

Copyright
by
Robert Blaine Elwell
2007

The Role of Frequency in Historical Change

by

Robert Blaine Elwell, B.A.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF ARTS

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2007

The Role of Frequency in Historical Change

APPROVED BY

SUPERVISING COMMITTEE:

Robert D. King, Supervisor

To my father.

Acknowledgments

I would like to take the time to extend my gratitude to those who have offered their support during my hectic time thus far as a graduate student:

- My family, for having supported and encouraged me throughout my life and my studies.
- Kayleigh A. Gaddor, for being endlessly patient with my various linguistic riffings.
- All those who had a part in my empirical interests—those in the UT Computational Linguistics group, and the Linguistics department in general.
- All those who I didn't mention and most probably deserve acknowledgment here.

The Role of Frequency in Historical Change

Robert Blaine Elwell, M.A.
The University of Texas at Austin, 2007

Supervisor: Robert D. King

Formal representations of historical change often make extreme implications about the evolution of language. This report seeks to frame language information-theoretically as an arbitrarily held codification of meaning that is constantly self-correcting, trying to find the most optimal characteristics given physical and cognitive constraints. Using this assumption, I show how frequency plays a major role in motivating or mitigating major systemic changes upon language, and give evidence from computational experiments and corpus data which translates to very clear theoretical and functional explanations.

Table of Contents

Acknowledgments	v
Abstract	vi
List of Tables	ix
List of Figures	x
Chapter 1. Introduction	1
1.1 Linguistic Encoding	2
1.2 Frequency and Change	3
1.2.1 Catalysis	3
1.2.2 Mitigation	5
1.2.3 Interplay	6
Chapter 2. Frequency as Catalysis for Historical Change	7
2.1 Phonological Change	8
2.1.1 Latin Post-Tonic Vowel Syncope	10
2.1.1.1 Introduction	10
2.1.1.2 Experiment	13
2.1.1.3 Results	15
2.1.1.4 Summary	16
2.2 Morphology	17
2.2.1 The Null Present Tense Morpheme in Bantu	17
2.2.1.1 Background and Hypothesis	17
2.2.1.2 Statistics and Experiment	18
2.2.1.3 Results and Discussion	19
2.2.1.4 Summary	20

2.2.2	The Null Singular Morpheme in English	21
2.2.2.1	Background and Hypothesis	21
2.2.2.2	Statistics and Experiment	22
2.2.2.3	Summary	24
2.3	Summary	24
Chapter 3.	Frequency's Role in Mitigating Regularity	26
3.1	The Romance Headedness Shift	27
3.1.1	Background	27
3.1.2	Portuguese	29
3.1.3	Spanish	29
3.1.4	French	33
3.1.5	Summary	34
3.2	Frequency and Irregular Inflection	34
3.2.1	English	35
3.2.2	Portuguese	38
3.2.3	French	40
3.2.4	Spanish	43
3.2.5	Swahili	43
3.2.6	Summary	46
3.3	Chapter Summary	48
Chapter 4.	The Interplay of Frequency-Based Mitigation and Catalysis in Historical Change: The Case of the Copula	49
4.1	Description	49
4.2	Investigation	51
4.3	Cross-Linguistic Observations	53
4.4	Conclusion	54
Chapter 5.	Conclusion	55
Bibliography		56
Vita		59

List of Tables

2.1	Entropy measurements given MLE probability distributions for tri-grams.	15
-----	---	----

List of Figures

2.1	Calculation of Entropy	7
2.2	Prosody-based analysis of vowel syncope. [13]	10
2.3	Entropy Measurements Per Data Set	20
2.4	Frequency and probability for noun tags in the Brown Corpus of Standard American English	24
3.1	Bigram tag statistics in the Floresta Synta(c)tica corpus of Portuguese	29
3.2	The ten most frequent adjectives in the Floresta Synta(c)tica corpus of Portuguese	30
3.3	The five most frequent adjectives in the HUB5 Spanish Transcripts corpus	30
3.4	The five most frequent adjectives in the Europarl Corpus – Spanish .	32
3.5	The five most frequent nouns in the Europarl Corpus – Spanish . . .	32
3.6	The ten most frequent adjectives in the Europarl corpus – French section	33
3.7	The five most frequent nouns in the Europarl corpus – French section	33
3.8	The ten most frequent verbs in the Brown Corpus of English	36
3.9	All verbal forms in English	39
3.10	Irregular verbal forms in English	39
3.11	Regular verbal forms in English	40
3.12	The 20 most frequent verbs in the Floresta Synta(c)tica Corpus of Portuguese	41
3.13	The 18 most frequent verbs in the Europarl Corpus – French	42
3.14	The 19 most frequent verbs in the Europarl Corpus – Spanish	44
3.15	The 14 most frequent verbs in the Hub5 Spanish Corpus	45
3.16	The 10 most common verb forms in the Helsinki Corpus of Swahili	46

Chapter 1

Introduction

Since the inception of the modern study of linguistics with Saussure, historical linguistics has been a driving force in the discipline. Diachronic change has been documented, analyzed, and commented upon, but rarely explained. As time has progressed, the pursuit has encountered data and discovered relations of languages and their past lives in other generations using very clear formalisms with little computational complexity. While the finite-state nature of generative phonology is a clear and accessible standpoint from which to analyze language, it does very little to explain why the change occurred. Movements in the middle of the 20th Century to use more complex mathematical approaches to language were gradually stemmed by the "Chomskyan Revolution", which regarded data-intensive linguistic methods as inherently flawed compared to using introspective data from a speaker's linguistic intuitions. Due to the rationalist approaches that would come from the generative phonological and syntactic frameworks, little explanation has still been given for linguistic change that has received broad attention.

The modern zeitgeist of linguistics has, however, slowly become more accepting of empirical approaches, and rediscovering what should have been great scholarly victories made by linguists during the height less data-intensive methods

that simply did not receive the proper acknowledgment. One goal of this report is to recognize these past achievements at this juncture in linguistics, where empiricism is again taking hold—and with good reason. A further goal is to continue in their tradition, focusing on a measurable concept which is rampant in a variety of aspects of language—frequency. Frequency measurements have become even more useful at this stage. While the unavailability of corpora and the computationalintensive-ness of utilizing frequency measurements in calculations during the '60s, '70s, and even '80s to an extent made such research a difficult proposition, the modern proliferation of digital texts and academically-led efforts to develop new corpora has significantly facilitated a move back to linguistics research focused on concrete, quantitative results that will truly serve as sources of insight.

1.1 Linguistic Encoding

In this work, I will be operating under a singular fundamental notion: that language is at its base a way to encode meaning in a manner that optimizes for both ease and eloquence. A speaker desires to transmit vocally some amount of information that is clear, coherent, and precise in the information it conveys, but also easy to pronounce, recall, and store (thus having at one's disposal as a word, phoneme, et cetera). However, they are working under fundamental constraints of general intelligence, complexity of codification, and physical abilities. Boersma [5] views this in Optimality Theory as a cycle of "eternal optimization", with fundamental functional principles consistently at odds with one another, spurring the process of historical change.

My argument given here is that speakers of languages utilize the linguistic data in the world around them and the unconscious knowledge of the language that they have developed to make the language easier for them to speak and easier to understand. These two ends are often at odds with one another, making this an excellent motivation for the underlying concept behind Optimality Theory. However, much of Optimality is treated in simply categorical terms, resulting in problems addressing synchronic opacity and diachronic chain shifts or historical changes in general. The mechanism as it stands can also be treated as finite-state, resulting in an improbable model of how individuals understand, process, and produce language.

I will also take time to address frequency and its role in syntactic and morphological change. For both of these, there are clear, logically motivated frameworks that I will be working in to make my case. However, in many cases, formal models of current theory in linguistics will not provide a sufficient backdrop upon which to couch the quantitative explanations given here. For these instances, I will provide the data and the explanation which best suits it, though this may be cause to reconsider certain theoretical underpinnings of a particular formalism.

1.2 Frequency and Change

1.2.1 Catalysis

Frequency affects linguistic change in two major ways that are provocatively at odds with one another. Through the frequency of a linguistic item (be it a word or a phoneme), we can arrive at the likelihood or uncertainty of its presence given

a specific context. Because this coincides so closely with general intelligence task of deductive reasoning, it is an advantageous position to take: the same underlying cognitive mechanism that could be used for word prediction may also be that which leads an individual to know that since they've been through so many storms in their lifetime, they know if their knee aches it will rain.

This deductive knowledge, given enough time in a system and the opportunity to make changes upon it, will result in the ability to make changes to the system to ease encoding without a loss of actual information. This has been treated by [16] as the noisy channel model, where one can probabilistically retrieve the intended input from output that has been altered through its actual production. In this model of any system where information is conveyed, there is a similar delicate balance between compression, or reducing the amount of encoding to be transmission by removing all redundancy from the encoding, and transmission accuracy, which encodes an amount of redundancy such that the intended input can be recovered even if it is altered by the noisy channel.

