Computational Explorations in Morphology

by

Nick Kloehn

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2018

The University of Arizona
Graduate College

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Nick Kloehn, titled Three Papers on the Internal Properties of Words and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Dr. Michael Hammond — Date: 4/27/18

Dr. Adam Ussishkin — Date: 4/27/18

Dr. Robert Henderson — Date: 4/27/18

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Director: Dr. Michael Hammond — Date: 4/27/18

Statement by Author

This dissertationhas been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Nick Kloehn

Acknowledgments

This dissertation was made possible by those who have guided me throughout my life. This journey, often fraught with self doubt and an inability to express myself was aided by those who have helped, pulled and pushed along the way. I give acknowledgment to all the teachers and students who helped me to develop and focus my own thoughts enough to externalize them, even though they often contained typos. The task of finishing this work has taught me that self expression is the truly the most essential and most difficult task. It should not be taken for granted.

Last, this work could not have been possible without the love and support of my parents who hold a special place on this page.

Contents

List of Tables

List of Figures

List of Figures

Abstract

This dissertation is composed of three papers that each summarize computational investigations in morphology. The first paper studies the relationship between nominal classification and complexity in language. It subsequently introduces a model for this relationship called the association model of nominal classification. This model claims that nominal classification aids in word storage in gendered languages. These claims are supported by data from German, French, English and Iraqw. The second paper investigates quantitative approaches to measuring the productivity of affixes in Swahili. It subsequently introduces a novel model of measuring productivity called the cumulative root ratio. This model gives a story for the variables that determine whether an affix is productive, and data come from corpus data of Swahili. The third paper studies the relationship between the semantic cohesion of derived words, and the meaning of the affixes which they contain. This study introduces the idea of semantic class coherence and argues that this is correlated with word decomposition in lexical access, and likewise is a prerequisite for affix productivity. This model is supported by data from English word vector space models, along with other corpus data from WordNet and Celex. These three papers each are examples of original research that study human language at the word level using data driven methods. This method of employing computational modeling and machines to investigate human language allows us to better understand the ways in which humans can interact with, acquire, and produce language.

Chapter 1

Introduction

In this chapter we introduce three papers that comprise this dissertation by describing their aim, hypothesis, data, basic methodology and results. Next, we give a brief summary of the how these papers are related. Then, we introduce general approaches to the study of morphology in order to couch these papers in their proper context in linguistic theory. Last, we give an outline of the structure of the dissertation.

1.1 Introduction to Computational Explorations in Morphology

This dissertation is composed of three individual papers which each study the internal properties of words using natural language processing, language modeling and other corpus linguistic techniques. The first of the three papers is entitled *Nominal Classification decreases the Entropy of Nominals in Gendered Languages*,

and it investigates the relationship between grammatical gender (Corbett 1991) and nominal complexity using measures from information theory (Shannon 1951). The second paper is entitled *Are affixes in Agglutinative languages always Productive?* and it investigates morphological productivity in Swahili, and studies how we are able to predict this effect from quantitative measures (Baayen 1992; Hay & Baayen 2002a). Last, the third paper is entitled *Is Word Derivation limited by Semantic cohesion?* and it investigates the meaning of derivational affixes in English, and how this meaning is subject to the tendency of derived forms to adhere to a semantically coherent set.

1.2 Paper One: *Nominal Classification decreases the Entropy of Nominals in Gendered Languages*

This is the first of three papers, and in it we use language modeling and information calculations to investigate complexity as it relates to nominal classification and the nominal lexicon.

1.2.1 Aim

The aim of this paper is to understand why grammatical gender persists in language when it is not a necessary component for all languages (Corbett 1991). While grammatical gender is licensed by the grammars of many languages, it does provide a

challenge to the language learner by increasing the amount of grammatical information that must be acquired (Anderson 2015). Therefore, this paper investigates whether there is an trade-off between grammatical gender and some other component of the language such that it allows this complex system to persist.

1.2.2 Hypothesis

Whereas previous research has argued that grammatical gender persists in German because it mitigates the entropy of nominals in lexical access (Dye *et al.* 2016), we introduce an alternative and broader solution. Based upon data from Iraqw (Mous 1993), and based upon results found in the psycholinguistic literature (Friederici & Jacobsen 1999), we propose that grammatical gender acts to reduce the entropy of nominals in the lexicon, and not at the point of lexical access. We call this approach the association model of nominal classification. If gender acts to reduce the amount of information in the nominal system of a language, then we submit that this effect should be measurable in the nominal systems of languages that vary in whether they contain grammatical gender. In order to test this hypothesis, we perform two studies that investigate the relationship between grammatical gender and language complexity in three different languages which vary in the number of genders that contain. We then investigate the role that complexity plays in each of these languages using measures from information theory (Shannon 1951). If grammatical gender reduces nominal complexity in the lexicon, then we expect measures of information to reveal an effect of gender, and for this effect to relate to nominal organization.

1.2.3 Data

To measure this effect, we look at three corpora (one of English, one of German, and one of French) which are similar in size and source; data come from 100 million word subsamples of the WaCky Corpus (M. Baroni & Zanchetta 2009). To calculate the complexity of each language, we create regular language bigram models of each language (Kleene 1951; Chomsky 1956) and measure conditional entropy, and mutual information (Shannon 1951) for all two word sequences. We then alter these language models to remove all gender distinctions where they exist and then recalculate these information measures.

1.2.4 Basic Methodology

For each language model and their de-gendered counterpart, we measured conditional entropy and mutual information in both the whole language, and in determiner and noun sequences alone. We then compared these calculations (conditional entropy and mutual information) between each language model and its degendered counterpart (German versus degendered German, French versus Degendered French, and English versus its lemmatized form), and then investigated this difference across the three languages (difference in German, difference in French, difference in English) in both of these two environments (overall language and determiner-noun sequences).

1.2.5 Results

In the determiner-noun sequences we find that languages with grammatical gender (German and French) have greater complexity in their degendered forms than in their gendered forms, and that in English there is no change unsurprisingly. Furthermore, the information in German and French gendered bigram sequences is roughly equivalent to the English sequences, but when gender is removed complexity is much greater. This result is true for both conditional entropy and mutual information measures.

For the overall language calculation, there is no such effect for conditional entropy, but there is one for mutual information. We find that the languages with gender have greater mutual information than their non-gendered forms, and that this effect is directly proportional to the number of gender divisions the languages have. However, we find that the conditional entropy values bear no relation to the number of genders a language has, and furthermore, that the non-gendered counterparts always have slightly less conditional entropy values which is the opposite of what we have predicted. This suggests that in the context of the overall language, word pairs have greater informativity when there is gender, but do not increase the overall contextual predictability of a language. We argue that these results support the association model based upon the mathematical properties of mutual information and based upon other follow up studies.

1.3 Paper Two: *Are affixes in Agglutinative languages always Productive?*

This is the second of three papers, and in it we use frequency calculations and corpus subsampling to model novel word formation in Swahili.

1.3.1 Aim

The aim of this paper is to investigate whether the tendency of an affix to occur in novel word forms (productivity) can be predicted by quantitative measures in a language with high rates of affixation. One model has been developed previously to predict the productivity of English affixes (Hay & Baayen 2002a), and it relies upon the fact the English derived forms (e.g. *reinsert* $\rightarrow$ *re*+*insert*) have related underived forms (e.g. *insert*) that exist as words in the language. However, this quality does not reflect the facts in agglutinative languages which tend to only have complex word forms (i.e. ones containing overt affixation). Given this difference, we show that these models make the wrong predictions for the productivity of affixes in the agglutinative language Swahili. Therefore, this paper examines whether affix productivity can be predicted by frequency patterns in word forms in Swahili using an alternative method.

1.3.2 Hypothesis

Previously, the frequency relation between derived forms (e.g. *reinsert*) and their underived counterparts (e.g. *insert*) have been argued to relate to the productivity

of a derivational affix (e.g. *re-*) in English (Hay & Baayen 2002a). However, given Swahili's rich morphological system, we must redefine the notion of an underived form since no such concept exists in the language. We redefine this notion by calculating the frequency of the underived form using the cumulative root frequency of the root of the word (e.g. the sum of all inflectional and derivational variants which contain the word *insert*) and compare that to the cumulative root frequency of the derived form (e.g. the sum of all inflectional and derivational variants which contain the word *reinsert*). We label this approach as the cumulative root ratio model. We expect that the ratio of our newly defined underived and derived forms should correlate with the number of novel occurrences in the language which we calculate using the established measures of (i) the number of hapax legomena, (ii) P, and (iii)$P*$ (Baayen 1993a) which do not rely on the frequency of underived words.

1.3.3 Data

In order to investigate the predictability of the cumulative root ratio model, we subdivided the 13.6 million token Helsinki Corpus of Swahili (Hurskainen 2004) into five evenly-sized, and randomly selected subsections.

1.3.4 Basic Methodology

We identified 48 different inflectional and derivational affixes split across the nominal and verbal domain. For each of these affixes, we calculated the cumulative root frequencies of all underived and derived forms of the words which contain the affix in each subsection. We plotted these frequencies in log space with the derived frequen-

cies on the x axis and the underived frequencies on the y axis. We then extracted three variables from this plot that we are associated with affix productivity (Hay & Baayen 2002a): (i) the proportion of forms containing an affix whose underived frequency is greater than its derived frequency (type ratio), (ii) the slope of a least trimmed squares linear model of the data, and (iii) the y-intercept of this model. We also calculate the number of hapax legommena, and the P, and $P*$ measures (Baayen 1993a) for these affixes in each subsection. Last, we calculated the correlation between each of the three cumulative root ratio variables and each of the three novel word-form variables in all of the non-overlapping subsections of the corpus.

1.3.5 Results

The results reveal significant positive correlations for all of these variables across the corpus. This indicates that cumulative root ratio is a plausible measure for the productivity of affixes in Swahili. We discuss the implications for this measure and its underlying assumptions throughout the paper.

1.4 Paper Three: *English Derivation and Semantic Class Coherence*

This is the third of three papers, and in it we use word vector space models and frequency measures to investigate the relationship between affix meaning, affix productivity, and the semantic coherence of an affix's derived set.

1.4.1 Aim

Often, English derivational affixes have obvious meanings (e.g. *im-* in *imperfect* means *not*, i.e. *not perfect*). The aim of this paper is to investigate the forces that contribute to the meaning of these derivational affixes in English. It has been shown previously that the meaning of an affix is related to its productivity (Siegel 1974; Hay 2004), such that more productive affixes have more coherent meanings. Meanwhile, affixes vary in the degree to which they are productive. Furthermore, this asymmetry is reflected in the fact that affixed word forms may not always have meanings that reflect their composition (Pelletier 1994), but rather have autonomous non-compositional meanings. In this paper, we investigate the relationship between affix productivity, compositionality, and meaning using word vector space models of English (Mikolov *et al.* 2013; Pennington *et al.* 2014).

1.4.2 Hypothesis

Words formed from English derivational affixes exhibit variation in the degree to which their semantic function can be easily defined. For example, the affix *-ist* in *pianist* has clearer definition than an affix like *be-* in *beguile*. We expect therefore that there should be more variation in the meanings of all the word forms that contain *be-* than those containing *-ist*, and that this variation should correlate negatively with calculations of semantic drift (a measure of non-compositionality), and positively with productivity (an alternative measure of semantic coherence).

1.4.3 Data

We take 79 derivational affixes of English, and for each affix we find all word forms which that affix along with their affixless counterparts. For all forms, we retrieve the 200-dimensional word vector representation from the GloVe corpus (Pennington *et al.* 2014). For all affixes, we calculate the average number of definitions of each word form in WordNet (Miller 1995; Hay 2004), and two variables associated with affix productivity from Celex (Baayen *et al.* 1993): (i) the type ratio (Hay 2004), and (ii) P (Baayen 1993a).

1.4.4 Basic Methodology

In order to calculate the semantic coherence of the set of derived words, we introduce a measure which we call the average cosine similarity. This measure calculates the variance in the cosine similarity (a measure of the similarity between two words in vector space) of all word forms as they relate to each other by measuring the average cosine similarity to their mean vector. We then argue that this value captures the semantic coherence of a set by showing that all derived forms of an affix have greater average cosine similarities than the underived forms. That is, all words containing an affix (e.g. *-ist*) are closer to each other in semantic space than the related word forms which do not contain that affix. This shows that (i) derivational affixes form a semantic grouping that are more coherent than the set of its derived forms, and (ii) that this measure can capture the difference between these two groupings which we assume to vary in their semantic coherence.

Given this effect, we measure average cosine similarity of all affixes and compare

them to the average dictionary entries of the derived forms of each affix, and the two productivity measures. If derivational affixes have meaning, and if this meaning varies as a function of the meanings of the members of the set of all words which contain the affix, then we predict that the average cosine similarity will negatively correlate with the number of dictionary entries, and will positively correlate with the productivity measures.

1.4.5 Results

The results reveal a significant negative correlation between the average cosine similarity calculation and the average number of dictionary entries, and a significant positive correlation between the average cosine similarity calculation and both productivity measures across all affixes. This suggests that the meaning of an affix is determined by the meaning and compositionality of its derived forms, mirroring the effect of productivity. In the paper, we end by proposing that semantic coherence is a precursor to productivity, and by relating affix meaning to the grammatical properties of affixes.

1.5 Relating the Three Papers to one another

These three papers each represent investigations into the internal structure of words in human language using computational methods and data driven analyses . More specifically, they focus on the effects that complexity, frequency, and semantic coherence have on the representation of word internal units, or affixes. Although the

first paper focuses on the German, French and English, the second paper focuses on Swahili, and the third paper focuses on English, these pieces of research are intended to inform linguistic theory in a general sense. They are each designed to relate to the human capacity for language regardless of the language spoken, and therefore are attempts to understand the relationship between cognition and language.

To do this, each piece of research employs computational methods to calculate quantitative measures associated with the different phenomena under investigation. In doing so, they each fall into the realm of computational linguistics and morphology.

The assumptions underlying each paper are as follows. First, each paper supposes that word level structure does not exist a priori, but rather that structure must be inferred by speakers from sets of words and consequently exhibit asymmetries in storage and productivity based upon lexical regularities. This perspective is couched in morphology more generally in the coming subsection.

Last, these papers are part of a sea change in the study of language in which there is enough data to begin creating models that interact with massive amounts of data. This fact means that these models encounter data approaching that of a human's experience, such that we can model an entire speaker's lexicon. This quality allows us to model language as a complex system using the techniques described above, which has become another proving ground for the testing of models of language and cognition.

1.6 The Internal Properties of Words and Linguistic Theory

In order to give context to these three papers as they relate to linguistic theory, we give a general introduction into what we mean by the internal properties of words.

Grammarians dating as far back as Panini (6-4th century BCE), De Courtenay, and De Saussure have noted that words are composed of smaller units that relate form and meaning field (Cardona 1997; de Courtenay 1972; De Saussure 2011). These units are often referred to as morphemes, and their covariation across words is the central focus of the field of morphology. It has been argued that these units are one of the essential elements of language although their form is made up of meaningless units (Hockett 1961). For example, the English plural form of the word *cradenza* is *cradenzas* such that there is an arbitrary yet meaningful relation between the suffix *-s* affix, the meaning of plurality, and the word *cradenza*.

In the last century, researchers in the field of morphology have sought to understand the role that these morphemes play in to the human capacity for language as it relates to grammar, and early insights argue that complex words are all composed of these units (Bloomfield 1933; Chomsky & Halle 1968; Matthews 1972; Aronoff 1976). These approaches are broadly known as morpheme based approaches and they argue that words are composed of morphemes in the same way that sentences are composed of words. More recent theories in this area have maintained this compositionality as the primary unit of word structure (Kiparsky 1982; Halle & Marantz 1993; Harley & Noyer 1999).

However, another class of theories have argued that complex words are not the result of the composition of smaller units, but are the result of the interaction between word schema that represent the features common to all morphologically related words (Zwicky 1985; Anderson 1992; Blevins 2006; Finkel & Stump 2007). Whereas morpheme based theories presuppose that words are composed of morphemes, word based theories argue that the covariation between these units are captured in the grammar by their similarities in form and meaning. For these realizational theories, the form of a word is an surface-level interaction between an underlying lexical form, and different environments (Bybee 1985; Aronoff 1994; Stump 2001).

Beyond these grammatical theories, another body of research has sought to understand the relationship between morphological structure and lexical access by performing behavioral studies in lab settings (Taft & Forster 1975; Taft 1979; Pinker 1997; De Jong IV *et al.* 2000; Baayen & Schreuder 2003). These models argue for models of the lexicon which vary on how these words are stored in the mental lexicons of native speakers of a language.

In this dissertation, each study touches on aspects of these bodies of research, and throughout we give descriptions and discussions of relevant research in these areas.

1.7 Outline of the Dissertation

The outline of the dissertation is as follows. The three papers described above are found in chapters two, three, and four respectively. Since these papers are au-

tonomous works, they each have a unique structure which we outline in the next few paragraphs.

Chapter two is comprised of six different sections. The first section introduces the idea of language complexity as it relates to nominal classification using real language examples. The second section discusses different theories of nominal classification and how they relate to the literature on lexical access, and then introduces the association model of nominal classification. Section three contains an introduction to regular language models and describes how we can use these models along with measures from information theory to investigate the relationship between complexity and nominal classification. Section four gives two different studies on how nominal classification impacts complexity: one at the determiner-noun level, and one at the full language level. Last it gives a follow up study which evaluates the effect of language model size on the conditional entropy and mutual information measures. Section five contains an argument for the adoption of the association model, and finally section five gives a brief summary of the paper's findings.

Chapter three contains five sections which outline an investigation into affix productivity in Swahili. Section one introduces the notion of productivity with examples from Swahili derived nominals. Section two outlines different models of affix productivity and explores how they relate to Swahili derived nominals. Section three describes the motivation behind the cumulative root ratio model of affix productivity and in doing so introduces different aspects of Swahili morphology. Section four proposes a corpus study to investigate the effectiveness of the cumulative root ratio to quantify affix productivity in Swahili before presenting the study results.

Last, section five summarizes these results and contextualizes the implications of the cumulative root ratio for theories of morphology.

Chapter four contains five sections which describe the investigation into the meaning of derivational affixes in English. Section one introduces the idea of affix meaning as it relates to compositionality and semantic drift. Section two summarizes different theories of affix composition as they relate to the lexicon. Section three outlines a new way to model the semantic coherence of a set of words using the average cosine similarity of vectors in word vector space. To do this, it gives a descriptions of word vector space models. Section four describes three studies that investigate the usefulness of modeling semantic class coherence using the average cosine similarity measures, and how this measure relates to semantic drift and affix productivity. Last, section five contains a discussion of the results of these studies and describes what they can tell us about the meaning of derivational affixes, and how this relates to morphological theory.

In addition to these three chapters, we include a concluding chapter at the end in chapter five. This chapter contains summaries of the findings of each of the three papers. Next, it discusses future steps for the research in each paper, and finally ties the three topics into a unifying vision.

Chapter 2

Nominal Classification Decreases the Entropy of Nominals in Gendered Languages

2.1 Introduction

Nominal classification is a grammatical property that is present in nearly half of the world's languages (Corbett 1991). More commonly referred as gender, nominal classification constitutes a family of phenomena that subdivide all nominals in a language into a number of categories. The categories that result from this subdivision must agree with words or affixes with some syntactic relation to the nominal (Corbett 1991). For example, in French, nominals are obligatorily divided into masculine and feminine genders (Price 2013). Articles and adjectival constituents that modify these

nominals must agree with them in gender.

(2.1) (a) Le petit livre

the.*masc* little.*masc* book.*masc*

'The little book'

(b) La petite table

the.*fem* little.*fem* table.*fem*

'The little table'

In (2.1a-b), each of the nominals are preceded by a determiner and an adjective. These words take different inflections based upon the gender of the nominal that they modify. For example, the definite determiner takes the form *Le* or *La* based upon whether it precedes *livre* or *table*. Given that all nominals are bound to a gender category, we can infer that the French learner must acquire these class labels along with the inflectional variants of the agreeing words. While French represents a binary classification (i.e. Masculine/Feminine), languages do vary in the number of nominal categories employed in classification (Comrie 1999). Crosslinguistically, nominal classification has several interesting properties:

(2.2) (a) It occurs in many languages spanning both families and typologies, but is not a necessary requirement to the grammar of all languages (Dryer *et al.* 2005; Corbett 1991); For example, Swahili (Niger-Congo) has 9 nominal classes (Mohamed 2001a) Welsh (Indo-European) two (Thorne 1993), Iraqw (Afro-Asiatic) three (Mous 1993), and Dyirbal (Pama–Nyungan) four (Dixon 1972).

(b) It can employ multiple criteria for classifying nominals: by the qualities of semantic similarity, phonological form, or can be totally arbitrary. And furthermore, it varies in the number of distinctions/labels in the system (Vigliocco *et al.* 2005; Kilarski 2007; Comrie 1999).

(c) It contributes to the overall complexity of a language, by increasing the number of inflectional variants of word forms (Anderson 2015; Hawkins 2004).

A few questions arise when we consider these properties. The fact in (2.2a) indicates that nominal classification is an enduring but non-essential quality of language that can vary in type and quantity (2.2b). This variation paired with the facts in (2.2c), leads us to ask why such a system would persist in a language. Specifically, we can ask which pressures would allow a non-essential system to not be evolutionarily selected against even though it increases the overall complexity of the language.

To be clear, when we use the term nominal classification we are specifically talking about a grammatical property of these languages. By grammar, we mean that native speakers of a gendered language must acquire gender in order to form grammatical nominal phrases. Given that this property is a part of the grammar but not necessary for grammars in all languages, then we propose that there must be some pressure outside of the grammar which influences the persistence of nominal classification systems. This pressure must therefore be enough to mitigate the issue of added complexity in order for nominal classification to persist in a given language.

In this paper we argue that nominal classification persists because it lowers the complexity of the nominals in a speaker's mental lexicon. We propose that the relation between gender and the nominal lexicon is tied to mental organization and memory. We define complexity by using measures from information theory (Shannon 1948) in order to formulate a quantitative value for natural language (Shannon 1951). Whereas other have argued that there is a performance based effect of nominal classification (Dye *et al.* 2016), we argue that the net effect of gender in a language cannot be based on contextual predictability alone, but must rather stem from pressures of lexical organization. We develop this argument by calculating the complexity of regular language models of languages that vary in the number of nominal classes (i.e. German, French, and English). Based upon the results, we propose that gender does not act only to disambiguate nouns in context, but rather is an associative mechanism (Paivio 1963). This mechanism aids in forming associations when acquiring nominals, and may also benefit the recall of nominals via redintegration (Lewandowsky & Farrell 2000; Jones & Farrell 2018).

The paper is organized as follows. In Section 2, we discuss previous nominal classification research and describe models of nominal classification. In Section 3, we introduce information theory and describe how we can calculate the complexity of nominal classification a language model. We then propose studies that investigate this impact in models of German, French and English. In Section 4, we present the results of two studies that investigate these effects and perform a follow up study that investigates corpus size. In Section 5, we expand these results to our association model more generally, and demonstrate why they are inconsistent with current models that argue for discrimination as the potential source for nominal classification (Dye *et al.* 2016; Baayen & Ramscar 2015).

2.2 Nominal Classification and the Lexicon

In this Section, we provide an account of the cognitive research on nominal classification both as it relates to the languages included in our study, and more generally. We then discuss candidate models for the pressure to maintain nominal classification, and discuss their cognitive implications. Last, we introduce our associative model.

2.2.1 Nominal Classification and Cognition

For as long as language has been studied, nominal classification has been of interest to language researchers (Kilarski 2007). Early studies into Indo-European grammatical gender argued that gender was unrelated to notions of biological sex (Wheeler 1899). Later research on gender expanded the study to a broader range of languages and in

doing so argued that while gender has synchronic grammatical properties common to all languages, its origins vary from language to language (Istvan 1959). More recently, the focus of research has been on classification's role in lexical access (Levelt 1993; Van Berkum 1996; Vigliocco *et al.* 1997), in syntactic agreement (Jelinek & Carnie 2003; Baker 2008), as well as in first language (Plaster *et al.* 2009; Arnon & Ramscar 2012), and second language acquisition (Rogers 1987; Guillelmon & Grosjean 2001; Sabourin *et al.* 2006). Meanwhile for languages with nominal classification, it is impossible to discuss grammar in any sense without making extensive reference to its role in the language (Greenberg 1960; Prins 1961; Heine & Reh 1984; Nsoh 2002).

In this paper, we focus on nominal classification in three different languages: English, French, and German. These three languages are chosen because they meet two specific requirements. First, they vary in the number of nominal classes required by the grammar (English one (Zandvoord 2013), French two (Price 2013), and German three (Durrell 2011; Bierwisch 1967)). Second, there are readily available equal-sized corpora that allow us to make comparable language models (M. Baroni & Zanchetta 2009). One major concession of this choice is that these languages are all members of the Indo-European language family, and have had extensive contact throughout their history (Bouckaert *et al.* 2012). Ideally, to test the role of gender we would prefer to compare languages of different families, however few comparable resources exist which are large enough to fit the size we require to compare lexical complexity. We leave this comparison for future work.

Although English has no grammatical gender, it does have subclasses of nominals that morphologically mark singular and plural agreement, and other non-

grammaticalized classes such as the mass-count distinction (Gillon 1999). However, unlike French and German, determiners do not have inflectional variants, but do mark the grammatical features of specificity and plurality in some cases (Smith 1964; Keizer 2007). French, and German on the other hand mark nominal classes in definite and indefinite determiners in both singular and plural forms. Although, French does not mark gender in plural articles (Price 2013).

In each of the three languages, nominal morphology has been examined from both grammatical and psycholinguistic perspectives. We focus on the psycholinguistic effects of gender in language comprehension and production as they relate to the languages in this study, but also in language generally. In order to study gender, most work employs either visual or auditory stimuli that are designed to understand the role that gender plays in lexical access or production. Here we present a few key studies.

One of the first studies into gender in French found that speakers commonly relied upon the endings of words to determine their gender when it was not already known. The major source of this being that sets of affixes in French commonly pair with either masculine or feminine gender (Tucker *et al.* 1968). These findings were challenged in later work which argued that gender was determined by lexical associations generally in addition these suffixes (Holmes & de la Bâtie 1999). More recently, the status of these lexical associations has been central to the study of gender. For example, experiments have studied how participants respond when some quality of the gendered stimulus is altered. Grosjean (1994) used gating and lexical decision to test how speakers access nominals in French. The stimuli had two conditions:

one where there was a gendered article preceding the nominal, and one where it was absent. They found that nominals were accessed faster in both modalities when the article was present, and slower when it was not included (Grosjean *et al.* 1994).

This finding was also found to be true in other languages. In Dutch, it was found that gender mismatches greatly hinder response latencies between articles and nominals (Schriefers 1993). Subsequent research in Italian has argued that this sort of effect involves a combination of facilitation and inhibition in word naming, gender monitoring, and grammaticality judgment tasks (Bates *et al.* 1996). These effects were also found in Dutch ERP studies using visual stimuli (Wicha *et al.* 2003).

Jacobsen (1999) found similar results in German for picture naming (visual), but only found inhibition effects of incongruency in the word naming task (auditory) (Jacobsen 1999). This inhibitory effect was also found to occur in German picture-word interference studies (Schriefers & Teruel 2000), as well as in spoken recognition tasks in German (Bölte & Connine 2004).

In 1999, two meta-studies were published focusing on gender perception and production. Friederici and Jacobsen (1999) review the perception literature and point out that the inhibitory effect is present across all modalities (both visual and auditory), but that the facilitation effects only occur in the visual modality. They claim that facilitation is an effect of semantics, and not an syntactic effect. This leads them to argue that gender shares a post-lexical relationship with nominals, and that it does not facilitate prelexical access (Friederici & Jacobsen 1999). Meanwhile, Schriefers and Jescheniak (1999) review the production literature. They conclude that the current studies in production suggested that gender is an abstract lexical

property, and that its processing occurs at an abstract grammatical level. From these studies, we can conclude that nominals are intimately linked to their nominal class, but that nominal class is not always a requisite for lexical access. This is reflected in more recent studies which show for example that grammatical gender is preserved in speech errors in German (Vigliocco *et al.* 2004), that it increases the ability of learners to acquire nominals in artificial language experiments (Arnon & Ramscar 2012), and strikingly that it activates specific portions of Broca's area that are different from regions in the brain associated with lexical access (Heim *et al.* 2002).

In summary, we can draw a few conclusions from this body of literature. First, we know that nominal classification is intimately tied to the words they modify, and that learners rely on gender when acquiring nominals. In perception, we see that speakers of gendered languages are in all cases inhibited by gender mismatches between nominals and their modifiers, but that only in visual studies do we see a facilitatory effect. In production, we see that gender has an abstract grammatical relationship to nominals. These behavioral studies reflect the constraints one should place on any model aimed at explaining the pressure to maintain nominal classification in the grammar of a language. In the next subsection, we present models of such pressures and discuss their assumptions and implications.

2.2.2 Models of Nominal Classification

Here, we review models of grammatical gender in language. First, we explore ones developed in the lexical access arena in a bit more detail before describing functional

approaches to gender. Last, we discuss our proposal for grammatical gender. Along the way, we identify assumptions and implications of the models.

First, we saw that the behavioral literature made reference mostly to models of lexical access. In these studies, when stimuli are presented they are manipulated to identify facilitatory effects when gender is present or inhibitor effects when it is not present. Broadly, the aim is to understand the levels at which nominal classification impact language whether they be prelexical or post-lexical. Under prelexical models, gender facilitates access of nominals by narrowing the options available for access (Bates *et al.* 1995). On the post-lexical or modular view, gender information is not used to access nominals but rather is only relevant after access occurs for syntactic congruency (Tanenhaus & Lucas 1987; Friederici & Jacobsen 1999). These models hold a connectionist view of the lexicon in which lexical access is associated with node activation, and the geometry and direction of such activations is under examination (Bechtel & Abrahamsen 2002).

Second, there is a body of research on the relationship nominal classification and its function in communication and discourse. For some, grammatical gender allows increased comprehension of referents in a discourse, for example in co-reference resolution (Zubin & Köpcke 1986). In this instance, gender should facilitate easier identification of referents in a discourse in a language like German because it alleviates possible ambiguities in potentially ambiguous syntactic configurations. Expanding on the idea of function in discourse, Dye et al (2016) develop the idea that gender in German has a specific role in managing the flow of information between speakers. Specifically, it reduces the uncertainty of possible candidate nouns which follow

gender-marked determiners in discourses (Dye *et al.* 2016). Under this view, words are seen as discrete units that must be isolated and inferred from a continuous speech signal. Given this task, some hold the view that language structure has evolved over thousands of years to optimize this process, while balancing the amount of information that is being conveyed (Baayen & Ramscar 2015). One way of balancing this is to redistribute the uncertainty in nouns to an increase in the uncertainty of determiners which precede them. Here, gender is a linguistic cue that acts to mitigate the uncertainty of later linguistic information at the point of discrimination (Baayen *et al.* 2011; Ramscar 2013). From this perspective, gendered determiners systematically narrow the set of possible candidates that they precede (Dye *et al.* 2016). In this way, nominals are more predictable in context precisely because they follow gendered markers in linear order, as in (2.3).

(2.3)

Lucas	sieht	ein	!	Halsbandpekari
Lucas	*see.3.sg.pres*	*indef.masc*		javelina

'Lucas sees a javelina'

Above, we see a simple transitive sentence of German in subject-verb-object order. As a German speaker hears this sentence, they would sequentially hear *ein* preceding the nominal *Halbsbandpekari*. The exclamation point reflects the point in which a nominal must be discerned following the gender marked cue. This effect would then indicate to the listener that the coming nominal is masculine (or neuter) effectively narrowing the logical set of potential candidates. In this way when people perceive a gendered classifier in a discourse, they are relying on the previous information as a

form of scaffolding to make nominals more predictable in that discourse. To be clear, Dye et al (2016) do not make the claim that this property is true for all nominal classification languages, but rather claim this to be the case only in German. Before describing how their model measures complexity, we present an alternative analysis. Upon doing this we describe different assumptions and implications for these two models.

2.2.3 Association Model

While others have suggested that nominal classification aids in perceptually navigating particularly difficult points in discourse, we present another proposal. Although it may be the case that gendered articles precede nominals in many language, we find that this generalization does not hold across all languages. For example in Iraqw, gender marking follows nominals as an affix as seen in (2.4). Iraqw contains three arbitrary nominal classes much like German, but varies in how it marks them inflectionally.

(2.4) (a) **Iraqw**

hhar-	-ta-	-ka
rope	*fem*	*indef*

'a rope'

(Mous 1993)

(b) **German**

ein	Seil
indef.neut	rope

'a rope'

In 2.4a, the suffix *-ta-* follows the root word *hhar-*. Here, we see that gender in Iraqw cannot reduce the uncertainty of this nominal in discourse simply because it does not precede it in linear order. This not the case in (2.4b), where gender marking on *ein* precedes *Seil*. Beyond the nominal syntax, we find the same ordering issue when we look at larger pieces of syntactic structure. Although Iraqw verbs do take gendered inflectional variants, the standard word order is subject-object-verb (Mous *et al.* 2002). An example of this can be found in (2.5).

(2.5)

an'ing	?	gitla-	-d'a	u-	-na	aahhiit
I		man	that	*obj.masc-*	*-past*	hate.*1sg*

'I hate that man.'

(Mous *et al.* 2002)

In this case, we see that the masculine verbal marker *u-* follows the nominal object

gitlad'a. The question mark indicates the location where nominal discrimination should occur in linear order. However, no gendered information has been given prior to it. Therefore in the Iraqw sentence, the syntactic order does not collude with gender to reduce the complexity of nominals in discourse. Rather, gender must have some other function in this language if it is to persist. Given the absence of corpora in Iraqw that would allow us to investigate the effect of gender, we cannot directly study the complexity of nominals in this language. Therefore, we can only keep this system in mind when evaluating nominal classification in our three languages.

In the sort of system found in Iraqw and more generally, we propose that nominal classification is a grammatical property that mitigates the complexity of nominals in the mental lexicon. In this case, classification is a method of increasing memorizability of nominals by giving the language acquirer a ready-made mechanism to aid in association mapping (Paivio 1963). Research has shown that while associations are taxing in terms of brain processing (Dennis *et al.* 2015), their creation enhances the likelihood that some item is kept in episodic memory (Paivio 1969). It has also been show that these sorts of associations, between an item and its name, is involved in areas related to language learning (Ranganath & Ritchey 2012).

Furthermore, it has been shown in artificial language learning experiments that grammatical gender helps learners acquire nominal systems that are completely novel for them (Arnon & Ramscar 2012). We feel that this quality is not an effect of disambiguating discourse, but simply an effect of aiding in rote memorization of word lists.

This proposal is informed by the fact that nominals are the largest and most com-

plex of lexical category in gendered languages, and therefore should be the aspect of language acquisition that requires a large amount of memorization. Such memorization would benefit from ready-made associations entailed by the grammar via nominal classification. Furthermore, redintegration effects in language suggest that this association may allow speakers to successfully recall associations from incomplete information (Lewandowsky & Farrell 2000; Towse *et al.* 2008; Jones & Farrell 2018). However, this use of long-term knowledge to facilitate recall presupposes that linguistic representations are recall-able. Likewise, the effects seen in perception and production experiments reflect that this sort of knowledge is accessed at a higher more abstract level. Under this perspective, there is pressure for gender to be preserved as a function of lexical organization in the minds of speakers, and not as an aid to discrimination in discourse.

2.2.4 Model Implications

In this subsection we discuss implications and predictions of the association and discrimination models of nominal classification.

For either model to be true, there needs to be evidence that nominal class reduces the complexity of nominals. This effect can be measured by taking a language model sourced from actual language and calculating complexity using information theory. In the next Section, we discuss this in greater detail, but for now we isolate the predictions that either model would place on complexity reduction.

