"	noun	8	232.22	72.20	14.71	25.16	37.60
seek, look for	quaero	buscar	5	Gaulish origin (most likely)	verb	11	909.38	119.80	354.77	649.91	370.58
seem	pareo	parecer	2	Sporadic additional syllable from VL	verb	17	234.66	281.80	565.70	1364.19	727.86
seize, grasp	prehendo	agarrar	5	Gaulish origin	verb	11	81.99	10.80	11.48	107.34	42.28
sell	vendo	vender	0	No change to stem.	verb	11	138.44	59.80	218.18	140.05	139.14
send	mitto	enviar	4	From the noun *via* "road"	verb	10	879.08	149.60	208.31	72.96	144.57
separate	separo	separar	0	No change to stem.	verb	12	48.41	199.20	68.50	153.88	140.50
servant	famulus	criado	4	From the verb *creare* "to create, produce"	noun	19	14.94	2.40	1.81	15.72	6.52
sew	suo	coser	4	Generalization: From *consuere* "to sew together"	verb	6	751.57	6.00	5.44	30.61	13.78
shade, shadow	umbra	sombra	3	From *sub umbra* "under shade"	noun	1	172.34	33.20	45.13	282.19	117.83
shake	tremo	estremecer	4	From *ex* "thoroughly" and *tremescere* "to begin to shake for fear"	verb	10	86.65	0.20	8.66	52.83	20.09
shame	verecundia	vergüenza	2	No reason for u > ue, or di > z	noun	16	50.20	0.60	11.68	84.91	31.63
share (distribute)	partior	repartir	4	Generalization: From *re* "again" and *partior* "to share"	verb	11	49.59	34.00	26.79	43.40	34.61
shark	pistrix	tiburón	5	Portuguese origin	noun	3	0.96	5.80	4.43	1.47	3.94
sharp	acutus	agudo	0	Intervocalic [k] voicing; intervocalic [t] voicing.	adjective	15	65.77	37.80	29.21	55.98	40.79
sheep	ovis	oveja	4	From the diminutive of *ovis* "sheep"	noun	3	39.80	32.80	3.63	18.24	18.26

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
shelf	pluteus	anaquel	5	Arabic origin	noun	7	8.61	0.40	0.81	8.18	3.05
shell	concha	concha	1	LA ending vowel added to SP stem.	noun	3	25.10	14.80	2.01	7.34	8.08
shield	scutum	escudo	1	e' insertion before word-initial 's' + consonant clusters; intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	20	35.14	36.80	8.06	26.63	23.82
ship	navis	nave	1	LA ending vowel added to SP stem.	noun	10	276.41	74.80	44.32	43.40	54.37
shiver	tremo	temblar	2	Sporadic addition of a syllable from VL	verb	4	86.65	0.80	6.85	119.08	41.13
shoe	calceus	zapato	5	Turkish origin	noun	6	8.13	9.60	23.37	133.76	54.43
shoemaker	sutor	zapatero	2	"zapato" + "ero"	noun	6	7.13	1.20	1.81	9.85	4.21
shoot	conicio	disparar	4	From *disparare* "to separate"	verb	20	117.13	23.00	49.96	61.01	44.39
shore	litus	costa	4	From *costa* "rib, side"	noun	1	33.82	291.80	85.62	57.65	146.54
short	brevis	corto	4	From *curtus* "shortened, mutilated"	adjective	12	292.22	81.20	40.49	61.43	61.08
shoulderblade	scapula	espaldilla	4	Probably from the diminutive of *scapula* "shoulderblade"	noun	4	5.54	0.00	0.00	0.21	0.07
shout, cry out	clamo	gritar	4	Generalization: From *quiritare* "to cry out for help, implore the aid of the Quirites"	verb	18	112.83	3.20	33.44	340.68	122.64
shovel	pala	pala	1	LA ending vowel added to SP stem.	noun	8	30.08	15.60	5.64	12.16	11.13
shriek, screech	exclamo	chillar	4	From *fistula* "reed pipe, tube"	verb	18	44.22	0.00	1.21	22.43	7.67
shut, close	claudo	cerrar	4	From *serrare* "to saw"	verb	12	64.14	42.00	121.48	317.20	157.87
sickness	morbus	enfermedad	4	From the adjective *infirmus* "feeble, not firm"	noun	4	364.53	288.40	181.31	67.51	180.81