If we view language as an arbitrarily shared encoding that is regularly both altered by speakers who have learned it and learned by other speakers contemporaneously, then we would see two different actual mental models of the language: one utilizing the noisy-channel model to extract the correct information from speech, and one that perceives the noisy output as the input. This speaker in turn will make his or her own changes to the newly cached noisy input and create an even noisier output. This analogizes language to a telephone game that spans all of human history, which actually makes an entertaining amount of abstract sense. Frequency

therefore *catalyzes* historical change, allowing changes to take place which could not if an individual were not able to make deductions about both the language they know and the words they hear.

In this report, I will be offering evidence of this very assumption through phonological, syntactic, and morphological changes using a series of quantitative experiments.

1.2.2 Mitigation

While we see that frequency assists in deductive reasoning that allows change to occur, it also performs a completely different task: impeding regular historical change. Indeed, regular change is not categorical change because words with a high enough frequency seem to be impervious to more sweeping changes. There are two ways to look at this phenomenon: either (a) more frequent items are impervious to regular change because their frequency makes them exceptional or (b) less frequent items are treated psychologically as a set of items all subject to the same rules, and more frequent items do not fall into this set because they are in some way exceptional, perhaps stored or recalled in a different manner than those other items. My preferred viewpoint is (b). This would make sense from a diachronic change perspective, because as these items change in predictable ways, a speaker is both able to interpret them through the noisy channel and learn them taken at face value.

Items that would not change because they are especially frequent therefore do not change because an innovation on it would not "stick". Imagine being the only person trying to make "be" more regular by saying "Stephen Hawking bes a

physicist”; not only would you not be easily understood, but you would possibly be ridiculed—both forms of discouragement that would easily prevent the change from taking place in even one individual’s grammar. Furthermore, because of the way that these forms are learned and possibly stored, attempting this change could even be awkward for the individual attempting it.

To support my argument, I will offer evidence that historically, many of the most infrequent items were not affected by regular alterations to the encoding of language. These can be seen in cases of phonological, morphological, and syntactic irregularity cross-linguistically, using comparative frequency measurements.

1.2.3 Interplay

One question this paper will address is, how do we capture the interaction of frequency-based motivations for historical change if they operate in inverse manners? There is a clear cross-linguistic example that I will present and account for using the viewpoints built in the following chapters. I show that what is proven in this work is the only possible explanation for a diachronic change that exists fractured by irregularity despite its frequency-based simplification in other senses.

Chapter 2

Frequency as Catalysis for Historical Change

Frequency's role in assisting in diachronic change can be framed in terms of entropy. Entropy was first coined by Shannon [16] as a measurement of uncertainty upon a probability distribution. This formula, shown in 2.1, returns a number that essentially describes the number of bits necessary to encode a set of items in binary given their assorted probabilities, which is calculated as frequency of the item over the probability mass, or summation of all frequencies of items in the list.

Shannon [17] used this to some success in his own research in the calculation of entropy in the English language, finding the entropy a 26-character alphabet contains as opposed to a 27-character system by including whitespace as a character. Intuitively, the usage of whitespace significantly reduces the entropy of written language, attributing to its usage by many groups. Shannon also calculated entropy of words, and, as such a measurement should assume, function words such as 'the' and 'of' are significantly less entropic than a given noun or verb, which is also fairly

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x)$$

Figure 2.1: Calculation of Entropy

intuitive, given the information that open-class word types versus closed-class word types carry.

Since this time, frequency has played a sporadic role in linguistic literature outside of computational linguistics. Mainstream theory in most fields of the discipline has been mainly influenced by the innateness hypothesis of language, causing formal work to generally avoid frequency as a source of answers to theoretical linguistic questions. In one sense, this is partially due to the fact that much of the fundamental underpinnings of this point of view on language immediately rule out the consideration of external data save in largely anecdotal examples; in another, it is because divulging that external data can be analyzed and drawn upon for learning is a threat to innate linguistic capability and would force one to re-examine acquisition as a question general intelligence exactly what those who view language as an instinct would prefer not to concede.

2.1 Phonological Change

Frequency—especially as it relates to entropy—and phonological change is a fairly easy abstraction to make if we are viewing language as a type of encoding. In the case of phonology, an individual has a desired output of sound that encodes semantic information that will be ultimately ordered in a syntax so as to assist correct combination of semantic propositions and arguments. Focusing on this first level, the sound pattern of a language as an encoding in and of itself is perhaps the least abstracted view of language in this manner.

In a linguistic event, an individual must take an output that has been made

noisy and in some way deduce the intended input, relating phoneme strings to semantic units. This output has been altered through the speaker's phonetic coarticulation and phonological attempts at reducing the complexity of the entire code itself while reducing the difficulty of pronunciation. One of these is a physical constraint, but the other is also a case where deduction has led the speaker to alter the the input for the sake of ease—constrained against, again, a hearer's ability to understand the intended message. If both sides of these processes of deduction are in some way probabilistic, then language change as being constant, fluid, and regular can be easily explained.¹

There have been several excellent past examples that have proven the perspicacity of incorporating frequency calculations in historical change. Fidelholtz [7] showed that word-initial vowel reduction was a function of frequency; more frequent words like 'astronomy' would undergo the process, while less frequent words such as 'gastronomy' would not. Similarly, Hooper [10] found that vowel elision is a function of frequency with words with a relatively high frequency being subject to the change, and therefore 'memory' being optionally tri- or bi-syllabic, but 'mammory' being mandatorily tri-syllabic. In this section, I will give a description of another frequency-induced historical change, and offer an information-theoretic explanation to account for all three of these changes.

¹The exceptions to regularity will be conveniently accounted for in the chapter on mitigation of historical change.


/manica/	FOOTTROCHEE	PARSE-SYL	MAX-SEGMENT
a. (má.ni.ca)	*!		
b. (má.ni)ca		*!	
c.  (mán.ga)			*

Figure 2.2: Prosody-based analysis of vowel syncope. [13]

2.1.1 Latin Post-Tonic Vowel Syncope

2.1.1.1 Introduction

Post-tonic syllabic peak deletion in Latin is a well-documented phenomenon that attributed to the historical direction which many Romance languages took—Spanish being one of these. According to an analysis by Lleó [13], this results from a series of optimality-theoretic constraint interactions, mainly focused around the prosodic well-formedness of footed or unfooted syllables in Latin. An example of this is the change from Latin MANICA, ‘sleeve’, to Spanish *manga*, where the first syllable is stressed and the following syllable loses its peak, thus causing its onset to become a coda on the preceding syllable. Lleó frames this as the domination of the workhorse constraints PARSE-SYL and FOOTTROCHEE (among others) over the constraint MAX-SEGMENT:²

While a syllable-structure-based analysis is of some theoretical interest, there are a few shortcomings in taking this approach. First, a change in a constraint ranking that originally allowed post-tonic syllables over a syllabic well-formedness rule assumes a great degree of change deriving from a single reranking. However, a set of rerankings carries its own set of questions, such as whether there is any

²For a more thorough description of the analysis, please see [13]

data to describe the sequence of constraint rerankings, or why so many constraints would rerank. If the change stemmed from contact, it would be difficult to believe that stress would be the first to change, as it is generally the case that single lexical borrowings will be forced to agree with the phonotactics of the borrowing language and not that of its native origins. Therefore, the clearest explanation would be that this is a case of internal phonological change.

Internal change can be best understood as the interaction of relexification with the conflict that exists between different functional principles that carry their own fixed or partly fixed constraint ranking hierarchies. These functional principles are eternally at odds at each other, and their constant reranking with respect to one another yields a cycle of “eternal optimization”. [5] This concept will be considered as the underlying postulation regarding what will be argued here. Functional principles that lead to fixed rankings for prosodic words will be considered as follows:³

- (2.1)
1. Minimization of articulatory effort.
 2. Maximization of perceptual contrast.
 3. Maximization of utilization of prosodic proclivities.

The constraint rankings which each of these conflicting principals yield would interact in the manner which ultimately resulted in the change from Latin to Old Spanish. (1), in the context of the prosodic word, can clearly be understood

³Adapted from the functional principles for obstruents in [5].

as a case of where a greater number of syllables would increase articulatory effort. (2), for the case of the prosodic word would, be such that every segment contributes to discerning words. For instance, a language consisting of the two sequences ‘xyz’ and ‘wyz’ would not be a proper maximization of perceptual contrast, while one consisting of ‘xyz’ and ‘abc’ would. (3) simply refers to the importance that every utterance be feasible within the precepts of the established sound pattern—especially with regards to stress.

Under this paradigm, Latin originally existed with the set of constraints that reflected (3) probably equally ranked with those reflecting (1). (2) would have been preventing deletion of vowels to maintain as much contrast as possible. The change stemmed from the continual striving for optimality, because over time, as the prosodic template became concretized, the information carried in words by post-tonic vowels was weakened, as they had the least presence in the word so as to contrast with the stressed syllable preceding it. This phonetic characteristic would lead to a minimized perceptual contrast in this vowel. Because of the emphasis on the preceding vowel, and such concepts as the obligatory contour principal mitigating the appearance of certain vowels (always stressed vowels, often long vowels), the number of different phonemic vowels was also reduced. A post-tonic unstressed vowel, therefore, is extremely more predictable than a vowel following an unstressed segment. This predictability directly correlates with the deletion of these vowels.