First, we saw that there is a difference between the requirements on the linear ordering of items in a nominal classification system in the two models. Dye et

al. (1999) argue that German gender markers reduce uncertainty in nominals only by preceding them. Their model and naive discrimination learning more generally claims that complexity reduction is a task that occurs in a linear order that we will label as asymmetric (Baayen & Ramscar 2015). In this model, for a nominal classification system to participate in complexity reduction the system must show this asymmetry. In the association model such linear order is unnecessary. Here, gender labels are learned as associations. Such a system has no limitations on the linear order of nominal classifiers and the nominals that they modify. This is because in this model class labels are features of a nominal that are acquired at the point of initial learning and therefore are linked with a given nominal but not in discourse time. Rather, the mental representations of nominals and their labels are associated with one another, and their access must not be asymmetric. On the contrary, we predict that such associations should inherently show properties of symmetry. In short, the discrimination view predicts that effects of complexity reduction should be asymmetric, and the association view predicts that they should be symmetric.

Second, these two models make different predictions when it comes to the overall complexity of the lexicon. For the discrimination model, classification reduces the uncertainty of an element in discourse and not in the lexicon. Therefore under this model, complexity reduction should only hold in discourses and not for all nominals in the language. Complexity reduction in upcoming nominals co-occur with other complexity reducing elements. For example when one is predicting an upcoming nominal, they would also be aided by the knowledge of who is speaking, the verb preceding the nominal, and the previous statement in discourse, for example. Classification would

then only be useful in reducing complexity in these already constrained contexts, and not in the overall language. On the other hand, the association model would predict that there must be some effect at the lexicon level. That is for all nominals in a language regardless of context, there would be some measurable reduction in complexity. Any measure that occurs in limited contexts should be calculable at any context. In summary, the discrimination model predicts that complexity reduction occurs in limited contexts, and the association model predicts that it must occur globally in the lexicon.

In this Section we have given a summary of literature on nominal classification and gender as it relates to cognition and lexical access. We then presented two models of nominal complexity reduction that aim to explain the pressure to maintain a nominal classification system in a language. Given their properties, we presented differences in the parameters of this complexity reduction. In the coming Section, we give a mathematical definition for complexity and then present language models that allow us to calculate such complexity. We will then see that these results suggest that we adopt the association model.

2.3 Regular Languages and Information Theory

In this Section, we introduce information theoretic ways of measuring complexity. We then describe the mathematical properties and implications of three different complexity measures. Next, we introduce regular language models of English, French, and German, and show how these models can reveal the effect of nominal classifi-

cation on complexity. Last, we describe how we can apply these measures to such models, and what they would reveal about discrimination versus association view.

2.3.1 Information Theory

Initially, information theory was developed as a means of quantifying the complexity of a message that would be sent from a sender to a receiver over a noisy communication channel (Shannon 1948). In order to quantify how complex a message may potentially be, Claude Shannon (1948) developed the notion of entropy. This measure is a bitwise value that tells us, given several possible messages how much information is needed to encode all possible messages of a language. Although the field of information theory was developed over 70 year ago, the notion of entropy and related calculations have pollinated over the years and have become increasingly used in language research (Moscoso Del Prado Martín *et al.* 2004; Bell *et al.* 2009; Ackerman & Malouf 2013).

According to information theory, we can quantify the complexity, or information contained within a single word by measuring the likelihood that a word occurs in some language model ($p(w)$). The inverse of this value is then log transformed in base two, as in (2.6). By using log base two, this formula converts complexity into a binary value which is essentially an information value that is composed of ones and zeros. The result is called a bitwise value that digitally encodes the amount of complexity of this individual word in a language model.

(2.6) $$I_w = \log_2(1/p(w))$$

One important quality to note here is that complexity is maximized when a word is maximally unpredictable. To understand this, let's consider a hypothetical example. Imagine a hat in which there are two pieces of paper with a different word written on each of them. Take for example the words, *coatimundi* and *javelina*. If one were to pull out a single piece of paper, what would the likelihood be that it would be *coatimundi*? Given that there is an equally likely chance of pulling out either, they each have a probability of occurrence of 1/2 ($p(coatimundi) = 0.5$). This value represents maximum unpredictability, and the entropy value would be 1 bit according to the formula in (2.6). This information value captures the intuition that one would have no way of predicting which word would be picked, and therefore we would need 1 bit to define the amount of information encoded when the two words are equally likely.

On the other hand, imagine that the hat contained three pieces of paper with the word *coatimundi*, and only one with the word *javelina*. Here, the probability of pulling out *coatimundi* would be 3/4 ($p(coatimundi) = 0.75$), and the amount of information in this word would be 0.41 bits.

Given this notion of information, we can quantify the complexity of all words in a given language by averaging the information calculations across all words that occur in that language. This value is known as entropy, and can be calculated with the formula in 2.7.

$$H(W) = -\sum_{w_i \in W} \mathrm{p}(w_i) \log_2 \mathrm{p}(w_i) \tag{2.7}$$

Entropy is a much more useful calculation than individual word complexity be-

cause it allows us to quantify the amount of complexity across all possible words in a language. To highlight this, let's consider another example. Recall that we made the claim earlier that nominals provide the greatest amount of complexity in gendered languages. Given a corpus of words, we can quantify such complexity by calculating entropy for all words in the language. To investigate the claim that nominals are more complex than any other lexical category, we can calculate the entropy of all words in each lexical category and compare them. To do this, we separate the entropy calculations from the overall language in (2.7) into the different lexical categories as seen in (2.8).

$$H(POS_j) = -\sum_{w_i \in POS_j} \mathrm{p}(w_i) \log_2 \mathrm{p}(w_i) \tag{2.8}$$

For all lexical categories in a language, we can quantify the difference in complexity between all categories by applying the formula above to a corpus. Here we do precisely this by applying the formula to the first 100 million words of the English, French, and German WaCky corpora (M. Baroni & Zanchetta 2009). Figure 2.1 gives the entropy calculation for each lexical category in each corpus. We see that indeed nominals contain a much greater amount of complexity than all other categories in each language.

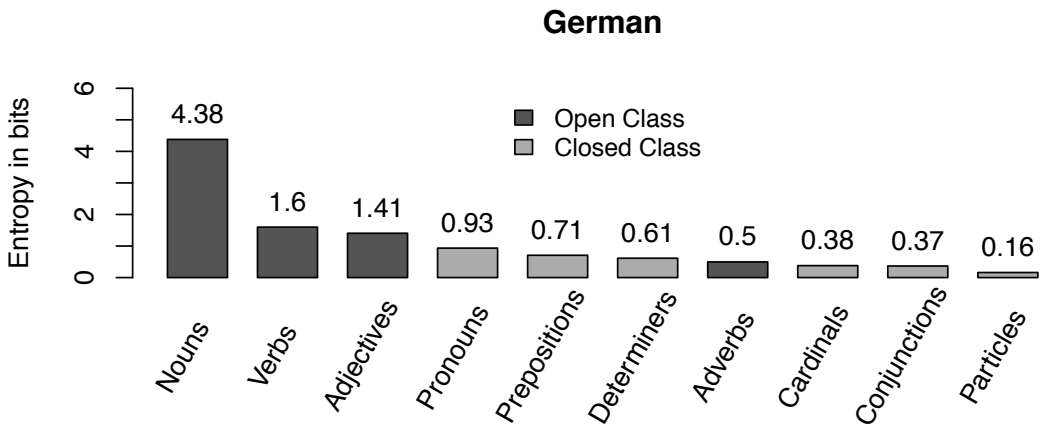

Figure 2.1: The Entropy calculation of all words in each lexical category in 100 million word subset of the German WacKy corpus.

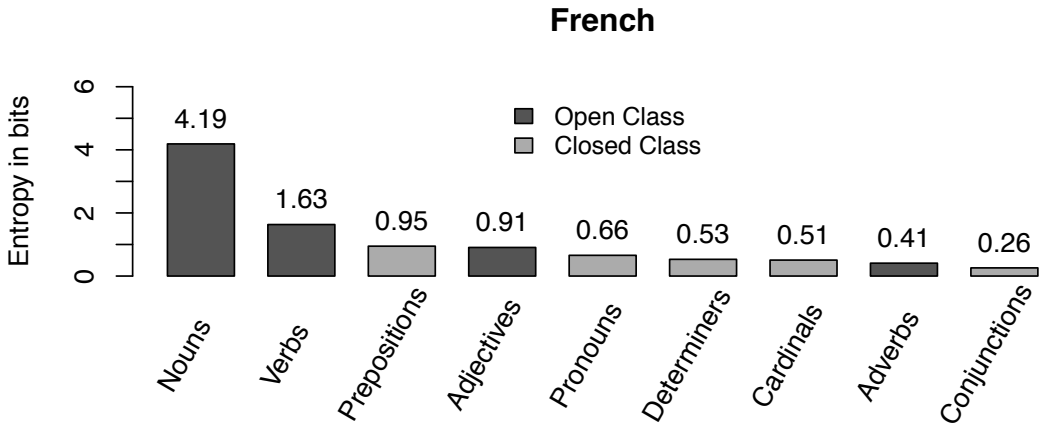

Figure 2.2: The Entropy calculation of all words in each lexical category in 100 million word subsets of French in WaCky.

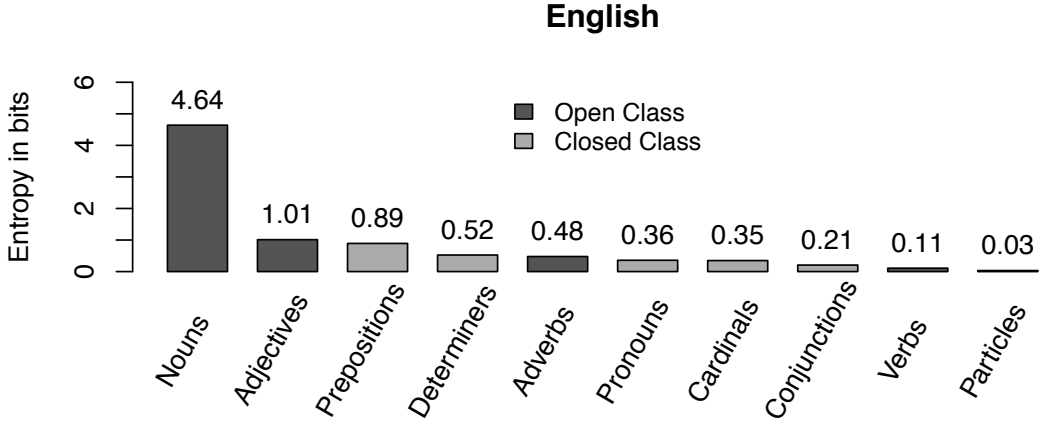

Figure 2.3: The Entropy calculation of all words in each lexical category in 100 million word subsets of English in WacKy. For each language, nouns contain around four times more entropy, or information than the next largest category.

From this data we can assess that in these language models nominals contribute the majority of information. This simultaneously seems trivially true, while being informative for our investigation. In any case, these languages tend to have lots of nouns relative to other lexical categories in the language. However, the entropy calculations give us the ability to quantify this notion in a model using a more informative calculation than a calculation like frequency.

So far, we've focused on the entropy of a single random variable by looking at single words in isolation. However, we are interested in measuring the relationship between nominal classifiers and their nominal counterparts. Thankfully, information theory provides mechanisms that enable us calculate the interaction of two random variables and how they influence each other. Given that gender marking in German and French is marked via inflection on determiners and adjectives that precede a

nominal, we need such methods to evaluate these multi-word collocations. To do this, we need to consider the relationship between two words by treating them as two random variables. Doing so will allow us to analyze the relationship of complexity between each word as it relates to the overall language.

(2.9) W_1 & W_2,

The schema above is simply a mathematically explicit way of saying that we want to look at all word pairs in a language and evaluate on average how the first word impacts the second and vice versa. We use the notation w_1 and w_2 to describe any single instance of these variables (i.e. individual members of the set W_1 and W_2). Note that all words in a corpus are members of both the set of W_1 and of W_2 as we iterate through it, but only relative to another word. Take for example the sentence in (2.10):

(2.10) *Lucas saw a javelina and a coatimundi.*

Given this sentence, we can extract all w_1 and w_2 which compose the variables W_1 and W_2 by iterating through the sentence and matching pairs of words. The result is eight word pairs, as seen in (2.11). We include a *START* symbol to denote the initial word in this sentence.

(2.11)

W_1	W_2
START	Lucas
Lucas	saw
saw	a
a	javelina
javelina	and
and	a
a	coatimundi
coatimundi	.

With this schema, we can calculate the probability of a two words co-occurrence. This is done by counting how many times a two word sequence occurs in a language. For example, the probability that *saw* & *a* occurs is in the sentence in (2.11) is 1/8.

This value is called the joint probability, and it tells us simply how likely this two word configuration is (2.12a). One difference between this calculation and the calculation for an individual word is that this probability is composed of the probability of both w_1 and w_2. For example in (2.10), the word *a* occurs twice as w_1 ($p(a_{w_1}) = 1/4$), but only occurs once with *coatimundi* as w_2 ($p(a + coatimundi) = 1/8$), and once with *javelina* ($p(a + javelina) = 1/8$). Therefore, the probability of *a* is composed of all of its joint probabilities. Notice that this means that the joint probability of a bigram sequence is a subset of $p(w_1)$ and $p(w_2)$, and equivalent to both when two words only ever occur with one another.

(2.12) (a) $p(w_1, w_2) = \frac{C(w_1,w_2)}{\sum_{w_i \in W_1} \sum_{w_j \in W_2} C(w_i,w_j)}$

(b) $p(w_2|w_1) = \frac{\frac{p(w_1,w_2)}{C(w_1)}}{\sum_{w_i \in W_1} C(w_i)}$

(c) $pmi(w_1; w_2) = \log(\frac{p(w_1,w_2)}{\frac{C(w_1)}{\sum_{w_i \in W_1} C(w_i)} \frac{C(w_2)}{\sum_{w_j \in W_2} C(w_j)}})$

In addition to joint probability, information theory provides another way to quantify the co-occurrence of two variables. A commonly used probability measure is conditional probability which quantifies the likelihood that some w_2 occurs given that some w_1 occurs (2.12b). For example, the chance that *saw* occurs in the position ofw_2 is one in all cases where w_1 is *Lucas* because these two words only ever co-occur (p(*saw* | *Lucas*) = 1). However, we can see that this likelihood is more complicated when we look at the example of the word *a*. Given that *a* occurs twice as a w_1, and in each case it has a different w_2, then the conditional probability of *coatimundi* given *a* cannot be one (p(*coatimundi* | *a*) = 0.5). This captures the intuition that you only see *coatimundi* half of the time when *a* is w_1. By calculating conditional probability (2.12b), we can quantify the change in probability for w_2, given the likelihood that w_1 has already occurred. The conditional probability of w_2 given w_1 is zero if these two words never co-occur ($p(w_1, w_2) = 0$), and one when w_2 only ever occurs after w_1 ($p(w_1, w_2) = p(w_1)$).

Last, there is another measure that allows us to quantify the complexity of the relationship between w_1 and w_2. This value is similar to conditional probability, but with the addition that it includes the likelihood for w_2 in the denominator and that it is commonly log transformed. We can calculate the point-wise mutual information

of w_1 and w_2 as in (2.12c). The point-wise mutual information is a way to measure the degree to which knowing w_1 changes $\mathrm{p}(w_2)$, but the value is critically not itself a likelihood, and values for it do not range between zero and one (McGill 1954). Lets again consider the example of *a* in 2.11. The point-wise mutual information for *a* and *coatimundi* is the probability of *a* + *coatimundi* divided by the product of the probability of *a* and *coatimundi* ($\mathrm{pmi}(a;coatimundi) = {}^{p(a+coatimundi)}/_{p(a)p(coatimundi)} = \log(4)$). These words share less mutual information than two words that always co-occur ($\mathrm{pmi}(Lucas;saw) = \log(8)$). These two values both tell us similar facts about the relationship between two random variables, but critically have a few mathematical differences. We will focus on these differences shortly.

These three measures give us a method for quantifying the relationship between the complexity of two words on their own and how it influences the distributions of one another. Given these methods, we can calculate the entropy values for each of these probabilities just as we did for single word sequences. In (2.13), we give the formulas for joint entropy, conditional entropy, and mutual information. These measures relate directly to the properties exhibited by their corresponding probabilities, as discussed above. These measures are an application of the equation in (2.6) to the probabilities in (2.12), weighted by the joint probability of the two variables.

$$\text{(2.13)} \quad \text{(a)} \quad \mathrm{H}(W_1, W_2) = \sum_{w_2 \in W_2} \sum_{w_1 \in W_1} p(w_1, w_2) \log_2 p(w_1, w_2)$$

$$\text{(b)} \quad \mathrm{H}(W_2|W_1) = \sum_{w_2 \in W_2} \sum_{w_1 \in W_1} p(w_1, w_2) \log_2 \frac{p(w_1, w_2)}{p(w_1)}$$

$$\text{(c)} \quad \mathrm{I}(W_1; W_2) = \sum_{w_2 \in W_2} \sum_{w_1 \in W_1} p(w_1, w_2) \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

With these measures, we can calculate the joint entropy (2.14), conditional entropy (2.15), and mutual information (2.16) of the sentence in 2.10.

$$
\begin{aligned}
(2.14)\quad \mathrm{H}(W_1, W_2) &= p(START, Lucas)\log_2 p(START, Lucas) \\
&+ p(Lucas, saw)\log_2 p(Lucas, saw) \\
&+ p(saw, a)\log_2 p(saw, a) \\
&+ p(a, javelina)\log_2 p(a, javelina) \\
&+ p(javelina, and)\log_2 p(javelina, and) \\
&+ p(and, a)\log_2 p(and, a) \\
&+ p(a, coatimundi)\log_2 p(a, coatimundi) \\
&+ p(coatimundi, .)\log_2 p(coatimundi, .) \\
&= \tfrac{1}{8}\log_2\tfrac{1}{8} + \tfrac{1}{8}\log_2\tfrac{1}{8} + \tfrac{1}{8}\log_2\tfrac{1}{8} + \tfrac{1}{8}\log_2\tfrac{1}{8} \\
&+ \tfrac{1}{8}\log_2\tfrac{1}{8} + \tfrac{1}{8}\log_2\tfrac{1}{8} + \tfrac{1}{8}\log_2\tfrac{1}{8} + \tfrac{1}{8}\log_2\tfrac{1}{8} \\
&= 3.00 \text{ bits}
\end{aligned}
$$

$$
\begin{aligned}
(2.15)\quad \mathrm{H}(W_2|W_1) &= p(START, Lucas)\log_2 \tfrac{p(START,Lucas)}{p(START)} \\
&+ p(Lucas, saw)\log_2 \tfrac{p(Lucas,saw)}{p(Lucas)} \\
&+ p(saw, a)\log_2 \tfrac{p(saw,a)}{p(saw)} \\
&+ p(a, javelina)\log_2 \tfrac{p(a,javelina)}{p(a)} \\
&+ p(javelina, and)\log_2 \tfrac{p(javelina,and)}{p(javelina)} \\
&+ p(and, a)\log_2 \tfrac{p(and,a)}{p(and)} \\
&+ p(a, coatimundi)\log_2 \tfrac{p(a,coatimundi)}{p(a)} \\
&+ p(coatimundi, .)\log_2 \tfrac{p(coatimundi,.)}{p(coatimundi)} \\
&= \tfrac{1}{8}\log_2 1 + \tfrac{1}{8}\log_2 1 + \tfrac{1}{8}\log_2 1 + \tfrac{1}{8}\log_2 \tfrac{1}{2} \\
&+ \tfrac{1}{8}\log_2 1 + \tfrac{1}{8}\log_2 1 + \tfrac{1}{8}\log_2 \tfrac{1}{2} + \tfrac{1}{8}\log_2 1 \\
&= 0.25 \text{ bits}
\end{aligned}
$$

$$
\begin{aligned}
(2.16)\quad \mathrm{I}(W_1; W_2) &= p(START, Lucas)\log_2 \tfrac{p(START,Lucas)}{p(START)p(Lucas)} \\
&+ p(Lucas, saw)\log_2 \tfrac{p(Lucas,saw)}{p(Lucas)p(saw)} \\
&+ p(saw, a)\log_2 \tfrac{p(saw,a)}{p(saw)p(a)} \\
&+ p(a, javelina)\log_2 \tfrac{p(a,javelina)}{p(a)p(javelina)} \\
&+ p(javelina, and)\log_2 \tfrac{p(javelina,and)}{p(javelina)p(and)} \\
&+ p(and, a)\log_2 \tfrac{p(and,a)}{p(and)p(a)} \\
&+ p(a, coatimundi)\log_2 \tfrac{p(a,coatimundi)}{p(a)p(coatimundi)} \\
&+ p(coatimundi, .)\log_2 \tfrac{p(coatimundi,.)}{p(coatimundi)p(.)} \\
&= \tfrac{1}{8}\log_2 8 + \tfrac{1}{8}\log_2 8 + \tfrac{1}{8}\log_2 8 + \tfrac{1}{8}\log_2 4 \\
&+ \tfrac{1}{8}\log_2 8 + \tfrac{1}{8}\log_2 8 + \tfrac{1}{8}\log_2 4 + \tfrac{1}{8}\log_2 8 \\
&= 2.30 \text{ bits}
\end{aligned}
$$

So far, we've given a description of three tools that will allow us to measure the information relations exhibited between two variables. However, we have done nothing to relate this to gender in natural language, or to our models of nominal classification. Here, we will briefly describe the application of these measures to our corpora, and in doing so describe our predictions.

2.3.2 Regular Language Models and Gender

In the previous subsection, we introduced a method to quantifying the complexity of a single random variable using the notion of entropy. We also introduced three ways of quantifying the complexity of the relationships between two random variables (joint entropy, conditional entropy, and mutual information). In this subsection we discuss how we can apply these measures to investigate how nominal classification impacts the complexity of nominals in natural languages. To do this, we first describe regular language models of English, French and German. We then show how we can evaluate the impact of gender on these models using the complexity calculations. We will see that both conditional entropy and mutual information will allow us to calculate the impact of gender on nominals, but joint entropy fails to be informative.

In order to calculate complexity in language, we need to create language models of gendered languages. One of the underlying assumptions in information theory is the idea that a natural language can be modeled by what is called a regular language (Shannon 1951). Essentially, a regular language is a finite formal language that can be represented by a regular expression or can be read by a finite state machine(Kleene 1951). This quality, among other things entails that such a model

is finite, even if the language it produces is not. Furthermore, such a language must adhere to a context dependent regular grammar (Chomsky 1956; Chomsky 1959). In a finite state machine which reads a regular language, the machine encodes the units of a language into nodes called automata, and represents multi-word collocations as transitions between these automata (Kleene 1951).

Here, we create regular language models of English, French and German that are able to read all two word sequences in each language. Commonly referred to as a bigram language model, this mechanism will allow us to see that conditional entropy and mutual information can inform the impact of gender marking on a nominal system.

In (2.17), we give an idealized example of bigrams in German, French, and English. This language is idealized because it (i) contains only determiner-noun pairs, and (ii) these pairs are evenly split across the number of nominal classes found in each language.

(2.17)

German	**French**	**English**
der Hund	le chien	the dog
der Drache	le dragon	the dragon
die Brücke	le pont	the bridge
die Frau	la femme	the woman
das Haus	la maison	the house
das Fenster	la fenêtre	the window

In these toy languages, nominals are divided across genders in order to highlight the relationship of gender division on nominals. The crucial difference between these

three languages is the fact that they each contain a different number of gender labels. Whereas English contains no grammatical gender distinction, French and German contain two and three gender divisions respectively. For German, there are three different articles of the definite determiner (der, die & das), in French two (le & la), and in English one (the)[1].

Here, we present finite state machines the describe each of these toy languages. In these models, each word is a state, and a transition (the line between states) indicate a bigram collocation. The *START* state is the starting point for any bigram collocation. We include this state to allow the model to show the likelihood of the first word occurring in any bigram sequence.

In these machines, the transitional probabilities between states are the likelihood of the next word occurring in sequence. For example in English, the likelihood that a bigram in the toy language starts with the word *the* is 1, because all words start with *the* in this language. This means that there is only one option for w_1. Next we see that the word *the* has six transitions to six different states. Each of these states represents a possible w_2. This is indicates that at this point there is a 1/6 chance of each individual word occurring in sequence.

In these machines, we can see the difference between gendered and non-gender marked languages. Notice that the first layer of states in each machine is larger when there are more gender labels. In gendered languages there is an increased number of states for w_1. For example German has three states in the first layer, French two, and English one. However, this effect is offset by the fact that once we have seen

[1]Again, we are simplifying here by ignoring case and phonologically conditioned alternations

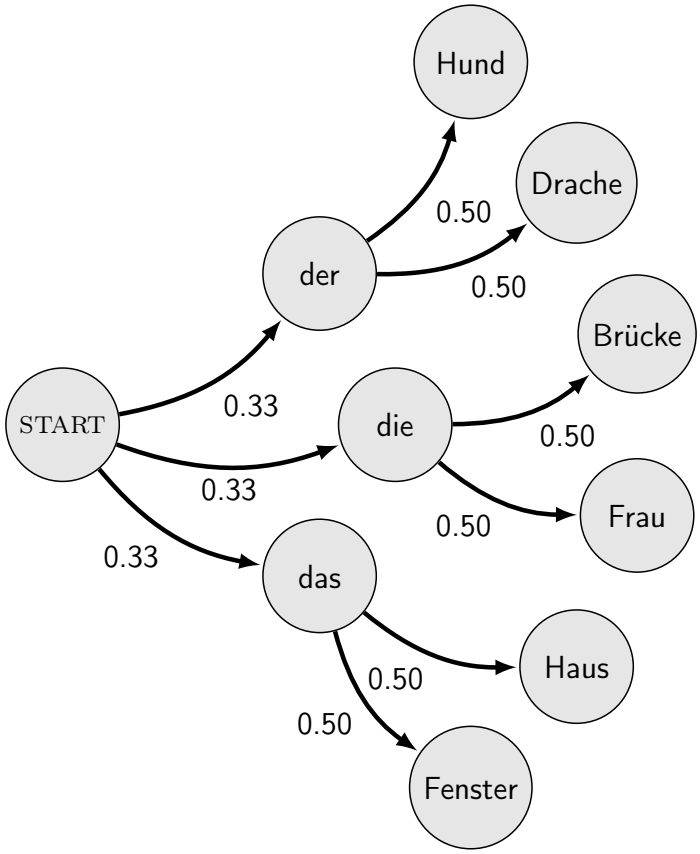

Figure 2.4: This figure presents a directed graph visualization of a Finite State Machine. This machine depicts a Regular Language that sufficiently describes all German Bigrams of our toy language.

w_1, the nominals in the second layer (w_2) have higher transition probabilities when a language has more gender divisions. In short, gender decreases predictability in the first layer, but increases it in the second layer as a function of the number of gender divisions in a language.

Now that we have these machines, we can see the effect of gender each language by calculating the measures described in 2.13. For each machine, we calculate joint entropy (H(W_1,W_2), conditional entropy (H($W_2|W_1$), and mutual information(

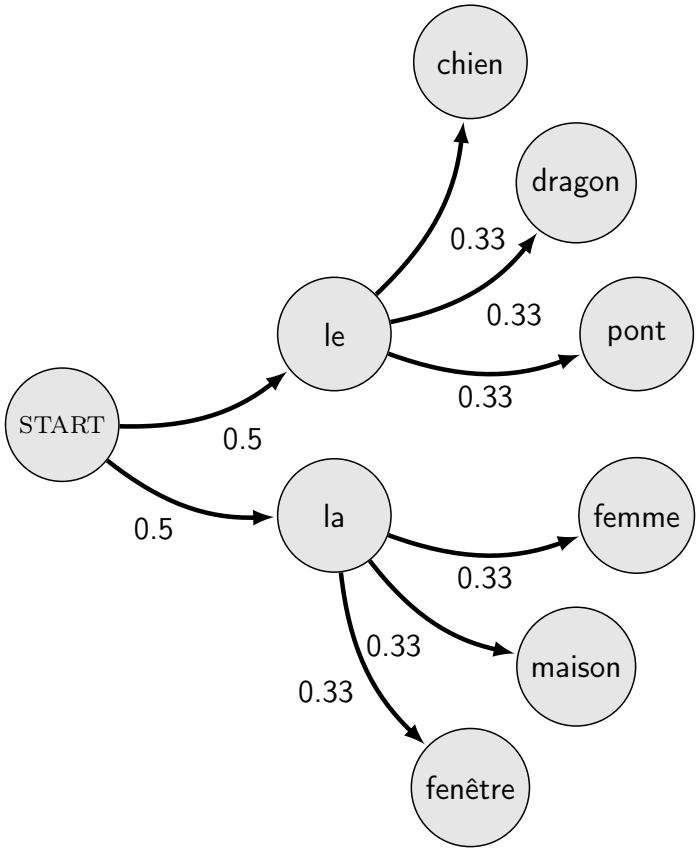

Figure 2.5: This figure depicts the Finite State Machine describing the French toy language.

I(W_1;W_2)). The results of these calculations are shown in 2.18. In addition to these calculations, we include entropy calculations for W_1 and W_2. These values reveal how each individual variable impacts the multi-word complexity calculations.

(2.18)

	Classes	$H(W_1)$	$H(W_2)$	$H(W_1, W_2)$	$H(W_2\|W_1)$	$I(W_1; W_2)$
German	3	1.58	2.58	2.58	1.00	1.58
French	2	1.00	2.58	2.58	1.58	1.0
English	1	0.00	2.58	2.58	2.58	0.0

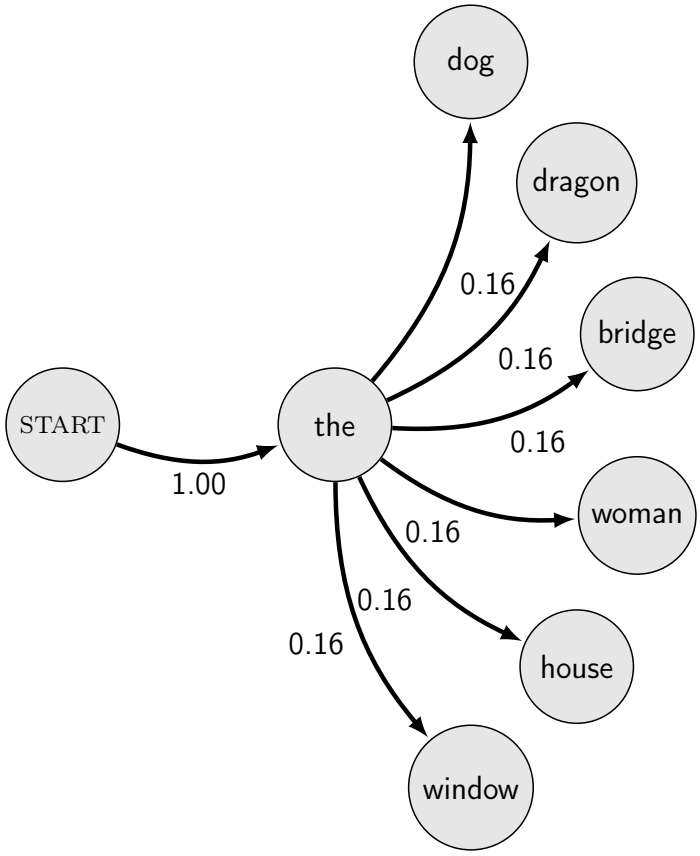

Figure 2.6: This figure depicts the Finite State Machine describing the English toy language

The results reveal a few interesting facts. First, we can see that joint entropy is the same for all languages (2.58 bits), and that this value is equivalent to the entropy of W_2. Meanwhile, the entropy calculation for W_1 increases in direct proportion to the number of nominal classes. This result is evidence that gender marking increases the complexity in W_1, as stated previously.

For conditional entropy, we see different values across the three languages. This variation can be attributed to the change in entropy values for W_1 across the lan-

guages. The result is that conditional entropy is greater in the non gendered system of English (2.58) than in French (1.58), which is yet higher than German (1.00). This has to do with the fact that gendered determiners narrow the options for w_2 to fewer options.

Likewise, we see variation in the mutual information calculation relative to the number of gender distinctions. However, mutual information grows in direct proportion to the number of genders found in each language. This means that for each language, W_1 and W_2 share an increasing amount of information between one another as a function of the number of grammatical distinctions in the language.

From these facts we can draw a few different conclusions. First, we know that the measure of joint entropy is constant across our three languages, and therefore doesn't reveal much about the impact of gender on nominals. Second, the conditional entropy and mutual information measures both expose differences in each language, so we are encouraged to adopt both of them. However, each measure has a few implications.

Recall that conditional entropy allows us to capture how predictable some word is (W_2) given the preceding word (W_1). Note that the likelihood of W_2 given W_1 is not equivalent to the likelihood of W_1 given W_2 in our bigram language. This is because we encode bigrams as a function of their order. This encoding then requires that the conditional probability calculation works from left to right in time. In this way, the calculation is asymmetric.

Mutual information however, does not have the same mathematical properties as conditional entropy. First, the term used to calculate the value is not itself a

probability, but rather a measure between different probabilities. Also, notice that the mutual information values in our model can be calculated by subtracting the marginal entropy of either word from the conditional entropy of that word[2]. Second, mutual information is equal to the marginal entropies of w_1 and w_2 minus their joint entropy[3]. The critical point that we can distill from these facts is that mutual information does not have the same rightward directionality that conditional entropy has. Rather, by including the overall, or marginal probability of w_2 in the denominator of the term, the measure bi-passes this directionality. Therefore, $I(w_1; w_2) = I(w_2; w_1)$ and $H(w_2|w_1) \neq H(w_1|w_2)$.

An original requirement in the development of information theory was the notion of symmetry (Shannon 1951). The entropy calculation of a model should not inherently be determined by the order of events in the system generally. However, when we calculate conditional entropy in sequence (i.e. W_1 and W_2), we create an asymmetric calculation. On the other hand, mutual information retains the symmetry found more generally in the theory.

In a more general sense, the addition of the value for $p(w_2)$ means that the point-wise mutual information term is simply a weighted version of the conditional probability of w_2 given w_1 as seen in 2.19. This indicates that for mutual information the measure of surprisal is relative to how surprising w_2 is generally, and not simply in terms of how surprising it is in relation to w_1, as is the case for conditional entropy.

[2]Here, $H(w_1|w_2) = 0$, because no w_2 (nominal) precedes a w_1 (determiner)
[3]$I(w_1; w_2) = (H(w_1) + H(w_2)) - H(w_1, w_2) = H(w_2) - H(w_2|w_1) = H(w_1) - H(w_1|w_2)$

$$
\begin{aligned}
pmi(w_1; w_2) &:= \frac{p(w_1, w_2)}{p(w_1)p(w_2)} = \frac{p(w_2|w_1)}{p(w_2)} \\
p(w_2|w_1) &:= \frac{p(w_1, w_2)}{p(w_1)} = \frac{p(w_2|w_1)}{1}
\end{aligned}
\tag{2.19}
$$

In short, conditional entropy tells us whether a word is more likely based on whether it follows another word, and mutual information tells us whether a word is more likely given some other word and given its own likelihood. Another way to say this is that mutual information tell us the degree to which the distribution of two independent words are reliant upon their co-occurrence. Conditional entropy does not make reference the size of the likelihood of the predicted word. On the other hand, measuring mutual information tells us, given the probability of some word, whether the likelihood for that word is made easier.

Here we have given a description of the mathematical properties of the conditional entropy and mutual information calculations. Next, we consider their application to language models and relate the effects of these measures to our models. We propose that if nominal classification reduces the entropy of nominals in a language, then this effect should be measurable in a corpus.