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
side	latus	lado	1	Intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	12	85.20	215.80	254.24	656.62	371.46
side of the head (temple)	tempus	sien	5	Germanic origin	noun	4	740.54	0.40	1.01	31.87	10.79
sieve	colum	colador	2	From verb for "sift" + derivational suffix	noun	5	12.07	0.40	0.60	1.68	0.88
silk	sericum	seda	4	From *seta* "bristle"	noun	6	2.03	48.80	7.25	60.80	38.69
silver	argentum	plata	5	Greek origin	noun	9	182.15	141.80	59.43	146.75	115.65
sing	cano	cantar	4	From *cantare*, frequentative of *canere* "to sing"	verb	18	171.52	50.40	164.39	181.35	149.73
sink	submergo	hundir	4	From *fundere* "to pour out"	verb	10	0.00	26.40	18.94	134.59	58.91
sister	soror	hermano	4	From *germanus* "having the same parents"	noun	2	188.72	121.00	154.12	411.33	226.15
six	sex	seis	0	[ks] > [i-sh] change (early), if word-final, results in depalatalization	number	13	81.63	144.00	263.11	139.00	182.51
skin	cutis	piel	4	From *pellis* "animal skin"	noun	4	143.66	118.40	36.06	260.80	136.76
skirt	gunna	falda	5	Germanic origin	noun	6	0.00	9.20	7.66	64.57	26.61
skull	calvaria	cráneo	5	Greek origin	noun	4	5.26	28.20	13.50	24.95	22.19
slave	servus	esclavo	5	From 'Slav' (LL)	noun	19	113.66	60.80	3.83	28.30	31.09
sleep	dormio	dormir	0	No change to stem.	verb	4	56.89	12.60	37.87	518.46	184.88
slip	labor	resbalar	4	Possibly from the adjective *varus* "knock-kneed"	verb	10	139.62	0.60	4.43	52.62	18.73
slow	lentus	lento	0	No change to stem.	adjective	14	68.09	50.80	31.83	96.65	59.25
smell (intrans)	oleo	oler	0	No change to stem.	verb	15	105.77	1.20	8.66	114.68	40.45

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
smell (transitive)	olfacio	oler	4	From the intransitive verb *olere* "to smell"	verb	15	0.00	1.20	8.66	114.68	40.45
smile	subrideo	sonreír	2	No reason for br > nr	verb	16	0.00	0.20	16.12	260.38	89.79
snail	cochlea	caracol	2	Derived irregularly through VL	noun	3	0.00	13.40	7.66	26.84	15.81
snake	colubra	culebra	2	No reason for vowels	noun	3	6.10	0.80	0.81	21.80	7.60
sneeze	sternuo	estornudar	4	From the frequentative of *sternuere* "to sneeze"	verb	4	2.15	0.20	0.60	5.45	2.04
snore	sterto	roncar	5	Greek origin	verb	4	5.98	0.00	0.20	15.51	5.09
snow	nix	nieve	2	No reason for long-i > ie	noun	1	37.53	26.80	11.08	33.96	23.82
soap	sebum	jabón	5	LA word is "tallow, grease", no word for soap. Germanic origin	noun	6	29.64	4.20	4.63	22.43	10.25
soft	mollis	blando	4	From *blandus* "pleasant, flattering"	adjective	15	181.67	32.20	7.25	37.32	25.45
soldier	miles	soldado	4	From *solidus* "solid"	noun	20	691.89	85.00	50.16	108.60	80.90
some	paulum	alguno	4	From *aliquis unus* "someone"	adjective	13	131.47	756.61	596.32	410.91	590.69
son-in-law	gener	yerno	2	Metathesis	noun	2	410.55	4.60	1.61	6.08	4.07
soon	mox	pronto	4	From *promptus* "prompt, ready, visible"	adverb	14	309.79	109.40	118.66	425.38	214.82
soul, spirit	anima	alma	2	No reason for n > l	noun	16	211.99	46.80	52.58	265.21	119.45
sound, noise	sonitus	sonido	1	Intervocalic [t] voicing; LA ending vowel added to SP stem.	noun	15	28.21	172.20	48.75	122.44	114.50
soup	ius	sopa	5	Germanic origin	noun	5	194.81	1.60	4.03	34.80	13.17
south	meridies	sur	5	French origin	noun	12	28.68	527.41	163.99	37.09	255.95
sow	porca	puerco	1	Diphthongization of short stressed 'o' > 'ue'; LA ending vowel added to	noun	3	1.67	0.60	1.61	4.19	2.10

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				SP stem.							
spade	pala	azada	4	From *ascia* "axe, mason's trowel"	noun	8	30.08	1.20	0.00	3.14	1.43
span (of hand)	palmus	palmo	1	LA ending vowel added to SP stem.	noun	12	16.85	0.40	2.82	8.39	3.80
speak, talk	loquor	hablar	4	From the noun *fabula* "conversation, story"	verb	18	547.31	215.80	457.51	1157.26	602.03
sphere, ball	sphaera	esfera	5	Semi-learned	noun	12	7.41	34.20	20.15	18.24	24.30
spider web	araneum	telaraña	3	From *tela* "web, net" and *araneus* "spider"	noun	3	6.51	0.80	2.22	19.50	7.33
spin	neo	hilar	4	From *filum* "a thread, string"	verb	6	678.63	4.40	1.41	3.56	3.12
spindle	fusus	huso	1	Word-initial 'f' > 'h' development; LA ending vowel added to SP stem.	noun	6	13.76	4.80	0.20	0.63	1.90
spine	spina	espina	1	e' insertion before word-initial 's' + consonant clusters; LA ending vowel added to SP stem.	noun	4	49.00	9.60	6.65	23.69	13.17
spit	spuo	escupir	4	Generalization: From *conspuere* "to spit upon, to spit upon in contempt"	verb	4	6.49	0.40	1.81	35.85	12.35
splash	aspergo	salpicar	2	"sal" + "picar"	verb	10	12.43	6.60	7.86	34.17	15.95
spleen	lien	bazo	4	Probably from *bayo* "reddish-brown"	noun	4	0.00	5.40	1.41	0.63	2.51
split	findo	hender	1	Must have come from a form in the paradigm with non-initial stress	verb	9	35.23	1.00	0.60	2.94	1.49
spoon	cochlear	cuchara	2	Derived irregularly through VL	noun	5	0.00	3.40	2.42	11.11	5.57
spread out	sterno	extender	5	Late borrowing from Latin	verb	9	76.63	348.00	77.36	135.43	188.01
spring	ver	primavera	3	From feminine of *primo vere* "in early spring"	noun	14	62.32	48.80	25.38	37.53	37.26