My hypothesis is that post-tonic vowels in Latin were lost because they carried little information and were by and large more predictable than other vowels.

This predictability was inversely related to the carrying of new information or perceptual contrast, rendering these vowels a weak link in the prosodic word. As the stress scheme of Latin changed to that of Old Spanish, losing nonfinality, in order to maintain all syllables in the prosodic word parsed in a trochaic, binary foot, these vowels were deleted for a minimal loss in actual perceptual contrast.

2.1.1.2 Experiment

To make my case, I will frame my argument within the concept of entropy. Using texts from the Latin Library⁴ annotated for syllable boundaries and vowel length, I will use a trigram model to calculate the maximum likelihood estimate (MLE) of all possible segments in the third position of a trigram. This probability distribution is what I will calculate the entropy for. Therefore, my hypothesis is that the entropy of the probability distribution of the MLE of a vowel position following a stressed vowel will be significantly lower than the entropy of probability distribution of the MLE of a vowel position following an unstressed vowel. Maximum likelihood estimate for trigrams are calculated as follows:

$$P_{MLE}(w_3|w_1, w_2) = \frac{C(w_1, w_2, w_3)}{C(w_1, w_2)}$$

Finding the entropy of the distribution of all outcomes for the third position of a trigram, therefore, is calculated as follows:

⁴I would like to take time to extend my gratitude to Lev Blumenfeld, whose work on Latin corpora and helpfulness made my research possible.

$$H(X_{MLE}) = - \sum_{x \in X} \frac{C(w_{x-2}, w_{x-1}, w_x)}{C(w_{x-2}, w_{x-1})} \log_2 \frac{C(w_{x-2}, w_{x-1}, w_x)}{C(w_{x-2}, w_{x-1})}$$

The corpus in use is 113,478 words long. For each word, I annotated the primary stress and word boundaries. The original work was annotated for length and syllable boundaries. My stressed version is only annotated for stress and length. The appearance of the Latin text is as follows:

#consvEtÚdinis@

#mIrÉtur@

#atrócitAs@

Using this intermediate corpus, I then created a set of trigrams using single segments. This yielded 860,695 trigrams at my disposal. The beginning and end characters were mainly used to separate different words to maintain correct trigram numbers; trigrams with either character in the middle position were discarded. These trigrams appear as follows:

c o

c o n

o n s

I then performed the above calculation for four different sets of trigram sequences:

1. Post-tonic segments ($\acute{V} C _$)
2. Post-tonic vowels ($\acute{V} C V$)
3. Post-atonic segments ($V C _$)
4. Post-atonic vowels ($V C V$)

2.1.1.3 Results

The calculations yielded the following results:

Trigram	Post-tonic		Post-atonic	
	$(\acute{V} C _)$	$(\acute{V} C V)$	$(V C _)$	$(V C V)$
Entropy	4.316	2.728	3.441	3.824

Table 2.1: Entropy measurements given MLE probability distributions for trigrams.

These results confirm my hypothesis with a slight irregularity: the entropy for all segments in which the first trigram position is a stressed vowel. This is due to the fact that not only vowels can appear two segments from a stressed vowel, but onsets after a coda as well. This is also compounded by the fact that stress is quantity-sensitive. However, with a more focused scope or simply looking at post-tonic vowels (and therefore ignoring cases of quantity sensitivity at the moment), the entropy is significantly lower in comparison. In information-theoretic terms, post-tonic vowels necessitate an entire bit less to encode than post-atonic vowels.

2.1.1.4 Summary

What is shown here is a clear-cut case where phonological output has been modeled against the noisy channel model. Phonological segments are perhaps the most obvious cases of units of encoding found in linguistics, and are an excellent starting point for this discussion at large. As is evident here, sounds prove to be little more than a highly complex system of units of encoding, grouped together in a manner which is intended to be as clear to understand and easy to pronounce as possible. Part of this goal is, of course, trimming of pronunciation, or compression of the transmission. The most achieved (and the most lost) in this particular endeavor is in regards to vowels, which is what makes the above case of such interest.

While the experiment here suggests that there are other factors in play than those mentioned in optimality-theoretic accounts, these insights could possibly be incorporated into the theory. Using probabilistic accounts to tie the entropy measurement would be useful, but it would assume a tacit, measurable understanding of frequency. While I argue here that frequency plays a role in language change, I in no way argue an exemplar-theoretic stance; speakers have a vague understanding of frequency, but it is closer to categorical rather than gradient understanding. All of these considerations would be necessary for a proper treatment in optimality theory, but it should suffice to say that the change occurred due to the perceived information a particular segment carried from generation to generation. Referring back to the earlier work of Hooper and Fidelholz, the case is quite the same: the more frequent forms of the particular environments described were those which exhibited instances of innovation.

2.2 Morphology

Due to the nature of morphology as a purveyor of atomic units of semantic information, there is an interesting trend of treating unmarked features as unrepresented in the morphology, but still assumed in the ultimate feature matrix of the word. Here, I will argue that these null morphemes which attribute to the meaning of words despite their segmental absence exist as a function of the low entropy of those features. Following this argument, we therefore can infer that a strong factor in the cross-linguistic phenomenon of the "emergence of the unmarked" has a strong tie to frequency and the resultant entropy as well.

2.2.1 The Null Present Tense Morpheme in Bantu

2.2.1.1 Background and Hypothesis

A remarkable example of an unmarked semantic feature emerging as unrepresented segmentally is the case of the present tense marker cross-linguistically in the Bantu language family. Nurse and Philipson[15] find that over 50% of the hundreds of Bantu languages represent the present or non-past tense with a zero morpheme, or null form. Furthermore, they assert that this is a trend widely seen in the Niger-Congo language family at large. I would like to assert that the present tense is unmarked because of its frequency in the language, which would yield comparatively low entropy. It would therefore be seen that there is a greater gain in convenience from not clearly encoding these forms than in the redundancy of stating the most predictable form.

To motivate this point, I have designed an experiment that should account

for this assertion. I use the language Swahili because of two interesting facts about the language. First, the present tense marker is not categorically null in this language, and therefore quantifiable against other markers. Second, cases where the marker is optionally null are the third person singular and plural the hypothetically most frequent and unmarked feature for person. Because of both the interaction of morpheme-dropping with specific markers and the historical Bantu tendency for a null present tense morpheme, it should be expected that the probability distribution of subject morpheme and present tense morpheme combinations should be significantly less entropic than the average distribution.

2.2.1.2 Statistics and Experiment

For this experiment, I will be using data from the Helsinki Corpus of Swahili (HCS).[8] In this corpus, there are of 2,028,832 verb tokens. Verbs with present tense morphemes make up 289,097 tokens. Verbs in the infinitive makes up 752,502 of verb tokens, but there is a different morphological position for the infinitive, and the purposes of nominalization render that specific morpheme a poor candidate for dropping due to frequency effects. The most frequent set of finite tense markers in the corpus are those associated with the past tense, at 448,694 tokens. However, for reasons to be discussed in the chapter on mitigation of historical change, we will see that this is an irreducibly complex form that is also not a candidate for deletion. Furthermore, it can be expected that the relative frequency of past tokens is partially attributed to the fact that the corpus is comprised of books and articles, which may be more prone to talk about concepts in the past.

The experiment at hand will measure the entropy of the probability distribution of the combination of the numerous possible subject morphemes in the verb with the present tense. This will be compared with the the probability distribution of all subject morphemes with any tense marker. If the entropy of the present tense morpheme set is lower than what is to be taken as the average entropy of every tense, then this should present it as both frequent and unmarked. This is the exact combination of phenomena that has resulted in the numerous related Bantu languages having a null present tense morpheme.

2.2.1.3 Results and Discussion

The data shows that the present tense is indeed significantly lower in entropy than that of the average. As seen in Figure 2.3, the entropy of the present tense data set is more than a whole integer lower than that of all tenses. In terms of encoding, this is an entire binary digit. What this signifies is that the present tense is an excellent candidate for dropping in these terms. Examining statistics simply related to tense marker frequency, the two variants of the present tense marker in Swahili make up 21.66% of the corpus. Morphemes consist of at least one syllable in Swahili, and the verb agglutinates for a great deal of semantic variables, so each item of information comes at a greater cost of effort. By rendering one fifth of the morphemes one would see in the tense position null, there is a great decrease in effort realized by the speaker. Because the morpheme would function as unmarked or default, the effort of the hearer to probabilistically decode the present tense from nothing would in fact be greater than a 1/5ths chance at deducing. Once this is

Data Set	All Tenses	Present Tense
Entropy	3.51	2.33

Figure 2.3: Entropy Measurements Per Data Set

learned as the standard manner of imparting present tense, there should be closer to 100% success at inferring present tense from no morpheme.