This position seems appealing if we consider our toy language above, but is problematic when we consider comparing two actual languages. First, even if we control for size and source of the document, every language uses different terms for lexical categories, and the part of speech tagging varies from language to language (Voutilainen 2003). Furthermore, languages employ different grammatical devices when it comes to issues like the orthographic expression of nominal compounding in German and English (Baroni *et al.* 2002). Whereas German tends to have compounds

represented as a single orthographic word, English would more likely treat it as a multi-word noun phrase. Therefore, we cannot directly compare entropy calculations between language models of the same size, but rather we employ an alternative method.

To get around these issues, we measure the complexity of a language, and then transform that language so that all gender-inflected variants of determiners are collapsed into a single category. We then measure the complexity of these de-gendered languages and compare them to their original gendered counterparts.

Take our toy languages above for example, this would effectively mean collapsing the layer of states in w_1 in German from three to one. Meanwhile, this transformation would maintain the probability distributions of nominals from the original language. If gender functionally reduces nominal entropy in language, then we would predict that a de-gendered language will have greater complexity than its naturally occurring counterpart. Furthermore, we would then expect for our experiment to have a stepwise effect. Specifically, the change in information and entropy between a gendered language and its de-gendered counterpart should change as a function of the number of classes present in a language. Therefore, we would expect the difference between the measures of the German models to be greatest (with three distinctions), then French next (with two distinctions), and finally for there to be no real difference for English (with no distinctions).

In this Section, we have introduced ways of measuring complexity in information theory, and then given examples of these measures in toy languages. We then described what the different can tell of about nominal classification and complexity.

In the next Section, we present two corpus studies where we de-gender the first 100 million words of the English, French, and German WaCky corpora (M. Baroni & Zanchetta 2009). These corpora contain samples of online news and media, are part of speech tagged, and contain lemmatized forms of all tokens in addition to the original token.

The first study investigates the claim that nominal classification reduces nominal complexity. We calculate this by measuring conditional entropy and mutual information in all determiner-noun sequences in each language. Here, we also provide evidence from adjective noun sequences, which also inflect for gender.

The second study explores the idea that this complexity reduction is meaningful in terms of the overall language. To do this, we calculate these same complexity measures on the full language models, and not only limited to specific lexical categories. Based upon the results, we discuss implications for possible models of gender.

The first study will tell us whether we can adopt a model of extra-grammatical pressure towards language generally, and the second study will help us investigate whether we should prefer the association or discrimination models.

2.4 Studies

In this Section we present corpus studies of each language at the determiner-noun level, and then present a study that investigates the all bigrams of each language. For each study, we measure conditional entropy and mutual information. We see that both measures reveal a reduction in nominal complexity in the first study. However,

this effect does not hold for both measures at the full language level. Based upon this, we then do some follow up experiments aimed at understanding why both measures do not succeed. These results lead us to argue for the association model.

2.4.1 Study One: Gendered Modifiers and Nominals

We created a bigram language model of all determiner-noun and adjective-noun sequences for German, French, and English based upon the first 100 million words of the WacKy corpus. We then created the same language models of de-gendered forms of each language. Given these models, we calculated conditional entropy and mutual information of all two word sequences. In order to test significance of the values, we ran each calculation over 100 subsamples of 100 million words in each corpus. Based upon these results, we calculate Cohen's D for effect size (Cohen 1992), and t-tests to see if all comparisons between the measures in each language are significantly different from one another. In addition to these calculations, we give Shapiro-Wilk tests which evaluate the normality of the distribution of all subsamples (Shapiro & Wilk 1965). This along with its respective p-value tells us whether the t-tests are merited.

First, we give the results for conditional probability in Figure 2.7. The corresponding significance tests are found in Table 1.1 and Table 1.2.

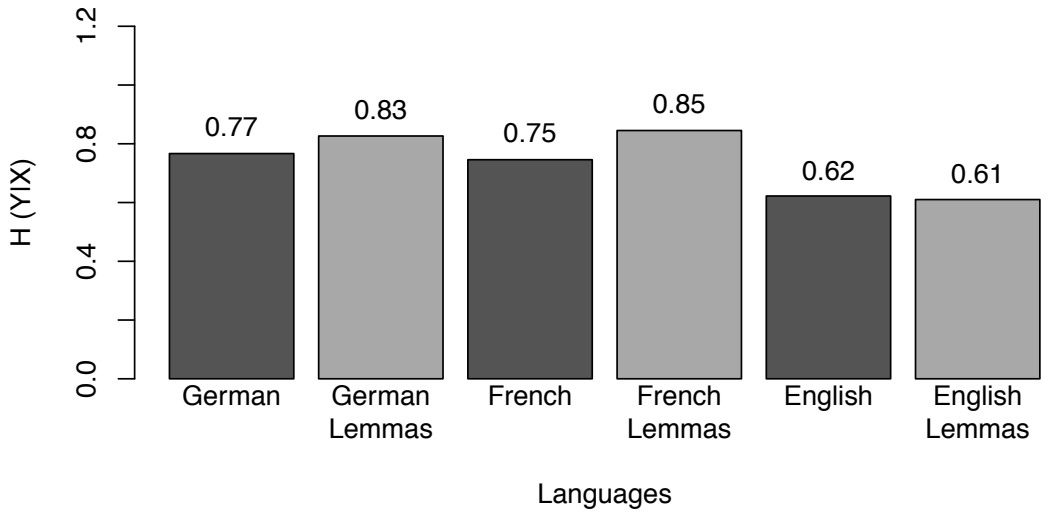

Figure 2.7: This bar plot gives the conditional entropy Calculation (H(Y|X)) for each language for all bigrams tagged as Determiner-Noun sequences. Overall probabilities are calculated with respect to the corpus overall, and not with respect to bigrams tagged as Determiner Noun. For each language, there is a value for its natural and lemmatized (de-gendered) form. Notice that gendered languages increase in conditional entropy when gender is removed, but English remains the same.

The results for conditional entropy in determiner noun sequences reveal two things. First, we can see that in both German and French, conditional entropy increases when gender is removed in these instances. For English, lemmatization in these sequences reduces conditional entropy, and does not increase as is the case in gendered languages. This means that gender marking reduces the complexity of nominals in these sequences. Second, we can see that the effect is not stepwise as we had predicted. German has an increase of 0.6 bits (0.83 - 0.77), and French has an increase of 0.1 bits (0.85-0.75). So, French (2 classes) had a greater change than

German (3 classes). We hold discussion of this effect to the next section.

The significance tests reveal that all comparisons are significant (t-tests), and have a large effect size (Cohen's D). This data is found in Table 1.1. Furthermore, the entropy calculations come from sufficiently normal distributions as seen in Table 1.2. A lack of significant p-values for Shapiro-Wilk shows that the distribution of conditional entropy in these languages is not significantly non-normal. The results is that the t-tests found in Table 1.1, and therefore our comparisons between languages are valid.

	German	German Lemmas	French	French Lemmas	English	English Lemmas
German		-104.7 ****	44.91 ****	-161.8 ****	309.2 ****	337.4 ****
German Lemmas	-15.03 ****		166.4 ****	-36.92 ****	419.4 ****	447.1 ****
French	6.449 ****	23.9 ****		-262.2 ****	347.3 ****	385.9 ****
French Lemmas	-23.23 ****	-5.301 ****	-37.64 ****		584 ****	622.1 ****
English	44.4 ****	60.22 ****	49.87 ****	83.86 ****		34.08 ****
English Lemmas	48.45 ****	64.19 ****	55.41 ****	89.34 ****	4.893 ****	

Table 2.1: Independent t-tests and Cohens-D effect size calculations for all comparisons, with significance levels. The t-tests are located in the upper-right half, and the Cohens-D values in lower-left half. For Cohens-D, four stars indicate a large effect size. All comparisons are significant.

	German	German Lemmas	French	French Lemmas	English	English Lemmas
Shapiro-Wilk	0.99	0.99	0.99	0.99	0.99	0.99
P-Value	0.38	0.41	0.45	0.60	0.68	0.65

Table 2.2: Shapiro-Wilk values for each language along with the corresponding p-value. These p-values are all not significant, and therefore all of the conditional probability calculations for these languages compose normal distribution.

In addition to conditional entropy, we calculated the mutual information values for these language models at the determiner-noun level. Figure 2.8 gives these values. Here, the effect is as have predicted, and is in line with the conditional entropy values. Recall that when mutual information is greater, there is more information shared between the two variables. Therefore, the amount information shared between w_1 and w_2 is greater in the naturally occurring forms of German and French than in their de-gendered counterparts. However, for English, there is no change. Effectively, in all cases lemmatization reduced the information shared between determiners and nominals in gendered language but did not alter the information shared between them in English. Again, we see a greater difference between natural and lemmatized French models than in the respective German models (0.13 versus 0.9 bits). Overall, these data corroborate the idea that in these sequences, Gender marking reduces nominal entropy.

Furthermore, in Table 1.3 and Table 1.4, we see the corresponding significance tests.

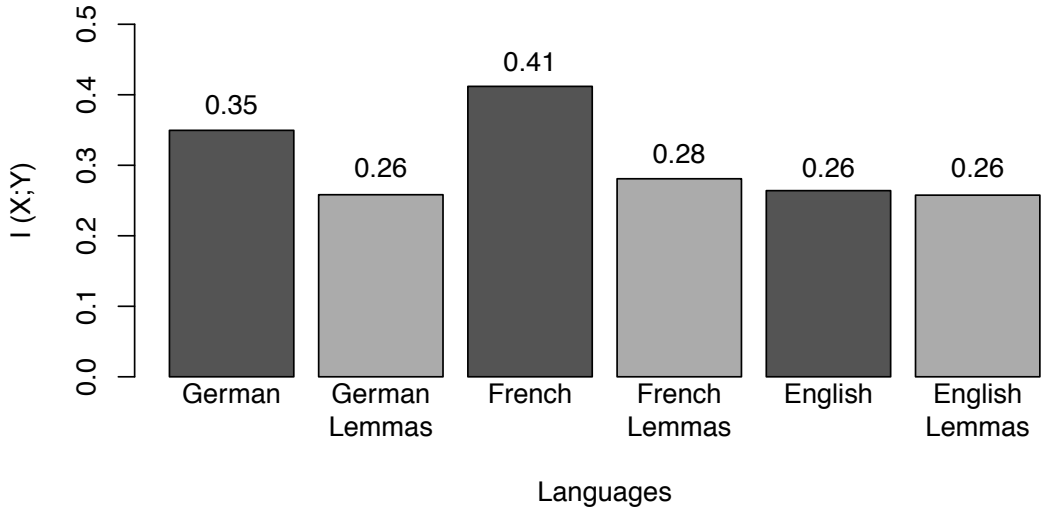

Figure 2.8: Mutual information calculation I(X;Y) for all languages for all bigrams tagged as determiner-noun. Probabilities are calculated with respect to the overall corpus.

	German	German Lemmas	French	French Lemmas	English	English Lemmas
German		603.6 ****	-372.7 ****	486.7 ****	592.1 ****	641 ****
German Lemmas	86.67 ****		-1123 ****	-220.2 ****	-54.01 ****	5.802 ****
French	-53.51 ****	-161.3 ****		1045 ****	1145 ****	1205 ****
French Lemmas	69.89 ****	-31.61 ****	150 ****		182.1 ****	255.9 ****
English	85.02 ****	-7.755 ****	164.4 ****	26.15 ****		66.82 ****
English Lemmas	92.04 ****	0.8331 ****	173 ****	36.74 ****	9.595 ****	

Table 2.3: Independent t-tests and Cohens-D effect size calculations for all comparisons, with significance levels. The t-tests are located in the upper-right half, and the Cohens-D values in lower-right half. For Cohens-D, four stars indicate a large effect. All comparisons are significant.

	German	German Lemmas	French	French Lemmas	English	English Lemmas
Shapiro-Wilk	0.99	0.99	0.98	0.99	0.99	0.99
P-Value	0.40	0.56	0.23	0.38	0.37	0.53

Table 2.4: Shapiro-Wilk values for each language along with the corresponding p-value. These p-values are all not significant, and therefore all of the conditional probability calculations for these languages compose normal distribution.

In this study, we saw that inflectional marking of nominal classification reduces the complexity of nominal systems in a these bigram models. For languages with gender, the removal of gender results in greater conditional entropy values and lower mutual information values in determiner-nominal sequences. This effect does not hold in English, unsurprisingly. Based upon these results, we have motivation for the claim that in some way nominal classification has the power to reduce the amount of information in the nominal system. In the next study, we analyze whether this effect holds for a language in total by measuring these calculations for all bigram sequences.

2.4.2 Study Two: Full Language Models

In this subsection we describe measurements of conditional entropy and mutual information for the full language models. The motivation for this is as follows. If nominal classification is powerful enough to mitigate nominal complexity, and nominals are the lexical category with the greatest amount of complexity, then we would expect the effect to be larger enough to be seen at the overall language level. This means that if we were to take all bigrams and de-gendered them, then we would predict that for each measure we would see the same effect as we saw at the determiner noun level. Here, we perform the same test as above, but with bigrams from all possible w_1 and w_2 combinations and ignore part of speech categories.

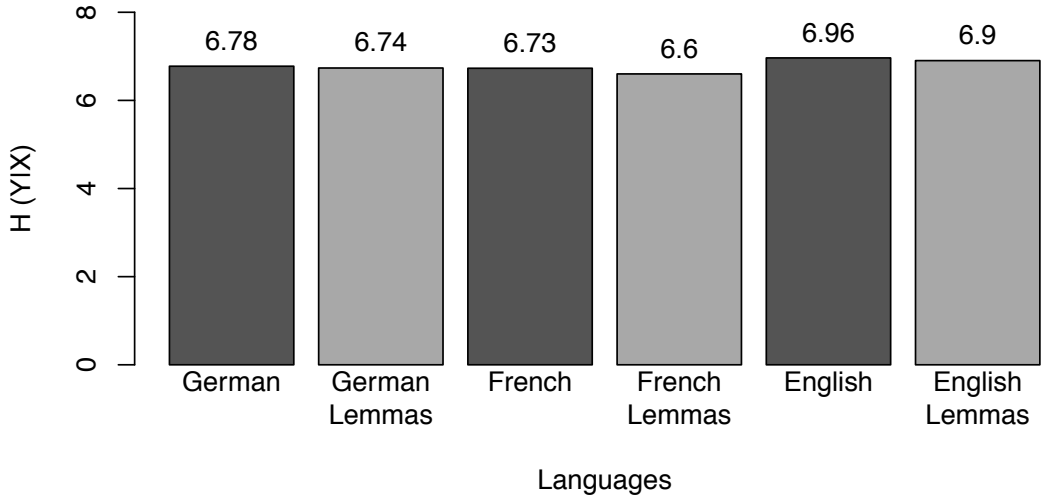

Figure 2.9: The total conditional entropy calculation for the conditional probability of all bigram sequences in a 100-million-word sample from the WaCky Corpus for German, French, and English

Figure 2.9 depicts the conditional entropy values for bigram language models and their de-gendered forms in all the first 100 million words in each corpus. We see that in all cases the conditional entropy value is slightly reduced in the de-gendered models relative to the models of the naturally occurring languages. This fact is also true for English. This result differs from what we saw in the previous study, where conditional entropy increased as a result of removing gender distinctions in gendered languages.

Tables 1.5 and 1.6 give the significance values for all comparisons. For English, the Shapiro-Wilks tests are significant for both the original and de-gendered corpora. This means that these comparisons may be a bit more tentative. Otherwise, all

comparisons are significant, have a large effect size, and are valid given the normality of distribution.

	German	German Lemmas	French	French Lemmas	English	English Lemmas
German		73.35 ****	78.3 ****	285.1 ****	-285.4 ****	-189.8 ****
German Lemmas	10.53 ****		9.644 ****	218.6 ****	-346.3 ****	-250 ****
French	11.24 ****	1.385 ****		198.2 ****	-338.5 ****	-246.9 ****
French Lemmas	40.94 ****	31.38 ****	28.46 ****		-510.1 ****	-418.9 ****
English	-40.98 ****	-49.73 ****	-48.6 ****	-73.25 ****		80.5 ****
English Lemmas	-27.25 ****	-35.9 ****	-35.45 ****	-60.16 ****	11.56 ****	

Table 2.5: Independent t-tests and Cohens-D effect size calculations for all comparisons, with significance levels. The t-tests are located in the upper-right half, and the Cohens-D values in lower-right half. For Cohens-D, four stars indicate a large effect. All comparisons are significant.

	German	German Lemmas	French	French Lemmas	English	English Lemmas
Shapiro-Wilk	0.99	0.99	0.99	0.98	0.91	0.91
P-Value	0.95	0.58	0.36	0.23	0.00	0.00

Table 2.6: The value and significance levels of Shapiro-Wilks tests on the mutual information calculations. For English, the results are significant and therefore these data may not be normally distributed.

Moving on, we now present the mutual information values for these language models. Here, we see different effect than in the conditional entropy data. Figure 2.10 reports these calculations for all bigrams.

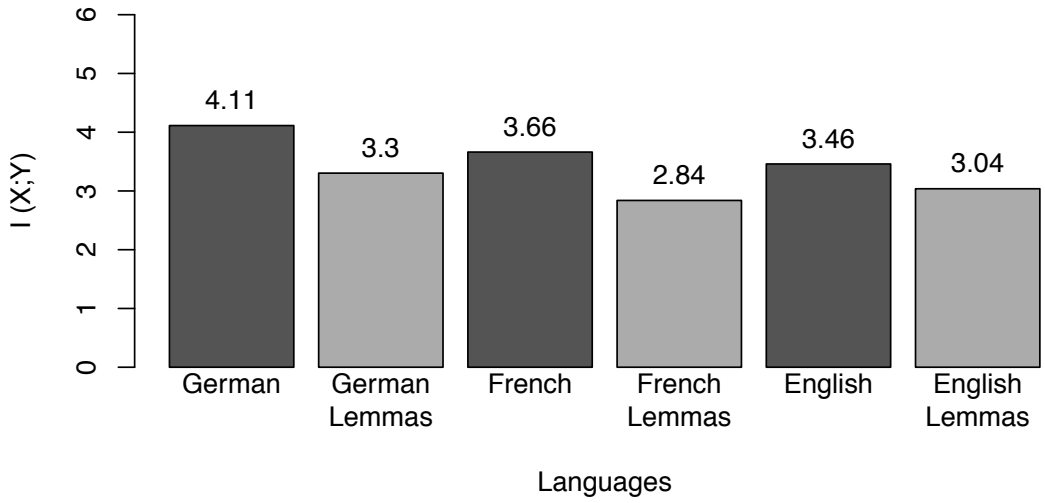

Figure 2.10: Mutual information calculation I(X;Y) for all languages for all bigrams. Probabilities are calculated with respect to the overall corpus.

In these models, mutual information decreases in all cases when the languages are lemmatized. German decreases by .81 bits, French by 0.82, and English by 0.42. This means that the gendered language values for mutual information decrease twice as much as the non-gendered language. Also, notice that the mutual information values of the original languages is directly proportional to the number of genders found in each language. German (three) has a value of 4.11, French (two) has a value 3.66, and last English (one) has a value of 3.04.

Tables 1.7 and 1.8 give the significance values for all comparisons. All tests are significant, effect sizes large, and the distributions are sufficiently normal.

	German	German Lemmas	French	French Lemmas	English	English Lemmas
German		839.5 ****	487.5 ****	1436 ****	661.1 ****	1116 ****
German Lemmas	120.5 ****		-401.2 ****	544.6 ****	-162.5 ****	285.3 ****
French	70 ****	-57.61 ****		1017 ****	220.3 ****	699.6 ****
French Lemmas	206.3 ****	78.19 ****	146 ****		-704.4 ****	-232.9 ****
English	94.93 ****	-23.34 ****	31.63 ****	-101.1 ****		440.4 ****
English Lemmas	160.3 ****	40.97 ****	100.5 ****	-33.44 ****	63.24 ****	

Table 2.7: Independent t-tests and Cohens-D effect size calculations for all comparisons, with significance levels. The t-tests are located in the upper-right half, and the Cohens-D values in lower-right half. For Cohens-D, four stars indicate a large effect. All comparisons are significant.

	German	German Lemmas	French	French Lemmas	English	English Lemmas
Shapiro-Wilk	0.98	1.00	0.99	0.99	0.98	0.99
P-Value	0.23	0.98	0.66	0.71	0.32	0.83

Table 2.8: Shapiro Wilks tests for the mutual information calculations of all languages. No values are significant, and are therefore sufficiently normally distributed.

In short, we see that conditional entropy calculations reveal no difference in gendered and non-gendered languages for all bigrams. We also see that mutual information does reveal an effect of gender in each language, and that this effect follows the pattern of entropy reduction that we predicted. This was not the case for conditional entropy in all languages for which entropy decreased.

In the next subsection we investigate the reasons for this difference by measuring the effect of corpus size on each calculation. We will find that the conditional entropy measures between W_1 and W_2 is inherently sensitive to the size of the lexicon, and mutual information is not.

2.4.3 Corpus Size and Entropy

In the previous subsection we saw that conditional entropy and mutual information measures showed different effects at the language level. For conditional entropy, all languages slightly decreased in complexity when they were lemmatized. This means that the complexity reduction from the determiner-noun study is effectively masked for this measure in the overall language. We interpret this fact as being a consequence of the low probabilities of the words in the corpus. As the corpus size increase, the average probabilities of W_1 and W_2 decrease. When these likelihoods are quite low the change in likelihood as measure by conditional entropy is insignificant relative to the overall probabilities. Another interpretation is that the reduction of complexity in the nominal domain may be offset by the complexity of the words which precede them.

At the same time, we saw that mutual information revealed an effect of gender

between W_1 and W_2 at the language level. When lemmatized, information shared between words decreases as a function whether gender exists in a language. This effect was in line with the effect found at the determiner-noun level.

A preliminary assertion we can make is that according to these data, the gendered language models share more information between words in accordance to the number of gender distinctions, and that this effect goes away when gender is collapsed. Furthermore, this effect is relative to the overall language, and not just for a limited number of sequences. Without making reference to part of speech, we can see this effect for mutual information but not conditional entropy.

Now that we've established the trends, let's look to understand how there can be an effect for mutual information but not for conditional entropy. Recall that conditional entropy makes no reference to the overall likelihood of the word being predicted, but only its predictability relative to W_1. The mutual information measure on the other hand, adjusts for this probability giving us a measure of complexity that is normalized by the overall probability of W_2 . Here, we investigate the effect of corpus size on both complexity measures in natural and de-gendered versions of each language. We do this by calculating conditional entropy and mutual information for W_1 and W_2 at increasingly large subsamples of the corpus.

First, we show how the conditional entropy measure is affected by corpus size. Figure 2.11 gives the conditional entropy values for natural and lemmatized subsamples of each language. We log transform the corpus size in order to focus on the space of effect.

Notice that as corpus size increases, the conditional probability value for both

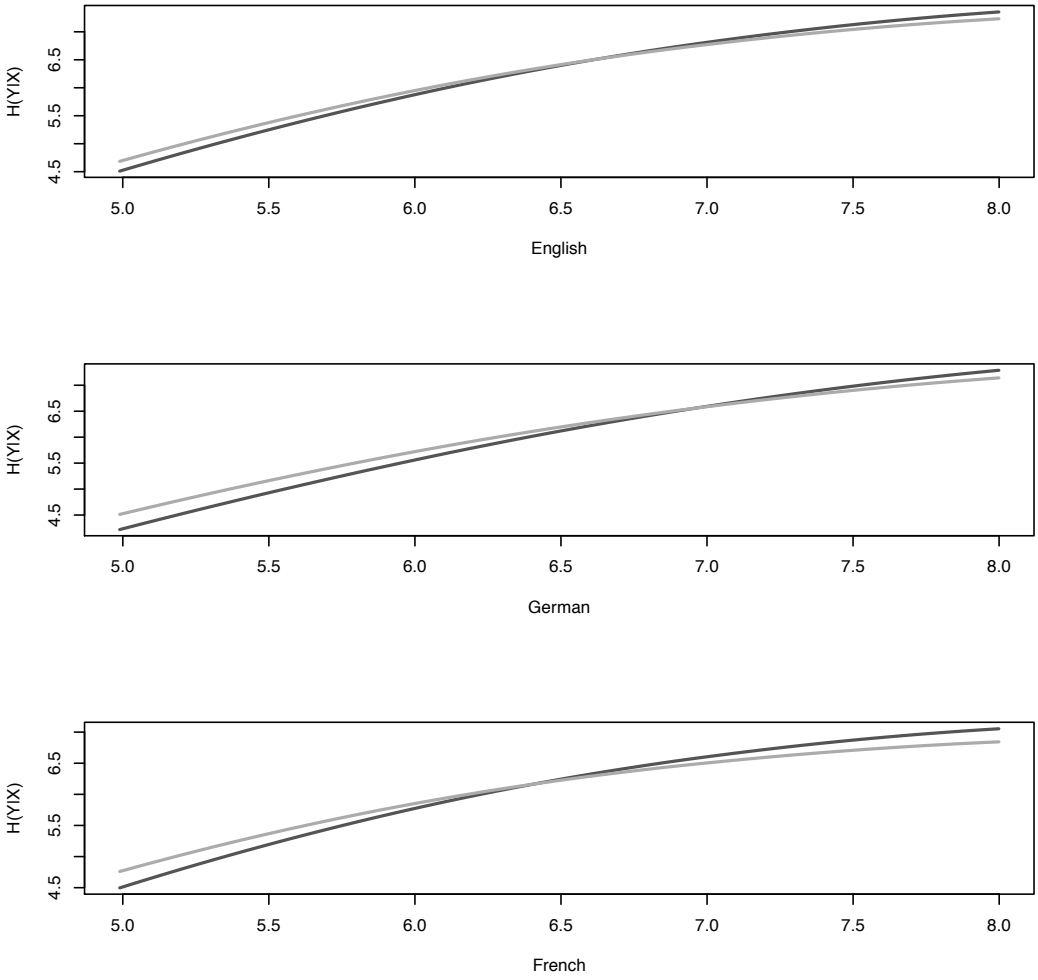

Figure 2.11: This figure illustrates the relationship between corpus size and the Conditional Entropy calculation for the lemmatized (grey) and naturally occurring (black) forms of the corpus. Corpus size is measured in log space ranging from 100,000 token subsamples to 100 million token subsamples. The curves are derived using Lowess Smoothing across eleven sample calculations.

languages increase in all languages in an arc. However, there is a difference in the rate of change in the conditional entropy measure for natural and lemmatized forms in all languages. Initially, the lemmatized form has greater conditional entropy, but as the corpus size increases, the natural form has increasingly greater conditional entropy. This difference continues for all observable corpus sizes. In effect, there is an interaction of corpus size and the degree to which nominal classification reduces nominal complexity. At smaller subsamples, inflection decreases complexity, but at larger samples it increases complexity. This interaction is problematic for a model of complexity reduction that employs this measure.

Next, we do the corresponding mutual information calculations on increasingly large subsections of the corpus. The results can be found in Figure 2.12.

In this Figure, we see that as corpus size grows the mutual information shared between W_1 and W_2 diminishes. This change reflects the lowering of probabilities of the individual words in the language as the vocabulary grows. Unlike conditional entropy, the mutual information calculations of natural and lemmatized forms of each language decrease at identical rates. In corpora of all sizes, the mutual information shared between W_1 and W_2 is greater in naturally occurring versions of each language than in the lemmatized form. Here, there is no interaction between corpus size and complexity.

In this Section we have given studies that evaluate the relationship between W_1 and W_2 in naturally occurring and lemmatized forms of each corpus. First, we saw that gender decreases the conditional entropy and increases the mutual information of nominals in determiner-noun sequences in gendered languages. In English, there

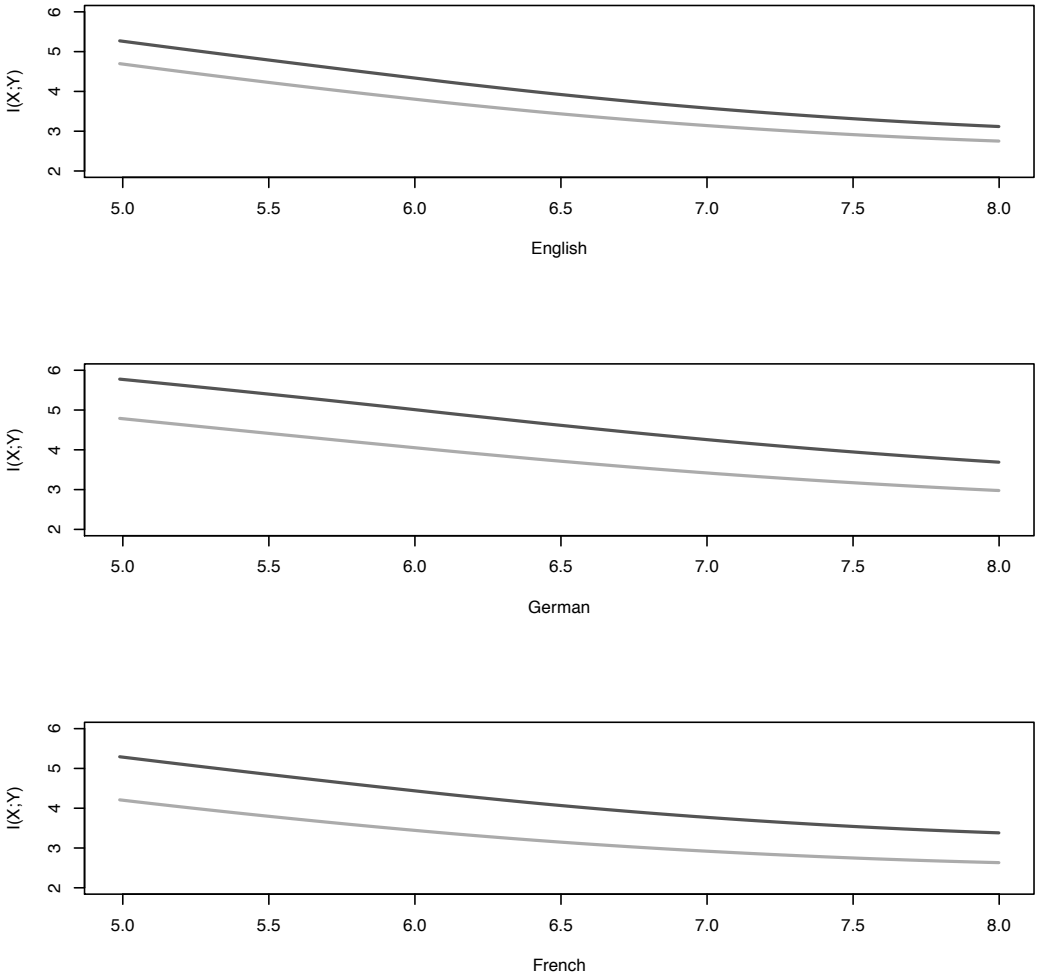

Figure 2.12: This figure illustrates the relationship between corpus size and the mutual information calculation for the lemmatized (gray) and naturally occurring (black) forms of the corpus. Corpus size is measured in log space ranging from 100,000 token subsamples to 100 million token subsamples. The curves are derived using Lowess Smoothing across eleven sample calculation.

was little or no change. When we evaluated the overall language models, we saw that conditional entropy did not reveal an effect of gender. However, mutual information did reveal such an effect. In addition to this, we saw that there was an interaction between corpus size and conditional entropy calculations, but not for mutual information calculations. In the next Section, we discuss these results as they relate to our original hypothesis, and discuss why we should adapt the associative model.

2.5 Discussion

In this Section, we discuss the results of our language model studies. Based upon these results, we give argument for adopting the association model.

2.5.1 Results and Model Selection

Recall that our initial hypothesis claimed that in order for nominal classification to persist in the grammar of a language, there must be some extra-grammatical pressure towards retaining it. We suggested that if classification somehow reduced nominal complexity, then we have a plausible source for this pressure. However, we saw that the mechanism for this pressure could come from two different sources.

We presented the discrimination perspective which argues that gender in German reduces nominal complexity in discourse. This perspective entailed that gender marking precedes the nominals that they modify. Furthermore, we suggested that this effect should be relevant for limited contexts. This was because under the discrimination view, other contextual cues would also narrow the search space of up

upcoming nominals in discourse. Therefore, only contextually relevant nominals would be considered. In short, this model predicts that gender marking reduces nominal complexity asymmetrically, and in narrow contexts.

Meanwhile, we argued for the position that nominal classification reduces the complexity of nominals in the mental lexicon. Given a list of nominals, gender labels would effectually give a label to these nominals that would aid in memorization via association. Given their connection, grammatical gender could be useful at a high level. Under this view, the reduction of nominal complexity would not necessarily be asymmetric. This is because such labels would not precede or follow a nominal in the mind, but would simply be an association. Rather, the relation would be symmetric. Furthermore, this effect would not be limited only to certain contexts, but would have to be a global effect.

Now that we have clear predictions, let's review the results of our studies. First, we saw that gender decreases the conditional entropy and increases the mutual information of nominals in determiner-noun sequences in gendered languages. In English, there was little or no change. When we evaluated the overall language models, we saw that conditional entropy did not reveal an effect of gender. However, mutual information did reveal such an effect. In addition to this, we saw that there was an interaction between corpus size and conditional entropy calculations, but not for mutual information calculations.

Let's now consider the two models as they relate to these results. The discrimination model matches the results of the conditional probability measures. Recall that this measure is inherently asymmetric. This is to say that this measure requires

directionality. We saw that in German, and French the determiners which preceded nominals did lower the complexity of those nominals in sequence. Furthermore, this model would predict that this effect works in narrow context. The language level effects are also in line with this assumption. While conditional entropy reveals only an effect at a these nominal-determiner junctions, it must not necessarily affect the whole lexicon overall to hold true.

The association model, however is enticing given the data. We saw that mutual information held at both the determiner-noun and overall language level. This measure furthermore, is symmetric such that there is no directionality required in evaluating the impact of the two variables on one another. If this effect holds at all levels and is in effect bi-directional, then we would be remiss not to adopt the associative model in some form.

However, the two models are not mutually exclusive. It is possible that nominal classification serves more than one purpose. While an association model may explain why nominal classification persists, the discriminative model can explain how it tends to be expressed in grammars generally. Imagine that a language employs gender as an associative mechanism to aid in reducing the nominal complexity for speakers. Given the direction of discourse, and given the ability of gender marking to reduce the search space for nominals in context, it would make sense that languages would tend to have gender marked elements preceding them. However, this perspective does not require that gender marking must precede nominals because its essential role is one that requires no such linearity. In the next subsection we provide some additional evidence for this view.

2.5.2 Further evidence

In the previous subsection, we argued for the associative model. This model would suggest that in gendered languages there is a closer relationship between determiners and nominals. One expression of this can be found in the fact that the number of determiners found in these languages is directly proportional to the number of nominal classes. In 2.20, we give the total number of determiners in a 100 million word subset of the WaCky corpora. We see that English has the fewest determiners, and French and German have increasingly more in a stepwise fashion. If nominal classification reduces nominal complexity, then it makes sense that a gendered language would require more determiners to help offset the amount of complexity found in the nominal lexicon.

(2.20) (a) **English** the **4,755,254**

(b) **French** le 1,578,443

la 2,166,721

les 1,413,565

l' 1,545,676

total **6,704,405**

(c) **German** der 2,390,458

die 2,239,248

das 600,368

des 603,568

den 812,417

dem 411,776

total **7,057,835**

What we are suggesting here is not that German, and French have a lower entropy than English, but rather that nominal classification allows the nominal lexicon in these languages to have a greater complexity than that of a non-gendered language. What nominal classification therefore does, is maintain a baseline level of complexity across these languages that would otherwise not exist without gender given their nominal complexity.

This nominal complexity may have different manifestations in different languages. For example, the tendency for German nominals to have orthographic compounds may reflect the ability of this lexicon to tolerate more complex nominal classes than

English. English on the other hand, tends to use noun phrase syntax to express the same concepts found in the German compounds. This fact is reflected when we consider the complexity of nominals which follow determiners in each language. Figure 2.13 depicts the basic entropy calculation for all nominals that follow determiners in each language. If gender allows these languages to have a greater nominal complexity, then we would predict that nominal complexity is greater following determiners in gendered languages than in non-gendered languages.