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
square	quadrus	cuadrado	4	From *quadra* "to square, make square"	noun	12	1.51	16.40	4.23	12.16	10.93
squeeze, wring	exprimo	exprimir	0	No change to stem.	verb	9	124.18	0.60	1.81	4.40	2.24
star	stella	estrella	2	Influence of *astrum*	noun	1	63.34	129.80	108.99	110.69	116.60
statue	statua	estatua	1	e' insertion before word-initial 's' + consonant clusters; LA ending vowel added to SP stem.	noun	9	77.45	32.00	13.30	46.12	30.27
steal	furor	robar	5	Germanic origin	verb	21	40.12	6.80	59.63	112.16	58.71
stepmother	noverca	madrastra	2	Analogy with *padrastro* 'stepfather'	noun	2	51.39	0.40	0.40	3.35	1.36
stingray	trygon	raya	4	From *radius* "wheel, ray (of light)"	noun	3	0.48	12.00	7.66	31.03	16.70
stir, mix	misceo	mezclar	2	Sporadic additional syllable from VL	verb	5	413.30	67.40	29.01	84.07	59.86
stocking	tibiale	media	4	From *medius* "half" and *calceus* "ankle-length shoe"	noun	6	0.00	9.80	22.16	52.20	27.69
stomach	stomachus	estómago	5	Semi-learned 'ch' > 'g' development indicates borrowing from Greek or Late Latin	noun	4	120.95	19.40	8.26	77.15	34.34
store, shop	taberna	tienda	4	From the verb *tendere* "to stretch"	noun	11	20.80	18.20	39.49	67.72	41.40
stove	focus	estufa	5	Greek origin	noun	7	58.80	3.60	3.42	7.55	4.82
stretch	tendo	tender	0	No change to stem.	verb	9	207.56	106.60	43.31	93.92	81.18
strife, quarrel	iurgium	pelea	4	From *pilus* "hair"	noun	19	0.00	4.20	24.58	37.53	21.85
strike (hit, beat)	verbero	golpear	4	From *colaphus* "blow with the fist"	verb	9	45.66	37.60	33.64	150.32	72.76
strong	fortis	fuerte	1	Diphthongization of short stressed 'o' > 'ue'; LA ending vowel added to SP stem.	adjective	4	453.45	198.20	166.61	202.10	188.82

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
study	studeo	estudiar	4	From the noun *studium* "study, eagerness"	verb	17	51.63	211.40	158.75	110.07	160.86
stupid	stupidus	estúpido	0	Regular 'e' insertion before word initial 's' + consonant clusters.	adjective	17	2.27	0.20	1.61	39.20	13.30
stutter, stammer	balbutio	tartamudear	4	Partly onomatopoeic (*tarta*), partly derived from *mutus* "silent, dumb"	verb	18	1.91	0.00	0.00	7.97	2.58
suck	sugo	mamar	4	From *mamma* "breast"	verb	5	4.42	1.60	0.81	5.87	2.71
sugar	saccharum	azúcar	5	Arabic origin	noun	5	0.00	97.20	38.08	39.41	58.57
summer	aestas	verano	4	From *ver* "spring"	noun	14	107.81	82.40	63.06	83.02	76.08
sun	sol	sol	0	No change to stem.	noun	1	273.84	133.60	88.24	407.14	206.87
supper	cena	cena	1	LA ending vowel added to SP stem.	noun	5	72.55	4.80	12.69	55.35	23.82
surprised, astonished	attonitus	asombrado	4	From *sub* "under" + *umbra* "shade"	adjective	16	0.00	0.00	0.00	0.00	0.00
surrender	dedo	rendir	4	From *reddere* "to give back, yield"	verb	20	227.03	34.20	55.00	36.48	41.94
suspect	suspicor	sospechar	4	From *suspicere* "to look up at, regard with awe"	verb	17	65.38	6.40	19.94	65.20	30.00
swallow	sorbeo	tragar	4	From *draco* "dragon"	verb	5	28.21	1.80	5.24	76.94	27.28
swamp	palus	ciénaga	4	From *caenum* "filth, mud"	noun	1	13.73	3.00	1.61	9.43	4.62
sweep	verro	barrer	2	No reason for e > a	verb	9	93.42	4.60	12.09	27.67	14.59
sweet	dulcis	dulce	1	LA ending vowel added to SP stem.	adjective	15	185.85	53.20	25.59	84.70	54.09
swelling	inflatio	hinchazón	4	From *inflare* "to blow into"	noun	4	26.06	2.40	0.20	4.19	2.24
swift, fast, quick	rapidus	rápido	0	No change to stem.	adjective	14	74.82	146.60	80.58	103.36	110.36
swim	nato	nadar	0	Intervocalic [t] voicing.	verb	10	79.40	11.80	8.26	29.56	16.36
sword	gladius	espada	4	Generalization: From *spatha* "blade	noun	20	88.20	20.20	9.27	31.03	20.02