An important fact to note is that the very subject morphemes that are the only candidates to allow the dropping of the present tense marker in Swahili—the third-singular and third-plural markers—make up 66% of the probability mass in the corpus for the present tense data set. Because the form of these subject morphemes (*/-a-/* and */-wa-/*), frequency and unmarkedness cannot be argued as the pure answer for this correlation; this could be a purely phonological process. However, it is food for thought, and the fact that this does not result in phonological lengthening does point to deletion not of the subject marker, but of the present tense morpheme.

2.2.1.4 Summary

This study has shown that the present tense morpheme in Bantu verbs is likely to drop because it is statistically frequent and semantically simple or unmarked. This has been shown in terms of entropy as a binary digit less to encode than on average for tense morphemes. Because of the relative non-complexity of the present tense, we can assume that learning default present tense from the presence of no tense marker is a result of probabilistic general-intelligence abilities of deduction that are a part of general linguistic intelligence.

2.2.2 The Null Singular Morpheme in English

2.2.2.1 Background and Hypothesis

English displays null morphemes for two semantic categories that would be considered default or unmarked: singular number in nouns and present tense in verbs. Here, I shall treat the null singular morpheme for nouns in English, as I have just outlined an approach to treat the null present tense for verbs. Many investigations regarding zero morphemes relate mainly to acquisition. In Combinatory Categorical Grammar, Hoeksema[9] posited the use of default semantic categories without null morphemes as a way to give necessary values to features in the unification structure. Aone and Wittenburg[1] offered an alternate method of using ‘zero’ morphemes which solve the issue of overriding semantic categories and feature values. These morphemes appear as unary rules, which surface from an algorithm which compiles the morphemes in a trial-and-error fashion, checking if the syntax of each morpheme is compatible to either side of every likely candidate in the sentence. Both approaches trade some amount of insight for the sake of ease or computation. These approaches both focus heavily on parsing as opposed to theory, and this affects the decisions made by both parties in formalization.

In fact, the proper approach should be somewhere in between both of these methods. Lexically specifying semantic features to then be overrode by certain morphemes is an awkward way of formalizing the phenomenon we see, but there is much to be said regarding the fact that candidates for null morphemes are *extremely* featurally uncomplex. No language has a zero morpheme representing the set all long, round things, or a zero morpheme representing a time longer ago than

yesterday but not longer ago than last week. Typologically, these are very simple, recurring features: third person, singular, possibly both, depending on the lexical category upon which the morpheme influences. Positing the unary rule with a trial-and-error parsing algorithm may be a proper model of parsing the correct feature where an opposed feature would be morphologically concatenated, but once it is learned that the zero singular morpheme only manifests itself to the right of a noun, and not when there is a plural morpheme, the amount of computation necessary to determine number in a noun should decrease dramatically. Such an algorithm should not be necessary in regular human parsing of speech.

To confirm this, it should be proven that this singular/plural dichotomy is extremely learnable. I argue here that the learnability of the difference between null singular and segmentally marked plural morphemes is directly related to the low entropy that should be expected in this case. As I have stated several times before, very frequent items that carry simple information are subject to dropping at a minimal increase of deduction for the hearer. Cases of minimal pairs of morphemes where one is dropped should show an extremely low entropy for the probability distribution of both morphemes, with the dropped morpheme carrying a large majority of the probability mass.

2.2.2.2 Statistics and Experiment

To confirm my hypothesis, I will be using statistics from the Brown Corpus of Standard American English. This corpus of about 1 million words has stood as a strong sampling of different manners in which American English has been

used, consisting of reportage, fiction, non-fiction, and scholarly material. There are 225,799 nouns in the corpus and 29 different tags related to nouns.

In this experiment, I simply acquire frequency statistics on each of the 29 noun-related tags, normalize each number to a probability in the standard manner, and derive an entropy measurement from the probability distribution. The two most common tags are `nn`, which refers to a simple singular noun, and `nns`, which refers to a simple plural noun. These two nouns make up roughly 67% and 24% of the probability mass respectively. More data can be seen in Figure 2.4. This figure also shows that when all tags related to singular nouns and all tags related to plural nouns are simplified into two groups, the percentage changes to 75% singular and 25% plural.

The entropy of the distribution of all 29 tags in the Brown Corpus is impressively low at 1.35. However, the more compelling number is found when all singular tags and all plural tags are composed into two categories. The entropy found here is .81, which means that it would take less than a single binary digit to encode the difference. This is a very strong indication that there is a quantifiable motivation behind this correlation. We are seeing in English exactly what the entropy of less than a single binary digit would reflect. The fact that the most common of the two features is represented by a zero morpheme is a reaffirmation of both the overall hypothesis of this work, and the hypothesis given for this specific phenomenon.

Tag	Frequency	Probability
nn	152470	.67
nns	55110	.24
All Singular	167471	.75
All Plural	58328	.25

Figure 2.4: Frequency and probability for noun tags in the Brown Corpus of Standard American English

2.2.2.3 Summary

This experiment clearly shows both the benefits and the motivation for a zero morpheme for the singular in English. By making the form that is the most highly represented the morpheme that is not the most explicitly stated, there is a great savings in effort. Furthermore, because of the statistics we see here, it is in fact more worthwhile to assume every noun to be singular rather than deduce this from a zero morpheme. This is, however, due to the ability of learning rather than parsing, as seen in the discussion above. There is no perfect model for how humans understand semantics, and there is no solution proposed here. However, a singular morpheme would require 75% more articulatory effort than found in English, and this extremely lopsided dichotomy is so predictable that it becomes *very* learnable. Were it not this learnable, there would certainly be an overt singular morpheme.

2.3 Summary

In this section, I have discussed multiple well-known examples of deletion of items on varying linguistic tiers due to a combination of high frequency and a

relatively low degree of information contributed to the higher domain of parsing. These cases all point to a very simple conclusion that can be easily drawn in an era where linguistic data is ever more available for empirical analysis: there are strong, principled ways that languages change, and this is in part due to tacit learning in the noisy channel and the general human deductive ability.

What makes the conclusions above so compelling is their perspicuity. It is unnecessary to formalize these assertions theoretically other than to support or stand in spite of specific theories related to these processes. The quantitative character of the results show clear correlations that maintain the validity of their functionalist explanations without drawing in large-scale theoretical abstractions. While these formalisms are necessary to make broader explanations about language, it is the duty of theory and formalism to conform to the data and conclusions made from it—not vice versa. I have shown here an aspect that requires linguistic theory to examine frequency for cases of regular deletion. Those theories that are more readily able incorporate this kind of knowledge into their formalisms should be found to be in a more advantageous position to explain language in general as a result.

Chapter 3

Frequency's Role in Mitigating Regularity

The goal of this chapter is to motivate, in essence, that despite the fact that items of high frequency are more prone to historical change, in many cases, very frequent items also can prevent systemic regularity. These extremely frequent items that refuse to follow suit with phonological, morphological, or syntactic change seem to be resistant to regularity because of the complex nature of what they represent—words of extremely frequent but still marked semantic meaning, and often of complex semantic representation. Here I will present an entirely different set of cases where it seems that, while the hearer learns these output version of these high-frequency words as their own input, he or she generally does not innovate upon them. This would lead us to believe there must be some factor which would discourage innovation, and the evidence given here is intended to point generally towards that mitigating factor.

My hypothesis is that certain linguistic items, on varying tiers of representation from phonological to syntactic¹, resist change based upon two criteria:

- The item is extremely high in frequency relative to other similar linguistic

¹And possibly beyond, given Grice's Maxims and how they can be related to the noisy channel model

items.

- The item is also semantically complex yet irreducible, such that it contributes information to a higher level of linguistic organization in a way that cannot be logically decomposed, but rather must be memorized.

These two criteria are necessary for irregular forms in historical change. Without the first criterion, even semantically complex items will undergo an amount of regular change such that the form can still be innovated upon. Without semantic complexity, the highly frequent form is more likely to delete, given what was seen in the previous chapter.

In this chapter, I will present two separate but well-known cases that can be used to support this argument. I will be using evidence from corpora spanning several languages to motivate my findings.

3.1 The Romance Headedness Shift

3.1.1 Background

For most individuals learning a Romance language as a second language, they are well aware of the generality that adjectives fall to the right of the noun, unless they are a certain set of regularly recurring adjectives. In which case, they are found to the left of the adjective. Lehman [12] was one of the first focusing on Indo-European historical syntax that identified several Romance languages as having alternating order for a “small set of very common adjectives”. Bauer [2] affirms

that from Proto-Indo-European to Old Latin, word order generally consisted of adjectives preceding their nouns. This changed in Latin, with a predictable point of cleavage between “descriptive” adjectives that preceded the noun and “distinctive” adjectives that followed. Romance languages, such as French, seemed to gradually evolve from somewhat arbitrary decisions on adjectival headedness to a majority of adjectives being left-headed.

However, there are adjectives that still trump this generality. What makes this interesting is the fact that these adjectives are extremely common and often semantically primal concepts expectable in any language. It can be expected that these concepts may have existed for thousands of years, into the time of Proto-Indo-European, which was a head-final rather than head-initial language. This allows for the possibility that the syntactic headedness of constructions using these adjectives has been preserved throughout history due to their frequency—possibly even earlier than the Old Latin source of directional ambiguity. This can be directly attributed to the high frequency and semantic irreducibility of the set of adjectives in question.