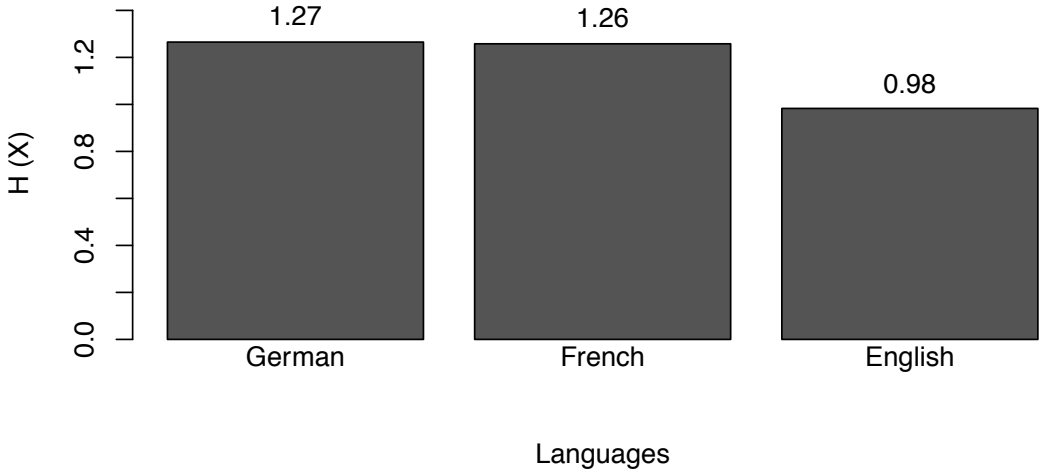

Figure 2.13: The entropy calculation (H(X)) for the unigram probability of nouns in a 100 million word sample from the WaCky Corpus for German, French, and English. Probabilities relative to the overall corpus.

In this bar plot, we see that this is in fact the case. Beyond this, we know that languages crosslinguistically vary in the number of noun classes they have. Furthermore, the ordering of the elements in this language may trend towards marking

classification before nominals, but it is not always necessary. To show this, we provide crosslinguistic data from WALS.

Table 1.9 show the relation between the tendency for grammatical nominal classification, and the ordering of determiners. We see that there is probably not a correlation between the number of genders found in a language, and the ordering of nouns and determiners. On the contrary, 14 languages which have more than five classes have determiners which precede nouns.

Genders	DN	ND	D Prefix	D Suffix	DN and ND	Mixed
None	74	45	0	0	2	5
Two	21	9	0	0	1	6
Three	19	4	0	0	0	0
Four	6	1	0	0	1	1
Five or more	6	14	0	0	0	1

Table 2.9: The relationship between gendered languages and Determiner-Noun ordering. The number of languages with a given configuration and a given number of gender distinctions is the value in any particular cell.

Aside from word order, we see in Table 1.10 that there is also no correlation between the number of nominal classes found in a language and whether that language employs prefixation. This means that in these languages not all gendered languages have class marking which precedes nominals. While it may be the case that all languages with nominal classification have either an inflected prefix or an inflected determiners that precedes nominals, we posit that the presence of languages like Iraqw support the association model. The trend to left leading marking may however suggest that the discrimination model also impacts the grammar.

In this section we argued for the association model based upon the results of our study. We also gave additional evidence which we feel support the idea that nominal

Genders	Little Affix-ation	Strong Suffix-ation	Weak Suffix-ation	Equal	Weak Prefix-ation	Strong Prefix-ation
None	21	59	15	17	9	3
Two	5	17	9	4	0	2
Three	1	13	2	3	2	1
Four	0	4	2	2	1	0
Five or more	1	2	1	3	3	7

Table 2.10: The relationship between gendered languages and Affix ordering. The number of languages with a given configuration and a given number of gender distinctions is the value in any particular cell.

classification lowers the complexity of nominals in the minds of speakers in gendered languages.

2.6 Summary

Given these results, we propose that nominal classification persists in language because it allows speakers to store more complex nominal systems than a language would otherwise allow. This pressure must be enough to persists in language over its evolution.

In the future, this model should be tested in the laboratory. We suggest employing this perspective in artificial language learning experiments such as was done previously, but in manipulating the ordering (Arnon & Ramscar 2012). Furthermore, we would like to expand the corpus studies in two ways. One, we would like to employ language models with greater magnitudes than two word collocations. And second, but more importantly, we would like to investigate the relationship between nominal classification, and nominal complexity in a language like Iraqw.

Chapter 3

Are Affixes in Agglutinative Languages Always Productive?

3.1 Introduction

Although the morphological structure of a word can be described as a static unit (e.g. the word *climber* contains a stem *climb* and affix *-er*), most linguists view the mental representation of this word as a result of a process where it is derived from a base form (e.g. the nominal *climber* is derived from *climb*). From this perspective, words are formed by means of a morphological process or rule. Various approaches to modeling these rules have been developed over the past century. For some, these rules are processes in which the individual units of a complex word (morphemes) are combined resulting in a complex word form (Bloomfield 1933; Chomsky & Halle 1968; Aronoff 1976). In this case, the complex word *climber* would be stored in the lexicon

as the combination of the morphemes *climb* and *er*. For others, complex words are not the result of the composition of smaller units, but are the result of word schema that represent the features common to all morphologically related words (Zwicky 1985; Anderson 1992; Blevins 2006; Finkel & Stump 2007). In this instance, the form *climber* would be stored as a whole unit in a morphological paradigm. For both morpheme-based (former) and word-based (latter) models, there is a challenge to explain the degree to which these these processes occur in novel words in a language. This phenomenon is known as morphological productivity (see (Haspelmath & Sims 2013) or (Bauer 2005) for an overview). The challenge for these theories is that productivity can vary for any given morphological process in a language (Schultink 1961; Booij 1977; Bauer 2001).

While it has been shown that there are phonological (Cutler 1980) and semantic (Aronoff 1976) restrictions on the productivity of morphological processes, there has been an influx of research over the past few decades which argue that productivity can be best understood in relation to word frequency effects (Baayen & Lieber 1991; Baayen 1993b; Bybee 1995a), and that these frequency effects are intimately tied to lexical processing (Frauenfelder & Schreuder 1992; Baayen & Schreuder 1999; Hay & Baayen 2002a).

In this paper, we present evidence from the Bantu language Swahili which challenge these models. We show that the frequency measures used to predict morphological productivity in the Parsing and Productivity Model of Hay and Baayen (2002) cannot explain the facts of productivity in Swahili. Although this model predicts Swahili affixes to always be unproductive, we show that there is reason to think

that some processes are productive, but others are not. In order to solve this problem, we propose an alternative method of quantifying the frequency of morphological processes which we label the Cumulative Root Ratio model. We show that this alternative method makes the same predictions in Swahili as the Parsing and Productivity model does in English (Hay & Baayen 2002a). However, the Cumulative Root Ratio model necessarily entails a critical adjustment to their model. Whereas the Parsing and Productivity model holds that speakers discern morphological processes by means of the surface frequency of words, the Cumulative Root Ratio model predicts that speakers must be using underlying representations to discern these processes. In the end, we relate these implications to the role that these models play in the word-based versus morpheme-based theories of morphology.

The paper is organized as follows. In section two, we present data from Swahili that challenge the Parsing and Productivity Model of morphological productivity, and along the way give an account of the productivity literature. In section three, we develop the CRR model using evidence from Swahili verbal and nominal derivation and then outline a corpus study that tests the validity of the Cumulative Root Ratio model. In section five, we present the results of this study, and in section six discuss the implications of these results.

3.2 Morphological Productivity

In this section, we introduce morphological productivity by giving examples from Swahili derived nominals. Upon doing this we introduce different quantitative meth-

ods for measuring the productivity of morphological processes, and give a brief description of why Swahili is problematic.

3.2.1 Productivity in the Swahili Nominal System

Swahili is a Bantu language from East Africa whose grammatical structure is primarily Bantu, but shares many lexical items with Arabic due to historical contact (Prins 1961). Although Swahili originated as the language of the peoples of the Swahili coast, it has become both the national language of Tanzania, and the lingua franca between people of various nations in East Africa (Whiteley 1969). The grammar of Swahili has a few famous features that have been widely discussed in the formal and typological linguistics literature. First, Swahili is known for having large and complex nominal classification system that is akin to grammatical gender in Indo-European, but varies in that it classifies nominals based upon semantic and phonological features (Dixon 1986; Moxley 1998; Mohamed 2001a). Second, Swahili has a rich verbal system in which complex verbal predicates can be composed of a single orthographic word (Maw 1976; Stump 1997; Mohamed 2001a). Last, Swahili is an agglutinative language such that there is a one to one correspondence between an affix and its grammatical or lexical function (Greenberg 1960; Mohamed 2001a).

Here, we introduce nominal derivation in Swahili. This process occurs when a verb is used to form a nominal which is semantically related to its original form (Schadeberg 2006). Swahili requires that this process be marked with a nominal derivational suffix. Examples of these derivational suffixes are given in Table 3.1.

Affix	Derived Nominal	Related Verb
-a	**mtapisha**	[**tapisha**]
	'emetic'	'cause to vomit'
-aji	**mwindaji**	[**winda**]
	'hunter'	'hunt'
-e	**lishe**	[**lisha**]
	'nutrition'	'nourish'
-fi	**ulafi**	[**kula**]
	'gluttony'	'eat'
-fu	**ulaanifu**	[**laania**]
	'cursing'	'curse at'
-i	**mpishi**	[**pikisha**]
	'(a) cook'	'make cook'
-o	**makumbusho**	[**kumbusha**]
	'souvenirs'	'remind'
-u	**fungu**	[**funga**]
	'password'	'lock/close'
-vu	**maumivu**	[**umia**]
	'aches/pains'	'ache'

Table 3.1: This Table lists the nominal derivational suffixes in Swahili. For each suffix there is an example of a derived nominal, and the related verb root. Crucially this root is not a surface form, as all word forms must take verbal agreement morphology. Historically, these suffixes were common in Proto-Bantu (Schadeberg 2006). However, their presence has fluctuated due to early influence from Arabic (Prins 1961), and due to its promotion to the national language of Tanzania (Whiteley 1969). Furthermore, the language has more recently seen an increase of influence from English (see (Barasa *et al.* 2010) or (Abdulaziz & Osinde 1997)).

The first column gives the orthographic form of the suffix, and in column two we see examples of nominals derived using these suffixes. Last, column three lists the verb from which this form is derived. What is critical here is that these verb forms are not surface forms, but are verb roots which we differentiate using brackets. Surface forms of these verbs must always contain verbal agreement morphology.

Notice that the nominalized form in column two is composed of the nominalizing suffix to its left, and the verb on its right with some changes due to phonological restrictions (e.g. *mtapisha* versus **mtapishaa*). Furthermore, in many cases there is an additional prefix which marks nominal class (e.g. *m-* in *mtapisha*). We will introduce this noun class system in the next section.

Although these suffixes have semantic and phonological restrictions to some degree, they all reflect a common derivational phenomenon (Schadeberg 2006). In this instance, we can say that these suffixes are in competition. Consider a scenario in which a native speaker of Swahili encountered a novel verb for which no derived nominal form exists. In this scenario, each of these possible suffixes could be used. This sort of competition is not unique to Swahili, but exists in Modern English where derivational affixes like *-ity* and *-ness* are in competition (Aronoff & Anshen 1998; Plag *et al.* 1999).

Given this competition, it has been argued that we can determine the productivity of these suffixes by calculating how often these suffixes occur in novel word forms (Baayen & Lieber 1991; Baayen 1992). The idea behind this argument is that for any given morphological process to be productive we would expect it to be used to create new words. If we can calculate how many new words are formed using each

affix, then we can get an idea of how likely we are to see each affix in a new word, or its morphological productivity.

One issue with this approach is that it is difficult to truly know when a word is an example of a novel occurrence. To get around this issue, Baayen (1991) measures novelty relative to a language corpus. When a word form only occurs once in a corpus, then it is called a hapax legomenon or hapax. Baayen (1991) argues that corpus hapax is akin to a novel coinage of a word form, and therefore is an approximation for word novelty.

Another way that we can characterize the hapax is to think of it in terms of type and token frequencies (Bybee 1995b). In a given corpus, the token frequency is the total count of words regardless of how many times we see a repeating word. Consider the sentence in 3.1. The token frequency of this sentence is four, because there is a total of four words. Meanwhile, there are only three unique words because *mwindaji* occurs twice.

(3.1) Mwindaji amemuona mwindaji mwingine.
hunter saw hunter another
'The hunter saw another hunter.'

Therefore, the form (type) *mwindaji* has a frequency (token count) of two, and all other forms have a frequency of one (i.e. C(*amemuona*) = 1, and C(*mwingine*) = 1). With these definitions, we can characterize a hapax as a type which has a token count of one.

With this definition, let's return to the issue of productivity in Swahili nominal

derivation. We can quantify the degree to which these suffixes are productive by identifying all of the types which contain each suffix, and for each type counting the number of them that have a token frequency of one. This will give us the total number of hapaxes in each suffix, and thereby give us an approximate value for their productivity.

Figure 3.1 gives the hapax counts for the nominalizing affixes in the Helsinki Corpus of Swahili (Hurskainen 2004). On the y-axis we see the count of hapaxes in log space for each suffix. The x-axis contains each suffix. These suffixes are ordered in descending rank order to highlight the variation in the hapax counts.

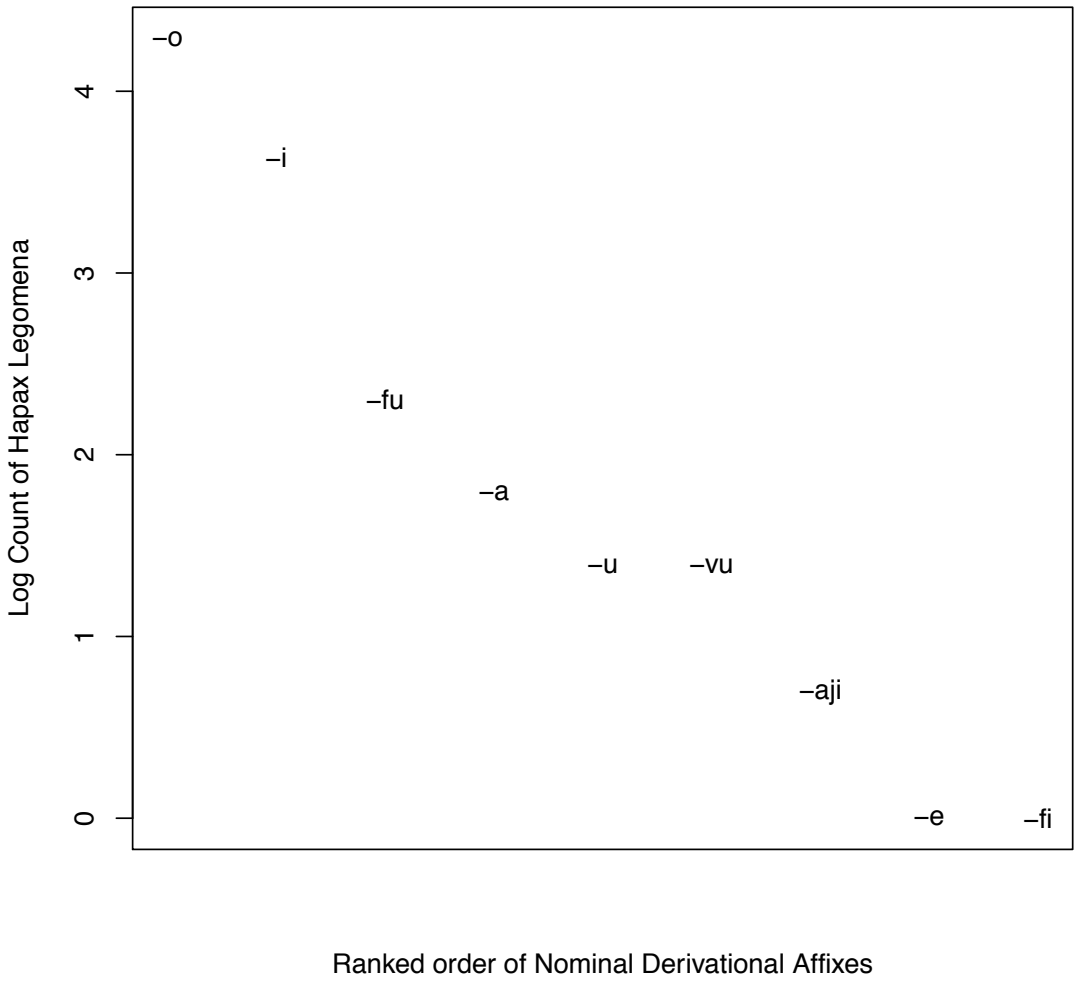

Figure 3.1: The Count of Log Hapaxes for each nominal derivational affix in the Helsinki Corpus of Swahili of 13.6 million words. The affixes are sorted in descending order, and exhibit an asymmetry suggesting that some affixes (left) are more productive than others (right).

Given this data, we can see that the number of hapaxes varies across the suffixes.

Whereas the suffix *-o* has a hapax log frequency of four, other affixes like *-e* have a log frequency of zero. Furthermore, this distribution impressionistically does not appear to be bimodal, but rather the counts are spread fairly evenly between these two values. A preliminary inference we can make is that these suffixes vary in the number of hapaxes in which they occur. Second, this variation is not a productive versus non productive division but rather is a continuous effect. Overall, this data show that these suffixes exhibit variation in the degree to which they occur in novel forms.

Although there may be reason to think that agglutinative languages should typologically not have unproductive affixation, simply due to the pervasive use of affixes in these types of languages. The data seen in the derivational suffixes seem to suggest that this is categorically not the case. At the same time, we will see that the Parsing and Productivity Model of morphological productivity predicts that these suffixes will never be productive. Before we get there, in the next subsection we present a more detailed description of affix productivity and ways of identifying when a process is productive or unproductive.

3.2.2 Models of Morphological Productivity

Early descriptions of morphological productivity defined the property in terms of a language user's ability to coin an infinite amount of new formations using some morphological process (Schultink 1961). Aronoff (1976) expanded on this idea by referencing the productivity of what he called Word Formation Rules, and argued that rules which are productive should have a greater numbers of possible words

than unproductive ones. He defines a possible word as a form that does not exist in the lexicon, but that may potentially exist in the future. In this way, for Aronoff productivity is a relation between actual words and possible words (Aronoff 1976; Aronoff 1983; Aronoff & Anshen 1998). When a Word Formation Rule has the potential to coin possible words its productivity is high. On the other hand, when a rule produces actual word forms who have high frequencies, then the rule is more likely to be unproductive. This analysis entails that productivity is related to the frequency of words, both existing and potential because the degree of productivity is differentiated by means of frequency values.

Another line of research has investigated the relationship between morphological productivity and the order in which affixes are concatenated with roots in English. Siegel (1974) shows that there are two classes of derivational affixes. One which can only occur on word roots, and another which can occur on both a word root and on other affixes. Under this idea, the second type tends to be more productive than the first type (Siegel 1974). This work inspired later models of lexical phonology that labelled this as an effect of level ordering (Kiparsky 1982).

As we noted earlier, more recent research has argued that productivity is a frequency based phenomenon in a vein similar to the earlier model of Aronoff (1976). In addition to the hapax counts which we have shown, Baayen(1991), (1992) and (1993) give more explicit calculations that describe the relative productivity of processes (Baayen & Lieber 1991; Baayen 1992; Baayen 1993a). Baayen (1991) argues that the calculation P gives degree to which an affix is productive in a given corpus. This value, given in (3.2) is the number of hapaxes of an affix (n_1) normalized by

the total token frequency of words containing that affix (N). This value gives the degree to which a process is productive in a given corpus.

(3.2) $P = \frac{n_1}{N}$

In this formula, when all types of an affix are hapaxes, then P will be one. When none of them are hapaxes, then P will be zero. Therefore, this method of normalization transforms the hapax count to a value between one and zero, and in doing so allows one to better compare the relative productivity of two (or more) processes.

Baayen (1993) expanded on this idea by developing the measure $P*$. He argues that this measure is better for comparing the relative productivity of two processes because it normalizes across the total number of hapaxes in the corpus for all possible word forms, regardless of whether they are simple (having no affix) or complex (having one or more affixes). This value therefore tells us what proportion of novel occurrences can be attributed to a given morphological process relative to all novel occurrences. This formula for $P*$ is given in (3.3).

(3.3) $P* = \frac{n_1}{h_t}$

In this value, the number of hapaxes of an affix (n_1) is divided by the total number of hapaxes of all types in a corpus (h_t). Once again this value is normalized to a space between one and zero, and tells us the proportion of novel occurrences that can be attributed to a given morphological process. While this measure informs of the degree of productivity, it does only that and not much more. Specifically, it is

simply a descriptor of novel production and does not provide a story as to why these affixes may exhibit asymmetric variation in novelty.

While this research mainly focuses on the process of derivation, others have suggested that inflectional processes are subject to the same types of frequency measures (Bybee 1988; Dressler 1997). In this work, we assume no difference between derivation and inflection, but do control for it. We do this by investigating the productivity of Swahili derivational and inflectional forms both separately, and combined.

Whereas these models of productivity describe how and when a process is productive or unproductive, Hay and Baayen (2002) attempt to explain the underlying process of productivity in terms of lexical processing (Hay & Baayen 2002a). They build on earlier ideas developed by Baayen (1993) which argue that complex words can be processed in two ways. This model, called the Dual Route Model (Baayen *et al.* 1997; Coltheart *et al.* 2001), argues that complex words can be processed either as a single unit, or as two separate units (i.e. a stem and an affix). For any given morphological process, therefore, types derived from this process can be processed either way. They claim that when complex word forms of a given suffix are processed as separate units, then this entails that the suffix be productive.

This research therefore draws a direct line between productivity, and how often an affix is parsed by a speakers in lexical access. In this way, productivity and lexical representation are intimately linked. This lexical representation is stochastically determined by how often the affix is parsed by an speakers. For them, parsing occurs when an affixed type (e.g. *googler* - 'a person who googles something') occurs less frequently than it's unaffixed counterpart (e.g. *google*). They suggest that when

this happens, the speaker will be more likely to first parse *google*, and then have a remaining affix -*er* left over to parse.

On the other hand, when an affixed form occurs more often than it's unaffixed counterpart, the affix will be parsed along with the root form (e.g. *pavement* would likely be parsed whole, since *pave* is less frequent than *pavement*). Given these two options, an affix would have its own representation when more than half of all affixed types are parsed by the speakers. On the other hand, if the majority weren't parsed, then the speaker would be less likely to consider the affix as an affix at all.

Productivity is calculated in this model by identifying all of the types of an affix in a corpus such that each type contains the affix (derived type). Then, the related base or root form lacking the affix is identified (underived type). Hay and Baayen (2002) then calculate the token frequencies of all forms in the set of those derived and underived types. These frequency values are then log transformed and plotted on a Euclidean plane. The x-axis contains the log derived frequencies, and the y-axis contains the log underived frequencies. The data is then analyzed in three ways.

First, an X=Y line is plotted. This line is simply a line in with a slope of 1, that passes through all points where the log frequency of the underived forms is equal to the log frequency of the derived forms. When an individual point sits above the X=Y line, then its underived frequency is greater than its derived frequency. In this case, this type would be more likely to be parsed in perception. When it's below the line, then the derived type is more frequent than the underived type. In this instance the affixed form would be less likely to be parsed in perception. When more than half of these types are above the line, then the affix should be productive. Hay and

Baayen (2002) quantify this value by calculating the ratio of types that are above the X=Y line and dividing it by the total number of types of the affix. We refer to this at the type-ratio. When the type-ratio is one, all affixes should be parsed, and when it is zero, no affixes should be parsed.

Second, they develop a linear model to describe the frequency ratio data. To do this, they employ a least trimmed squares robust regression model. This linear model is chosen because it models the log frequency ratio data without being unduly affected by outliers (Rousseeuw & Leroy 2005). The slope and y-intercept of this line reflects the rate of parsing of these types. When the intercept is high, and when the slope is steep, then these affixes should have higher rates of parsing in perception. These two variable, along with the type-ratio are correlated with a high hapax count, and should overall give a prediction for the relative productivity of morphological processes.

This model is enticing in that it has a psychological motivation for explaining productivity. More recent approaches have furthermore shown that looking at the degree to which these suffixes order reflect similar measures of productivity (Ingo Plag & Harald Baayen 2009; Sims & Parker 2015).

In the next subsection we will describe why Swahili is problematic for the Parsing and Productivity Model of productivity as described above.

3.2.3 The Parsing Ratio and Swahili

In this subsection we present reasons as to why Swahili morphology provides a challenge for the Parsing and Productivity Model. We do this by showing that word

roots in Swahili hardly ever occur in isolation. On the contrary, there is almost always overt affixation. This fact means that complex word forms never have surface underived forms (surface forms in which the root occurs in isolation). We will see that this fact assumes that Swahili affixes will in all cases be unproductive.

The Parsing and Productivity model is appealing in its ability to assign a story to productivity, and we would like to evaluate whether such a theory can be validated when tested in languages other than English and Dutch. However, when we look to Swahili we run into a large issue. First, in Swahili the derivational processes we've described only ever co-occur with multiple affixes, as seen in Table 3.2.

Table 3.2 depicts affix by affix breakdowns of a few derived nominals in Swahili. Column one contains the nominal class marker (similar to grammatical gender) of the nominal. These affixes denote the class of the derived nominal and this classification is obligatory for all nominals. Column two gives the verbal root of the derived nominal, which is consistent across all nominals (i.e. *piga*). Column three gives derived verbal forms which contain verbal derivational affixes along with the verbal root. Notice that the presence of these affixes alters the semantics of the derived nominal. Finally, we see the nominalizing suffixes described in the previous section. These glosses of complex forms are intended to demonstrate that we have a lot of additional variation to deal with if we consider the relation between say *-aji* and *piga*.

These facts are not surprising considering Swahili's designation as an agglutinative language. However, these facts present a problem for the Parsing and Productivity Model. Here, we present a concise summary as to why. First, recall that Hay

Nominal Class	Verbal Root	Verbal Derivation	Nominal Derivation	
m-	**-pig-**	-	**-aji**	**mpigaji**
CL1	*hit*		N_{AGENT}	'kicker/hitter'
mi-	**-pig-**	-	**-o**	**mipigo**
CL4	*hit*		N_{INST}	'kicks/blows'
m-	**-pig-**	**-an-**	**-aji**	**mpiganaji**
CL1	*hit*	REC	N_{AGENT}	'fighter'
mi-	**-pig-**	**-an-**	**-o**	**mipigano**
CL4	*hit*	REC	N_{INST}	'battles'
u-	**-pig-**	**-an-**	**-aji**	**upiganaji**
CL14	*hit*	REC	N_{AGENT}	'rivalry'
u-	**-pig-**	**-an-**	**-o**	**upigano**
CL14	*hit*	REC	N_{INST}	'contest'

Table 3.2: This Table contains six different derived nominals from a single verbal root (i.e. *piga* - 'hit'). These nominals vary in the nominal affix, verbal derivational affix, and nominal class.

and Baayen's (2002) analysis compares the frequency of affixed forms to their affixless counterparts in order to determine whether a form is parsed. In a language like English or Dutch, non-affixed forms tend to occur as inflected forms with no overt marking. If we recall our earlier example of *googler*, we can see that in English, *google* occurs as both a verb and nominal without any affixation. However, in Swahili, given the facts in Table 3.2, we have a problem.

If we were to analyze these forms in the strictest of terms, then we would want to compare the frequency of a nominal like *mpigaji* to a from like *mpiga*, which itself does not occur as a surface form. Logically, it follows then that for all derived forms of nominal derivational affixes, the frequency of affixless forms ($f = 0$) would always be less than the frequency of affixed forms ($f > 0$). From here, all affixes would then

go unparsed, and therefore would be predicted to be not productive. However, we know this to not be the case given the facts in Figure 3.1.

In the coming Section we provide an adjustment to this parsing measure that will allows us to accurately predict the facts regarding the number of hapaxes of each affix, and in doing so, will extend the parsing story to fit a language like Swahili and others which have similar morphological properties.

3.3 The Cumulative Root Ratio Model of Morphological Productivity in Swahili

In this section we introduce the Cumulative Root Ratio model of morphological productivity. In order to do so we give a more in depth description of Swahili nominal derivational patterns and how they interact with the verbal derivation and nominal classification systems. We then show why the Parsing and Productivity Models fail to account for Swahili. Last, we introduce the Cumulative Root Ratio.

3.3.1 Properties of Swahili Morphology

In this subsection we introduce the Swahili nominal classification system in addition to the verbal derivation and inflection systems. We do this by elaborating on the nominalizing suffixes found in the previous subsection. This data will set the groundwork for the Cumulative Root Ratio model.

First, we show that derivational processes in Swahili are not as straightforward as in English. In Figure 3.2, we see a directed graph describing the pathways which

underlie the creation of the word forms seen in Table 3.2 in the previous section. In this directed graph we see an explicit description of the path that the derived nominal can take. At the top we see a verb root (i.e. piga). This root can either combine with a verbal derivational affix (i.e. *pigana*, or remain as a simple root. From here, it can either undergo nominalization by combining with an affix like -*aji*, or it can become a verb, but in doing so must combine with over inflectional affixation (i.e. *ku-*).

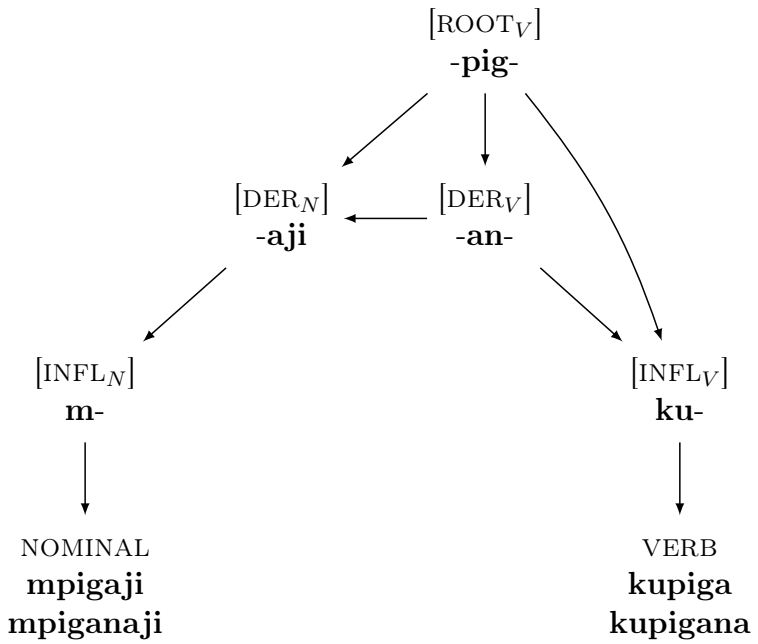

Figure 3.2: This figure depicts a direct graph describing the processed associated with verbal roots. A verbal root may only occur as a surface form given the processes described here. At each node, there is a feature label, and in bold an example of that feature. The bracketing denotes that these labels occur only as underlying forms. Where there is a lack of bracketing the label denotes a surface form.

These facts have some interesting consequences for productivity of derivational

processes in Swahili. If we were to consider the productivity of the affix *-aji*, then we would have to consider a few different things. First, we can see that derivational affixes can combine with verbs which themselves may have derivational affixes. Second, in any case, derived nominals must always be a member of a nominal class, and this nominal class can vary based upon the nominal being derived. As evidenced in Table 3.2 above, this classification is not absolute, but rather works similar to derivation. Take for example, the minimal pair of *mpiganaji* ('fighter') and *upiganaji* ('rivalry'). Here the only difference between the two words is their nominal class. Therefore, in considering the relative productivity of *-aji*, we would have to make a decision as to how we would compare the nominal forms. Should we look at each complex type individually, or should we look at all complex types combined? In the coming subsection we address this issue. Before doing this, we give a more in depth description of the nominal and verbal agreement systems, and also describe the verbal derivation system.

Swahili Nominal Classes

The Swahili nominal class system evokes for many an example of a complex nominal system that is rarely seen crosslinguistically. Here, we give a brief description of it in order to give due credit to the its degree of complexity. Table 3.3 gives a summary of nominal classes found in Swahili. In column one, we see the class number. Where there is a split, the number on the left indicates the singular class and the number on the right indicates the plural class. Although the affixes have different forms, the singular and plural forms of the same class have the same semantic characteristics

(Moxley 1998).

The second column gives the prefix associated with each class. Where the prefix takes the form *N*, the class does not always contain a prefix. Often, the words contained within this class are borrowings, and do not take prefixes (e.g. *meza* - 'table'). When they are not borrowings, they tend to have an alveolar nasal, hence *N*. Take for example, the form *ndege* ('bird') which contains the nasal *n*-. The semantic distinction is found in column three. While these classes have some semantic characteristics, it is not always the case that all members of these classes hold these characteristics. In fact, some research shows that as Swahili has grown, these semantic classes hold less and less coherence (Barasa *et al.* 2010). Finally, in column four, we give examples of nominals in each class along with their affix by affix breakdown and translations.

It's important to note that the classification used here follows a tradition of characterizing classification that is itself to a typological description of the language. For example, whereas many languages would not describe the infinitive as a nominal, in Swahili, the infinitive (class 15) is identical to the gerund form of a verb which triggers agreement in an manner identical to all nominals. Likewise, locatives are also typologically argued to occur as nominal forms (i.e. classes 16-18).

Noun Class	**Prefix SG/PL**	**Semantic Distinction**	**Example SG/PL**	
1/2	m-/wa-	*humans*	m- -swahili CL1 swahili 'swahili person'	wa- -swahili CL2 swahili 'swahili people'
3/4	m-/mi-	*plants*	m- -nazi CL3 coconut 'coconut tree'	mi- -nazi CL4 nazi 'coconut trees'
5/6	ji-/ma-	*groups* *augments*	ji- -cho CL5 eye 'eye'	ma- -cho CL6 eye 'eyes'
7/8	ki-/vi-	*tools* *diminutives*	ki- -atu CL7 shoe 'shoe'	vi- -atu CL8 eye 'shoes'
9/10	N-/N-	*animals* *loan words*	m- -buga CL9 savannah 'savannah'	m- -buga CL10 savannah 'savannahs'
11	u-	*extension*		u- -nywele CL11 hair 'hair'
14	u-	*abstraction*	u- -moja CL14 one 'unity'	
15	ku-	*infinitive*	ku- -ongea CL15 talk 'to talk/ talking'	
16	pa-	*specific location*	pa- -le CL16 there 'over there'	
17	ku-	*general location*	ku- -le CL17 there 'over there'	
18	mu-	*internal location*	mu soko- -ni CL18 market LOC 'in the market'	

Table 3.3: This Table contains the Noun Classes

We see that this nominal classification system has a few interesting properties. First, notice that there are many more classes than are common in languages generally. Second, what is of special interest here is that the nominal classes have overt semantic distinctions that operate in the way that derivation operates in other languages. Recall our earlier example of *upiganaji* ('rivalry') and *mpiganaji* ('fighter') which take Class 14 and Class 1, respectively. Here, the nominal class relates to different semantic form, but does not denote a different lexical category.

This property has implications for the role of productivity as we have suggested earlier. If we step away from derived nouns, we can see that this fact is true also in nouns generally. The nominals in (3.4) are examples of nominals that are not derived from other forms (i.e. simplex nominals). Here we see that the same simplex nominal root *moja* can occur in at least three different nominal classes to denote different nominals. The examples each contain an affix and root breakdown of the nominal, along with corresponding glosses and translations. Notice that the the Class 9 form of the nominal does not contain an overt affix. Here the underspecified classifier N surfaces as the nasal m, but is not repeated as is the case in Class 1.