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
				of a sword"							
tailor	sartor	sastre	5	Catalan origin	noun	6	0.00	1.20	2.22	8.81	4.00
tall	altus	alto	0	No change to stem.	adjective	12	17.02	533.61	445.02	343.20	442.12
taste	gusto	gustar	0	No change to stem.	verb	15	27.25	12.20	124.30	443.41	189.57
tattoo	stigma	tatuaje	5	French origin	noun	6	1.91	0.60	0.60	5.87	2.31
tax, tribute	tributum	tributo	1	LA ending vowel added to SP stem.	noun	11	11.42	19.60	17.93	3.98	13.98
teach	doceo	enseñar	4	From *in* "in" + *signare* "to mark, seal"	verb	17	462.06	48.20	44.72	130.19	73.57
tear	scindo	rasgar	4	From *resecare* "to cut loose, cut off"	verb	9	36.09	0.60	2.62	17.82	6.86
tell story	narro	contar	4	From *computare* "to reckon, compute, sum up"	verb	18	139.66	317.80	358.20	551.59	407.10
temple, church	templum	iglesia	4	From *ecclesia* "assembly of citizens"	noun	22	311.23	130.20	46.74	31.87	70.25
ten	decem	diez	0	Loss of word-final m, then [ki] > [ts], which results in orthographic z	number	13	46.93	86.20	186.75	180.09	150.47
tent	tentorium	tienda	4	From the verb *tendere* "to stretch"	noun	7	8.37	18.20	39.49	67.72	41.40
testicle	testiculus	testículo	1	LA ending vowel added to SP stem.	noun	4	21.04	8.80	2.62	5.45	5.63
thatch	culmus	paja	4	From *palea* "chaff"	noun	7	9.08	10.20	5.24	30.40	15.07
thick (in dimension)	grossus	grueso	0	Geminate simplification; diphthongization of short stressed 'o' > 'ue'	adjective	12	1.31	35.00	14.51	95.39	47.65
thief	fur	ladrón	4	From *latro* "mercenary soldier, brigand"	noun	21	23.66	1.80	2.01	38.58	13.78
thigh	femur	muslo	4	From *musculus* "muscle"	noun	4	41.71	3.00	2.22	59.33	20.97

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
thin (in dimension)	tenuis	delgado	4	From *delicatus* "dainty, charming"	adjective	12	195.53	38.00	32.03	67.30	45.47
thing	res	cosa	4	From *causa* "reason, cause, motive"	noun	11	1672.04	86.40	389.42	1217.43	554.65
think (= reflect)	cogito	pensar	4	From *pensare* "to weigh, ponder, consider"	verb	17	227.34	85.40	380.56	1315.34	583.02
third	tertius	tercer	4	From *tertiarus* "of a third"	number	13	159.36	195.00	219.79	75.26	164.59
thirst	sitis	sed	2	No reason for i > e	noun	5	55.71	6.00	6.04	0.21	4.14
thousand	mille	mil	0	Word final LL > L in O. SP (after loss of word-final e)	number	13	530.42	44.00	493.58	199.38	245.76
thread	filum	hilo	1	Word-initial 'f' > 'h' development; LA ending vowel added to SP stem.	noun	6	48.64	54.60	22.16	84.07	53.21
threaten	mino	amenazar	2	Presence of initial 'a' likely analogical	verb	18	91.65	68.60	91.06	58.49	72.89
three	tres	tres	0	No change to stem.	number	13	411.14	635.41	850.97	585.13	691.75
thresh	tero	trillar	4	From *tribulare* "to press"	verb	8	173.26	0.00	1.21	1.05	0.75
threshing floor	area	era	1	a > e under the influence of following r + yod	noun	8	33.22	0.00	0.00	0.00	0.00
throat	fauces	garganta	4	From *gurges* "whirlpool, gulf"	noun	4	109.00	12.80	7.86	82.81	33.80
throw	iacio	lanzar	4	From *lancea* "lance"	verb	10	0.00	116.40	170.84	185.12	156.99
thunder	tonitrus	trueno	4	From the verb *tonare* "to thunder"	noun	1	14.34	2.80	2.82	26.42	10.45
tide	aestus	marea	5	French origin	noun	1	96.93	28.20	8.06	15.09	17.17
tie, bind	ligo	atar	4	From *aptare* "to fit, adapt, apply, fasten"	verb	9	33.82	6.00	4.84	31.87	14.05