Before the era of the proliferation of machine-readable linguistic data, frequency could only be identified via intuitions, as seen from Lehman’s claims. While there must be some vague understanding of what constructions or linguistic items are more frequent than others, it should further strengthen the observations above to acquire concrete correlates.

I will use the statistics from various Romance-language corpora to prove that the most frequent adjectives are those which display irregular syntactic constructions with respect to directionality to the phrasal head. In doing so, I will also

Bigram	Number of Instances
ADJ N	1886
N ADJ	4164

Figure 3.1: Bigram tag statistics in the Floresta Synta(c)tica corpus of Portuguese

be reaffirming both the assertion made at the beginning of this chapter as well as the large-scale assertions I have made regarding frequency effects on historical change and its role in the noisy-channel learning schema previously described.

3.1.2 Portuguese

The Floresta Synta(c)tica is a tagged Portuguese corpus of news articles consisting of 169,216 tokens and 26,628 types. There are 41 tags in all. Those of interest in this investigation are ADJ, for adjective, and N for noun.

Results of frequency calculations show that the top ten most frequent adjectives in the corpus all are found predominantly if not exclusively in ADJ N order in the corpus, as opposed to N ADJ. This is shown in Figure 3.2. While the assertion stands that those constructions where the adjective precedes the noun consist of more frequent adjectives, the regularity of the head-initial generality in Portuguese results in a greater amount of adjective-noun constructions in all in the corpus, as seen in Figure 3.1.

3.1.3 Spanish

The 1997 HUB5 Spanish Transcripts corpus is an unannotated collection of 20 transcribed telephone conversations in Spanish. The corpus contains approxi-

Count	Adjective	Gloss	Construction Type
107	<i>maior</i>	‘greater’	ADJ N
97	<i>primeiro</i>	‘first’ (m)	ADJ N
89	<i>novo</i>	‘new’ (m)	ADJ N
88	<i>primeira</i>	‘first’ (f)	ADJ N
87	<i>grande</i>	‘big’	ADJ N
62	<i>nova</i>	‘new’ (f)	ADJ N
58	<i>grandes</i>	‘big’ (pl)	ADJ N
47	<i>ltima</i>	‘last’	ADJ N
45	<i>segunda</i>	‘second’	ADJ N
45	<i>melhor</i>	‘greater’	ADJ N

Figure 3.2: The ten most frequent adjectives in the Floresta Synta(c)tica corpus of Portuguese

Count	Adjective	Gloss	Construction Type
317	<i>bueno</i>	‘good’	ADJ N
29	<i>poco</i>	‘little’	ADJ N
25	<i>súper</i>	‘super’	ADJ N
24	<i>terrible</i>	‘terrible’	ADJ N
24	<i>grande</i>	‘big’	ADJ N

Figure 3.3: The five most frequent adjectives in the HUB5 Spanish Transcripts corpus

mately 46,977 tokens and 3,955 types. While this is a relatively small corpus, its frequency statistics do support the correlation between frequency and syntactic irregularity. The only adjective in the top 100 most frequent words in the corpus is *bueno*, meaning ‘good’, with a count of 317. As Figure 3.4 shows, the top five adjectives in this corpus all follow this correlation.

The value of this particular corpus is that it is an example of spoken language. While much of the corpora used here consist of either written language or

prepared speeches, it is valuable to show that these correlations and assumptions exist outside the written sphere of language as well. Furthermore, this affirms the utility of treating written varieties language as a suitable sample of a language as a whole.

An interesting fact to note is the inclusion of ‘súper’ among the top adjectives. For a perceivably uncommon noun to be both very frequent as well as syntactically irregular points to the strong possibility that this is an English borrowing, both due to the English usage of ‘super’ coinciding with that of the Spanish usage as well as the regular syntax of English noun phrases. Unfortunately, there is no data about the location or origin of the speakers in the corpus, but it would be interesting to note whether this is a case of code-switching where syntactically, the usage fits, or whether regular contact with English speakers—for instance, as a speaker of a minority language in the United States or a Spanish speaker in an American possession—has caused a lexification of both the word as well as its syntactic composition.

The European Parliament Proceedings Corpus (Europarl) is a multilingual aligned corpus in several European languages. I will be using statistics from it for not only Spanish, but French as well. While there may be some questions as to whether this is appropriate, the concept of these adjectives being semantically primal means that they should be highly frequent in most languages. What is being scrutinized in this case is whether these adjectives all follow irregular rules of syntactic composition.

The Europarl corpus follows well with the data previously presented. How-

Count	Adjective	Gloss	Construction Type
74527	<i> europea </i>	'European' (f)	N ADJ
48994	<i> europeo </i>	'European' (m)	N ADJ
22957	<i> gran </i>	'big' (m)	ADJ N
20651	<i> otros </i>	'other' (pl)	ADJ N
20001	<i> mismo </i>	'same' (m)	ADJ N
18969	<i> primer </i>	'first' (m)	ADJ N
14655	<i> otra </i>	'other' (f)	ADJ N
14442	<i> otro </i>	'other' (m)	ADJ N
11179	<i> último </i>	'last' (m)	ADJ N
11098	<i> segundo </i>	'second' (m)	ADJ N

Figure 3.4: The five most frequent adjectives in the Europarl Corpus – Spanish

Count	Noun	Gloss
134,532	<i> comisión </i>	'commission'
73390	<i> unión </i>	'union'
71930	<i> parlamento </i>	'parliament'
70835	<i> presidente </i>	'president'
61921	<i> consejo </i>	'advice'

Figure 3.5: The five most frequent nouns in the Europarl Corpus – Spanish

ever, the top two adjectives do not follow the stated pattern. This is a case of arena of discussion, however, as seen in Figure 3.5. The fact that words meaning 'European' would be the two most frequent adjectives in the proceedings of the European Parliament is perhaps as unsurprising as the fact that most frequent nouns all relate to political interaction. To bolster this argument, I have included a list of the top nouns in Figure 3.5 to show how the sphere of discussion has somewhat distorted the sample of language in use.

Count	Adjective	Gloss	Construction Type
77775	<i>européenne</i>	‘european’ (f)	N ADJ
46991	<i>européen</i>	‘european’ (m)	N ADJ
43609	<i>même</i>	‘same’	ADJ N
19749	<i>important</i>	‘important’	N ADJ
19514	<i>autres</i>	‘other’	ADJ N
18751	<i>économique</i>	‘economic’	N ADJ
18748	<i>certains</i>	‘certain’	ADJ N
15277	<i>grande</i>	‘big’	ADJ N
14388	<i>peu</i>	‘little’	ADJ N
12437	<i>nouveau</i>	‘new’	ADJ N

Figure 3.6: The ten most frequent adjectives in the Europarl corpus – French section

Count	Noun	Gloss
135308	<i>commission</i>	‘commission’
70266	<i>parlement</i>	‘parliament’
68643	<i>rapport</i>	‘report’
64627	<i>l’union</i>	‘the union’
62658	<i>conseil</i>	‘council’

Figure 3.7: The five most frequent nouns in the Europarl corpus – French section

3.1.4 French

The French section of the Europarl corpus consists of 381,365 tokens and 14,126 types. Again, while the statistics seen in Figure 3.6 show an amount of intrusion from specialty words in the arena of discussion, the majority of the ten most frequent adjectives are head-final rather than head-initial. Once again, to show the jargon-heaviness of the data, Figure 3.7 has been included to support my argument.

3.1.5 Summary

This section has shown that there is a notable correlation between frequency and syntactic irregularity in Romance languages. The strongest evidence shown here is in Portuguese. This could be simply due to the format of each corpus; the Floresta Synta(c)tica was the only corpus used here specifically tagged for part of speech. Furthermore, newspaper articles consist of multiple topics, while basing data entirely parliamentary proceedings will be somewhat skewed. Usage of a spoken language corpus supplied interesting affirmative results, but would optimally have involved a great deal more data.

Given the results above, it is very safe to make the assumption that frequency is somehow tied to syntactic irregularity with regards to headedness. This data-based assumption allows us to speculate that frequency is in fact the cause of this. While the statement, ‘an adjective follows a noun, unless it is very frequent in use’ explains the data quite well, frequency is not something a language learner can use to check each adjective before delivering an utterance. Thinking abstractly, this would cause us to assume that this syntactic alternation is memorized for each form individually at a very early stage, and is reaffirmed through practice and experience.

3.2 Frequency and Irregular Inflection

Irregular verbal forms are problematic in easy, regular analysis of language and are, in fact, proof of the role in frequency as preventing regularity across language as a result of the way human beings learn language. This is a very clear case

of both mitigating criteria described above being met and having an incontrovertible effect on the systemicity of regularity in language. In this section, I will use data from corpora of several languages to show the direct correlation between frequency and irregularity. Furthermore, I shall use evidence from English to show that this correlation is evocative of work from Mandelbrot and Zipf on word frequency distribution in general.