(3.4) (a) m- -moja

CL1 *one*

'one person'

(b) u- -moja

CL14 *one*

'unity'

(c) moja

CL9.*one*

'the number one'

This property is problematic when we consider calculating the productivity of these affixes. This is because nominals in all forms must occur in some nominal class, and in that sense there is no such thing as an underived nominal. Here we have seen that both the nominal classification and derivation system have a complexity unseen in a language like English. Now, we turn to the verbal system where we will see the same sort of issue.

Swahili Verbal Derivational System

In addition to the nominal system, there is a rather large verbal derivation system in Swahili. Complex predicates in Swahili are marked with overt derivational affixes in cases where a language like English uses other means such as multi-word verbal

phrases or unrelated forms. This can be found in the verb in (3.5). In this example, dashes indicate affix boundaries and spaces indicate word boundaries. The verb *kuambukizwa* ('to be transmitted') is a single orthographic word in Swahili, and can be broken down into a simple verb root *ambuka*, a passive marker *-w-*, and a verbal derivational affix denoting causation *-iz-*. Together these markers form a complex verb that in English is derived by combining different word forms[1].

(3.5) Ma- -gonjwa haya- -wezi **ku- -ambuk- -iz- -w- -a** ...
CL6 *disease* CL6NEG *can* CL15 *awake* CAUSE PASS FV ...
'(The) diseases cannot be transmitted...'

This example demonstrates a sample of the verbal derivational morphology available to the Swahili speaker. In Table 3.4, we give a list of these affixes which are commonly referred to as verbal extensions (Mohamed 2001b). The first column gives the name of the extension type, and the second column gives the possible surface variants of the suffixes. These suffixes are subject to phonological conditions of vowel harmony (Mohamed 2001b)[2]. The third column gives the semantic distinction of each suffix, and the fourth column includes examples of alternations of verbs containing these suffixes. In these examples we see that these affixes occur with verb roots, and in doing so can (i) alter the semantic denotation of the verb, or (ii) alter its syntactic adicity (i.e. the number of nominal arguments it takes) (Mohamed 2001b). For example, the causative form of *kula* ('eat') changes the meaning of the verb

[1]Note that *transmitted* in English does contain something of a prefix in *trans-*, as well as an overt passive marker *-ed*. However, there is no such verb a *mit* in English, whereas *ambuka* in Swahili is quite common.

[2]For example, the stative takes the form *ika* when the verb root contains an initial *a*, *i*, or *u* vowel, and takes the form *eka* when it contains an initial *e*, or *o* vowel.

kulisha ('feed'), and the stative makes a transitive verb *vunja* ('break (something)') intransitive *vunjika* ('be broken'). Often, these extensions can stack, or co-occur and derived complex verbs, as seen in (3.5).

Extension	Form	Semantic Distinction	Example Simplex/Complex	
Simple	-a,-i -u,-e	*base form*	chez- -a *play* FV 'play'	
Stative	-ka,-ika -eka,-lika -leka, -uka	*intransitive*	vunj- -a *break* FV 'break'	vunj- -ika *break* STAT 'be broken'
Applicative	-ia,-ea -ilia,	*action applied*	lip- -a *pay* FV 'pay'	lip- -ia *pay* APPL 'pay for'
Causative	-isha,-esha -iza, -eza -sha, -za	*cause to*	kul- -a *eat* FV 'eat'	kul- -isha *eat* CAUS 'feed'
Augmentative	-ua,-oa -za, -liza -leza	*intensiveness*	song- -a *press* FV 'press'	song- -oa *press* AUG 'press out'
Reciprocal	-na,-ikana -ekana	*reciprocated action*	pig- -a *hit* FV 'hit'	pig- -ana *hit* REC 'fight'
Reversive	-ua,-oa -ia	*reversed action*	fung- -a *close* FV 'close'	fung- -ua *close* REV 'open'
Static	-ma,-mana	*static (in)action*	fich- -a *hide* FV 'hide'	fich- -ama *hide* STA 'be in hiding'

Table 3.4: This Table contains the verbal derivational suffixes.

This system pairs with deverbal nominal suffixes that we first introduced in the paper, but differs in a few ways. First, they are similar in that they are suffixes that alter the semantics of the word which they modify. However, verbal suffixes do not change the lexical category of the verb which they modify, and they furthermore can only occur on roots inside the nominalizing suffixes, as seen in the example *upiganaji* ('rivalries'), where the reciprocal *-an-* occurs between the root *piga* and the nominalizing suffix *-aji*. The reverse scenario is unattested (i.e. **upigajiana*).

Swahili Verbal Inflectional System

In addition to derivation, the verbal system overtly marks agreement with nominal subjects across the different nominal classes. This agreement can be seen in (3.6) where the nominal subject *Madirisha* must agree with a verbal inflectional prefix in nominal class (i.e. *ya-*). This verbal agreement is required on lexical verbs when no model or auxiliary occurs (Mohamed 2001b), and pairs precisely with the gender system. Therefore, for every subject in Swahili there is a verb that contains a corresponding nominal class prefix.

(3.6) **Ma-** -dirisha **ya-** -me- -vunj- -ik- -a

CL6 *window* CL6 *have* *break* STAT FV

'(The) windows have been broken...'

These facts indicate that the verbal forms require inflectional marking in all surface examples. Since verbs agree with nominal class, the complexity of verbal agreement mirrors the complexity found in the nominal inflection system. We can see

an example of this system in Table 3.5. This table is very similar to Table 3.3 which describes the nominal class system. The first column contains the class number. The second column gives the form of the verbal inflectional prefix which is similar to but distinct from the nominal prefix. One difference is that the person agreement system all falls under class one. In the third column, we see the semantic or grammatical distinction for each class, and in column four there are singular and plural examples of each class. Notice that class 15 indicates infinitives, and that this marking is identical to the nominal prefix for infinitives. Furthermore, classes 16-18 denote locative subjects. These forms are much more restricted in their usage because they modify places and receive a sort of existential reading (Mohamed 2001b).

Class	Prefix SG/PL	Semantic Distinction	Example SG/PL	
1/2	ni-/tu-	1^{st} person	ni- -na CL1.1SG have 'I have'	tu- -na CL2.1PL have 'we have'
	u-/m-	2^{nd} person	u- -na CL1.2SG have 'you have'	m- -na CL2.2PL have 'you(all) have'
	a-/wa-	3^{rd} person	a- -na CL1.3SG have 'she/he has'	wa- -na CL2.3PL have 'they have'
3/4	u-/i-	*plants*	u- -na CL3 has 'it has'	i- -na CL4 have 'they have'
5/6	li-/ya-	*groups* *augments*	li- -na CL5 have 'it has'	ya- -na CL6 have 'they have'
7/8	ki-/vi-	*tools* *diminutives*	ki- -na CL7 have 'it has'	vi- -na CL8 have 'they have'
9/10	i-/zi-	*animals* *loan words*	i- -na CL9 have 'it has'	zi- -na CL10 have 'they have'
11	u-	*extension*		u- -na CL11 have 'it has'
14	u-	*abstraction*	u- -na CL14 one 'it has'	
15	ku-	*infinitive*	ku- -na CL15 have 'to have'	
16	pa-	*specific location*	pa- -na CL16 have 'there is (specific)'	
17	ku-	*general location*	ku- -na CL17 have 'there is (generic)'	
18	mu-	*internal location*	mu- -na CL18 have 'there is (inside)'	

Table 3.5: This Table contains the Verbal Markers

So far we have introduced the nominal and verbal affixation systems. We have left out facts concerning other verbal internal affixes including tense, relative markers, and object agreement. We do this to keep our study as simple as possible while evaluating all aspects of Swahili nominalization.

Summary of Swahili Morphology

In this subsection we gave descriptions of the Swahili nominal classification system as well as the verbal derivation and inflection systems. Doing this serves two purposes. First, it gives insight into the complexity of the agreement and derivation systems in Swahili generally, and second sets the groundwork for showing why Swahili is problematic for the Parsing and Productivity Model. Namely, these facts show that surface forms of nouns and verbs almost always have overt affixation, and therefore neither verbal nor nominal roots occur in isolation as a surface form.

In the next subsection, we give hapax counts for each of the affixes represented in these systems. Doing this demonstrates that productivity asymmetries not only exist in the nominal derivational suffixes, but that it also exists in each of these systems. These facts will lead us to develop a model which can explain the facts of all these asymmetries.

3.3.2 Productivity asymmetries in Swahili

In this subsection, we give the hapax counts for the affixes in the verbal derivation system, as well as in the nominal and verbal classification systems. We will see that these paradigms each exhibit the same hapax asymmetry that we saw in the nominal

suffix affixes. In the next section we will discuss the implication of this data in detail as it relates to the Parsing and Productivity Model, and in doing so provide an alternative model.

The verbal derivation affixes represent a morphological process that could plausibly be productive in some cases, and not be productive in others. For example, some of the affixes have similar functions both syntactically and semantically. Consider the stative form and the passive form. They both reduce the adicity of the verb, and in doing so describe a condition or state. Without control for the broader semantic and syntactic variations between the two, from a morphological perspective there is little difference between the two forms. Consider the verb *vunjika* ('be broken'). Here the stative blocks the formation of a passive form **vunjwa* ('break' + *-w-*) or even a passive form **vunjikwa* ('broken' + *-w-*). These facts suggest that these two may be in competition for possible related forms, much like we saw in the case of derived nominals.

To investigate this, we calculated the hapax counts for each of these affixes in the Helsinki Corpus of Swahili (13.6 million tokens). The results are shown in Figure 3.3. This Figure is similar to Figure 3.1 because it gives the log hapax counts for each affix in descending ranked order. There are two things to notice in this plot. First, notice that the affixes once again show variation in the relative number of hapaxes. Whereas *-w-* has a hapaxes log count of ten , *-an-* has a hapax log count of around eight. Second, notice that these values are much higher than those found in the nominal derivational affixes. For these affixes, recall that the greatest log hapax count was around four. This is a difference of two orders of magnitude (i.e. $log(100) = 4.61$

and $log(10,000) = 9.21$). This large difference can be attributed to the fact that verbs have considerably more affixation than nominals due to marking things like tense, aspect, objects and subject agreement, amongst other things. However, what is crucial here is that these affixes still exhibit a similar asymmetry in these hapax numbers relative to the other members of this class.

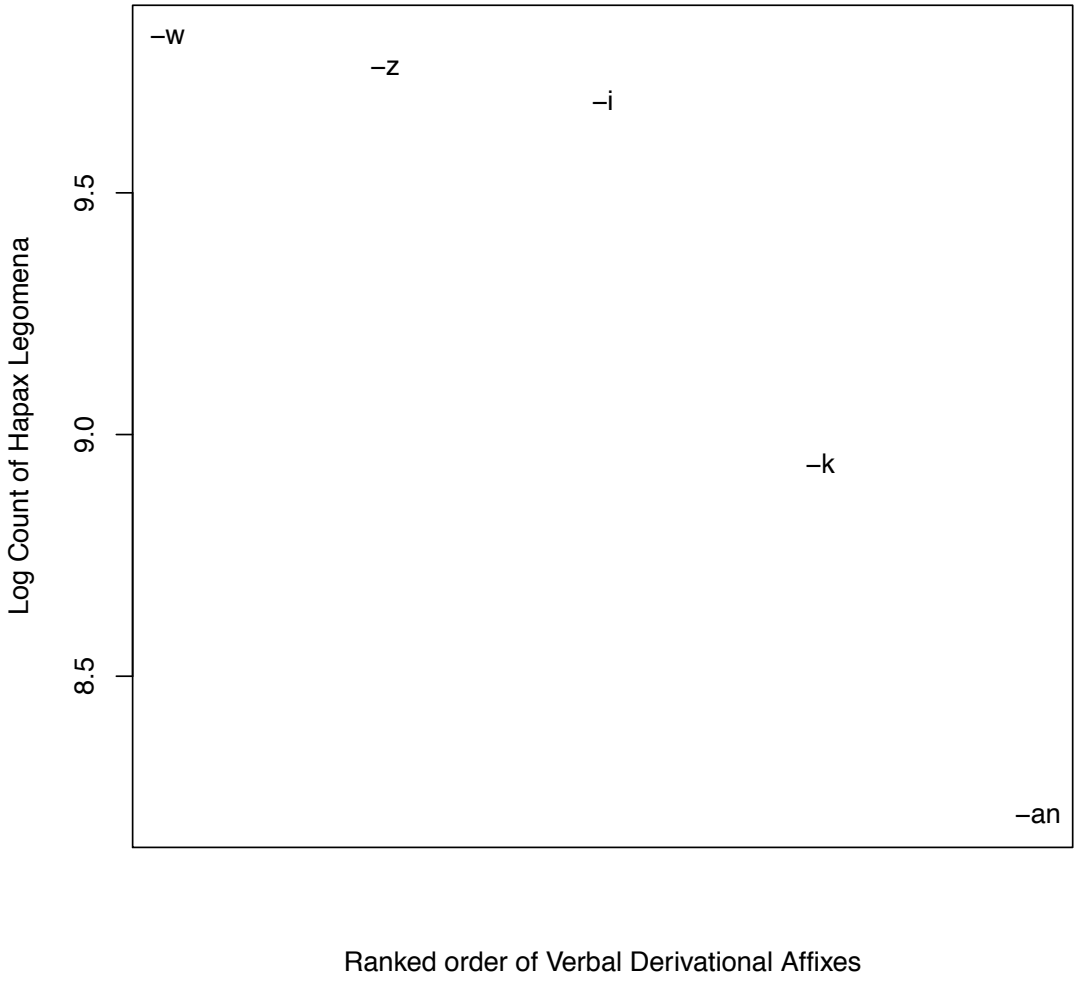

Figure 3.3: The Count of Log Hapaxes for each verbal derivational affix in the Helsinki Corpus of Swahili of 13.6 million words. The affixes are sorted in descending order, and exhibit an asymmetry suggesting that some affixes (left) are more productive than others (right).

This results indicate that much like the nominal derivational suffixes, we would

expect there to be productivity asymmetries in the verbal derivational affixes. While we have only focused on derivation so far, we can ask whether these productive asymmetries exist for the inflection system. We propose that since nominal classification can denote functions similar to derivational processes, then the inflectional system should act like derivational systems in other languages. Recall the examples found in (3.4) where a single nominal had three different surface forms in three different nominal classes.

Therefore, we predict there to be productive asymmetries in this system. Since class 9/10 is typically reserved for borrowings, we expect this class to be more productive in both the verbal and nominal domains. Here, we see that for both verbal and nominal classifications there are hapax asymmetries between the affixes. This facts can be found in Figure 3.4 and Figure 3.5 respectively.

In Figure 3.4, we see that in fact the Class 9 affix *n-* has the greatest log hapax count, and that the locatives and infinitives are much lower, approaching zero. Furthermore, the other nominal classes occupy the space between these two values, with a gap between the more typical nominals and locatives.

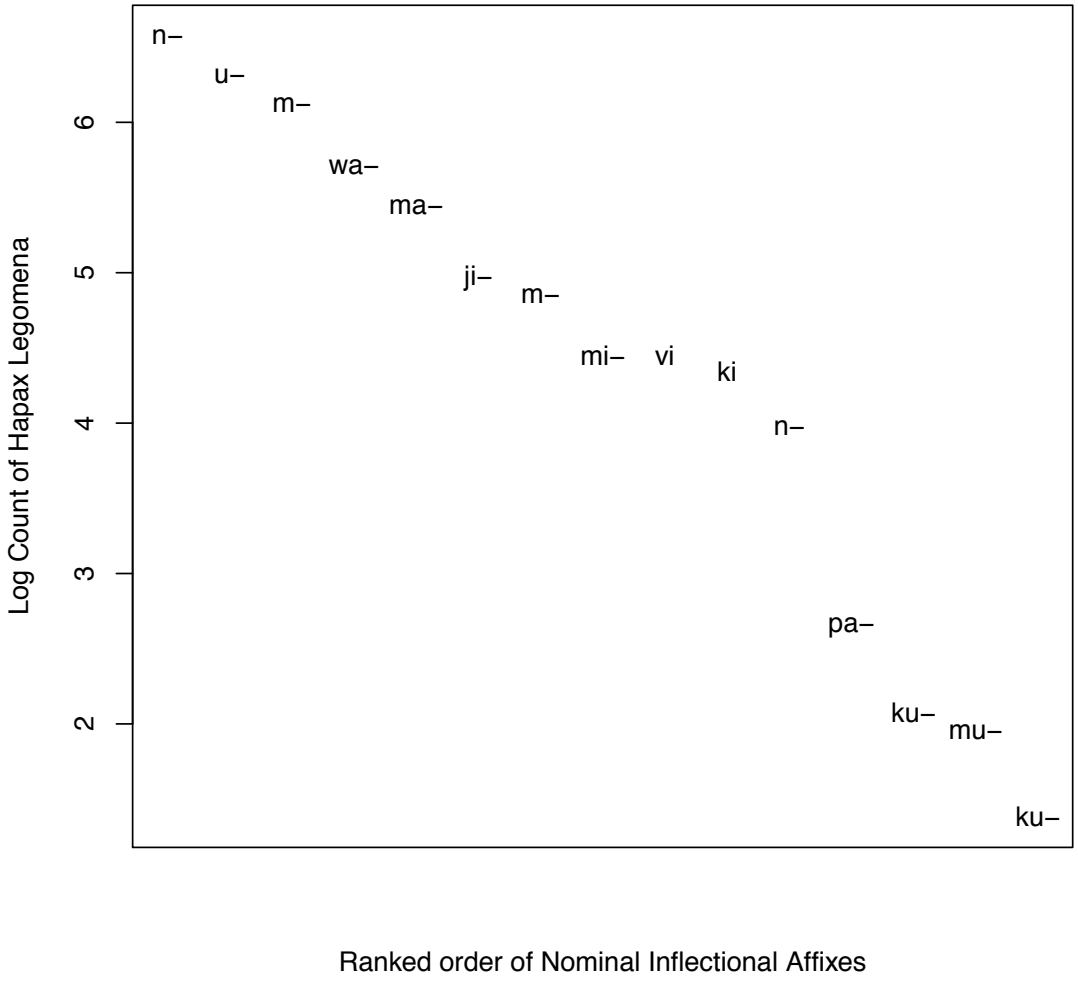

Figure 3.4: The Count of Log Hapaxes for each nominal inflectional affix in the Helsinki Corpus of Swahili of 13.6 million words. The affixes are sorted in descending order, and exhibit an asymmetry suggesting that some affixes (left) are more productive than others (right).

Figure 3.5 gives the log hapax counts for the verbal prefixes. Again, we see an

even spread of log hapax counts across the nominal classes similar to the log hapax counts for nominals. Notice again however that the counts are about two orders of magnitude greater than in the nominal prefixes. This quality mirrors the value change between nominals and verbs more generally.

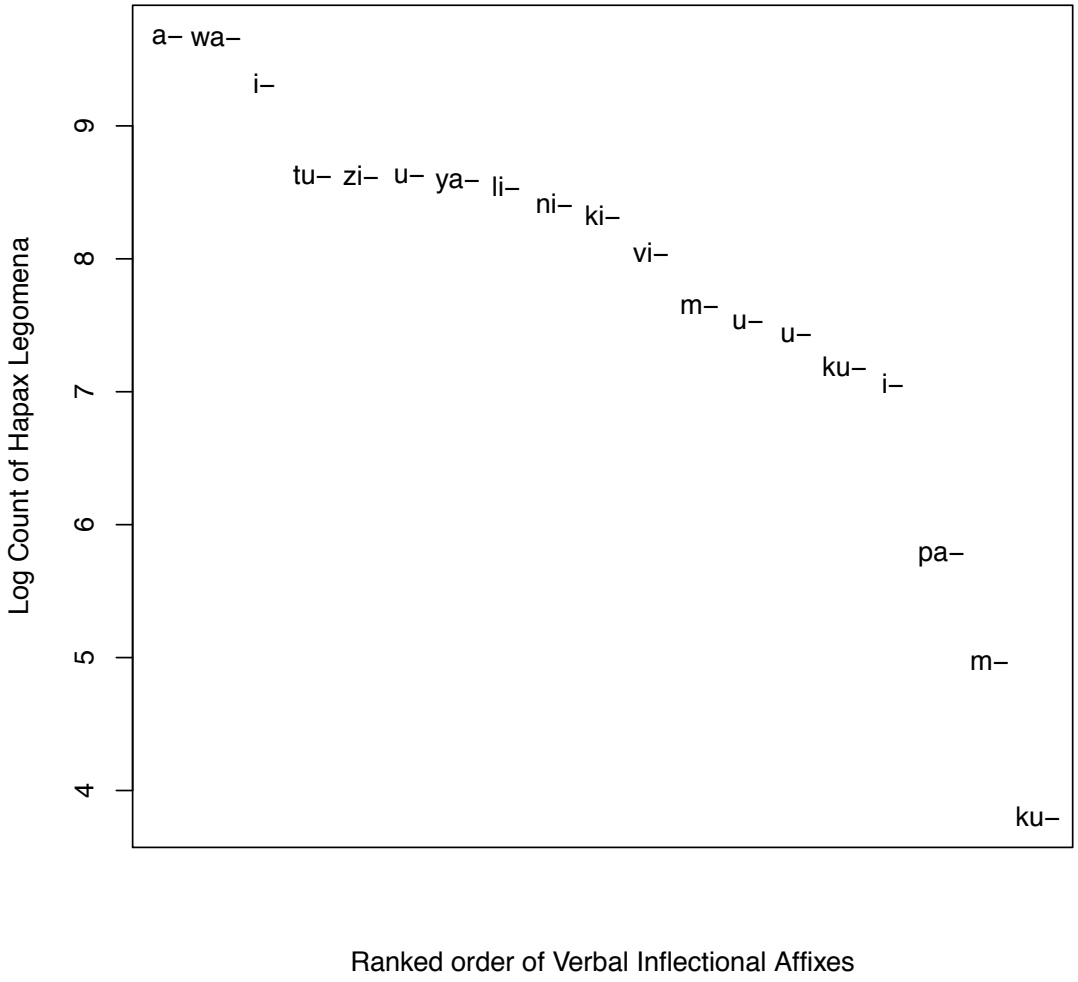

Figure 3.5: The Count of Log Hapaxes for each verbal inflectional affix in the Helsinki Corpus of Swahili of 13.6 million words. The affixes are sorted in descending order, and exhibit an asymmetry suggesting that some affixes (left) are more productive than others (right).

To recap, we first saw that Swahili derived nominal affixes exhibit asymmetrical

variation in the degree to which they occur in novel forms in a corpus. Furthermore, we saw that these forms are also contain verbal derivation and nominal classification processes. This nominal classification is reflected in the verbal agreement system, and is required for any surface verb form. Last, we saw that each of these systems exhibit log hapax count asymmetries that would suggest that certain affixes within each paradigm should be more productive than others. Next, we evaluate this claim in terms of the Parsing and Productivity Model. We will see that this model cannot predict the asymmetry seen in the hapax data. We therefore propose the Cumulative Root Ratio model, which we develop in the next section.

3.3.3 Cumulative Root Ratio

In this section we introduce the Cumulative Root Ratio to explain productive asymmetries in Swahili. We do this first by showing why the Parsing and Productivity Model fails to account for productive asymmetries in Swahili verbal and nominal affixes. To do this, we briefly recap the underlying format of the Parsing and Productivity model. Then we describe a way to get around the issue of multiple affixes using cumulative root frequency, and not surface type frequency to determine the relative productivity of morphological processes.

In the Parsing and Productivity Model of Hay and Baayen (2002), an affix's productivity is determined by a bottom-up stochastic relation between all word forms that contain the affix, and their non-affixed counterparts. Let's consider the nominal derivation affix -*er* in English. For them, the representation of this affix and therefore the productivity of its use is determined by taking into account all of the types that

contain this affix (e.g. *catcher*, *farmer*, *talker* etc.). For each of these types, they have a related word form from which they are derived, referred to as their underived form (e.g. *catch*, *farm*, *talk* etc.). For each of these types, the English speaker is then tacitly comparing the underived and derived frequencies relative to one another. When the underived forms have a greater frequency on average, then the affix should have an autonomous representation in the lexicon, and would therefore be productive.

For a language like English, where non-affixed forms occur as surface forms, this make sense. However, we have seen that in the literal sense, this cannot be true in a language like Swahili. Given the complex word *msemaji* - 'a spokesperson', we would imagine that the Swahili speaker would parse this form given the nature of it composition, as seen in (3.7). In this case, it seems attractive to imagine that the prefix denoting a human *m-*, and the affix denoting an agent *-aji* would have autonomous representations. This idea is furthermore bolstered by the notion that the verb *sema* - 'speak' is much more common.

(3.7) m- -sem- -aji

CL1.HUMAN *speak* N_{AGENT}

'speaker/spokesperson'

Unfortunately, as we have seen with the verbal inflection system, the root *sema* never occurs in isolation. Rather, it always co-occurs with either verbal inflection, and optionally with verbal derivation. Therefore, the actual frequency of a root like *sema* is split across several inflectional and derivational forms. To adjust for this, we propose that rather than comparing the token frequencies of a non-existent

underived surface form (*m-sem*, or *sem*), the Swahili speaker would do the following, as in (3.8). In this formula, we see that for an equation that describes the Cumulative Root Ratio. Here, we propose that Swahili speakers compare the cumulative root frequencies (CRF) of the derived forms to the cumulative root frequency of the underived forms of each type (t) for an affix (T).

(3.8) $CRR := \sum_{t \in T_{[\text{AFFIX}]}} CRF(Underived_t) : CRF(Derived_t)$

This implies that the Swahili speaker would construct their representation for *-aji* from the set of all surface forms containing *-aji*. For the word *msemaji*, the likelihood of parsing would then be a function of the cumulative root frequency of [*sema*] to that of [semaji][3]. We calculate cumulative root frequency as the sum of the token frequencies of all word forms that contain each root, as in (3.9) [4].

(3.9) $CRF_{[\text{ROOT}]} := \sum_{w \in W_{[\text{ROOT}]}} C(w)$

In (3.9), we calculate cumulative root frequency as the sum of the counts (C) all of the surface frequencies of all word forms containing some root ($w \in W$.

Given this definition, the cumulative root frequency of each 'root' is composed of various derived forms, which then are split across multiple inflections. This adjustment therefore shifts the stochastic construction of affix representations from a relation between a set of underived and underived surface forms (Parsing and

[3]the adjusted calculation is [*sema*] = [*sema*]-[*semaji*], since [*sema*] contains the derived form

[4]The notion of cumulative comes from previous work from the psycholinguistic literature, where the frequency of all morphologically related forms impacts lexical access. (Colé *et al.* 1989; Moscoso Del Prado Martín *et al.* 2004)

Productivity Model), to a relation between constructed underived and derived representations whose quality is subject to the frequencies of their various surface forms (Cumulative Root Ratio Model). One result of this shift is that parsing should occur significantly more readily in Swahili than in a language like English. However, this sort of trend is precisely what one would predict given the degree of affixation in the agglutinating language. But, how can we test this claim?

Recall that Hay and Baayen (2002) calculated the degree of representation of an affix by graphing the log frequencies of all types containing and affix (derived on the x axis), against the frequencies of those forms without the affix (underived on the y-axis). They then perform a least trimmed squares r^2 best fit line of the data. The slope of this best fit line, and the Y-intercept are both directly proportional to the degree of representation, and therefore productivity of the affix. In addition to this, they create an X=Y line, with a positive slope of 1. This line delineates the types which should be parsed, against those that should not be parsed.

Let's consider if we were to do this with all affixes containing -*aji*. If we were to plot the derived and underived frequencies of all forms containing -*aji*, we would have a problem. First, all derived forms would have at or greater than a log frequency of zero. However, as seen with *msemaji* in (3.7) there is no surface underived form in the strictest sense. This means that the underived from has a value of -infinity (i.e. $log(0)$). Therefore, all derived forms would have a greater frequency than underived forms and all affixes would be predicted to be unproductive.

Based upon the hapax data that we saw previously, we know that this is not the case. We propose that employing the Cumulative Root Frequency will allow us

to characterize the relationship of relative productivity between an affixes derived forms and its underived forms. This means that we will adjust the calculation of the parsing line from derived and underived surface forms to calculating them using the Cumulative Root Frequency, labelling this approach as the Cumulative Root Type Ratio.

Given this proposal, we have a quantitative measure for calculating the tendency for affixes to have autonomous representations. We can then use these measures as the dependent variables in a study with a more certain measure associated with affix productivity. For our purposes, we measure the adjusted Cumulative Root Ratio calculations for Swahili.

In this subsection, we have introduced the Cumulative Root Ratio model of productivity in Swahili. In the next subsection, we describe a study for evaluating the validity of this model using the Helsinki Corpus of Swahili (Hurskainen 2004). We do this by evaluating the relationship of the Cumulative Root Ratio values to the hapax data in non-overlapping subsections of the corpus.

3.3.4 Evaluating the Cumulative Root Ratio in Swahili

In this subsection we describe a study to test the validity of the Cumulative Root Ratio model. This study uses corpus statistics from the Helsinki Corpus of Swahili to evaluate the relationship between the Cumulative Root Ratio and hapax counts for each of the affixes. We do this by calculating whether for each affix the raw frequency of hapaxes, and Baayen's (1991) P and (1993) $P*$ can be predicted from variables extracted from the Cumulative Root Ratio model.

To do this, we investigate the affixes found in each of the systems described above. Specifically, we took the affixes found in (i) the verbal derivation system (ii) the nominal derivation system, in the (iii) verbal inflection system, and in (iv) the nominal inflection system. This amounts to 48 affixes in total.

For each of these affixes, we extracted all types and their associated surface forms. For the inflectional system, this was a single form, but for the derivational form, a type was determined by counting the inflected forms associated with a single derivational type. Then we extracted affixless forms, and calculated all surface forms that contained the root. Based upon these frequencies, we were able to calculate relative frequencies for our underived and derived forms.

Since we were working with limited data, we split the Helsinki Corpus of Swahili into 5 randomly selected sub-units. We then performed a k-folds analysis on these non-overlapping subsections. We did this by iterating through the k number of folds of the corpus (five). In the one fold (n) we calculated, for each affix (i) the proportion of derived forms whose frequencies located them above the parsing line, and (ii) the slope and (iii) y-intercept of a least trimmed squares regression line. Recall that these variable are all associated with the degrees of affix productivity in the Parsing and Productivity model. In the next subsection of the corpus ($n + 1$), we measured the log number of hapaxes, as well as the two other descriptors of Productivity (P, and P* from (Baayen & Lieber 1991; Baayen 1993a)[5]).

Data were all collected using the existing tagset in the Helsinki Corpus of Swahili. We were unable to isolate three of the verbal derivational affixes, as they are not

[5] $P = \frac{the count of Hapaxes of an affix}{token frequency of all forms of the affix}$, and $P* = \frac{the count of Hapaxes of an affix}{total count of hapaxes in the corpus}$

tagged in the corpus (Hurskainen 1996). Using the tagset, we extracted the derived types for each affix. We then extracted the underived forms by using various regular expressions. This gave us our underived root and derived substrings, which themselves should not have surface forms but are split across different inflection and derivational forms. In order to capture the cumulative root frequency, we did substring matches to see if the underived and derived roots were contained within a given word form. If so, we added the frequencies of these word forms as part of the underived and derived frequencies.

Given these data, our predictions are as follows. If the Cumulative Root Ratio model is on the right track, then we would expect a significant correlation between the variables extracted from fold n and those found in fold $n + 1$. This means that the ratio extracted using the Cumulative Root Ratio would accurately correlate to the hapax data in another non-overlapping section of the corpus. If the contrary were true, then we would need some other way of explaining the data.

3.4 Cumulative Root Ratio Study

In this section we present the results of the k-folds study. Recall that we expect to see a correlation between the variables in fold n to the variables in fold n + 1 if the Cumulative Root Ratio model is on the right track. We first present data from the nominal and verbal derivational suffixes, and second present the results of the nominal and verbal inflectional prefixes. Last, we show the data for all affixes combined. Each of these results are displayed in correlation matrices, which include correlation

outcomes for the variables from each fold. Last, we give individual correlation plots for all affixes to highlight the results.

3.4.1 Derivational Suffix Correlations

In order to analyze our data, we present a correlation matrix containing the values of the variables found in fold $k = n$, and $k = n+1$ for all five folds. First we present the derivational affixes, found in Figure 3.6. In this figure, we see a correlation matrix between all six variables (i.e. the cumulative root ratio of parsed types, the slope, and y-intercept in fold $k = n$, and the log hapax count, P, and $P*$ in fold $k = n+1$). The correlation matrix gives the r^2 correlation between each variable across the different cells of the matrix. The cells which are outlined in black represent the intersections of the variables from the different folds. Those outside the black cells represent the correlations between the variables in the same folds, and are therefore of less interest [6]. The color associated with each correlation is given in the index below the matrix, and the degree of significance is given by the number of stars in each cell[7].

If we look at the cells within the black boundary, we see the correlations between the variables in fold n and $n + 1$. Here, the Cumulative Root Ratio, and the slope and y-intercept represent the first, second and third columns. The log hapax count, P and $P*$ represent the three rows. We see that in all instances there is a positive r^2 correlation between all variables between 0.1 and 0.3. There are significant correlations for the log hapax counts for all three variables in fold n. For P, we see a

[6] We would definitely expect P and $P*$ to correlate in the same corpus, and therefore the results are not terribly interesting.

[7] These plots were created using the Corrplot Package (Wei & Simko 2017) in R (R Core Team 2016).

significant correlation only with the slope, of the line and correlations approaching significance for the ratio and y-intercept. For $P*$ all variables in fold n are significant.

These results suggest that for the nominal and verbal derivational suffixes, the Cumulative Root Ratio model does correlate to the hapax data in non-overlapping subsections of the corpus.

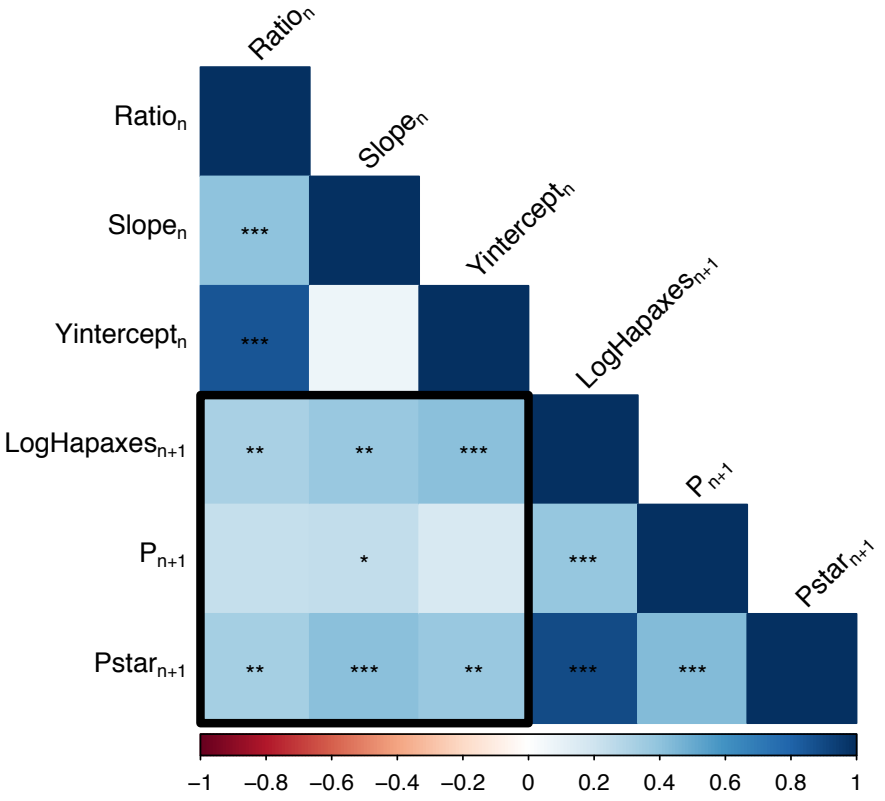

Figure 3.6: This figure contains a correlation matrix of the variables associated with the CRR from fold n, and the descriptors of productivity from fold n+1, for the derivational affixes. The color relates to the value of the r^2, and the asterisks denote the degree of significance. The black rectangular box surrounds the variables that are of interest, since they stem from the different folds.