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
time	tempus	tiempo	1	Diphthongization of short stressed 'e' > 'ie'; LA ending vowel added to SP stem.	noun	14	740.54	778.61	868.29	1344.27	991.95
tired	fatigatus	cansado	4	From *campsare* "to sail around a headland"	adjective	4	0.00	4.20	15.51	108.39	41.74
to plow	aro	arar	0	No change to stem.	verb	8	103.36	2.80	3.83	4.19	3.60
tomorrow	cras	mañana	4	From *mane* "in the morning"	noun	14	7.65	11.60	162.38	362.69	176.06
tongs	forceps	tenaza	4	From *tenax* "tenacious, holding firmly"	noun	5	0.00	0.20	0.60	5.03	1.90
tool	instrument um	herramienta	4	Generalization: From *ferramentum* "iron tool"	noun	9	36.33	54.40	23.97	16.35	31.83
touch	tango	tocar	4	Onomatopoeic origin (De Silva)	verb	15	190.39	59.20	145.25	395.61	197.10
towel	gausapina	toalla	5	Germanic origin	noun	6	0.48	0.20	1.81	30.82	10.66
tower	turris	torre	1	LA ending vowel added to SP stem.	noun	20	94.18	43.80	55.40	42.98	47.44
trade, barter	cambio	cambiar	0	No change to stem.	verb	11	0.00	153.80	221.20	334.81	235.11
trap	irretio	atrapar	5	French origin	verb	20	0.96	24.80	26.79	61.01	37.19
trap	laqueus	trampa	5	Germanic origin	noun	20	26.06	6.60	12.89	44.24	20.90
tree	arbor	árbol	2	Sporadic r > l change	noun	8	271.55	134.00	48.75	233.76	137.58
tribe	tribus	tribu	1	LA ending vowel added to SP stem.	noun	19	106.25	89.60	10.07	19.08	39.98
trousers	bracae	pantalón	5	French origin	noun	6	1.35	7.80	15.92	120.55	47.04
true	verus	verdadero	2	"verdad" + "ero"	adjective	16	2039.47	106.80	147.67	158.91	137.44

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
trumpet	tuba	trompeta	5	French origin	noun	18	45.54	5.40	4.63	17.61	9.09
try, attempt	conor	intentar	4	From *intentare* "to stretch out, extend towards"	verb	17	131.00	243.20	173.46	222.65	213.05
turn around	gyro	volver	4	From *volvere* "to roll, turn, unroll"	verb	10	4.66	262.00	450.87	1526.25	734.92
turn over	verto	volver	4	From *volvere* "to roll, turn, unroll"	verb	10	309.51	262.00	450.87	1526.25	734.92
twelve	duodecim	doce	2	No reason for disappearance of 'u'	number	13	44.70	41.80	62.65	71.70	58.51
twenty	viginti	veinte	2	No reason for long-i > e, assumed analogy with other numbers	number	13	92.51	34.40	80.18	121.18	77.92
twin	geminus	gemelo	4	From *gemellus*, diminutive of *geminus* "twin"	noun	2	132.19	4.80	6.25	13.63	8.14
two	duo	dos	2	No reason for the disappearance of 'u'	number	13	521.82	1424.62	1584.68	1393.12	1468.35
uncle	patruus	tío	5	Greek origin	noun	2	21.15	14.00	14.30	238.37	86.74
under	sub	bajo	5	Greek origin	preposition	12	710.90	611.01	358.20	530.20	499.67
untie	solvo	desatar	4	From *aptare* "to fit, fasten"	verb	9	350.79	6.00	32.03	29.14	22.26
up, above	super	sobre	2	Influence of *supra*	preposition	12	25.38	1515.42	1604.22	1720.17	1611.62
urinate	mingo	orinar	4	From the noun *urina* "urine"	verb	4	4.90	1.00	0.81	18.87	6.72
vegetables	holus	verdura	4	From *verde* "green"	noun	5	0.00	15.20	5.44	13.84	11.47
village	vicus	aldea	5	Arabic origin	noun	19	79.80	15.80	18.13	20.76	18.19
vine	vitis	vid	0	Intervocalic [t] voicing.	noun	8	217.04	6.80	0.81	1.89	3.19
vomit	vomo	vomitar	1	From the past participle form *vomitus*	verb	4	77.09	0.40	0.81	24.53	8.35
vulture	vultur	buitre	1	v > b is irregular in spelling only	noun	3	7.13	6.20	2.01	6.71	4.95