Irregular verbs occur, following the concept of the noisy channel as a model of language learning, because probabilistically, there is such a proliferation of tokens of these forms in comparison to regular forms. Attempts at innovation or deductive reasoning are invariably overcome by the regular exposure to the popularly accepted form along with the semantic irreducibility acting as a constraint upon innovation. While all verbs generally consist of complex semantic constructions, infrequent verbs are stored as lemmata and combine with morphemes in a regular manner. Irregular forms, however, are stored lexically, containing all morphological information within a single simplex form.

3.2.1 English

English is well known for its fairly impoverished agreement system short of irregular verbs. The data to be given here shows that this is indicative of a frequency-related historical irregularity; more elaborate agreement paradigms exist only in irregular verbs in English because of their nature of being extremely frequent and semantically irreducible. It is well known that English originally had a much more elaborate system of case and agreement, and these very frequent forms

Count	Verb	Irregular?
10065	<i>is</i>	✓
9806	<i>was</i>	✓
4372	<i>are</i>	✓
3925	<i>have</i>	✓
3281	<i>were</i>	✓
2710	<i>would</i>	✓
2470	<i>been</i>	✓
2430	<i>has</i>	✓
1043	<i>did</i>	✓
883	<i>should</i>	✓

Figure 3.8: The ten most frequent verbs in the Brown Corpus of English

are the few remainders from that point in the language's evolution.

The Brown Corpus is a popular language resource comprised of various text sources to reflect English usage in a variety of written settings. The corpus contains 9,069 verb types and 157,955 verb tokens. Of the verb tokens, 77,698 inflect regularly, while 80,257 inflect irregularly. However, there are only 483 irregular types, compared to 8586 regular types. Irregular vowels contain almost exactly half the probability mass at .508. The entropy of determining whether a verb is regular or irregular, therefore, would be close to 1.0.

Figure 3.8 shows that the ten most frequent verbs in English are all irregular. Assuming that forms which cannot be morphologically decomposed are lexical, and assuming that there is a morphological composition process beyond the lexical stage which requires some extra degree of processing, the evidence shown here fits quite well with work from Zipf [18], known as the Principle of Least Effort. This concept assumes that individuals will opt for the least costly strategy in terms of

exertion. His work with regards to word frequency was later refined by Mandelbrot [14]. Both show that there is a constant relation between number of tokens and rank in order of frequency for any corpus. Given the above assumptions, there is an incentive for lexicalizing forms that will be the most frequently reused. Rather than constantly performing the same processes or morphological composition, all information morphological items would give to these irregular forms are stored, and therefore each form for an inflection paradigm.

While one of the reigning views in syntax regarding the treatment of the lexicon is one of Lexical Integrity [6], there are some drawbacks to this point of view. The lexicalization of all morphologically derived forms from all lemma in a language would be a veritable explosion in storage with a great deal of redundancy. I have shown in other sections that a major part of language faculty is deduction, and it makes a clear deal more sense that language learning involves deducing regular morphological rules and acting upon them. Work is forthcoming in the area of Combinatory Categorical Grammar that affirms this, and shows that morphology is in fact a highly restricted combinatory process that can be included in the syntax at a comparatively low finite state computational complexity. The finite-state character of morphology has been affirmed extensively through the work of Karttunen and Beesley [3]—with languages ranging in morphological strategies from inflecting, to agglutinative, to templatic.

Assuming this, the Principle of Least Effort, in this context, would place the weight of morphological combination on about half of the verbs used in English—the multitudes of infrequent verbs. The comparatively few verb forms that are fre-

quently used are stored for ostensibly faster and more convenient use. Given this, Least Effort actually contributes greatly to the current working model of language learning proposed in this work, and how frequency relates to historical change through it.

Taking this perspective is reaffirmed by examining the relation to frequency and rank. Figure 3.9 shows the relation between frequency and rank for all verbs in the Brown Corpus. Comparing this to Figures 3.10 and 3.11, which relate frequency and rank to all regular verbs and irregular verbs in the corpus respectively, it is clear that while both forms have the same hyperbolic slope, the first portion of Figure 3.9 is almost completely irregular forms. Irregular tokens consist of 96.4% of the first third of the verbs in the corpus ranked by frequency. This results in an entropy of 0.22 for determining whether a verb form is irregular among the 3000 most frequent verbs of the corpus.

What this evidence suggests is that frequency and irregularity in verbal forms are inextricably linked. If a language has irregular verbal forms, it will necessarily be among those forms that are extremely frequent. English is not anomalous in this matter, as will be shown below.

3.2.2 Portuguese

The verb frequencies in Figure 3.12 from the Floresta Synta(c)tica Corpus of Portuguese also reflect the assertion made using the Brown data in English. All of the 20 most frequent verbs from this corpus are irregular. The similarities in glosses to the most frequent verbs in English are unmistakable; there is an abundance of

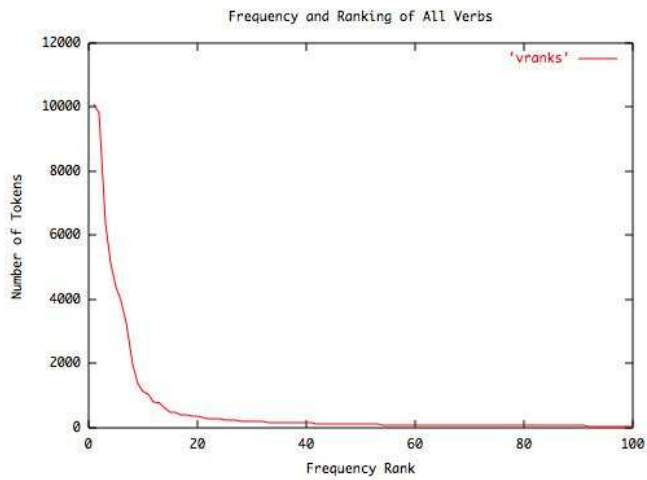


Figure 3.9: All verbal forms in English

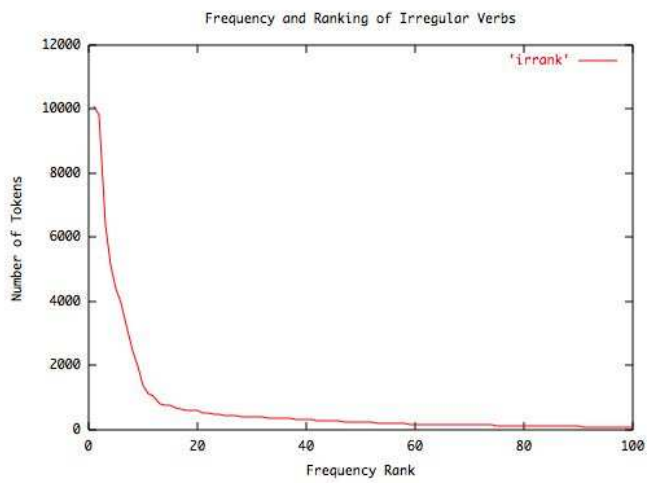


Figure 3.10: Irregular verbal forms in English

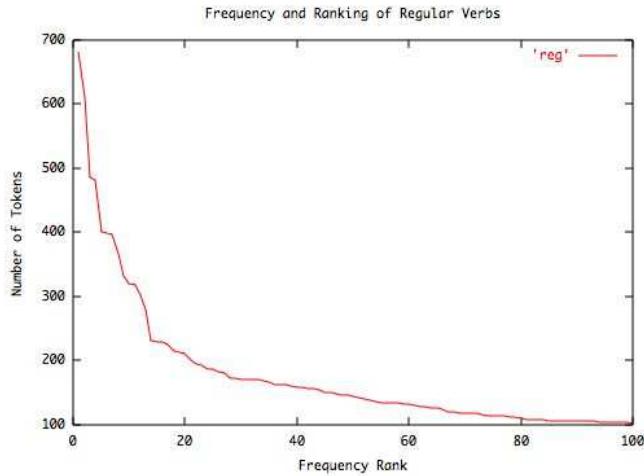


Figure 3.11: Regular verbal forms in English

copula forms as well as modal verbs.

3.2.3 French

French also displays this tight correlation, as seen in Figure 3.13. The many recurring forms are unsurprising: 'be', 'do', modals, and verbs that express very basic cross-cultural concepts. What is of interest to note is that among verbs, there is no intrusion of special-area jargon from the Europarl Corpus. This shows the relative strength of the verb category and its somewhat recursive character, as seen from the prevalence of modals among the most frequent verbs.