3.4.2 Inflectional Prefix Correlations

Let's now, turn to the inflectional system. The correlation matrix of the inflectional verbal and derivational prefixes can be found in Figure 3.9. As in the derivational system, we see a significant positive correlation between the log number of hapaxes, P and $P*$ in fold $k = n+1$, and each of the variables associated with the Cumulative Root Ratio model in fold $k = n$. Again, these correlations are found in the cells in the black box, in a fashion identical to the previous figure.

This result suggests that in the inflectional system, there is a direct relationship between the hapax asymmetries that we saw previously and the Cumulative Root Ratio.

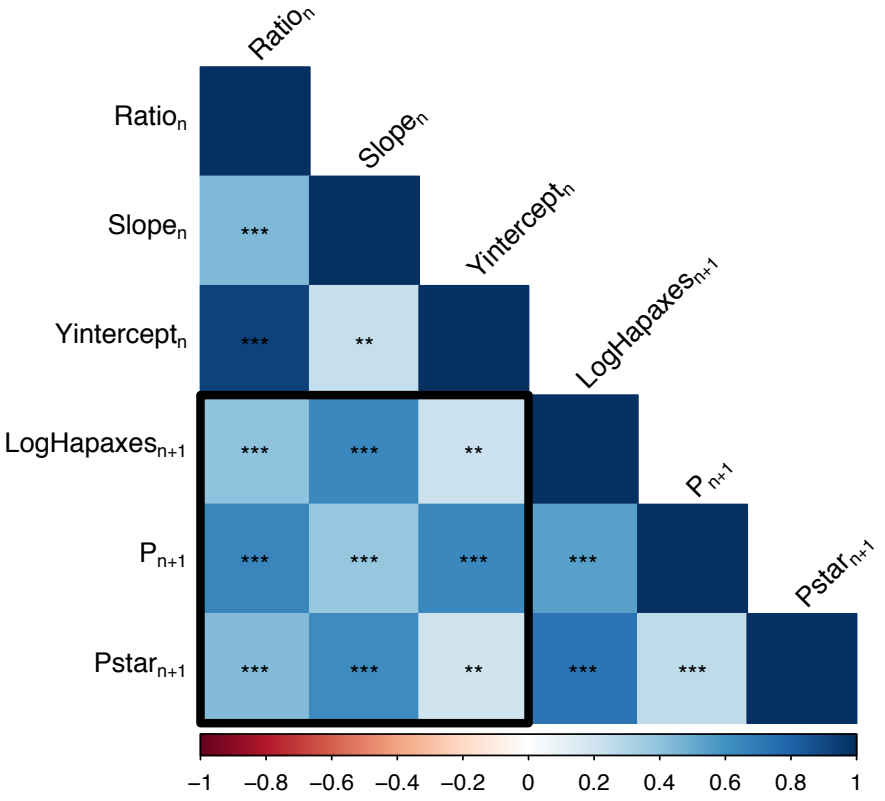

Figure 3.7: Again, This Figure contains a correlation matrix of the variables associated with the CRR from fold n, and the descriptors of productivity from fold n+1, but for the inflectional affixes.

3.4.3 Correlations for all Affixes

Given that this is true for both derivation and inflection individually, we now examine whether this correlation holds for all affixes across the two systems. In Figure 3.8 we give the results of all affixes combined.

Here, we again see a positive significant correlation for all variables in fold $k = n$ and fold $k = n + 1$. These results suggest that both inflectional and derivational affixes are held to the same productivity effects. This indicates that there is no categorical difference between the productivity of derivational and inflectional affixes, but rather that their productivity rates occur on a continuous spectrum. This supports the idea that the productivity of morphological processes are subject to the same frequency restrictions, regardless of function. This result furthermore, confirms that the Cumulative Root Ratio analysis of Swahili has potential for modeling the productivity of these processes.

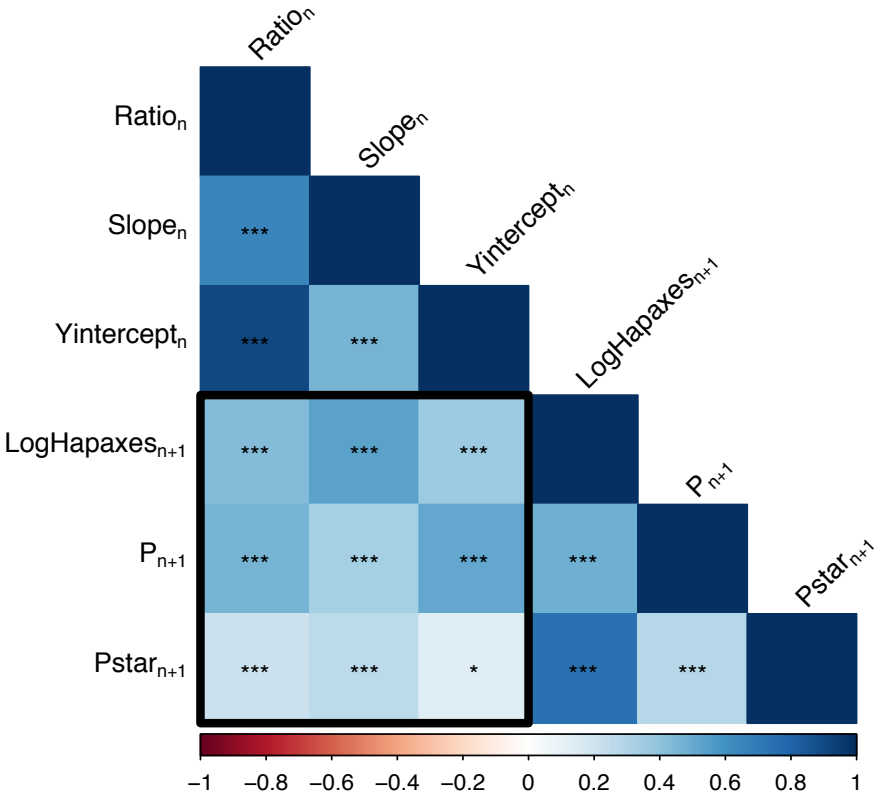

Figure 3.8: This Figure contains a correlation matrix of the variables associated with the CRR from fold n, and the descriptors of productivity from fold n+1, for all affixes.

Now that we have shown the individual positive correlations, we evaluate the correlations of the individual cells. Here, we give correlation plots for the log hapax

counts of all affixes, and all other variables. Figure3.9 contains the correlation plot of the log hapax count of all affixes in fold $k = n + 1$, and the Cumulative Root Ratio in fold $k = n$. Here, we see the plot equivalent of the top left cell in the boxed portion of Figure 3.8. The x-axis contains the Cumulative Root Ratio values for each affix across all folds, and the y-axis contains the log hapax values for those same affixes across the corresponding folds. There is a significant r^2 value of 0.1839 ($F(1, 240) = 134.1, p < 2.2e16$). This indicates that the Cumulative Root Ratio value for all affixes can positively account for over 18% of the variance of the log hapax counts in a non-overlapping but equal sized portion of the corpus.

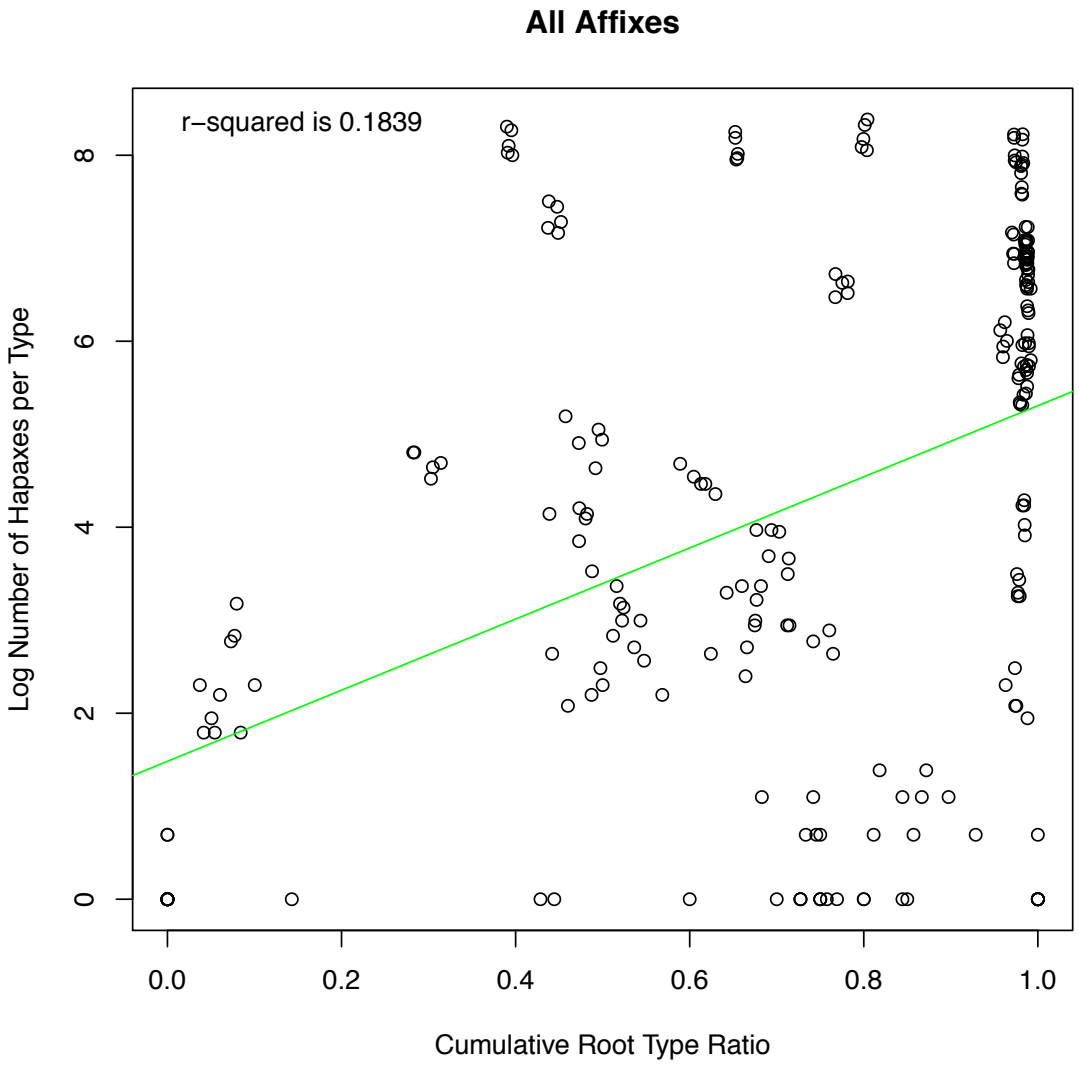

Figure 3.9: The Cumulative Root Frequency Ratio is positively correlated with the number of hapaxes in a non overlapping, but equal-sized subset of the corpus, r-squared = 0.1839, F(1,240) = 134.1, p < 2.2e16.

In addition the to Cumulative Root Ratio correlation, in Figure 3.10 we give the correlation plot of the log hapax counts of each affix and the y-intercept of the Cu-

mulative Root Ratio least trimmed squares line. Here, we see the plot corresponding to the top middle cell of the correlation matrix in Figure 3.8. The x-axis gives the y-intercept of the best fit line for the Cumulative Root Ratio in all folds, and the y-axis depicts the log hapax counts in the corresponding folds. Again, we see a significant r^2 correlation, although slightly weaker ($r^2 = 0.129, F(1, 240) = 134.1, p < 2.2e16$).

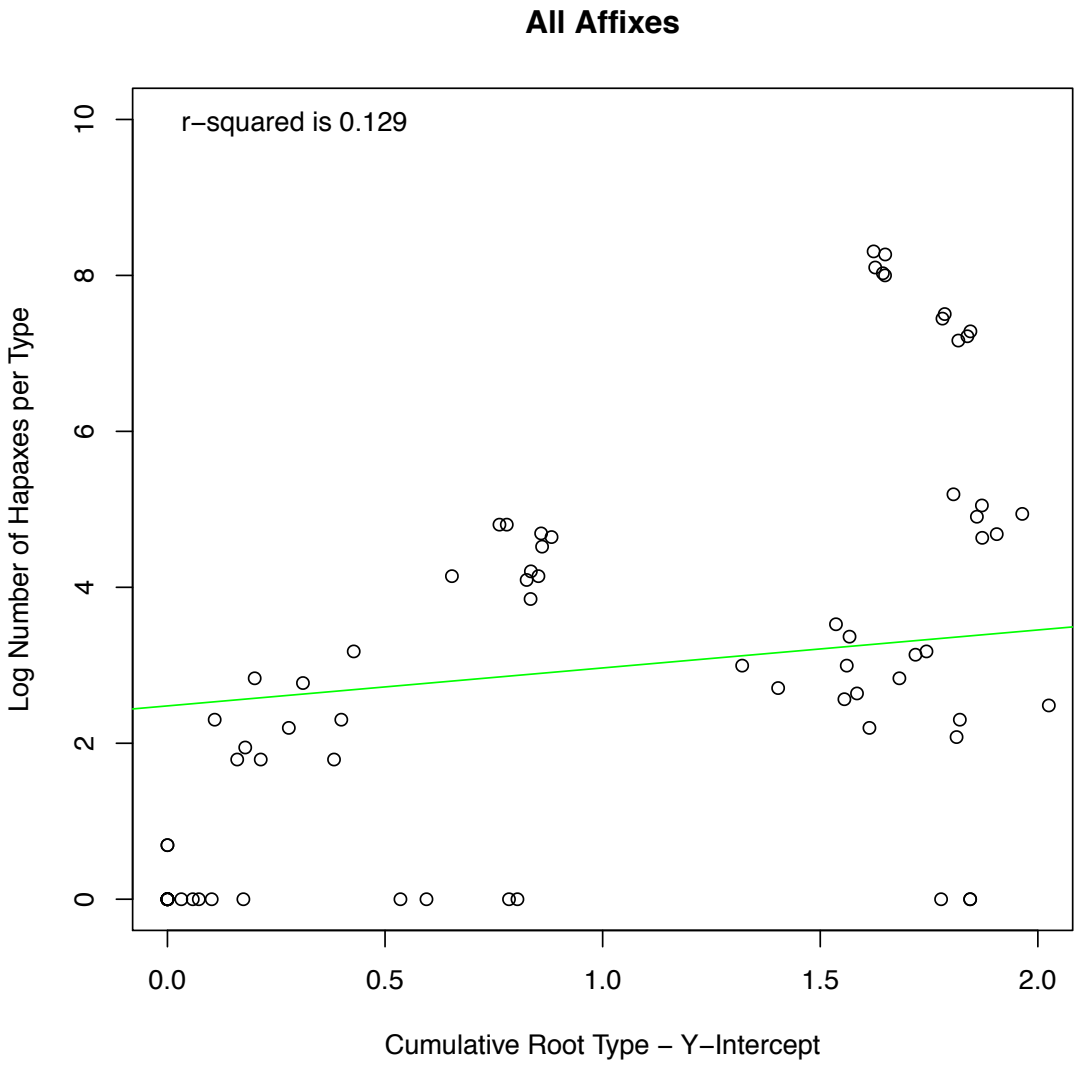

Figure 3.10: The Cumulative Root Frequency Ratio is positively correlated with the number of hapaxes in a non overlapping, but equal-sized subset of the corpus, r-squared = 0.129, F(1,240) = 134.1, p less-than 2.2e16.

Last, we give the correlation plot of the log hapax counts and the slope of the best fit Cumulative Root Ratio line for each affix. This plot can be found in Figure 3.11.

The x-axis gives the slope of the Cumulative Root Ratio best fit line for each affix across all folds, and the y-axis gives the log hapax counts for these affixes in the next fold. We see once again a positive correlation between these two variables ($r^2 = 0.2794, F(1, 240) = 134.1, p < 2.2e16$).

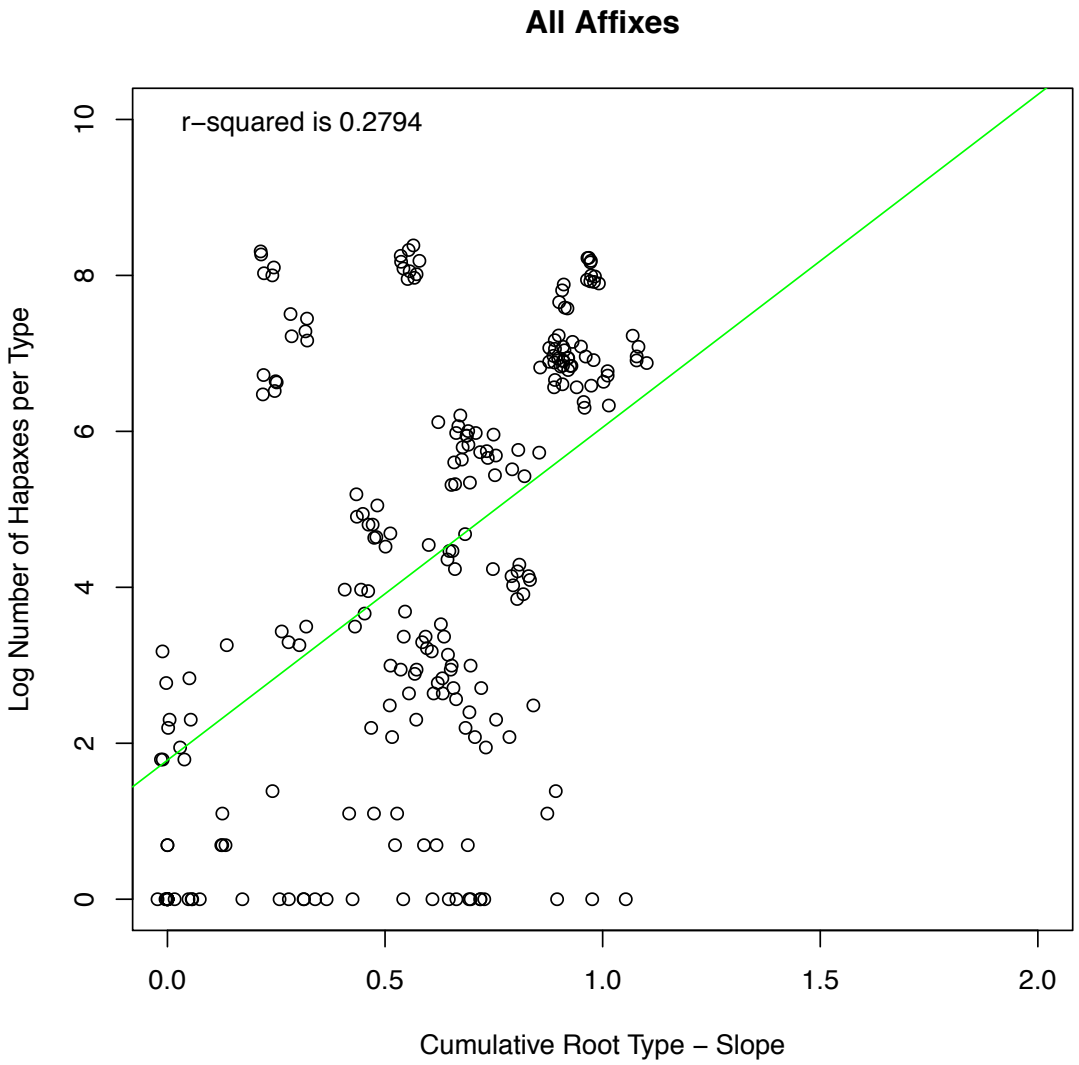

Figure 3.11: The Cumulative Root Frequency Ratio is positively correlated with the number of hapaxes in a non overlapping, but equal-sized subset of the corpus, r-squared = 0.2794, F(1,240) = 134.1, p less-than 2.2e16.

In this section we saw the results of the corpus study designed to investigate the plausibility of the Cumulative Root Ratio model. We did this by creating five

random non-overlapping subsections of the HCS corpus of 13.6 million words. We then extracted three variables from the Cumulative Root Ratio across derivation and classification classes from each subsections, or fold. We then calculated alternative variables associated with productivity that are not dependent upon parsing ratios from each subsection (hapax count, P, and $P*$). We then calculated the correlations between each of these variables between different subsections of the corpus. The results revealed that these variables are correlated across all subsections, and for all morphological systems. In the next section, we discuss these results, and discuss implications for theories of morphology.

3.5 Discussion

In this section, we summarize the findings in the Swahili study and in doing so discuss the implications of the results. Next we give a more detailed description of the characteristics of the Cumulative Root Ratio model and discuss what it can tell us about theories of morphology. Last, we discuss the implications for these results for other languages, and then propose ideas for future work.

3.5.1 Summary of Results

In this subsection, we summarize the findings of the Swahili study. Recall that Swahili is an agglutinative language with a high degree of affixation. This affixation comes from a high degree of agreement on affixes, and from nominal and verbal derivation. We then introduced the notion of morphological productivity and gave

different methods of identifying it in a language. We used the notion of the hapax to characterize the degree of productivity of a given morphological process. We saw that for each of the Swahili morphological systems, the hapax counts revealed asymmetries in predicted productivity. This prediction was problematic for a Parsing and Productivity Model which predicts that none of these processes would be productive.

We then introduced the Cumulative Root Ratio model as an alternative. Here the representation of an affix is a relation between the cumulative surface frequencies of (i) all derivational and inflectional variants containing a derived form, and (ii) all derivational and inflectional variants containing the underived form. This model therefore entails that speakers would be using intermediate representations to weigh whether to parse an affix that are constructed from multiple surface forms.

In order to investigate this model, we proposed a corpus analysis in which we generated three variables from the Cumulative Root Ratio model (Cumulative Root Ratio, and the slope and y-intercept of the best fit line describing that data). These three variables were correlated with three variables associated with productivity that do not entail a parsing perspective, but simply on the count of hapaxes of affixes. We found these variables to significantly correlate, therefore showing that the Cumulative Root Ratio predicts the degree of productivity of an affix in a non overlapping subsection of the corpus. These results imply the the Cumulative Root Ratio is a plausible model for morphological productivity in Swahili. In the next subsection, we discuss these implications in more detail.

3.5.2 Implications for Theories of Morphology

We see that the Cumulative Root Ratio predicts measures associated with productivity in non overlapping sections of a corpus. This provides us with the working hypothesis that speakers of a language like Swahili can stochastically construct affixes from parsing, so long as they are considering the cumulative frequencies of the underlying representations of these affixes, and not strict surface forms.

This data bolsters the plausibility of a systems in which a speaker relies upon the frequencies of the surface form to develop a root representation. In the area of psycholinguistics, the notion of morphological family size and root frequency has been shown to affect processing (De Jong IV *et al.* 2000; Moscoso Del Prado Martín *et al.* 2004). We propose that speakers use these data to stochastically parse word forms across all languages. In a language like English, this process is not terribly difficult due to low levels of affixation, however in Swahili where there is a considerably greater degree of affixation, this process is much more common.

In Figure 3.12 we give a schema for how this process would work. In this figure we see a bracketed form of the Swahili nominalizing suffix [-*aji*], which never occurs in isolation. We propose that the representation of this affix is a function of the set of all underlying word forms which contain the affix (=: $\{d|d \in D_{-aji}\}$). For -*aji*, this means that its representation, and likelihood of being parsed is subject to the representations of all of its types, as they compare to the representations of all corresponding underived types (=: $\{u|u \in U_{-aji}\}$).

In turn, these representations are composed of all the surface variants which contain these types. This is shown in Figure 3.13. Here, the representation for

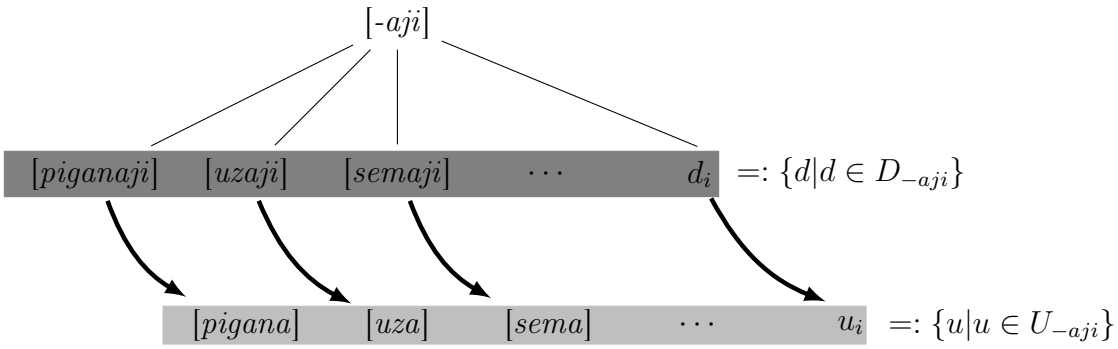

Figure 3.12: This graph shows the relation between underived and derived forms of the affix -aji in Swahili The representation for [-*aji*] is composed of all underlying forms containing the affix *aji* (D_{aji}). For each of these "derived" wordforms, there is a corresponding "underived" form. We call this set U_{aji}. The representation of this affix is subject to the relative representations of these derived and underived underlying forms.

piganaji is composed of the frequencies of the surface forms of all of its inflectional variants.

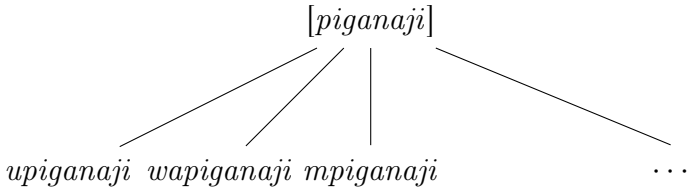

Figure 3.13: This graph shows the relation between an underlying forms and its surface forms. The representation of an underlying form is a function of the frequencies of all of its inflectional variants that occur as surface forms.

Under this model, we extend the Parsing and Perception models to deeper levels of affixation and in doing so claim that speakers of morphologically complex languages cannot simply be parsing surface forms in perceptions, but must be weighing the relative representations of underlying forms.

3.5.3 The Cumulative Root Ratio model across languages

Beyond Swahili, we propose that the Cumulative Root Ratio model is advantageous because it can be extended to other languages. First, let's reconsider the English data. Recall that English does have surface forms that are underived, and this supports the idea that speakers are weighing surface forms to determine affix representations. However, these forms in English are inflected for lexical category, but are simply lacking in overt morphology. We propose that if we were to recalculate the English data using the Cumulative Root Ratio, then we would get similar effects to what we found in Swahili. Namely, if we considered the *google* to *googler* example, then we propose calculating the parsing ratio by taking into consideration the frequencies of all inflectional variants of *google* (i.e. $google_N$, $google_V$, *googles*, *googling*, etc.) to all inflectional variants of *googler* (i.e. *googler* and, *googlers*).

3.5.4 Implications for theories of Morphology

If the Cumulative Root Ratio correlates with novel word formation, then we need to consider this implication for theories of word storage. Recall that we described three different perspectives regarding the production and storage of morphologically complex word forms. First, we saw that in rule based theories, a word and its root should be stored in all cases as separate morphemes (Bloomfield 1933; Chomsky & Halle 1968; Aronoff 1976). Second, we saw theories which suggested that words are stored as a single unit in a paradigm (Zwicky 1985; Anderson 1992; Blevins 2006; Finkel & Stump 2007). Last, we saw a model which suggested that storage was based upon frequency relations between a complex form and its simple related form

(Baayen *et al.* 1997; Coltheart *et al.* 2001). Given, these different models, what does the Swahili indicate?

Recall that the Cumulative Root Ratio is based upon groupings of surface forms that are morphologically related, and not specific underived and derived types. If this model is correct, then it would indicate that the representation of affixes in Swahili are determined based upon frequency ratios between various complex forms. Whether these word forms are stored whole or as multiple units, would then be subject to frequency relations, as is the case in the Dual Route Model. However, these relations are not as simple as previously thought, but are relevant at each level of affixation.

3.5.5 Future Work

In this paper we have proposed the Cumulative Root Ratio model in order to explain productive asymmetries in Swahili morphological processes. We have shown that this model predicts these asymmetries in a natural language corpus. We think this model can be evaluated and extended in four ways. First, we would like to perform a more in depth analysis by having a more complete tagset in Swahili, and thereby including all verbal derivational suffixes. Second, we would like to extend this model to predict novel word formation in the wild, for example on word forms in Twitter. Third, we would like to expand this model by evaluating it in other language like Arabic, and English. Last, we would like to test this using behavioral data.

Chapter 4

Is Word Derivation Limited by Semantic Cohesion?

4.1 Introduction

In English, words are often composed of a root and other subword units that are commonly referred to as affixes (Bloomfield 1933; Chomsky & Halle 1968; Aronoff 1976). These affixes can occur on the left edge of a word as a prefix (e.g. *re-* in *redo*), on right edge of a word as a suffix (e.g. *-al* in *survival*, or in a few limited cases within a word as an infix (e.g. *fucking* in *asbo-fucking-lutely*) (McMillan 1980). These affixes denote an operation that affects the root of the word by either altering its lexical category, or its semantic denotation. This relation is typically referred to as derivation (Siegel 1974; Seidenberg & Gonnerman 2000), and can be identified when some word has a related counterpart which lacks the affix.

In basic terms, we define derivational affixes as subword units that correspond a form field to a semantic or syntactic function such that it alters either (i) the syntactic category of the root word, or (ii) the semantic quality of the root word. An example of the first case would be English *-ness*. Roots that have affixed forms of *-ness* are adjectives when bare, and nouns when combined with the affix, e.g. *sad* is an adjective, and *sadness* a noun. An example of the second case would be *anti-* which denotes a reversal in the semantic quality of the root which it modifies, but does not change its syntactic category, e.g. *apartheid* versus *anti-apartheid*.

However, these words which we label as as *complex*[1] words may vary in the degree to which their meaning is directly related to the combination of their root and affix, even at times being unrelated to their form. This quality has been identified as an area of interest from the perspective of modeling the grammar of a language (Siegel 1974; Aronoff 1976; Kiparsky 1982; Hay 2001), and directly corresponds to the degree to which the principle of compositionality holds for a given form[2] (Pelletier 1994).

In order to define this issue with real language examples, let's consider the two complex words that vary in the degree to which they adhere to the principle of compositionality: *ledger* and *dancer*. Both these complex words contain the affix *-er* and both have a related *root* form that stands on its own, i.e. *ledge*, and *dance*. To demonstrate the difference in meaning, we first give the definitions for *ledge* and *ledger* in (ledge.ex), and then give the definitions for *dance* and *dancer* in (4.2). Definitions for these words come from the online lexical semantic database WordNet

[1]We define a *complex* word as a word that contains a derivational affix, cf. a *simplex* word that only contains a single word root. e.g. *complex* painter cf. *simplex* paint.

[2]This principle states that the meaning of a complex unit is a function of the meaning of its parts and the rules used to combine them.

(Miller 1995).

(4.1) (a) **ledger**: *a record in which commercial accounts are recorded*

(b) **ledge**: *a projecting ridge on a mountain or submerged under water*

We can see that the definitions of *ledger* and *ledge* have no ostensible relationship based upon their definitions. While the definition of these two words are historically related and can be traced back to early forms of the term *lay*, they seemingly lack any synchronic similarity (Harper *et al.* 2001). Therefore, it does not seem to be the case that the meaning of *ledger* is a function of the meaning of its subunits, and so does not adhere to the principle of compositionality. However in (4.2), we see that *dance* and *dancer* have related meanings such that their meanings are completely predictable from each other and the meaning of dancer clearly reflects the meaning of its individual parts.

(4.2) (a) **dancer**: *a performer who dances professionally*

(b) **dance**: *an artistic form of nonverbal communication*

These two examples represent two idealized extremes regarding the relationship between the meaning of a complex word and the meanings of its parts. Whereas some complex forms have a direct semantic relation to their non-affixed counterpart, others seemingly have absolutely no synchronic relation.

However, it is also the case that a word can have two different definitions that reflect these extremes, as seen in (4.3). Here, *reader* has two definitions that relate

to the word *read*, but one deviates from the compositional definition to a greater degree than the other.

(4.3) (a) **reader**:

(i) *a person who enjoys reading*

(ii) *one of a series of texts for students learning to read*

(b) **read**: *interpret something that is written or printed*

These examples highlight the notion that a word's composition must not always directly correspond to its meaning. At one extreme, there is a clear semantic relation between a word's meaning and the meaning of its components. At another extreme, there is no relationship between between the meaning of a word, and the surface level meanings of its components.

When we look within the meaning of individual complex words, what at all can we say about the meaning of the affixes themselves? When we consider the meaning of an affix, it is critical to note that affixes never occur in isolation, but only occur as members of complex forms. Meanwhile, many derivational affixes in English are associated with a clear meaning. Consider for example the prefix *re-* which means something like 'do x again', the suffix *-y* which means something like 'x like', and the prefix *non-* which means something like 'not x.' In these examples we see that these affixes have clear semantics. These meanings would be obvious enough if they were to occur on novel word forms, then speakers of English would likely be able to infer the meaning of novel complex forms which contain these affixes. This is because

these resulting forms would likely adhere to the principle of compositionality, but again we are left to wonder how this meaning is inferred and to furthermore wonder what the underlying mechanisms for this process are.

In this paper, we argue that the meaning of an affix can be modeled as a stochastically inferred semantic form made from the set of words containing that affix. We argue that this meaning is determined by the degree to which the members of this set adhere to the principle of compositionality (Pelletier 1994), which itself is directly related to whether these forms are stored in the lexicon as whole units or as complex forms (Baayen 1993b; Hay 2002). When the meaning of a word adheres to the principle of compositionality, we predict that this form is stored as two separate units in the lexicon. We label the degree of adherence to the principle of compositionality as the degree of semantic class coherence.

We propose that a high degree of semantic class coherence is a prerequisite for affix productivity. This effect is mirrored in the fact that affix productivity is directly related to the degree to which the members of the affix set are stored as complex forms in the lexicon (Hay 2002). These propositions entail that the meaning of derivational affixes are subject to two properties. First, the meaning of an affix is subject to the word superiority effect(Baron & Thurston 1973; Taft 1979; Segui *et al.* 1982), which claims that the frequency of a word impacts how it is stored and accessed. Second, these proposals entail that the meaning of an affix is subject to the distributional properties of language (Wittgenstein *et al.* 1953; Givón 1986), such that representations of word forms are dependent upon the linguistic environments in which they are found.

Given this claim, we predict word forms that occur with a derivational affix to share some semantic quality with one another. For example, in order for *non-* to have an autonomous semantic denotation, it must be the case that all words which contain *non-* share the meaning of 'not x'. We refer to the idealized quality of these word to contain a unified semantic denotation as semantic class coherence. This quality is an idealized one in which a derivational affix represents a set of words in which the meaning of each member adheres to the meaning of the affix. This quality entails that the meaning of each member in this set has a compositional meaning, such that all words containing the affix share the function it describes. This perspective represents an idealized perspective in which an affix has a semantic denotation in the same way that a word has a semantic denotation, but instead of denoting a set of referents, it denotes a functions which relates different sets of referents whose meanings are compositional.

One issue that can affect the semantic class coherence hypothesis is the idea that certain members can exist in this set which do not have compositional meaning, as was the case above for *ledger*. Often this notion is referred to as semantic drift or change, and occurs naturally through dispersion or change in the semantic quality of a word over time (Fauconnier 1994; Fortson 2003). Generally, word forms tend to undergo semantic drift regardless of whether they are *complex*, such that the resulting word is not always predictable from its parts (Traugott 1989).[3] This force would work against the semantic class coherence for any given affix, because it would generate members of the set of words containing an affix which do not share the same

[3]Another example of this is the musical instrument called a *recorder*, which is not an object (*-er*) that retains information to permanent record (*record*).

semantic quality as the other members in the set. If it were the case that semantic drift occurred to a large degree for the word forms of some affix, then we would expect the semantic class coherence to be relatively low. The antithetical end of the spectrum for affix representation occurs when semantic drift is ubiquitous, and a given affix has no consistent meaning association between word forms that contain it. We argue here that the representation of an affix can tend to either extremes. When semantic class coherence is low, semantic drift is common, and vice versa.