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
waist	cingulum	cintura	4	From *cingere* "to gird"	noun	4	2.15	6.60	6.65	70.44	27.28
wake up	expergiscor	despertar	4	Generalization: From *de* "completely" + *expergisci* "to become awake"	verb	4	11.71	26.60	53.99	301.69	124.88
war	bellum	guerra	5	Frankish origin	noun	20	652.63	902.01	226.64	210.28	450.54
wash	lavo	lavar	0	No change to stem.	verb	9	91.21	17.80	30.62	98.74	48.32
wasp	vespa	avispa	2	Influence of *abeja*	noun	3	6.21	11.60	1.01	6.50	6.38
waterfall	cataracta	cascada	5	Italian origin	noun	1	1.79	5.20	7.86	11.32	8.08
wave	fluctus	ola	5	Arabic origin	noun	1	139.84	35.00	31.83	49.06	38.48
we	nos	nosotros	3	From *nos alteros* "we others" (emphatic)	pronoun	2	1713.30	52.00	194.81	397.29	211.90
weak	debilis	débil	0	No change to stem.	adjective	4	20.08	40.20	32.03	39.62	37.26
weather	tempestas	tiempo	4	From *tempus* "time"	noun	1	152.98	778.61	868.29	1344.27	991.95
week	hebdomas	semana	4	From *septem* "seven"	noun	14	0.00	74.60	564.29	222.86	287.57
weigh	pendo	pesar	4	From *pensare,* frequentative of *pendere* "to weigh"	verb	11	104.42	113.20	176.08	161.01	149.86
west	occidens	oeste	5	Old English origin (via French)	noun	12	8.78	190.00	19.74	11.11	74.73
whale	balaena	ballena	5	Presence of 'll' in Spanish indicates semi-learned or borrowed origin, likely from Latin	noun	3	0.72	28.40	5.84	6.08	13.57
where?	ubi	donde	3	From *de unde* "from whence"	adverb	17	1147.14	702.61	908.58	1115.33	905.62
which?	quis	cual	4	From *qualis* "of what kind, how constituted"	pronoun	17	6525.85	664.21	1236.56	2982.47	1607.55

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
whirlpool	vertex	remolino	4	From *mola* "millstone"	noun	1	122.63	3.00	1.61	20.76	8.28
whisper	susurro	susurrar	0	No change to stem.	verb	18	0.90	0.00	3.22	41.51	14.52
white	albus	blanco	5	Germanic origin	adjective	15	72.32	167.80	201.66	414.06	258.93
wide, broad	latus	ancho	4	From *amplus* "wide, ample"	adjective	12	85.20	38.80	11.48	86.38	45.00
widow	vidua	viuda	2	Metathesis	noun	2	8.09	9.60	17.12	31.66	19.28
widower	viduus	viudo	2	Metathesis	noun	2	15.02	0.60	1.21	3.14	1.63
wife	uxor	esposa	4	From *sponsus* "betrothed man, groom"	noun	2	220.63	44.00	104.36	133.76	93.39
window	fenestra	ventana	4	From *ventus* "wind"	noun	7	17.09	19.60	28.61	321.39	120.34
witch	saga	bruja	5	Celtiberian origin	noun	22	1.93	7.20	6.85	26.21	13.24
witness	testis	testigo	4	From *testificari* "to testify"	noun	21	114.70	18.80	77.56	56.61	50.84
wood	lignum	madera	4	From *materia* "tree trunk, timber, material, matter"	noun	1	72.19	249.60	47.34	152.00	149.86
work	laboro	trabajo	4	From *tripalus* "having three stakes"	noun	9	204.65	400.00	598.13	356.82	452.77
worship	adoro	adorar	0	No change to stem.	verb	22	27.73	11.00	8.26	32.71	17.10
wound	plaga	herido	4	From *ferire* "to strike, beat, cut"	noun	4	30.93	12.20	37.87	69.39	39.37
wrinkled	rugatus	arrugado	2	Word initial 'a' by analogy (with 'arare'? to plow)	adjective	15	0.00	0.00	0.00	0.00	0.00
wrist	pugnus	muñeca	5	Pre-Roman (possibly pre-Indo-European) origin	noun	4	8.37	7.00	6.45	73.80	28.44
write	scribo	escribir	0	Regular 'e' insertion before word initial 's' + consonant clusters.	verb	18	648.50	526.21	317.30	414.69	419.79
wrong	falsus	falso	0	No change to stem.	adjective	16	84.76	22.80	56.81	53.88	44.32
yard, court	area	patio	5	Provencal origin	noun	7	33.22	27.80	33.64	190.78	82.53

Meaning	Latin	Spanish	Code	Coding Notes	PoS	Sem Cat	LA-norm	SP-ACAD norm	SP-NEWS norm	SP-FICT norm	SP-ALL norm
yawn	oscito	bostezar	3	From *boca* "mouth" and *oscitare* "to yawn"	verb	4	4.54	0.00	0.60	17.61	5.90
yellow	flavus	amarillo	4	From *amarus* "bitter"	adjective	15	44.82	37.20	26.79	108.81	56.88
yesterday	heri	ayer	2	Appearance of word-initial 'a' by analogy with other adverbs, e.g. *afuera* "outside"	noun	14	4.18	0.20	1.21	4.82	2.04
young woman (adolescent)	puella	joven	4	From *juvenis* "young, young person"	noun	2	163.26	100.00	228.05	257.24	194.05
zero, nothing	nihil	nada	3	From *nulla res nata* "no thing born"	noun	13	1867.60	5.40	5.64	27.25	12.56

Appendix B.