Count	Verb	Gloss	Irregular?
1006	<i>é</i>	'are'	✓
433	<i>foi</i>	'were'	✓
370	<i>ser</i>	'to be'	✓
257	<i>tem</i>	'have'	✓
245	<i>está</i>	'are'	✓
225	<i>são</i>	'are'	✓
203	<i>ter</i>	'to have'	✓
170	<i>disse</i>	'said'	✓
153	<i>há</i>	'have'	✓
150	<i>foram</i>	'had been'	✓
146	<i>vai</i>	'go'	✓
122	<i>pode</i>	'can'	✓
110	<i>era</i>	'were'	✓
109	<i>fazer</i>	'to make'	✓
108	<i>estão</i>	'are'	✓
103	<i>diz</i>	'say'	✓
95	<i>será</i>	'will be'	✓
91	<i>têm</i>	'have'	✓
91	<i>sido</i>	'been'	✓
79	<i>tinha</i>	'it had'	✓

Figure 3.12: The 20 most frequent verbs in the Floresta Synta(c)tica Corpus of Portuguese

Count	Verb	Gloss	Irregular?
200786	<i>est</i>	'is'	✓
84802	<i>sont</i>	'are'	✓
66767	<i>être</i>	'to be'	✓
64153	<i>c'est</i>	'it is'	✓
63789	<i>été</i>	'been'	✓
62798	<i>fait</i>	'do'	✓
59630	<i>ont</i>	'have'	✓
45777	<i>faire</i>	'to do'	✓
39673	<i>avons</i>	'have'	✓
33633	<i>peut</i>	'can'	✓
31098	<i>voudrais</i>	'would like'	✓
30442	<i>faut</i>	'lack, need'	✓
29683	<i>n'est</i>	'isn't'	✓
27685	<i>devons</i>	'owe, should'	✓
24727	<i>dire</i>	'to say'	✓
23566	<i>j'ai</i>	'I have'	✓
21901	<i>sommes</i>	'are'	✓
21789	<i>suis</i>	'am'	✓

Figure 3.13: The 18 most frequent verbs in the Europarl Corpus – French

3.2.4 Spanish

The Europarl corpus of Spanish in Figure 3.14 maintains the strong correlation between irregular verbal forms and frequency. However, there is a regular verb form among the within the top ten most frequent forms. This raises a number of questions regarding whether this is specialty-area intrusion upon actual frequencies, or if not, whether regular verb forms can also be lexicalized if they are frequent enough or must necessarily be treated combinatorially, despite the assumed cost in processing or computation.

Figure 3.15 does not require such questions to be asked, and as a record of spoken language, it is much closer to what would be expected from practiced extemporaneous language. The value of this data is its non-textual nature; not only does it continue to reflect observations gained from examining statistics within text-based corpora, but it is even stricter. Those forms seen in the top 14 only consist of modals, the copula forms, and extremely basic words with similar cross-cultural meanings throughout all languages.

3.2.5 Swahili

Swahili is a language with comparatively few irregular forms. As evident in Figure 3.16, the ten most frequent verb forms are predominantly infinitive. This provides an interesting counterpoint in that frequency and irregularity are not necessarily inseparable. However, there are interesting facts to note regarding Swahili that seems to strengthen and not weaken the arguments given in this work. Swahili is a highly agglutinating language with verbal morphemes for tense, aspect, mood,

Count	Verb	Gloss	Irregular?
133133	<i>ha</i>	'has'	✓
56900	<i>han</i>	'they have'	✓
55943	<i>estados</i>	'been'	✓
43111	<i>está</i>	'is'	✓
40425	<i>ser</i>	'to be'	✓
36752	<i>debe</i>	'should'	
36030	<i>puede</i>	'can'	✓
32602	<i>hemos</i>	'we have'	✓
28402	<i>hacer</i>	'to do'	✓
28327	<i>tiene</i>	'to have'	✓
28252	<i>creo</i>	'I believe'	✓
27937	<i>hecho</i>	'done'	✓
27860	<i>decir</i>	'to say'	✓
22587	<i>he</i>	'I have'	✓
22442	<i>están</i>	'they are'	✓
20917	<i>hace</i>	'does'	✓
20025	<i>sea</i>	'were' (subj)	✓
18899	<i>dicho</i>	'said'	✓
18154	<i>estamos</i>	'we are'	✓

Figure 3.14: The 19 most frequent verbs in the Europarl Corpus – Spanish

Count	Verb	Gloss	Irregular?
141	<i>voy</i>	'I go'	✓
108	<i>sea</i>	'were'	✓
100	<i>tiene</i>	'has'	✓
98	<i>dice</i>	'says'	✓
92	<i>tengo</i>	'I have'	✓
90	<i>dijo</i>	'said'	✓
84	<i>ver</i>	'to see'	✓
83	<i>hay</i>	'there are'	✓
83	<i>hacer</i>	'to do'	✓
81	<i>estoy</i>	'I am'	✓
75	<i>ir</i>	'to go'	✓
74	<i>fue</i>	'did/was'	✓
70	<i>dije</i>	'I said'	✓
69	<i>digo</i>	'I say'	✓

Figure 3.15: The 14 most frequent verbs in the Hub5 Spanish Corpus

discourse, and argument structure, as well as subject, object, and relative verbal morphemes that agree with over ten separate noun classes. The odds of a form being frequent enough to be paradigmatically irregular, therefore, is very low. This is affirmed in Finnish, a language of a completely separate family that is also highly agglutinative and displays significantly fewer irregularly inflecting verbs than the data seen above.

Using statistics on verbs from each corpus, the most frequent verb type in the HCS accounts for only 3.5% of all verb tokens in the corpus, compared to the inflectionally impoverished English, where the most common verb type in the Brown Corpus accounts for 6.4% of all verb tokens. Conversely, a more inflectional language would have something in between; the most common verb type in the Flo-

Count	Verb	Gloss
71270	<i>alisema</i>	'he/she said'
34319	<i>kuwa</i>	'to be'
13811	<i>kufanya</i>	'to do'
11456	<i>kutoa</i>	'to put out'
11244	<i>alikuwa</i>	'he/she was'
10183	<i>amesema</i>	'he/she said'
9458	<i>akasema</i>	'and he/she said'
8435	<i>kwenda</i>	'to go'
8149	<i>kupata</i>	'to get'
7111	<i>kutumia</i>	'to send or use'

Figure 3.16: The 10 most common verb forms in the Helsinki Corpus of Swahili

resta Synta(c)tica Corpus of Portuguese comprises 4.2% of the tokens. This would suggest that there not only exists a correlation between frequency and irregularity, but that this relation is modulated by the morphological strategy of the language: the more possible inflected forms stemming from one lemma, the less probable the most frequent types are. These probabilities must in some way be utilized in language learning.

3.2.6 Summary

Irregular verbal forms have an undeniably strong link to frequency. These forms consist mainly of words with logical forms that can be expected to be in some way universal to all human experience. This is not affirmative to nativist assumptions of language, but rather other basic biological proclivities—motion, action, and interaction. The best explanation for the data above is that which stems from the Principle of Least Effort, with the following underlying assumptions:

- Morphology and syntax are not separate, but rather morphology is a highly restricted combinatory environment working at finite-state speed within the level of syntax.
- While morphology at the level of syntax has low computational complexity, there is a processing cost for combining forms rather than using a fully lexical form.
- Lexicalized forms are faster and easier to use than morphologically inflected forms, but there is a space cost for the lexicon.

The Principle of Least Effort compromises these considerations by lexicalizing only those verbal forms which are the most common. Because of this tendency, there is a vicious cycle throughout a language's evolution. Highly frequent forms that act in this manner change very little, and because of this, over time, any ability to deduce the morphology of these forms will be rendered impossible in light of the grander scheme of regular verbs, which has evolved at a separate rate. Social convention will further complicate these irregular verbal forms. Evidence of this can be seen in the merging of English *beon* and *wessan*, as well as the merging of the singular 2nd person with the plural.

Despite the complication of irregular forms for making regular predictions about language on the surface, viewing them in the context of frequency adds a great deal of explanation to this problem. Frequency serves in part as the answer to why a form would be irregular at all, and the Principle of Least Effort along with

the underlying assumptions stated above accounts for why in the most plausible manner possible.

3.3 Chapter Summary

In this chapter, I have described two prolific ways in which high frequency mitigates historical change. As opposed to the cases of catalysis, where items with high frequency and a low degree of information that cannot be otherwise deduced are easily innovated upon or deleted, these items are irreducibly semantically complex. Introducing The Principle of Least Effort to the discussion, the benefits gained from this specific approach are clear: a practical balance between storage and processing. Furthermore, the data described and interpreted above shows a converse effect to innovation on frequent items: the persistence of linguistic ‘traditions’ or ‘conventions’, despite their clear illogical nature to even the speaker.

Chapter 4

The Interplay of Frequency-Based Mitigation and Catalysis in Historical Change: The Case of the Copula

The previous two chapters have established that frequency can affect diachronic change in two interesting and opposed manners. Those linguistic items that are weak in information and high in frequency can be deleted at a minimal loss to understanding. Those that are complex in information and similarly high in frequency tend to resist historical change, creating systemic irregularity. In this chapter, I will investigate the well-known cross-linguistic phenomenon of copula-dropping and use the insights I have provided in the previous two chapters to motivate a principled cross-linguistic explanation in terms of entropy and the Principle of Least Effort.

4.1 Description

Copula-dropping can be described as a process which occurs in multiple languages under roughly the same criteria:

- The copula is only dropped in the present tense.
- Inflections for marked (non-present) tenses of the copula are irregular.