In addition to this relationship, we argue that affix productivity is naturally related to an affix's semantic class coherence. We define productivity as the tendency for some affix to occur on a novel word form (Schultink 1961; Booij 1977; Bauer 2001). Here, we propose that productivity has two requirements, first that the affix have some autonomous form representation in the lexicon, and second that this representation relates to a semantic or syntactic function such that its members on average adhere to the semantic class coherence. This requirement does not invoke subword notions of an affix, or suggest that derivational affixes exist a priori. Rather, it entails that individual language speakers invoke these representations based upon the statistical properties of the meanings of forms which contain the affix, and the degree to which they on average correspond to some function or meaning[4]

In this paper, we define the forces that contribute to affix meaning by investigating the relationship between affix representation, root representation and complex word representation as they relate to the semantic class coherence and measures of

[4]The give and take between the semantic class coherence and semantic drift has been argued to correspond to the tendency for a affix to be productive previously(Hay 2001; Hay 2004; Cotterell & Schütze 2017). Hay (2004) observed that when an affix is relatively productive, its word forms have fewer definitions on average.

productivity. Our first step into defining this relationship is to introduce a method of quantifying the variation between the semantic coherence of two sets. We show that word vector space models (Mikolov *et al.* 2013; Pennington *et al.* 2014) (automatically derived semantic representations of words based upon their context) allow us to capture semantic and syntactic relations shared between derivational affixes (Cotterell & Schütze 2017). While vector space models have been used for countless applications in NLP, this paper employs these models in order to investigate the lexicon as it relates to derivational affixes in English. Next, we relate these vector space models to affix representations by showing how they relate to semantic drift and productivity. We propose that derivational affixes should have greater semantic class coherence than non-productive affixes, in line with earlier studies which evaluate the relationship between productivity and semantics (Hay 2001), and other which show that productivity is related to semantic similarity in word vector space (Cotterell & Schütze 2017).

The structure of this paper is as follows. In section two, we give a description of our model of the lexicon. In section three, we develop a model for the semantic class coherence by using semantic measures in a word vector space model. In section four, we describe three studies: one which establishes the the efficacy of using word vector space models to model the semantic class coherence, a second which relates the semantic class coherence to semantic drift, and finally a third which relates the semantic class coherence to measures of productivity. In section five, we give a discussion of the results of these three studies, and describe how it relates to models of the lexicon.

4.2 Models of the Lexicon

In this section we give a description of a few models of the lexicon as they relate to complex word forms, and how they relate to the semantics complex words. In the end, we adopt a model of the lexicon that allows for the meaning of complex words to be optionally compositional, or non-compositional.

Generally, linguists argue for three positions regarding the mental lexicon. First, generative models argue that *complex* words are composed of stems and morphemes, and that their composition is rule based (Bloomfield 1933; Chomsky & Halle 1968; Aronoff 1976; Chomsky & Halle 1991). Essentially, this position holds that *complex* words are not stored in the lexicon. Rather, a complex form is generated via the application of a word formation rule that combines an affix morpheme and a root morpheme. Therefore, a word such as *reader* is formed by means of applying the morpheme *-er* to the morpheme *read*. More modern theories have moved away from rule based models, but still model a complex word as being a composition of individual morphemes (Halle & Marantz 1993; Harley & Noyer 1999). This general position holds that the meaning of a complex form is directly related to the principle of compositionality, and that when a form does not adhere to this principle then it undergoes what is often referred to as semantic bleaching and is lexicalized such that it is no longer complex (Sweetser 1988).

A second perspective argues that words exist in the lexicon as atomic units, and that morphological relations are discriminative differences between the words that exist in a paradigm of features (Zwicky 1985; Anderson 1992; Blevins 2006; Finkel & Stump 2007). This means that words in these models are not split up

into morphemes, but rather are stored in word schemas which relate the form and function of complex words to each other. In this way a schema is an abstraction from words that share some morphologically related quality, and is itself akin to a lexical entry. For example, in this case the prefix *re-* would be a lexical entry that is abstracted from the set of all words that contain the prefix. These words would be related to this schema in the lexicon, and its meaning would be subject to the uniformity of the schema.

Another vein of research in the area of word base models, argues that whether or not a complex word (e.g. happiness) is stored as one unit (happiness) or as multiple units (happy+ness) has to do with whether or not the component morphological units are parsed when accessed (Baayen & Lieber 1991; Baayen 1993a; Hay & Baayen 2002a). If one actively decomposes these forms during perception, then both units will be stored in the lexicon (as happy and -ness), and if one does not actively decompose them, then the form will be stored as a single unit (as happiness). Crucially, they claim that parsing is a function of the frequency ratio of the individual units. Specifically, parsing occurs when the frequency of the stem (happy) is greater than the frequency of the complex form (happiness), and does not occur when the ratio is the reverse. The model encompassing these concepts has been termed The Dual Route Model (Baayen *et al.* 1997; Coltheart *et al.* 2001) This tendency to parse has been argued to be hierarchically ordered from more separable units, which can freely attach to a base or stem (-ness), to less separable affixes which may only attach to a base (-th) (Siegel 1974; Kiparsky 1982; Hay 2004; Ingo Plag & Harald Baayen 2009) pending semantic or

syntactic selectional restrictions.

In morpheme based models, all complex words are stored as a individual units, and in the word based models complex words are stored as whole units. In a variation of the word based model, whether an affix is stored separate or as a part of a word is probabilistic based upon the frequencies of the complex words containing that affix. In this paper, we propose that semantic class coherence is related to whether complex words are stored in a similar way to relative frequencies. Namely, we propose that when a set of complex words share an affix, and this set has a relatively high semantic class coherence, then an affix will tend to have an autonomous lexical representation. When semantic class coherence is low, then the words will tend to be stored as whole units. This proposition entails that there is no a priori representation for affixes, but whether they are actually viewed as affixes is a function of the consistency of their semantics, and overlapping surface representation.

In this section we have given a brief description of a few models of the lexicon, and have opted for a model which coheres with a model in which affixes do not exist a priori, much as we do not assume that affixes in all word forms have autonomous meanings.

4.3 Modeling the semantic class coherence

The aim of this section is to introduce the notion of the word vector and how it relates to natural language. Next, we provide a mechanism that allows us to measure the semantic class coherence using these models.

4.3.1 Word Vector Space Models

Word vector models have increasingly become more and more used in Natural Language Processing application over the past five years since their initial development to help in Information Retrieval applications (Collobert & Weston 2008; Mikolov *et al.* 2013; Goldberg & Levy 2014; Pennington *et al.* 2014). Informally, word vector representations are created by generating a representation for words based upon their context in very large corpora in order to derive a discrete yet quantified representation for that word. Whereas word vector representations are derived automatically by projecting the context of some word into a neural network, human annotated semantic databases like WordNet (Miller 1995) require hours of hand coding much like a dictionary. Therefore, these models have been extremely successful in the area of NLP because they take little time to develop and nearly as good in relating word meaning (Nematzadeh *et al.* 2017). This means also that these representations can be trained on any corpus and are thus reproducible and useful for any language given some corpus.

Recently, more research has come about that employs word vector space models to model the semantics of language in order to investigate language itself, and not to develop language based applications. One area of research has investigated the relationship between word vector space models and semantic drift by investigating how the meanings of individual words have changed over time (Hamilton *et al.* 2016a) as well as investigating gender biases in word usage (Garg *et al.* 2018; Hamilton *et al.* 2016b).

One recent paper also attempts to model the degree to which word internal com-

positionality impacts meaning using word vector space models (Cotterell & Schütze 2017). They use recurrently neural nets to show that this processes is different for individual tokens in a language, and that for productive affixes, individual examples of derived forms are closer to one another in vector space. In this work, we expand on this idea by considering the closeness of the overall class, and by relating it to semantic drift and productivity more generally. Whereas Coterell (2017) does not go as far as to make claims about the human lexicon, we expand on his work to argue that these meanings are related lexical storage and semantic class coherence.

We employ these model in order to evaluate semantic class coherence because the vector representations allow us to make semantic comparisons between words beyond classical definitions such as synonymy and antonymy (Stanojevic 2009). Word vectors are coordinates in multidimensional space that are distilled as the principle components of the values of different layers in a neural network that is trained on a corpus with word context as the features. These multidimensional representations preserve the essential components of context, and allow us to relate words in terms of their semantics in two ways.

First, we can take a word vector and find the other vectors that are most closely related to it. It has been shown that when we consider the most similar vectors, we find that the nearest neighbors are semantically related words. For example pennington (2014) shows that for the term *frog*, the most closely related vectors are the ones that represent the terms *toad*, *litorea* and *leptodactylidae* (Pennington *et al.* 2014). This indicates that synonymy is captured in these models.

Second, it has been shown that when a pair of word share a semantic relation, this

relation is consistent across word forms of which share a similar semantic relation (Mikolov *et al.* 2013). For example, the relation between *man* and *woman* is roughly equivalent to the relation between *king* and *queen*. This characteristic is show in Figure 4.1. In this Figure, each term is represented as a node in the three dimensional space. The relation between *man* and *woman* is depicted as a line between the two. We see that although *king* and *queen* exist in a different are in the three dimensional space, the angle and direction of their difference is roughly equivalent to that of *man* and *woman*. This is represented by the two parallel lines that describe their difference in this space.

4.3.2 Cosine Similarity

So far, we have given a brief description of how these models work as well as a description of the sort of relations held within them. Here, we describe how these relations are calculated.

In word vector space models, the similarity or difference between words is quantified by calculating the cosine similarity between their vectors. That is, for each word there is a list of 200 coordinates that describe their placement in vector space, and we compare the distance between a these coordinates to those of another word by calculating how similar these two sets coordinates are.

We calculate cosine similarity using the formula in (4.4). In this formula, X and Y represent two vectors. The cosine similarity is a measure that calculates the cosine of the angle between the two vectors X and Y. This measure calculates the orientation of X and Y and not their magnitude by normalizing their length across the different

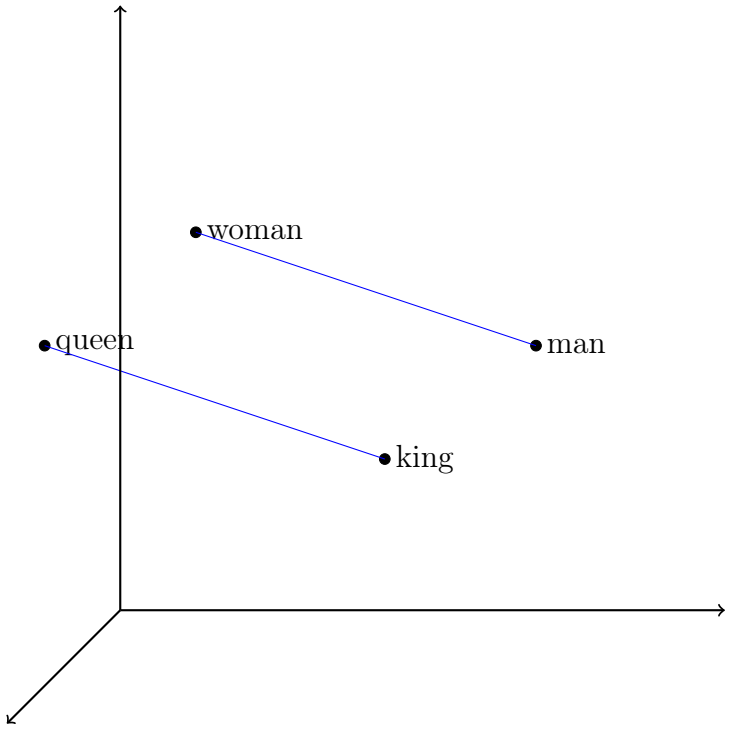

Figure 4.1: This graph depicts a three dimensional representation of 200 dimensional word vectors for the terms *man*, *woman*, *king* and *queen*. In this vector space, we can see that the distance and direction between *man* and *woman* is represented as a blue line, and that this line is parallel to the one which depicts the distance and direction of the vectors *king* and *queen*. This example serves to show that the semantic relation between these terms (i.e. masculine versus feminine) are captured in the word vector space.

dimensions.

$$\text{cosin(X,Y)} = \frac{X \cdot Y}{\sqrt{(X \cdot X)(Y \cdot Y)}} \tag{4.4}$$

The value for (4.4) is calculated by finding the dot product for X and Y (numerator), and normalizing it by the product of their magnitudes (denominator). In this denominator, each vector is projected onto one another to normalize their length and magnitude. Using this measure, we can calculate the similarity between two vectors.

Figure 4.2 gives grafts which depict the sorts of cosine relations that hold between two vectors X and Y. When two vectors are identical in orientation (i.e. their representations are identical) then the cosine similarity is equal to one. When two vectors are completely orthogonal in orientation (i.e. their representations are completely unrelated) then the cosine similarity is equal to zero. When two vectors are completely opposite in orientation (i.e. their representations are totally opposite) then the cosine similarity is equal to negative one.

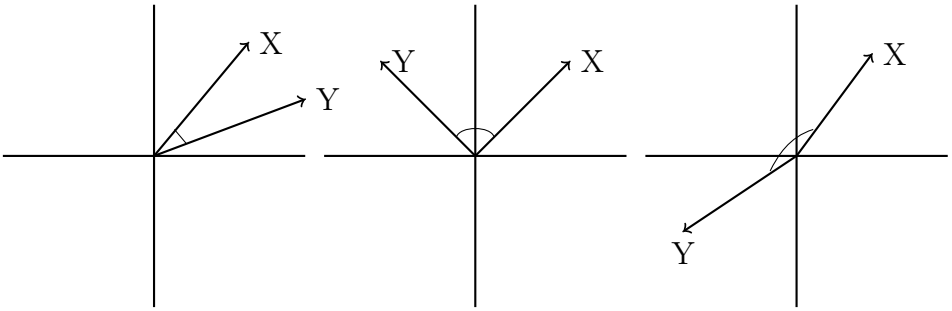

Figure 4.2: These plots depict three different cosine similarity relations. In the leftmost ploy, we see two vectors (X and Y) whose cosine similarity is approaching 1. Two words which share this relation will be synonyms. In the center, we see two vectors which are completely orthogonal because their angle of difference is 90 degrees. Consequently, the cosine similarity between X and Y is zero. Two words which share this relation should have no semantic relationship. Last, on the right we see two vectors which are in direct opposition. Therefore, the cosine similarity between X and Y in this instance is approaching -1. In this case, X and Y would be antonyms.

The plots in Figure 4.2 show three different plots depict the case when two words are similar (left), when they are related (center), and when they are in opposition (right). When these happen, the cosine similarities are close to one (left), close to zero (center), and close to -1 (right). This value allows us to compare how similar two words are in the semantic word vector space.

Although this calculation gives us a way to compare two words, our hypothesis attempts to measure an effect that is distinctively more subtle than those described above. In order to test the semantic class coherence we cannot simply calculate cosine similarities between vectors, but must rather develop a method of comparing sets of vectors. Recall that the semantic class coherence is a claim about the coherence of a set of words, and therefore requires us to calculate an average similarity between a set of words. In the next subsection, we present a way to measure the coherence of a set of vectors.

4.3.3 Average Cosine Similarity

In this subsection, we present a method for measuring the semantic similarity of a set of vectors in a word vector space model.

In order to investigate the semantic class coherence, we propose the following method. For each affix a_i, we collect all the vectors of each word w_j that contain the affix. The formula in (4.5) gives a description of this set. This formula is a mathematically precise way of modeling the set of all affixes, and the subsequent word forms which are members of the sets that form these affixes.

(4.5) $w_j^i \quad := \quad$ word j for affix i

For each affix, we take all the vectors within this set and calculate the mean point between them. We do this by calculating the mean point between all these vectors as seen in (4.6). In this formula, we sum all of the vectors of the words within a given affix set an normalize them by the size of the set.

$$\bar{w}^i_j := \frac{1}{a_i}\sum_j w^i_j \tag{4.6}$$

Based upon the formula in (4.6), for each affix we have a mean vector and the set of vectors in the affix set. Now, we can measure the coherence of this set, or semantic class coherence, by calculating the average cosine similarity of the set. We do this by taking every vector of an affix set and measuring the cosine similarity to the mean vector. This value is then averaged across all the members of each affix set, as seen in (4.7).

$$\text{cosin}(w^i_j,\bar{w}^i) = \frac{\frac{1}{a_i}\sum w^i_j \cdot w^i_k}{||w^i_j|| \quad ||\bar{w}^i||} \tag{4.7}$$

Using the formula in (4.7), we have a mechanism for calculating the semantic class coherence which we call the average cosine similarity. The resulting values will be between negative one and one. The greater the value for the average cosine similarity, the more similar the vectors within a set. When average cosine similarity is low, the vectors of a given set are less coherent. This measure captures how close these words are in vector space [5].

In this section we have introduced the idea of the word vector space model and described the cosine similarity measure. Next, we introduced the average cosine similarity measure of a vector set. We propose that this value is a way to model the semantic class coherence of a set of words. When average cosine similarity approaches one the semantic class coherence totally coherent, and as the average cosine similarity approaches zero the semantic class coherence is totally incoherent.

[5]This measure does however make the assumption that all vectors are contained within a single cluster

4.4 Three Studies

So far, we have claimed that the denotation of a derivational affix is determined by the semantic coherence of the set which contains that affix. An affix set is coherent when word forms within that set tend to have clear compositional semantics. We then suggested that this notion, which we labeled semantic class coherence can be measured using the average cosine similarity of the word vector representations of all forms within an affix set. In this section, we test these claim using three studies. First, we present a study that investigates the efficacy of the semantic class coherence measure. Second, we present a study which investigates the relationship between semantic drift and semantic class coherence, and last we present a study comparing two productivity measures with the semantic class coherence. We will see in study one that the average cosine similarity does capture the degree of coherence, in study two that semantic class coherence is inversely related to semantic drift, and in study three the semantic class coherence is directly related to two productivity measures of an affix.

In all studies we us 79 different derivational affixes of English. These affixes are split between prefixes and suffixes, and are identical to the ones used in prior studies (Hay & Baayen 2002b). For each study, we measured the Average Cosine Similarity of each affix using pre-trained word vectors sources from GloVe (Pennington *et al.* 2014). These word vectors were trained on six billion tokens sources from Wikipedia, and the Gigaword corpus (Moore & Lewis 2010; Napoles *et al.* 2012) and are represented by 200 dimensions.

4.4.1 Study One: average cosine similarity and semantic class coherence

In this study, we investigate the usefulness of the average cosine similarity measure for modeling semantic class coherence. We do this by investigating the average cosine similarity for two different sets of words which we predict to have different semantic class coherence values. These two sets are the derived forms which contain an affix, and the related underived forms. We predict that the derived forms of an affix should have a greater semantic class coherence than the underived forms, and that this difference should be reflected in the average cosine similarity. Beyond showing the use the average cosine similarity measure, this study also serves to investigate whether affixes have a unified semantic quality in vector space.

In this study, we calculate the average cosine similarity values for all types for each of the 79 affixes. We used regular expressions to isolate the underived forms of each word form in each affix set. This entailed removing the affix from each complex word, and identifying the word form related to it if one existed. If no underived form existed, this derived type was excluded from the study. These underived forms were hand checked, and pairs that contained spurious matches were excluded. We then identified the vectors of these underived forms in the GloVe pre-trained vectors, and calculated their average cosine similarity.

The result is that for each affix we have (i) the average cosine similarity calculation of all words which compose an affix set, and (ii) the average cosine similarity calculation of all simplex roots which are related to these complex forms. If the average cosine similarity can capture the semantic class coherence, then we would

expect for the average cosine similarity value of derived forms to be greater than the average cosine similarity value of underived forms. This means that all word forms which contain an affix should on average have greater coherence than the set of their underived counterparts. This coherence would mean that the spread of the set of derived forms is smaller than the spread of the underived forms in in 200 dimensions.

In other words, if this model captures some notion of an affix (e.g. the affix -*ness*), then all the words containing that affix should be closer to one another than all their corresponding affix-less forms (e.g. the space between *happiness*, *sadness*, *crassness* should be less than the space between *happy*, *sad*, and *crass*. This shrinking of the space would reflect more similar semantics in the derived versus underived sets. On the other hand, if the average cosine similarity calculation is unrelated to the semantic class coherence or the semantic class coherence were not meaningful, we would expect their to be no difference between the average cosine similarity and the set of derived and underived vectors.

4.4.2 Study One Results

Here, we present the results of the first study. Figure 4.3 presents a graph of the average cosine similarity values for the underived and derived sets for all affixes. The x-axis depicts the average cosine similarity value of all underived forms, and the y-axis depicts the average cosine similarity values of all derived forms.

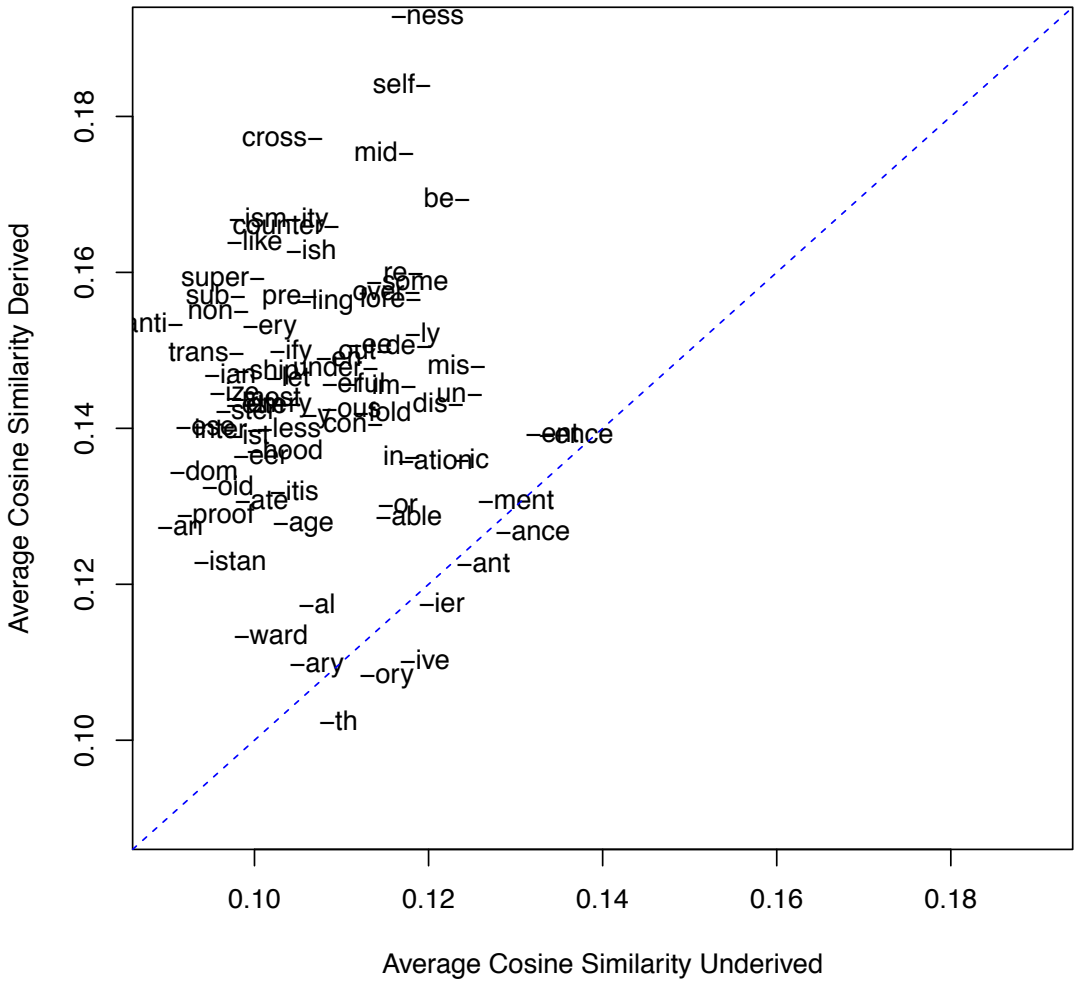

Figure 4.3: The average cosine similarity calculation for all underived word vectors is plotted against the average cosine similarity calculation of all derived vectors. For each affix above the X=Y line, the average cosine similarity is greater for derived forms than for underived forms. We conclude that derived word forms are closer to their mean, and therefore are more semantically similar than underived forms.

We see that the vast majority of affixes have a greater average cosine similarity

value for the derived set than the underived. There are around six affixes which sit on the X=Y line or slightly below it. These results indicate that for the English derived affixes, the average cosine similarity values reflect the notion that a set of derived word forms compose a more coherent set than the set of their underived counterparts. These facts suggest that the average cosine similarity measure can capture the semantic class coherence. Namely, this is because we see one set which we presuppose to be more coherent (derived) has a greater average cosine similarity than related sets which we presuppose to be less coherent (underived).

Here, we have provided results which suggest that the average cosine similarity value is an efficacious method of calculating semantic class coherence. Next, we test the relation between the semantic class coherence and semantic drift.

4.4.3 Study Two: The semantic class coherence and Semantic Drift

In this study, we evaluate the relationship between the semantic class coherence and semantic drift. Recall that semantic drift was one of the factors which we claimed would work against semantic class coherence. The more pervasive the semantic drift, the lower the semantic class coherence of the affix set.

In order to calculate the semantic drift of the affix set, we use the average number of dictionary entries of an affix set. Since we cannot directly quantify semantic non-compositionality, we are forced to use another measure of drift. The average number of dictionary entries have been used prior by others who have argued that they are negatively correlated with productivity measures (Hay 2004).

The reasoning behind using this measure is as follows. If a word is subject to semantic drift, it should have multiple meanings: one that potentially reflects its compositional meaning and one or more that represent its extension to another domain. Recall the case we saw for *reader*, as in (4.3). On the other hand, words which have compositional meaning will tend to have a single meaning. This is more like our earlier example of dance (4.2), in which there was a single compositional meaning.

In this study, we calculate the average number of dictionary entries using WordNet (Miller 1995). We did this using the by taking all derived forms found in the GloVe corpus for each affix, and looking up their definitions in WordNet[6]. We then calculated the average number of definitions for each affix set.

We predict that if the meaning of an affix is a function of the meanings of its set, then we should expect to see a frequency based relation between the count of dictionary entries and the semantic class coherence. If this is the case, then the average cosine similarity values should be negatively correlated with the average number of dictionary entries.

4.4.4 Study Two Results

In this subsection, we present the results of the semantic drift study. Figure 4.4 gives a linear model for the average cosine similarity as a function of average dictionary entry. The x-axis depicts the average number of dictionary entries per derived type, and the y-axis depicts the average cosine similarity values for the derived

[6]using the Python NLTK API (Bird & Loper 2004)

affix set. There is a significant negative correlation between these two variables ($r = -0.258, df = 77, p < 0.025$).

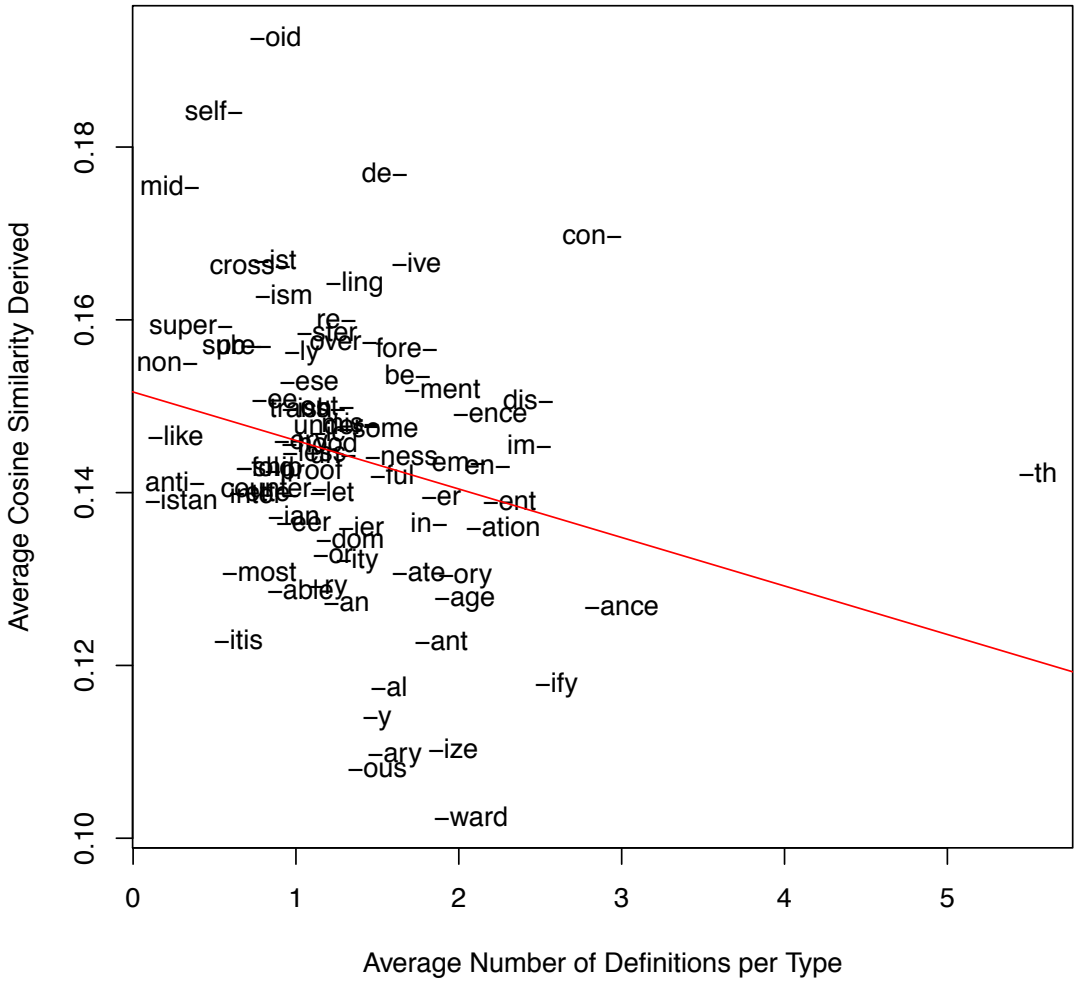

Figure 4.4: This plot depicts the relationship between the Average number of Definitions per Type, and the Average Derived Cosine Similarity to the Mean vector for all affixes. The red line depicts the r^2, with a value 0.05426. A Pearson's Correlation test reveals significant correlation (r = -0.258, df = 77, p < 0.025).

These results indicate that as the average number of dictionary entries increases, the average cosine similarity decreases. This suggests that when semantic drift is greater, then semantic class coherence is less cohesive. In the next section we evaluate the relationship between the semantic class coherence and productivity measures.

4.4.5 Study Three: The semantic class coherence and Productivity

In this subsection, we evaluate the relationship between the semantic class coherence and affix productivity. We propose that semantic class coherence and productivity are correlated for a few reasons. First, it has been proposed that productive affixes tend to have coherent semantics in prior research (Siegel 1974; Aronoff 1976; Hay 2004; Ingo Plag & Harald Baayen 2009; Cotterell & Schütze 2017). Second, when an affix is applied to a novel form, the we propose that the affix must have some coherent semantic quality in order to be available in the lexicon of a speaker. If a speaker were to use the affix -*ness* in a novel word form, then we would predict that a speaker of English must have some semantic function for this transformation, or in other words must have some semantic denotation for the word that it may possibly form.

This would suggest that whether an affix has a coherent form and meaning representation is intimately linked with productivity. Here, we test this claim by evaluating the relationship between two measures of productivity and the semantic class coherence.

Our first measure of productivity is the Productivity value P which we source

from Celex (Baayen *et al.* 1993; Hay & Baayen 2002b). This value is the number of hapax legommena of an affix divided by the number of tokens in a corpus. A hapax legomenon or hapax is an affix type that only has a single token, and has been argued to be correlated with novel word formations (Baayen & Lieber 1991).

The second calculation which we use is sourced from the parsing ratio in the Parsing and Productivity model (Hay & Baayen 2002b). This model argues that when the majority of types have a greater underived frequency than derived frequency, an affix is more often parsed in perception is therefore more likely to have an autonomous representation. The productivity of an affix is directly tied to its autonomy. One value that they use measure to calculate this value is the ratio of types that have a underived frequency greater than its derived frequency. When this value is one, all affixes are above the X=Y line, and are more likely to be productive, and when it is zero, then all affixes are below the X=Y line, and the affix is therefore more likely to be unproductive. We use this ratio, calculated using Celex as our second measure for productivity.

4.4.6 Study Three Results

In this subsection, we present the results of linear models between the two productivity measures and the average cosine similarity of the derived word forms for each affix.

In Figure 4.5, we see the linear model which depicts the relationship between the productivity value P and the average cosine similarity. The P values are depicted on the x axis, and the average cosine similarity values of the derived forms are depicted

on the y axis. While the P values mostly cluster between 0 and 0.1, we do see a significant positive correlation between the two variables ($r = 0.2365$, $df = 77$, $p < 0.04$).

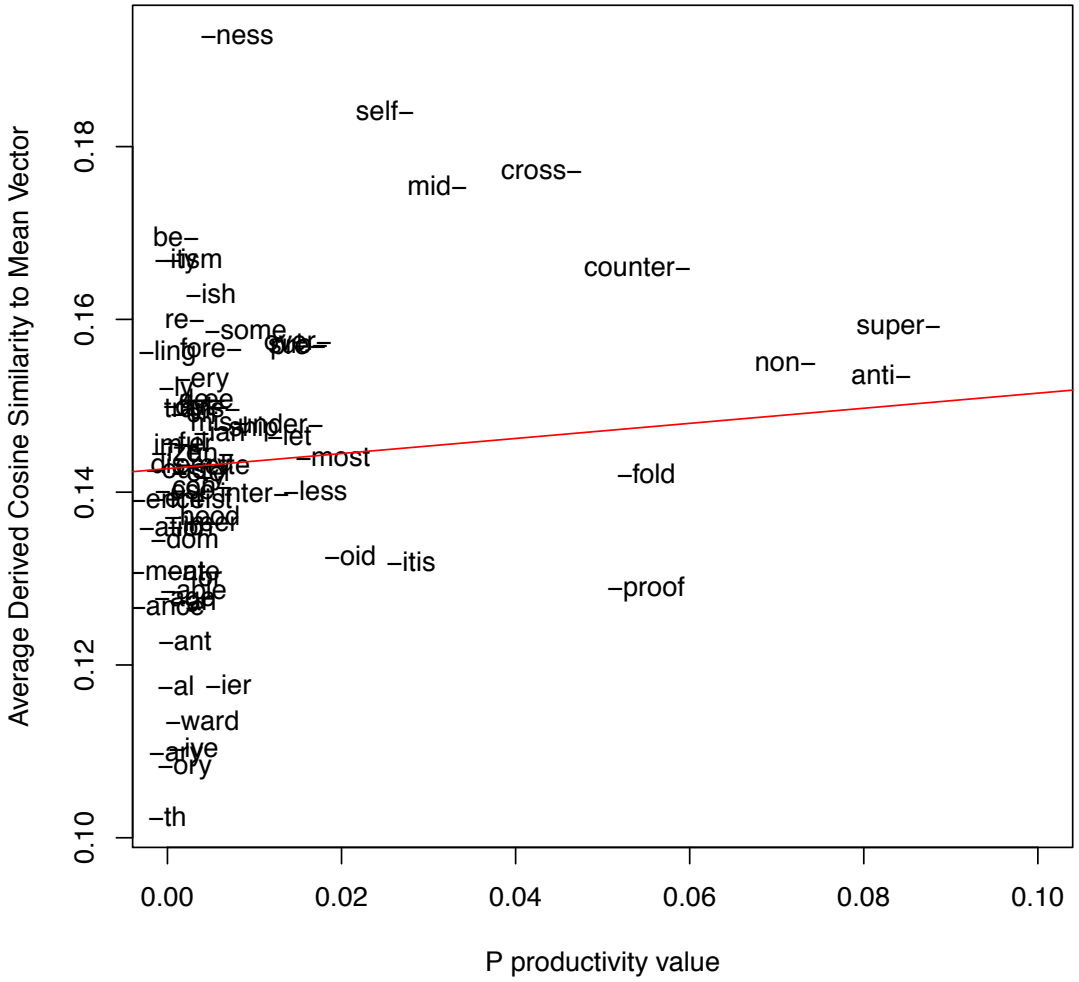

Figure 4.5: This plot depicts the relationship between the P productivity Value from Celex, and the Average Derived Cosine Similarity to the Mean vector for all affixes. The red line depicts the r^2, with a value 0.04351. A Pearson's Correlation test reveals significant correlation (r = 0.2365, df = 76, p < 0.04).