List of Latin Texts Used (accessed from the Perseus database)

Author	Title	No. of words	Est. date
Caesar Augustus	Res Gestae Divi Augusti	2615	63 BCE-14 CE
C. Julius Caesar	De bello civili	32339	100-44 BCE
C. Julius Caesar	De bello gallico	51295	100-44 BCE
C. Valerius Catullus	Carmina	12857	84-54 BCE
M. Tullius Cicero	Academica	5176	106-43 BCE
M. Tullius Cicero	Brutus	30473	106-43 BCE
M. Tullius Cicero	De amicitia	9366	106-43 BCE
M. Tullius Cicero	De divinatione (muller)	27540	106-43 BCE
M. Tullius Cicero	De fato	5540	106-43 BCE
M. Tullius Cicero	De finibus bonorum et malorum	58771	106-43 BCE
M. Tullius Cicero	De inventione	33488	106-43 BCE
M. Tullius Cicero	De legibus	18836	106-43 BCE
M. Tullius Cicero	De natura deorum	36054	106-43 BCE
M. Tullius Cicero	De officiis	34225	106-43 BCE
M. Tullius Cicero	De optimo genere oratorum	1953	106-43 BCE
M. Tullius Cicero	De oratore	62166	106-43 BCE
M. Tullius Cicero	De partitione oratoria	11057	106-43 BCE
M. Tullius Cicero	De republica	22317	106-43 BCE
M. Tullius Cicero	De senectute	8355	106-43 BCE
M. Tullius Cicero	Epistulae ad familiares	120835	106-43 BCE
M. Tullius Cicero	Letters to and from Brutus	9789	106-43 BCE
M. Tullius Cicero	Letters to and from Quintus	18672	106-43 BCE
M. Tullius Cicero	Letters to Atticus	127209	106-43 BCE
M. Tullius Cicero	Lucullus	22313	106-43 BCE
M. Tullius Cicero	Orationes, cum senatui gratias egit, cum populo gratias egit, de domo sua, de haruspicum responso, pro sestio, in vatinium, de provinciis consularibus, pro balbo	78712	106-43 BCE
M. Tullius Cicero	Orationes, divinatio in Q. Caecilium, in C. Verrem	132170	106-43 BCE

Author	Title	No. of words	Est. date
M. Tullius Cicero	Orationes, pro milone, pro marcello, pro ligario, pro rege deiotaro, philippicae I-XIV	93406	106-43 BCE
M. Tullius Cicero	Orationes, pro P. Quinctio, pro Q. roscio comoedo, pro A. caecina, de lege agraria contra rullum, pro C. rabio perduelliones reo, pro L. flacco, in L. pisonem, pro C. rabiro postumo	66836	106-43 BCE
M. Tullius Cicero	Orationes, pro sex. roscio, de imperio cn. pompei, pro cluention, in catilinam, pro mureno, pro caelio	90438	106-43 BCE
M. Tullius Cicero	Orationes, pro tullio, pro gonteio, pro sulla, pro archia, pro plancio, pro scauro	41643	106-43 BCE
M. Tullius Cicero	Orator	22860	106-43 BCE
M. Tullius Cicero	Paradoxa stoicorum ad m. brutum	4639	106-43 BCE
M. Tullius Cicero	Timaeus	4353	106-43 BCE
M. Tullius Cicero	Topica	7438	106-43 BCE
M. Tullius Cicero	Tusculanae disputationes	73979	106-43 BCE
Q. Tullius Cicero	Essay on running for consul	4356	102-43 BCE
Q. Horatius Flaccus	Carmina	13292	65-8 BCE
Q. Horatius Flaccus	Satyrarum libri	14372	65-8 BCE
Q. Horatius Flaccus	De arte poetica liber	3090	65-8 BCE
Q. Horatius Flaccus	Carmen saeculare	313	65-8 BCE
Q. Horatius Flaccus	Epodon	3006	65-8 BCE
Q. Horatius Flaccus	Epistles	9905	65-8 BCE
Titus Livius	The history of rome (1-10) (WEISSENBORN)	159186	59 BCE- 17 CE
Titus Livius	The history of rome (21-30) (WEISSENBORN)	152951	59 BCE- 17 CE
Titus Livius	The history of rome (31-38) (WEISSENBORN)	107251	59 BCE- 17 CE
Titus Livius	The history of rome (39-40) (WEISSENBORN)	29532	59 BCE- 17 CE
Titus Livius	The history of rome (41-45) (WEISSENBORN)	56156	59 BCE- 17 CE