- If the copula only drops for one person-number combination, it is for third person singular.

This is the case for many different languages across the globe: Japanese, Russian, Hungarian, Arabic, and African American Vernacular English (AAVE). Given their strong similarities and the presence of a mainstream dialect which can provide data with over examples of the copula for comparison, I will be focusing primarily on African American Vernacular English for investigation, and offer corollaries where necessary for these other languages.

African American Vernacular English has been a topic of discussion in linguistics through much of the 20th Century. Empirical work in the 1960s and '70s from Labov [11] sought to describe and offer a logical formal description for this variant of English. More recently, work from Bender [4] has sought to use modern empirical theoretical viewpoints to explain such variation. In understanding copula dropping, there are three kinds of 'linguistic knowledge' involved, according to this work. These include "knowledge of social meaning attached to linguistic forms, direct knowledge of a grammatical structure that is computable from more basic signs already in the grammar, and knowledge of the frequentistic, non-categorical grammatical constraints on variation." The second two kinds of knowledge are those which are the most interesting to this investigation, because they directly coincide with how frequency plays a role in catalysis of historical change. The third type of knowledge would be directly related to entropy, while the second type of knowledge would be directly related to the hearer's task of decoding noisy-channel output.

Only if the hearer can deduce the input from the noisy output is linguistic innovation viable.

4.2 Investigation

Because of the nature of the copula as a null element, it is inherently difficult to obtain the proper corpus data to quantitatively describe it. However, because AAVE is a variant of a language which does use an overt copula in all cases, this is a language upon which statistical analysis for the copula can be performed with relatively little trouble. Here, I will use linguistic data from the Brown Corpus to arrive at a quantifiable explanation for why copula-dropping is such an acceptable convention in so many different languages.

As seen in Figure 3.8 in the previous chapter, the two most frequent verbs in the Brown Corpus of English are ‘is’ and ‘was’. Their frequency numbers are quite close—10,108 tokens versus 9,815. The frequency is so close, in fact, that it violates to some extent the Zipfian laws of how frequency and rank relate in language. I would like to motivate that if the speaker can syntactically deduce that a verbal category is necessary for a proper sentential output, then probabilistically, they will opt for a copular reading of the verb based on frequency.

The present and past tense third person singular copulas carry the most of the probability mass of any verb in the Brown Corpus, both representing roughly 6% of all verbal tokens. Comparatively, the next three verbs in the top five measure at 4%, 3.2%, and 2.8%. The tenth most frequent verb carries 0.7% of the probability mass. Of the ten most frequent verbs in English, four are some inflection of ‘be’.

Adding all their probabilities together, we arrive at 19.1% of the probability mass, and this is only four of the several inflections for the verb.

Given the above statement that 'is' and 'was' are the two most frequent words in the corpus and Zipf's observation that there is a constant relationship between frequency and rank, it can be stated quite clearly that there is no verb that contains nearly the same amount of probability mass. For instance, if any verb were to compete with 'be', it would be 'have', which has two forms in the ten most frequent verbal form tokens. However, the probabilities of these verbs combined only arrive at .052, both affirming the Zipfian observations of frequency as well as confirming beyond a doubt that 'be' is the most probable—and therefore least entropic and easiest to deduce—verb form in English. We can assume that this is the case for most languages, given the data discussed in the previous chapter and the regular reappearance of the copula in many verb form frequency lists.

Here, we have established the high probability of predicting 'be' as the reading of a deduced verbal category from a grammatical combination which could otherwise not be syntactically parsed. The interesting question at this point is as follows: if 'be' can be deduced as the reading of a deduced verbal category, then why does only the present tense third person inflection drop, and not the past tense? Given what we have seen in the previous two chapters, there is only one viable interpretation of the data.

As in the discussion of the present tense morpheme in Bantu, there can be only one default tense, and it would be the most probable of all tenses involved. While the present tense is intuitively 'unmarked', I used quantifiable data to show

that we can frame this concept of markedness or unmarkedness in terms of frequency as well. Furthermore, in the chapter on the mitigation, I showed that if a form is irreducibly complex or especially marked, we can expect that this form would resist historical change. The past tense would be considered less probable, and therefore more marked. A speaker would deduce present tense rather than past, and therefore, if the speaker wants to impart a past tense, they must actively pronounce the verb despite its highly probable semantic meaning.

4.3 Cross-Linguistic Observations

In AAVE, non-third person inflections in the present tense require an overt inflected ‘be’. This varies among languages that exhibit copula-dropping, and future work will be necessary in arriving at a quantifiable explanation for this variation. However, it can be assumed that this would be associated with the degree of deducibility within the various variables in the syntax and grammar. Russian may be freer for copula-dropping in any person in the present tense because case forces certain agreement readings. Japanese, on the other hand, only allows copula-dropping if the copular complement is an adjective. It could be assumed that because of the nature of adjectives as specifying nouns and the character of the copula as a relation function, there is some amount of redundancy that allows dropping which would otherwise be difficult to deduce—especially given the animacy bifurcation in Japanese verbs for existence.

4.4 Conclusion

Using concepts established earlier in this work and data from English, I have shown that there is a strong probabilistic incentive to drop the copula in specific environments where the necessary variables for a grammatical sentence such as tense and occasionally agreement are easy to deduce. Other strategies are taken in cases where there is either a different verbal predicate or a marked feature structure. Copula dropping shows the delicate balance which natural language utilizes in frequency-based historical change. This alternation between deletion of the unmarked and conservation of the marked is a consternating scheme which can only be *explained* through quantitative analysis, such as given above.

Chapter 5

Conclusion

In this work, I have described several cases where we can use frequency along with quantitative models and principles that extend beyond the realm of linguistics to describe both regular diachronic change as well as irregular diachronic conservation. Each concept which I have framed in this manner shows the value of a functional, quantitative explanation to underlie any theoretical account for the data.

With these explanations now available for the data given, it is now of great import to examine what available theories have to offer, and which are most suited for all of the findings given here. It will show quite clearly that much of this information stands counter to several reigning opinions across the discipline. However, I have taken great lengths to take a cross-linguistic and cross-sub-field approach to show that these general traits of encoding not only exist in language, but shape its past, present, and future.

Bibliography

- [1] C. Aone and K. Wittenburg. Zero morphemes in unification-based combinatory categorial grammar. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 188–193, University of Pittsburgh, Pittsburgh, 1990.
- [2] Brigitte L. M. Bauer. *The Emergence and Development of SVO Patterning in Latin and French*. Oxford University Press, 1995.
- [3] Kenneth R. Beesley and Lauri Karttunen. *Finite State Morphology*. CSLI Publications, Stanford, CA, 2003.
- [4] Emily M. Bender. *Syntactic Variation and Linguistic Competence: The Case of AAVE Copula Absence*. PhD thesis, Stanford University, Palo Alto, California, 2000.
- [5] P. Boersma. The odds of eternal optimization in Optimality Theory. In *Optimality Theory and language change*, pages 31–65. Kluwer Publishers, 2003.
- [6] Joan Bresnan and Sam A. Mchombo. The lexical integrity principle: Evidence from bantu. *Natural Language and Linguistic Theory*, 13:181–254, 1995.

- [7] James Fidelholtz. Word frequency and vowel reduction in English. *Chicago Linguistic Society*, 11:200–213, 1975.
- [8] HCS. Helsinki corpus of Swahili, 2004. Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC - Scientific Computing.
- [9] Jack Hoeksema. *Categorial Morphology*. Garland Publishing, Inc., New York and London, 1985.
- [10] J. Hooper. Word frequency in lexical diffusion and the source of morphophonological change. pages 96–105. North Holland.
- [11] William Labov. Contraction, deletion, and inherent variability of the english copula. *Language*, 45(4):715–762, 1969.
- [12] Wilfred P. Lehman. *Proto-Indo-European Syntax*. University of Texas Press, 1974.
- [13] C. Lleó. Some interactions between word, foot, and syllable structure in the history of Spanish. In D. Eric Holt, editor, *Optimality Theory and Language Change*, pages 249–283. Kluwer Academic Publishers, 2003.
- [14] Benoit Mandelbrot. Structure formelle des textes et communication. *Word*, 10:1–27, 1954.
- [15] Derek Nurse and Gerard Philipson. Common tense-aspect markers in Bantu. *Journal of African Languages and Linguistics*, 27:155–196, 2006.

- [16] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–432, 623–656, 1948.
- [17] Claude E. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, pages 50–64, January 1951.
- [18] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.

Vita

Robert Blaine Elwell was born in Schenectady, New York on 5 June 1984, the son of Robert C. Elwell and grandson of Maria L. and Robert D. Elwell. He received the honors-track Bachelor of Arts degree in Linguistics from the State University of New York at Albany in 2005, finishing in three years. From there, he was directly accepted into the University of Texas at Austin as a PhD student in the Linguistics program.

Permanent address: 2501 Wickersham Ln #1132
Austin, Texas 78741

This report was typeset with \LaTeX^\dagger by the author.

[†] \LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's \TeX Program.