This result suggests that as the P value grows for a given affix, then there is a

greater Semantic class coherence. Next, we give the values for the type parsing ratio.

In Figure 4.6, we see the linear model which depicts the relationship between the type parsing ratio and the average cosine similarity values of the derived forms. The type parsing values are depicted on the x axis, and the average cosine similarity values of the derived forms are depicted on the y axis. This variable has a greater variance than the P value seen previously, and we see a significant positive correlation between the two variables ($r = 0.3729$, $df = 77$, $p < 0.001$).

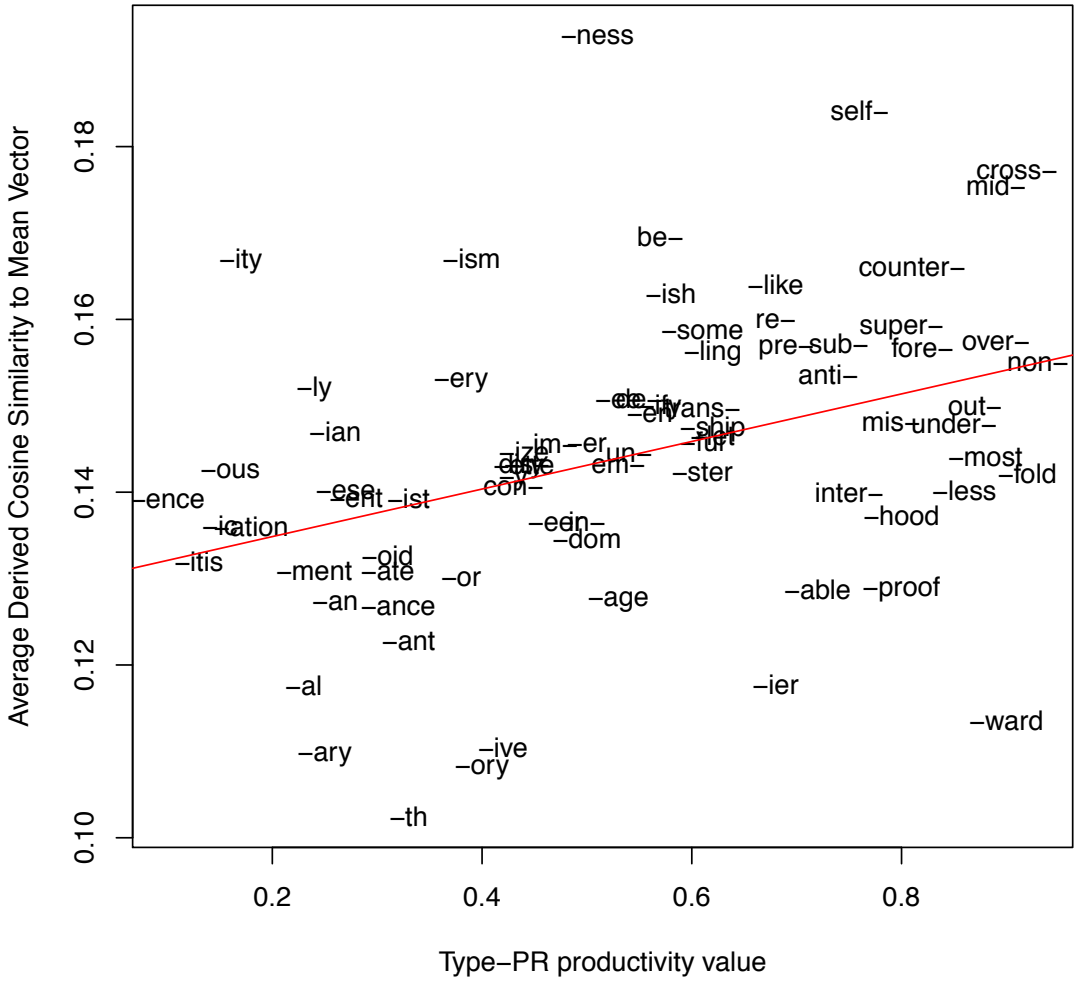

Figure 4.6: This plot depicts the relationship between the average number of definitions per type, and the derived average cosine similarity to the mean vector for all affixes. The red line depicts the r^2, with a value 0.1277. A Pearson's Correlation test reveals significant correlation (r = 0.3729, df = 76, p $<$ 0.001).

In all, these two correlations show that the semantic class coherence significantly

correlates with productivity of a given affix.

4.5 Discussion

In the previous section we saw three studies which investigated the meaning of derivational affixes in English. These studies evaluated the semantic class coherence of all the derived forms using the measure of average cosine similarity. We saw that semantic class coherence negatively correlated with the average number of dictionary entries of an affix, and that it positively correlated with two separate measures of affix productivity. These results have a few implications which we discuss in the next subsection.

4.5.1 A model of Semantic Class Coherence

Earlier, we claimed that semantic class coherence is related to decomposition rates for complex forms as they relate to lexical storage, and that this effect is mirrored in the degree to which complex word forms adhere to the principle of compositionality. Furthermore, we suggested that when semantic class coherence was high, an affix would be more productive and the set of word forms containing the affix would on average be stored as multiple units, and would have compositional meaning. When semantic class coherence was low, this would be reflected in whole word storage, and words being stored as single units. Here we describe a model that reflects this facts.

We propose that when a native speaker acquires a language, words generally are learned and stored whole units. As the speaker acquires complex forms which have

regularities in meaning in form that are separate from the meanings of individual forms, they stochastically construct representations for affixes that allow them to group these word forms semantically. As these words are acquired, the quality and consistency of their semantic coherence is reflected in the autonomy of their unified representation. This representation is determined by the compositionality of the meanings of these words, and is causally related to their use in novel environments. That is, we propose that productivity is the result of semantic class coherence at the level of the individual speaker. When a speaker has an autonomous representation of an affix employ in novel word formation, then this representation must have been inferred from a semantically consistent set which are stored as complex units.

The intermoving parts of this model are reflected in the results of the study which show correlations between compositionality, productivity, and the degree of semantic class coherence. These results suggest that the semantic class coherence is plausible effect and predicts that speakers should be sensitive to the semantic coherence of set of words which share an affix. Under this model, we propose that affixes do not exist a priori, but are representations inferred from consistent regularities in meaning and form. Speakers then construct these forms based upon these regularities. If this were not to be the case, then other models must explain the facts related to affix meaning, productivity, and semantic cohesion.

4.5.2 Semantic Class Coherence and Morphology

Earlier, we saw three perspectives of morphology which proposed three different methods of word storage and computation. The semantic class coherence model is

very much tied to the idea that subword phenomena do not exist prepackaged in the grammar, but rather that they are probabilistic forms inferred from large sets. This perspective is in line with word based models which relate storage to frequency effects.

4.5.3 Future Work

In future work, we would like to develop a ways to quantify semantic drift using cosine similarity, and then test the semantic class coherence using this model. We would also like to develop a formal analysis of the facts above in which we can describe the variance of sets in a formal ontology.

Chapter 5

Discussion

In this chapter we discuss the results and implications of the three papers, and outline the future steps for the research outlined within them. Last, we give a unifying vision of these three papers as it relates to methodology, style of analysis and linguistic theory.

5.1 Conclusions for all Papers

In this section we give a brief summary of the conclusions of each of the three papers in the dissertation.

Paper One: ***Nominal Classification decreases the Entropy of Nominalsin-GenderedLanguages***

The first paper investigated the relationship between nominal classification and language complexity, and proposed that nominal classification reduces the complexity

of nominals in gendered languages. It contained arguments for the perspective that this complexity reduction occurs in the lexicon when acquiring a language via the association model.

Paper Two: ***Are affixes in Agglutinative languages always Productive?***

The second paper investigated quantitative measures of affix productivity in the agglutinative language Swahili. We saw that the previous models which argue that affix productivity is a related to word decomposition failed to make the correct prediction for productive asymmetries in Swahili. In order to model these asymmetries, we proposed the cumulative root ratio model. This model requires that affix productivity for the Swahili speaker is a function of root frequencies, and not surface frequencies. We saw this this measure significantly correlated with other quantitative measures of productivity that do not rely upon parsing ratios. These results suggest that the productivity and therefore representation of affixes is based on regularities in the frequency relations between all inflectional and derivational variants of competing forms.

Paper Three: ***Is Word Derivation limited by Semantic cohesion?***

The third paper investigated how speakers of a language determine the meaning of derivational affixes. It contains evidence from English word vector space models which are able to model the lexical semantics of word forms. This paper proposed that semantic coherence is a prerequisite for affix productivity, and that it is determined by the degree to which the members of an affix's derived set adhere to

the principle of compositionality. To evaluate this proposition, it includes a novel measure of semantic class coherence in word vector space by measure the average cosine similarity of a set of word vectors. It then contained a comparison between this measure and semantic drift (non-compositionality of meaning), and measures of affix productivity. Average cosine similarity negatively correlated with semantic drift, and positively correlated with semantic class coherence.

In this section we have given summaries of the findings of each of the three papers in the dissertation. In the next section we given the logical next steps for the research outlined in each paper.

5.2 Next Steps

In this section, we give the future steps for the line of research described in each of the three papers. In doing this, we give brief discussions on the sorts of studies that will help to confirm or deny the findings in each of these three papers.

Paper One: ***Nominal Classification decreases the Entropy of Nominals in Gendered Languages***

This paper introduces a new model of nominal organization in which classification is used to increase the learnability of a nominal system. This association model proposes that speakers of a language tacitly employ classification as a means of organizing their mental lexicons. This proposition can be explored in a few concrete ways. First, we propose that this model be tested in artificial language learning

studies which manipulate word order and nominal classification. The model would predict that languages with nominal classification should be easier to learn and recall than those without classification for human subjects, and that this effect should hold regardless of the syntactic order of gender marking and nominals. Second, the language models in these studies should be expanded in two separate ways. These models should be expanded to a much greater number of languages which contain grammatical gender in order to investigate this effect in language more generally, and furthermore the complexity of this models can be increased in order to understand this effect in greater detail. Specifically, we propose the use of neural language models which employ machine learning to explore the role between complexity and form in classification in these languages.

Last, one thing that we did not consider in this study is the relationship between grammatical gender and other aspects of the morphological structure. For one thing, many gendered languages have affixes or groups of affixes which are strongly associated with specific genders. Meanwhile, many languages which employ nominal classification also often have semantic categories associated with this classification, as is the case in Swahili (Mohamed 2001a). We therefore propose using this method as a feature in the study of more complex systems of classification in order to understand the relationship between classification, morphological form, and semantic classes.

Paper Two: ***Are affixes in Agglutinative languages always Productive?***

This paper proposes a new model of affix productivity and in doing so makes a claim about how affix productivity correlates to the lexicon. We propose that the cumulative root ratio must be confirmed in English, and that it must be extended to other languages which have properties similar to the morphological system found in Swahili, such as Semitic language. We predict that this model will hold for these languages and will show that affix productivity in a general sense is a frequency relation related to the probabilistic nature of affix representation. Furthermore, this study should be confirmed with additional data in Swahili. We propose two different sources of data to confirm such a model.

First, novel word formation in other domains such as in Twitter data should correspond to productivity rates as predicted by the cumulative root ratio. One study could measure these ratios for affixes purely from Twitter data, and then be correlated with non overlapping examples of novel occurrences of affixed forms. In addition to monolingual data, we propose that affix usage in code-switching environments should be considered as a possible proving ground for productivity.

Last, these rates of productivity should be tested with behavioral data. One such study could explore novel word formation in the lab setting using artificial language learning techniques to see whether human subjects can be coerced to employ novel affixes in nonce forms whose frequencies are controlled throughout the experiment.

Paper Three: ***Is Word Derivation limited by Semantic cohesion?***

This paper proposes a model of morphology in which affixes and their meaning are stochastically inferred from regularities in the form and meaning of complex forms. If affixes have no autonomous representation a priori, then variation in semantic class coherence should reflect other effects in storage and novel formation. We propose that this semantic class coherence can be more accurately modeled by using more sophisticated mathematical techniques that are able to capture more subtle variation between word vectors. For example, where we use the mean vector to get a general idea of the size of the semantic space of word vectors, we could employ a more data driven approach. One such approach would be to employ gaussian mixture models (Zivkovic 2004) to model a more accurate vector describing the data, and then calculate average cosine similarity using this vector. Another such approach would be to employ the k-means clustering algorithm (Hartigan & Wong 1979) to calculate the number of groupings one can make from a class of vectors, and understand how this impacts semantic class coherence.

Last, we propose studying this effect in the lab using human subjects. If semantic class coherence is a precursor to productivity, then novel affixes that have a greater semantic class coherence should be more readily employed by human subjects than another affix whose forms do not compose a semantically coherent set.

In this section we have outlined the next possible steps in the line of research described in each of the three papers. In the next section, we describe the aspects which unify these papers regarding methodology, analysis, and linguistic theory.

5.3 A Unifying Vision

In this section, we introduce the unifying aspects of the research in these three papers. We relate them in terms of their methodology, their style of analysis, and last in what they combined can reveal about linguistic theory.

5.3.1 Methodology

These three papers each represent investigations into the internal structure of words in human language using computational methods and data driven analyses . More specifically, they focus on the effects that complexity, frequency, and semantic coherence have on the representation of word internal units, or morphemes.

Although the first paper focuses on the German, French and English, the second paper focuses on Swahili, and the third paper focuses on English, these pieces of research are intended to inform linguistic theory in a general sense. They are each designed to relate to the human capacity for language regardless of the language spoken, and therefore are attempts to understand the relationship between cognition and language.

5.3.2 Analysis

Each piece of research employs computational methods to calculate quantitative measures associated with the different phenomena under investigation. In doing so, they each fall into the realm of computational linguistics and morphology.

5.3.3 Linguistic Theory

The assumptions underlying each paper are as follows. First, each paper supposes that word level structure does not exist a priori, but rather that structure must be inferred by speakers from sets of words and consequently exhibit asymmetries in storage and productivity based upon lexical regularities. This perspective is couched in morphology more generally in the coming subsection.

Last, these papers are part of a sea change in the study of language where we have enough data to begin creating models of language that increasingly are able to interact with linguistic data approaching that of a human's experience with language over the acquisition time of a language. This quality allows us to model language as a complex system using the techniques described above, which has become another proving ground for the testing of models of language and cognition.

In summary, this dissertation describes three different computational inquiries into issues related to lingusitic morphology. These studies each represent future lines of research that will be built upon in the future using both the methods employed in this dissertation, and those discussed which can inform the phenomena explored.

Bibliography

Abdulaziz, Mohamed H, & Ken Osinde. 1997. Sheng and engsh: Development of mixed codes among the urban youth in kenya. *International Journal of the Sociology of Language* 125.43–64.

Ackerman, Farrell, & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89.429–464.

Anderson, Stephen R. 1992. *A-morphous morphology*, volume 62. Cambridge University Press.

Anderson, Stephen R. 2015. Dimensions of morphological complexity. In *Understanding and measuring morphological complexity*, ed. by Matthew Baerman, Dunstan Brown, & Greville G. Corbett, 11–26. Oxford: Oxford University Press.

Arnon, Inbal, & Michael Ramscar. 2012. Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition* 122.292–305.

Aronoff, Mark. 1976. Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass.* 1–134.

——. 1983. Potential words, actual words, productivity and frequency. In *Proceedings of the 13th international congress of linguists*, 163–171.

——. 1994. *Morphology by itself: Stems and inflectional classes*. Number 22. MIT press.

——, & Frank Anshen. 1998. *Morphology and the lexicon: Lexicalization and productivity*. Wiley Online Library.

Baayen, Harald. 1992. Quantitative aspects of morphological productivity. In *Yearbook of morphology 1991*, 109–149. Springer.

——. 1993a. On frequency, transparency and productivity. In *Yearbook of Morphology 1992*, 181–208. Springer.

——. 1993b. On frequency, transparency and productivity. In *Yearbook of Morphology 1992*, 181–208. Springer.

——, & Rochelle Lieber. 1991. Productivity and english derivation: a corpus-based study. *Linguistics* 29.801–844.

——, & Robert Schreuder. 1999. War and peace: Morphemes and full forms in a noninteractive activation parallel dual-route model. *Brain and language* .

Baayen, R Harald, Ton Dijkstra, & Robert Schreuder. 1997. Singulars and plurals in dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37.94–117.

——, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix, & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review* 118.438.

——, Richard Piepenbrock, & Rijn van H. 1993. The {CELEX} lexical data base on {CD-ROM}.

——, & Robert Schreuder. 2003. *Morphological structure in language processing*, volume 151. Walter de Gruyter.

Baayen, RH, & Michael Ramscar. 2015. Abstraction, storage and naive discriminative learning. *Handbook of cognitive linguistics* 39.100–120.

Baker, Mark C. 2008. *The syntax of agreement and concord*, volume 115. Cambridge University Press.

Barasa, Sandra Nekesa, & others. 2010. *Language, mobile phones and internet: a study of SMS texting, email, IM and SNS chats in computer mediated communication (CMC) in Kenya*.

Baron, Jonathan, & Ian Thurston. 1973. An analysis of the word-superiority effect. *Cognitive psychology* 4.207–228.

Baroni, Marco, Johannes Matiasek, & Harald Trost. 2002. Predicting the components of german nominal compounds. In *Proceedings of the 15th European Conference on Artificial Intelligence*, 470–474. IOS Press.

Bates, Elizabeth, Antonella Devescovi, Arturo Hernandez, & Luigi Pizzamiglio. 1996. Gender priming in italian. *Perception & Psychophysics* 58.992–1004.

——, Antonella Devescovi, Luigi Pizzamiglio, Simona D'amico, & Arturo Hernandez. 1995. Gender and lexical access in italian. *Perception & Psychophysics* 57.847–862.

Bauer, Laurie. 2001. *Morphological productivity*, volume 95. Cambridge University Press.

——. 2005. Productivity: theories. In *Handbook of word-formation*, 315–334. Springer.

Bechtel, William, & Adele Abrahamsen. 2002. *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks*. Blackwell Publishing.

Bell, Alan, Jason M Brenier, Michelle Gregory, Cynthia Girand, & Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language* 60.92–111.

Bierwisch, Manfred. 1967. Syntactic features in morphology: general problems of so-called pronominal inflection in german. *To Honour Roman Jakobson* 239270.

Bird, Steven, & Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, p. 31. Association for Computational Linguistics.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42.531–573.

Bloomfield, Leonard. 1933. *Language history: from Language (1933 ed.).*. Holt, Rinehart and Winston.

Bölte, Jens, & Cynthia M Connine. 2004. Grammatical gender in spoken word recognition in german. *Attention, Perception, & Psychophysics* 66.1018–1032.

Booij, Geert Evert. 1977. *Dutch morphology: A study of word formation in generative grammar*. Number 2. Peter de Ridder Press.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, & Quentin D Atkinson. 2012. Mapping the origins and expansion of the indo-european language family. *Science* 337.957–960.

Bybee, Joan. 1995a. Regular morphology and the lexicon. *Language and cognitive processes* 10.425–455.

——. 1995b. Regular morphology and the lexicon. *Language and Cognitive Processes* 10.425–455.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*, volume 9. John Benjamins Publishing.

——. 1988. Morphology as lexical organization. *Theoretical morphology* 119141.

Cardona, George. 1997. *Pānini: A survey of research*. Motilal Banarsidass Publ.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on information theory* 2.113–124.

——. 1959. On certain formal properties of grammars. *Information and control* 2.137–167.

——, & Morris Halle. 1968. The sound pattern of english.

——, & Morris Halle, 1991. The sound patter of English.

Cohen, Jacob. 1992. A power primer. *Psychological bulletin* 112.155.

Colé, Pascale, Cécile Beauvillain, & Juan Segui. 1989. On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and language* 28.1–13.

Collobert, Ronan, & Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, 160–167. ACM.

Coltheart, Max, Kathleen Rastle, Conrad Perry, Robyn Langdon, & Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review* 108.204.

Comrie, Bernard. 1999. Grammatical gender systems: a linguist's assessment. *Journal of Psycholinguistic research* 28.457–466.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Cotterell, Ryan, & Hinrich Schütze. 2017. Joint semantic synthesis and morphological analysis of the derived word. *arXiv preprint arXiv:1701.00946* .

Cutler, Anne. 1980. Productivity in word formation. In *The Sixteenth Regional Meeting, Chicago Linguistic Society*, 45–51. CLS.

de Courtenay, Jan Niecisław Baudouin. 1972. *A Baudouin de Courtenay anthology: the beginnings of structural linguistics*. Indiana University Press.

De Jong IV, Nivja H, Robert Schreuder, & R Harald Baayen. 2000. The morphological family size effect and morphology. *Language and cognitive processes* 15.329–365.

De Saussure, Ferdinand. 2011. *Course in general linguistics*. Columbia University Press.

Dennis, Nancy A, Indira C Turney, Christina E Webb, & Amy A Overman. 2015. The effects of item familiarity on the neural correlates of successful associative memory encoding. *Cognitive, Affective, & Behavioral Neuroscience* 15.889–900.

Dixon, Robert MW. 1972. *The Dyirbal language of north Queensland*, volume 9. CUP Archive.

——. 1986. Noun classes and noun classification in typological perspective. *Noun classes and categorization* 105–112.

Dressler, Wolfgang U. 1997. On productivity and potentiality in inflectional morphology. *Cross-Language Aphasia Study Network (CLASNET) Working papers* 7.3–22.

Dryer, Matthew S, David Gil, Bernard Comrie, Hagen Jung, Claudia Schmidt, & others. 2005. The world atlas of language structures.

Durrell, Martin. 2011. *Hammer's German grammar and usage*. Routledge.

Dye, Melody, Petar Milin, Richard Futrell, & Michael Ramscar. 2016. A functional theory of gender paradigms. *Morphological paradigms and functions. Leiden: Brill* .

Fauconnier, Gilles. 1994. *Mental spaces: Aspects of meaning construction in natural language*. Cambridge University Press.

Finkel, Raphael, & Gregory Stump. 2007. Principal parts and morphological typology. *Morphology* 17.39–75.

Fortson, Benjamin W. 2003. An approach to semantic change. *The Encyclopedic Dictionary of Applied Linguistics: A Handbook for Language Teaching* 648–666.

Frauenfelder, Uli H, & Robert Schreuder. 1992. Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In *Yearbook of morphology 1991*, 165–183. Springer.

Friederici, Angela D, & Thomas Jacobsen. 1999. Processing grammatical gender during language comprehension. *Journal of psycholinguistic Research* 28.467–484.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, & James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* .

Gillon, Brendan S. 1999. The lexical semantics of english count and mass nouns. In *Breadth and depth of semantic lexicons*, 19–37. Springer.

Givón, Talmy. 1986. Prototypes: Between plato and wittgenstein. *Noun classes and categorization* 77–102.

Goldberg, Yoav, & Omer Levy. 2014. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* .

Greenberg, Joseph H. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics* 26.178–194.

Grosjean, François, Jean-Yves Dommergues, Etienne Cornu, Delphine Guillelmon, & Carole Besson. 1994. The gender-marking effect in spoken word recognition. *Perception & Psychophysics* 56.590–598.

Guillelmon, Delphine, & François Grosjean. 2001. The gender marking effect in spoken word recognition: The case of bilinguals. *Memory & Cognition* 29.503–511.

Halle, Morris, & Alec Marantz. 1993. Distributed morphology and the pieces of inflection.

Hamilton, William L, Kevin Clark, Jure Leskovec, & Dan Jurafsky. 2016a. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, p. 595. NIH Public Access.

——, Jure Leskovec, & Dan Jurafsky. 2016b. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, p. 2116. NIH Public Access.

Harley, Heidi, & Rolph Noyer. 1999. State-of-the-article: distributed morphology. *GLOT International* 4.3–9.

Harper, Douglas, & others, 2001. Online etymology dictionary.

Hartigan, John A, & Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.100–108.

Haspelmath, Martin, & Andrea Sims. 2013. *Understanding morphology*. Routledge.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford University Press on Demand.

Hay, Jennifer. 2001. Lexical frequency in morphology: is everything relative? *Linguistics* 39.1041–1070.

——. 2002. From Speech Perception to Morphology: Affix Ordering Revisited. *Language* 78.527–555.

——. 2004. *Causes and consequences of word structure*. Routledge.

——, & Harald Baayen. 2002a. Parsing and productivity. In *Yearbook of Morphology 2001*, 203–235. Springer.

——, & Harald Baayen. 2002b. Parsing and productivity. In *Yearbook of Morphology 2001*, 203–235. Springer.

Heim, St, Bertram Opitz, & Angela D Friederici. 2002. Broca's area in the human brain is involved in the selection of grammatical gender for language production: evidence from event-related functional magnetic resonance imaging. *Neuroscience letters* 328.101–104.

Heine, Bernd, & Mechthild Reh. 1984. *Grammaticalization and reanalysis in African languages*. Buske Helmet Verlag Gmbh.

Hockett, Charles F. 1961. Linguistic elements and their relations. *Language* 37.29–53.

Holmes, Virginia M, & B Dejean de la Bâtie. 1999. Assignment of grammatical gender by native speakers and foreign learners of french. *Applied Psycholinguistics* 20.479–506.

Hurskainen, Arvi. 1996. Disambiguation of morphological analysis in Bantu languages. *Proceedings of the 16th conference on Computational linguistics* - 1.568.

——. 2004. Helsinki corpus of swahili. In *Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC – IT Center for Science.*.

Ingo Plag, & Harald Baayen. 2009. Suffix Ordering and Morphological Processing. *Language* 85.109–152.

Istvan, Fodor. 1959. The origin of grammatical gender. *Lingua* 8.186–214.

Jacobsen, Thomas. 1999. Effects of grammatical gender on picture and word naming: Evidence from german. *Journal of Psycholinguistic Research* 28.499–514.

Jelinek, Eloise, & Andrew Carnie. 2003. Argument hierarchies and the mapping principle. *Formal approaches to function in grammar* 265–296.

Jones, Timothy, & Simon Farrell. 2018. Does syntax bias serial order reconstruction of verbal short-term memory? *Journal of Memory and Language* 100.98–122.

Keizer, Evelien. 2007. *The English noun phrase: The nature of linguistic categorization*. Cambridge University Press.

Kilarski, Marcin. 2007. On grammatical gender as an arbitrary and redundant category. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 3* 112.24.

Kiparsky, Paul. 1982. Lexical morphology and phonology. *Linguistics in the morning calm: Selected papers from SICOL-1981* 3–91.

Kleene, Stephen Cole. 1951. Representation of events in nerve nets and finite automata. Technical report, RAND PROJECT AIR FORCE SANTA MONICA CA.

Levelt, Willem JM. 1993. *Speaking: From intention to articulation*, volume 1. MIT press.

Lewandowsky, Stephan, & Simon Farrell. 2000. A redintegration account of the effects of speech rate, lexicality, and word frequency in immediate serial recall. *Psychological Research* 63.163–173.

M. Baroni, S. Bernardini, A. Ferraresi, & E. Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43.209–226.

Matthews, Peter Hugoe. 1972. *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*, volume 6. CUP Archive.

Maw, Joan. 1976. Focus and the morphology of the swahili verb. *Bulletin of the School of Oriental and African studies* 39.389–402.

McGill, William J. 1954. Multivariate information transmission. *Psychometrika* 19.97–116.

McMillan, James B. 1980. Infixing and interposing in english. *American Speech* 55.163–183.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Miller, George A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38.39–41.

Mohamed, Mohamed Abdulla. 2001a. *Modern Swahili Grammar*. East African Publishers.

——. 2001b. *Modern Swahili Grammar*. East African Publishers.

Moore, Robert C, & William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, 220–224. Association for Computational Linguistics.

Moscoso Del Prado Martín, Fermín, Aleksandar Kostić, & R. Harald Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94.1–18.

Mous, Maarten. 1993. *A grammar of Iraqw*, volume 9. Buske Verlag.

——, Martha AS Qorro, & Roland Kiessling. 2002. *Iraqw-English dictionary: With an English and a thesaurus index*, volume 18. Rüdiger Köppe.

Moxley, Jeri L. 1998. Semantic structure of swahili noun classes. *Language history and linguistic description in Africa* 2.229.

Napoles, Courtney, Matthew Gormley, & Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, 95–100. Association for Computational Linguistics.

Nematzadeh, Aida, Stephan C Meylan, & Thomas L Griffiths. 2017. Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 859–864.

Nsoh, Avea E. 2002. Classifying the nominal in the gurene dialect of farefare of northern ghana. *Journal of Dagaare Studies* 2.83–95.

Paivio, Allan. 1963. Learning of adjective-noun paired associates as a function of adjective-noun word order and noun abstractness. *Canadian Journal of Psychology/Revue canadienne de psychologie* 17.370.

——. 1969. Mental imagery in associative learning and memory. *Psychological review* 76.241.

Pelletier, Francis Jeffry. 1994. The principle of semantic compositionality. *Topoi* 13.11–24.

Pennington, Jeffrey, Richard Socher, & Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pinker, S. 1997. Words and rules in the human brain. *Nature* 387.547–548.

Plag, Ingo, Christiane Dalton-Puffer, & Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language & Linguistics* 3.209–228.

Plaster, Keith, Maria Polinsky, & Boris Harizanov. 2009. Noun Classes Grow on Trees: Noun Classification in the North-East Caucasus. *Scholar.Harvard.Edu* 1–16.

Price, Glanville. 2013. *A comprehensive French grammar*. John Wiley & Sons.

Prins, Adriaan Hendrik Johan. 1961. *The Swahili-Speaking Peoples of Zanzibar and the East African Coast: Arabs, Shirazi and Swahili*, volume 12. International African Institute.

R Core Team, 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramscar, Michael. 2013. Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija* 46.377–396.

Ranganath, Charan, & Maureen Ritchey. 2012. Two cortical systems for memory-guided behaviour. *Nature Reviews Neuroscience* 13.713.

Rogers, Margaret. 1987. Learners difficulties with grammatical gender in german as a foreign language. *Applied Linguistics* 8.48–74.

Rousseeuw, Peter J, & Annick M Leroy. 2005. *Robust regression and outlier detection*, volume 589. John wiley & sons.

Sabourin, Laura, Laurie A Stowe, & Ger J De Haan. 2006. Transfer effects in learning a second language grammatical gender system. *Second Language Research* 22.1–29.

Schadeberg, Thilo C. 2006. Derivation. In *The Bantu Languages*, 71–89. Routledge.

Schriefers, Herbert. 1993. Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19.841.

——, & Encarna Teruel. 2000. Grammatical gender in noun phrase production: The gender interference effect in german. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26.1368.

Schultink, Henk. 1961. Produktiviteit als morfologisch fenomeen. In *Forum der letteren*, volume 2, 110–125.

Segui, Juan, Jacques Mehler, Uli Frauenfelder, & John Morton. 1982. The word frequency effect and lexical access. *Neuropsychologia* 20.615–627.

Seidenberg, Mark S, & Laura M Gonnerman. 2000. Explaining derivational morphology as the convergence of codes. *Trends in cognitive sciences* 4.353–361.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27.379–423.

—— 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 30.50–64.

Shapiro, Samuel Sanford, & Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52.591–611.

Siegel, Dorothy, 1974. *Topics in English Morphology*.

Sims, Andrea D., & Jeff Parker. 2015. Lexical processing and affix ordering: cross-linguistic predictions. *Morphology* 25.143–182.

Smith, Carlota S. 1964. Determiners and relative clauses in a generative grammer of english. *Language* 40.37–52.

Stanojevic, Maja. 2009. Cognitive synonymy: A general overview.

Stump, Gregory T. 1997. Template morphology and inflectional morphology. In *Yearbook of Morphology 1996*, 217–241. Springer.

——. 2001. *Inflectional morphology: A theory of paradigm structure*, volume 93. Cambridge University Press.

Sweetser, Eve E. 1988. Grammaticalization and semantic bleaching. In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, 389–405.

Taft, M. 1979. Recognition of affixed words and the word frequency effect. *Memory & cognition* 7.263–272.

Taft, Marcus, & Kenneth I Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior* 14.638–647.

Tanenhaus, Michael K, & Margery M Lucas. 1987. Context effects in lexical processing. *Cognition* 25.213–234.

Thorne, David A. 1993. *A comprehensive Welsh grammar*. Blackwell.

Towse, John N, Nelson Cowan, Graham J Hitch, & Neil J Horton. 2008. The recall of information from working memory: Insights from behavioural and chronometric perspectives. *Experimental Psychology* 55.371.

Traugott, Elizabeth Closs. 1989. On the rise of epistemic meanings in english: An example of subjectification in semantic change. *Language* 31–55.

Tucker, G Richard, Wallace E Lambert, Andre Rigault, & Norman Segalowitz. 1968. A psychological investigation of french speakers' skill with grammatical gender. *Journal of verbal learning and verbal behavior* 7.312–316.

Van Berkum, Jos JA. 1996. *The psycholinguistics of grammatical gender: Studies in language comprehension and production*. Nijmegen: Max Planck Instituut voor Psycholinguïstiek.

Vigliocco, Gabriella, Tiziana Antonini, & Merrill F Garrett. 1997. Grammatical gender is on the tip of italian tongues. *Psychological science* 8.314–317.

——, David P Vinson, Peter Indefrey, Willem JM Levelt, & Frauke Hellwig. 2004. Role of grammatical gender and semantics in german word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.483.

——, David P Vinson, Federica Paganelli, & Katharina Dworzynski. 2005. Grammatical gender effects on cognition: implications for language learning and language use. *Journal of Experimental Psychology: General* 134.501.

Voutilainen, Atro. 2003. Part-of-speech tagging. *The Oxford handbook of computational linguistics* 219–232.

Wei, Taiyun, & Viliam Simko, 2017. *R package "corrplot": Visualization of a Correlation Matrix*. (Version 0.84).

Wheeler, Benj Ide. 1899. The origin of grammatical gender. *The Journal of Germanic Philology* 2.528–545.

Whiteley, Wilfred Howell. 1969. *Swahili: The rise of a national language*, volume 3. Methuen.

Wicha, Nicole YY, Eva M Moreno, & Marta Kutas. 2003. Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in spanish. *Cortex* 39.483–508.

Wittgenstein, Ludwig, Gertrude Elizabeth Margaret Anscombe, & Ludwig Wittgenstein. 1953. *Philosophical Investigations... Translated by GEM Anscombe.(Philosophische Untersuchungen.) Eng. & Ger*. oxford.

Zandvoord, RW. 2013. *A handbook of English grammar*.

Zivkovic, Zoran. 2004. Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, 28–31. IEEE.

Zubin, David, & Klaus-Michael Köpcke. 1986. Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. *Noun classes and categorization* 139–180.

Zwicky, Arnold M. 1985. How to describe inflection. In *Annual Meeting of the Berkeley Linguistics Society*, volume 11, 372–386.