Author	Title	No. of words	Est. date
Lucretius	De rerum natura	49028	99-55 BCE
Cornelius Nepos	Vitae	28128	110-25 BCE
P. Ovidius Naso	Amores, epistulae, medicamina faciei feminae, ars amatoria, remedia amoris	41390	43 BCE- 18 CE
P. Ovidius Naso	Ex ponto	21481	43 BCE- 18 CE
P. Ovidius Naso	Fasti	31610	43 BCE- 18 CE
P. Ovidius Naso	Ibis	4032	43 BCE- 18 CE
P. Ovidius Naso	Metamorphoses	78098	43 BCE- 18 CE
P. Ovidius Naso	Tristia	23590	43 BCE- 18 CE
Sextus Propertius	Elegies	4384	50-15 BCE
Tibullus	Elegiae	12364	55-19 BCE
P. Vergilius Maro	Aeneid	63719	70-19 BCE
P. Vergilius Maro	Eclogues	5757	70-19 BCE
P. Vergilius Maro	Georgicon	14183	70-19 BCE
Vitruvius Pollio	De architectura	57619	80-15 BCE
Aulus Cornelius Celsus	De medicina (spencer)	102500	25 BCE- 50 CE
Lucius Junius Moderatus Columella	Res rustica, books I-IV	49850	4-70 CE
Lucius Junius Moderatus Columella	Res rustica, books V-IX	51045	4-70 CE
C. Valerius Flaccus	Argonautica	50596	pre-90 CE
Lucius Annaeus Florus	Epitome rerum romanorum	30049	74-130 CE
Juvenal	Satires	26564	late 1st early 2nd c. CE
M. Annaeus Lucanus	Pharsalia	51066	39-65 CE
Martial	Epigrammata	58314	40-104 CE
Persius	Satires	5769	34-62 CE
Petronius	Satyricon, fragmenta, and poems	34460	27-66 CE
Phaedrus	Fabulae aesopiae	11562	15 BCE- 50 CE
Pliny the Elder	Naturalis historia	404690	23-79 CE
Pliny the Younger	Letters	66670	61-112 CE
Quintilian	Institutio oratoria (preface, books 1-3)	45271	35-100 CE
Quintilian	Institutio oratoria (4-6)	41913	35-100 CE
Quintilian	Institutio oratoria (7-9)	47435	35-100 CE
Quintilian	Institutio oratoria (10-12)	42080	35-100 CE
Quintus Curtius Rufus	Historiae alexander magni	74333	~41-79 CE

Author	Title	No. of words	Est. date
Seneca the Elder	Controversiae	65738	54 BCE- 39 CE
Seneca the Elder	Fragmenta	144	54 BCE- 39 CE
Seneca the Elder	Suasoriae	10128	54 BCE- 39 CE
Seneca	Ad lucilium epistulae morales	128089	4 BCE - 65 CE
Seneca	Agamemnon	5583	4 BCE - 65 CE
Seneca	Apocolocyntosis	2989	4 BCE - 65 CE
Seneca	De beneficiis	46092	4 BCE - 65 CE
Seneca	De brevitate vitae	6406	4 BCE - 65 CE
Seneca	De clementia	8312	4 BCE - 65 CE
Seneca	De consolatione ad helvium	6871	10 BCE- 65 CE
Seneca	De consolatione ad marciam	8845	10 BCE- 65 CE
Seneca	De consolatione ad poylbium	5808	10 BCE- 65 CE
Seneca	De constantia	5334	10 BCE- 65 CE
Seneca	De ira	22675	10 BCE- 65 CE
Seneca	De otio	1984	10 BCE- 65 CE
Seneca	De providentia	4132	10 BCE- 65 CE
Seneca	De tranquillitate animi	7887	10 BCE- 65 CE
Seneca	De vita beata	7517	10 BCE- 65 CE
Seneca	Hercules furens	7653	10 BCE- 65 CE
Seneca	Hercules oetaeus	11325	10 BCE- 65 CE
Seneca	Medea	5693	10 BCE- 65 CE
Seneca	Octavia	5259	10 BCE- 65 CE
Seneca	Oedipus	5938	10 BCE- 65 CE
Seneca	Phaedra	7206	10 BCE- 65 CE
Seneca	Phoenissae	4112	10 BCE- 65 CE
Seneca	Thyestes	6315	10 BCE- 65 CE
Seneca	Troades	6834	10 BCE- 65 CE
Silius Italicus	Punica	76675	28-103 CE
P. Papinius Statius	Achilleis	7598	45-96 CE
P. Papinius Statius	Silvae	28828	45-96 CE
P. Papinius Statius	Thebais	66886	45-96 CE
C. Suetonius Tranquillus	De vita caesarum: divus julius, divus augustus, tiberius, caligula, divus claudius, nero, galba, otho, vitellius, divus vespasianus, divus titus, domitianus	70338	69-122 CE
Cornelius Tacitus	Annales	88905	56-117 CE
Cornelius Tacitus	De origine et situ germanorum	5513	56-117 CE

Author	Title	No. of words	Est. date
	liber		
Cornelius Tacitus	De vita iulii agricolae	6740	56-117 CE
Cornelius Tacitus	Dialogus de oratoribus	9309	56-117 CE
Cornelius Tacitus	Historiae	51495	56-117 CE
Valerius Maximus	Facta et dicta memorabilia	80125	14-37 CE

Appendix C.

Awk Script

Script used for extracting frequency data from the PDLP.

```
#!/bin/sh

htmlfile="${1:-caesar100table.html}" ## The default html file is "caesar100table.html", in the current directory.

keywordfile="${2:-keywords.txt}" ## The default keywords file is "keywords.txt", in the current directory.

< "${keywordfile}" \

tr '\r' '\n' | \

while read keyword

  do

    printf "${keyword},"

    < "${htmlfile}" \

    tr '\r' '\n' | \

    grep -A 3 ">${keyword}<" | \

    tail -n 1 | \

    sed \

      -e 's/.*<td>//' \

      -e 's/<.*//'

  done
```